On Syntactic Anonymity and Differential Privacy

Chris Clifton and Tamir Tassa

Abstract

Recently, there has been a growing debate over approaches for handling and analyzing private data. Research has identified issues with syntactic approaches such as k-anonymity and ℓ-diversity. Differential privacy, which is based on adding noise to the analysis outcome, has been promoted as *the* answer to privacy-preserving data mining. This paper looks at the issues involved and criticisms of both approaches. We conclude that both approaches have their place, and that each approach has issues that call for further research. We identify these research challenges, and discuss recent developments and future directions that will enable greater access to data while improving privacy guarantees.

1 Introduction

In recent years, there has been a tremendous growth in the amount of personal data that can be collected and analyzed. Data mining tools are increasingly being used to infer trends and patterns. Of particular interest are data containing structured information on individuals. However, the use of data containing personal information has to be restricted in order to protect individual privacy. Although identifying attributes like ID numbers and names can be removed from the data without affecting most data mining, sensitive information might still leak due to linking attacks that are based on the public attributes, a.k.a *quasi-identifiers.* Such attacks may join the quasi-identifiers of a published table with a publicly accessible table like a voter registry, and thus disclose private information of specific individuals. In fact, it was shown in [74] that 87% of the U.S. population may be uniquely identified by the combination of the three quasi-identifiers birthdate, gender, and zipcode. This has lead to two related research areas: privacy-preserving data mining (PPDM) [3] enables the learning and use of data mining models while controlling the disclosure of data about individuals; privacy-preserving data publishing (PPDP) focuses on anonymizing datasets, in order to allow data disclosure without violating privacy.

The Official Statistics community has long recognized the privacy issues in both data publishing and release of statistics about data; for overviews see [38, 83]. Statistical Disclosure Limitation has primarily focused on tabular statistics, where a cell represents either a count of individuals matching that row/column (e.g., age range and income level), or a sum/average (e.g., years of education by race and state). Methods such as suppression (e.g., eliminating cells that reflect fewer than, say, five individuals), generalization by rounding values, or noise addition have been used to prevent individual identification [27]. There has been extensive work for ensuring that the combinations of values from such tables cannot be "solved" to reveal exact values for individuals, e.g. [81]. Such a privacy-aware release of statistics can be considered as PPDM.

This community has also worked on PPDP, specifically on the generation of privacy-preserving public use microdata sets. Many techniques were proposed in this context, including sampling,

suppression, generalization (particularly of geographic details and numeric values), adding random noise, and value swapping. There has been work on showing how such methods can preserve data utility; for example, value swapping maintains univariate statistics, and if done carefully, it can also maintain controlled approximations to multivariate statistics [61]. The state of practice is based on standards for generalization of certain types of information (e.g., any disclosed geographic unit must contain at least 10,000 [38] or 100,000 individuals [39]). Following such standards for generalization of specific types of data, the U.S. Healthcare Information Portability and Accountability Act (HIPAA) safe harbor rules [26] detail the types and specificity of data generalization that are deemed to make the data safe for releasing. A problem with this prescriptive approach is that each new domain demands new rules (e.g., due to different perceptions of the risk associated with re-identification and disclosure of data of different types, such as census data vs. health data vs. educational data). The proliferation of domains where data is being collected and may need to be published in a private manner makes this prescriptive approach impractical in the new big data world.

Moreover, even this prescriptive approach does not provide a guarantee of individual privacy, but only an expectation of privacy. For example, the HIPAA safe harbor rules allow the disclosure of the year of birth and the first three digits of the postal code (typically a region of roughly a county); if, by some strange anomaly, a county only has one person born in 1950, then that individual is revealed even though the rules are followed. The result is that these prescriptive approaches are often very conservative, resulting in lower utility of the data. The fact that such standards exist, given the knowledge that they do not provide perfect privacy, suggests that PPDM and PPDP do not need to provide an absolute guarantee of privacy; adequate privacy (which may vary by domain) can be sufficient.

The Official Statistics research community has developed numerous methods for generating privacy-protected microdata, but this has not resulted in a standard approach to PPDP. One difficulty is that much of the work emphasizes methods to produce microdata sets, often for a particular domain. This makes the work difficult to generalize. There has recently been an explosion of attempts in the computing research community to provide formal mathematical definitions that either bound the probability of identification of individuals, or the specificity of information released about individuals. While much of the earlier (and current) work in Statistical Disclosure Limitation is highly relevant, a comprehensive survey and comparative analysis of those methods is beyond the scope of this paper. Herein, we focus only on the recent definitions offered by the computing research community, and indicate claims or interpretations that we perceive as misunderstandings that are impacting the progress of research in this field.

Probably the first formal mathematical model to achieve wide visibility in the computing research community was k-anonymity, proposed by Samarati and Sweeney [69, 70, 75]. This model requires that each of the released records be indistinguishable from at least $k-1$ other records when projected on the quasi-identifier attributes. As a consequence, each individual may be linked to sets of records of size at least k in the released anonymized table, whence privacy is protected to some extent. This is accomplished by modifying table entries. The above seminal studies, and the majority of the subsequent studies, modify data by generalizing table entries. However, other techniques have also been suggested to achieve record indistinguishability (see more on that in Section 2). All those techniques first partition the data records into blocks, and then release information on the records within each block so that the linkage between quasi-identifier tuples and sensitive values within a given block is fully blurred.

Several studies have pointed out weaknesses of the k-anonymity model and suggested stronger measures, e.g., ℓ-diversity [56], t-closeness [53], or β-likeness [10]. Other studies attempted to enhance the utility of such anonymized tables, e.g., [33, 78, 48, 86]. Those models, which we describe in Section 2, are similar to k-anonymity in that they (typically) generalize the database entries until some syntactic condition is met, so that the ability of an adversary to link a quasi-identifier tuple to sensitive values is restricted.

Despite the enhanced privacy that those models offer with respect to the basic model of k-anonymity, they are still susceptible to various attacks. As a result of those attacks, it seems that part of the research community has lost faith in those privacy models. The emergence of differential privacy [18], a rigorous notion of privacy based on adding noise to answers to queries on the data, has revolutionized the field of PPDM. There seems to be a widespread belief that differential privacy and its offsprings are immune to those attacks, and that they render the syntactic models of anonymity obsolete. In this paper we discuss the problems with syntactic anonymity and argue that, while all those problems are genuine, they can be addressed within the framework of syntactic anonymity. We further argue that differential privacy too is susceptible to attacks, as well as having other problems and (often unstated) assumptions that raise problems in practice.

While criticisms of syntactic anonymity stem from its shortcomings in providing full privacy for the individuals whose data appear in the table, it is imperative also to discuss the second aspect of PPDP: the utility of the sanitized data for legitimate (non-privacy-violating) purposes. As we see in the news on a regular basis (such as the changes to privacy policies and practices of Google and facebook), without regulation utility trumps privacy: if the choice is between a method that provides privacy but fails to adequately support data analysis, or sharing data at a greater risk to privacy, the choice will be to share the data. Another example comes from the U.S. HIPAA Privacy Rule [26], which provides a clear syntactic mechanism to anonymize data to meet legal standards. This safe harbor mechanism appears to have received little use in practice. Instead, (limited) datasets are disclosed only under data use agreements. (Such restrictions on use of de-identified data seem widespread. For example, even though the NIH database of Genotypes and Phenotypes (dbGaP) only accepts de-identified individual level data, the (de-identified) individual level data can only be accessed with an approved Data Use Certification [64].) These agreements can increase privacy risk, as data can be shared in identifiable form; but also constrain what can be done with the data. Any new use demands a new agreement, even if that use poses no additional privacy risk. This results in less freedom in use of the data, as well as greater potential threats to privacy. As for differentially private data, a study of its utility is still in order. Until it is clarified how useful it is for practitioners of data mining, differential privacy has still not reached the maturity to replace other existing models of PPDM.

Throughout this note we concentrate on data in the form of a table of records, in which each record is a multidimensional tuple that provides information about a single entity in the underlying population. Other types of data, such as graph data (as in social networks), introduce additional complications that we will only discuss briefly.

1.1 Issues in using privacy technology

The key tradeoff in analysis of privacy-sensitive data is between the risk of privacy violations and utility of the data. Privacy violations are typically expressed in terms of *individual identi-*

fiability, a concept enshrined in most privacy law [23, 26]. A simple interpretation of immunity against individual identifiability is that for any individual, one should not be able to identify or learn private information about that individual with probability or confidence greater than some threshold. (Smaller settings of the threshold imply better immunity.)

As for utility, it typically refers to two features of the processed data: the ease of use of that data for data mining and other analysis purposes; and the correctness of conclusions drawn from the processed data by such analysis. Users of data typically feel that they are getting full utility out of the original data, and lower utility after privacy-preserving measures have been taken, but quantifying this loss is difficult. Several measures were proposed in the literature for the utility of syntactic anonymization, but they do not seem to be effective in comparing different anonymization approaches (or even different algorithms within the same approach) to evaluate the impact on data mining [65]. Evaluating the impact of noise addition techniques on utility would seem more obvious, as the noise can often be calculated in direct terms of impact on the results (e.g., a Gaussian distribution around the true answer). However, this evaluation is often non-trivial, as noise addition may occur on the data itself (as with data modification methods for public use microdata sets), or on intermediate results (such as noise-based approaches to statistical disclosure control for frequency counts, or differentially private histograms). Noise addition seems to run into considerable resistance among practitioners, who are not willing to accept data which was subjected to noise addition (even though the original data itself may contain some errors).

There are other important factors that should be weighed in when choosing a privacy-preserving procedure; e.g., the efficiency and scalability of the algorithm implementing the selected procedure; or the suitability of the selected solution to the capabilities of the data custodian (the party that holds the data) and the expected demands of the data users (we elaborate on that in Section 4). However, those factors are not necessarily in tradeoff with either risk to privacy or utility.

The tradeoff between privacy risk and utility makes analysis and publishing of privacy-sensitive data a challenge. Some of the specific issues that arise in privacy preservation techniques, which, in particular, distinguish between syntactic and noise-based models, are listed below.

Managing privacy policy. Privacy policy for syntactic models can generally be defined and understood based on the data schema; parameters have a clear privacy meaning that can be understood independent of the actual data, and have a clear relationship to the legal concept of individual identifiability of data. On the other hand, while ε-differential privacy does relate to individual identifiability, the privacy parameter ε does not have such a clear relationship [49]. An appropriate setting of ε requires an extensive analysis of the query, the data, and the universe as a whole. (This is discussed further in Section 6.3.) What this means is that there is no clear way to set a general policy for a value of ε that provides sufficient privacy.

Open-ended vs. compact distribution. Syntactic models typically provide a compact distribution (e.g., a generalized value or a set of possible values) for a given anononymized data value. Perturbation-based models typically give an open-ended distribution (e.g., Gaussian or Laplacian distributions.) This has significant implications for utility: analysis of compact distributions can provide guarantees on correctness of results; open-ended distributions provide probabilistic bounds that, even if tighter, may not provide the level of certainty desired [25].

Application to multiple uses of data. Models based on the release of anonymized data can safely be used for as many distinct uses as desired. Methods that release only query results require tracking the results: early uses of the data can affect the quality of later uses, or even result in a threshold beyond which no new queries can be permitted on the data. While noise can be applied to a data release (e.g., the techniques used in Public Use Microdata Sets [62]), most recent research has concentrated on answering queries against the data. It has long been known that care must be taken to ensure that multiple queries do not violate privacy [17]. Differential privacy does address this, as differentially private answers to queries are composable, with each consuming a portion of the "privacy budget". For example, in order to achieve ε-differential privacy over two queries, the answer to each query can be made noisier so that each complies with $\varepsilon/2$-differential privacy. However, if the query stream is not known in advance, adding too little noise to early queries can prevent reasonable answers to later queries.

There is also an issue of determining how to set a "privacy budget". Assuming public access to a dataset, any privacy budget could quickly be exhausted. An alternative is to assign individual privacy budgets, but this requires ensuring that individuals do not collude, limiting dataset access to individuals who can be authenticated and vetted. This poses interesting policy challenges for use of differential privacy.

Results with correlated values. Correlation among data values raises a number of issues, particularly with noise addition techniques. For example, publishing a dataset with noise added to attributes `age` and `height` is challenging: a 5 year-old with a positive noise added to age and a negative noise added to height would indicate that the true age is probably less than the published age, and the true height is greater than the published height, since those two attributes are correlated [42]. The basic definition of differential privacy, that limits the probability of the noisy outcome belonging to a particular set, accounts also for multi-dimensional outcomes. However, when attributes are correlated, it is not sufficient to independently add noise to the different attributes using (for example) the Laplacian mechanism of differential privacy. Correlated values in multi-dimensional outcomes require a careful mechanism design rather than a simple composition of mechanisms for scalar-valued outcomes.

Syntactic approaches to anonymity must also account for correlation. As a simple example, suppose that a record of a 5-year old child was generalized to a record with the generalized age [5-15], but except for that single 5-year old child, all individuals whose records were generalized in the same manner are teenagers. Assume, in addition, that the sensitive attribute is `disease`, and that the diseases in all other records in the same group are usually found only in teenagers. Then it may be possible to infer that the remaining disease in that group afflicts the 5-year old child. More subtle issues with correlation are discussed in Section 5.1.

1.2 Outline

We start with an overview of the two types of privacy models which are at the center of this study: in Section 2 we describe the main syntactic models of anonymity, and in Section 3 we discuss differential privacy. We then commence our discussion of these two approaches, towards our conclusion that the two approaches have their place, one alongside the other, and that both should be further studied, explored and improved.

In Section 4 we explain that those seemingly competing approaches are targeting two different

scenarios of privacy-preserving usage of data: PPDP and PPDM. We discuss those two scenarios and highlight the differences between them. In particular, we show that syntactic models are designed for PPDP, while differential privacy is more suitable to PPDM; hence, the comparison between the two approaches is an apples-and-oranges comparison. In particular, one approach cannot replace the other.

In Section 5 we discuss the most important criticisms of syntactic anonymization approaches, and presents recent work and research directions to deal with these concerns. Section 6 does the same for noise addition approaches, with a particular focus on differential privacy.

We conclude in Section 7 with the above mentioned conclusion, and suggest future research directions in both approaches.

2 Syntactic models of anonymity

Herein we survey some of the main models of syntactic anonymity. Most of those models are based on generalizing table entries. Such data distortion preserves the truthfulness of data, in the sense that a generalized value defines a group of possible original values. (Generalization also includes, as a special case, the operation of suppression; suppression also defines a group of possible original values since usually the dictionary of possible values for each attribute is known.) Those models provide privacy for the data subjects by rendering some sort of record indistinguishability.

k-Anonymity and most of the models that evolved from it are based on partitioning the database records to blocks and then anonymizing the records so that those that appear in the same block become indistinguishable. In k-anonymity, all blocks are required to be of size at least k, and the records within each block are replaced with their closure, being the minimal generalized record that generalizes all of them. Each such block is called a QI-block (Quasi-Identifier block). All generalized records within the same QI-block agree in their quasi-identifier attributes, whence, they are indistinguishable to an adversary, under the assumption that the quasi-identifiers are the only attributes that could be used by an adversary in a linking attack. An example is given in Table 1; in that example, there are two quasi-identifiers (`age` and `zipcode`), and one sensitive value (`disease`).

age	zipcode	disease
28	10145	measles
21	10141	hepatitis
21	12238	hepatitis
55	12256	flu
53	12142	angina
48	12204	angina

age	zipcode	disease
$[21-28]$	1****	measles
$[21-28]$	1****	hepatitis
$[21-28]$	1****	hepatitis
$[48-55]$	12***	flu
$[48-55]$	12***	angina
$[48-55]$	12***	angina

Table 1: (a) A table (left); (b) a corresponding 3-anonymization (right).

Several studies have pointed out weaknesses of the k-anonymity model and suggested stronger measures such as ℓ-diversity [56], t-closeness [53], or p-sensitivity [79]. The main weakness of k-anonymity is that it does not guarantee sufficient diversity in the sensitive attribute within each QI-block. Namely, even though it guarantees that every record in the anonymized table is indistinguishable from at least $k-1$ other records, when projected on the subset of quasi-

identifiers, it is possible that most (or even all) of those records have the same sensitive value. Therefore, an adversary who is capable of locating a target individual in that block of records will be able to infer the sensitive value of that individual with probability that might be higher than what is desired. For example, an adversary who targets a 21 year old person in zipcode 12238 may deduce, from the 3-anonymization in Table 1, that the person has hepatitis with probability 2/3.

Machanavajjhala et al. [56] proposed the security measure of ℓ-diversity. They suggested that the sensitive attribute in each QI-block will have at least ℓ "well represented" values. They offered two interpretations of that measure. In one interpretation, the entropy of the distribution of sensitive values in that attribute in every QI-block should be at least $\log \ell$, for some predetermined value of the parameter ℓ. In practice, a simpler interpretation of ℓ-diversity is usually applied [84]. According to that interpretation, an anonymization is considered ℓ-diverse if the frequency of each of the sensitive values within each QI-block does not exceed $1/\ell$. Table 2 shows an alternative 3-anonymization of Table 1(a). This anonymization respects 3-diversity.

age	zipcode	disease
$[21-53]$	1****	measles
$[21-53]$	1****	hepatitis
$[21-55]$	122**	hepatitis
$[21-55]$	122**	flu
$[21-53]$	1****	angina
$[21-55]$	122**	angina

Table 2: A 3-anonymization of Table 1(a) that respects 3-diversity.

The notion of t-closeness is stronger than ℓ-diversity since it demands that the distribution of the sensitive values within every QI-block would be sufficiently close to its general distribution in the table. The notion of p-sensitivity, on the other hand, relaxes the notion of ℓ-diversity as it only requires each QI-block to have p distinct sensitive values, but does not impose any condition on their distribution.

It is important to understand that these definitions do not replace k-anonymity. They offer essential *enhancements* to k-anonymity in the sense that one must require them *in addition* to k-anonymity. In accord with this, Truta et al. [79] propose algorithms that generate tables that are both k-anonymous and p-sensitive, and Wong et al. [84] consider the conjunction of k-anonymity with the last interpretation of ℓ-diversity (they call this conjunction of conditions $(1/\ell, k)$-anonymity).

We clarify that point using the example of ℓ-diversity. The diversity of a table is bounded from above by the number of possible sensitive values (equality holds if and only if the distribution of the sensitive values is uniform). The diversity of any anonymization of the table is bounded from above by the diversity of the entire table (equality holds if and only if the distribution in each QI-block equals the global distribution). Therefore, if the table has a sensitive attribute with a small number of possible values, all of its anonymizations will respect ℓ-diversity with ℓ that does not exceed this number. For example, in the case of a binary sensitive attribute, one can aim at achieving ℓ-diverse anonymizations with $\ell \leq 2$ only. In such a case, if one imposes only ℓ-diversity, the blocks of indistinguishable records could be of size 2. Such small blocks do not provide enough privacy for the individuals in them, because if an adversary is able to learn the sensitive value

of one of those individuals, it may infer that of the other one as well. If, on the other hand, we demand that such ℓ-diverse anonymizations are also k-anonymous, for a suitable selection of k, then the adversary would have to find out the sensitive values of at least $k/2$ individuals before it would be able to infer the sensitive value of the target individual.

Some variants of the above basic models were also suggested in the literature. For example, the so-called Anatomy model [87] also starts by partitioning the table records. But instead of generalizing the quasi-identifiers of all records in the same QI-block to their closure, it leaves them unchanged and, instead, randomly shuffles the sensitive values within each such block. Table 3 illustrates the Anatomy anonymization of Table 1(a) that uses the same partitioning into QI-blocks as the anonymization in Table 2. The privacy guarantee remains the same, since an adversary who locates a target individual in some QI-block, can still infer that the target's sensitive value is one of the values that appear in that block.

age	zipcode	disease
28	10145	hepatitis
21	10141	angina
53	12142	measles
21	12238	flu
55	12256	hepatitis
48	12204	angina

Table 3: An Anatomy anonymization of Table 1(a) that uses the same partitioning into QI-blocks as the anonymization in Table 2.

Another variant [2] suggests to publish cluster centers and radii. Such a variant is applicable when the quasi-identifiers take values in a metric space, as is the case with the two quasi-identifiers in our running example.

Gionis et al. [33, 78] proposed a novel approach that suggests achieving anonymity without basing it on partitioning the data records into QI-blocks. Their non-partition models of anonymization extend the corresponding partition-based models. They showed that this extension of the anonymization framework allows achieving similar levels of privacy with smaller information losses. The recent studies [48, 86] further explored that idea, suggested corresponding anonymization algorithms, and demonstrated the advantages offered by such models.

3 Differential privacy

In the middle of the previous decade, the research community began exploring new privacy notions that are not based on a syntactic definition of privacy, most prominent among which is differential privacy [18]. Differential Privacy is a formal definition relating uncertainty at an individual level to the noise or randomization used in a privacy mechanism.

Let $\mathcal{A}$ be a randomized algorithm that accepts as an input a database D and outputs an answer to some query on the database. $\mathcal{A}$ is said to provide ε-differential privacy if for any two databases, D_1 and D_2, that differ in only one entry, and for any subset S of values in the range of outputs of $\mathcal{A}$, $\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(D_2) \in S]$. Namely, a randomized algorithm satisfies ε-differential privacy if it is ε-hard, in the above probabilistic sense, to distinguish between two

databases that differ in *any* single entry. (The slightly weaker notion of (ε, δ)-indistinguishability [66] adds an additive small term to the upper bound on the first distribution in terms of the other distribution.)

Owing to its rigorous approach and formal privacy guarantees, differential privacy has started to be adopted by a growing part of the academic community as the only acceptable definition of privacy, sometimes to the extent that it is viewed as rendering previous privacy models obsolete.

A key value of differential privacy is that it is proof against an attacker with strong background knowledge. The strong attacker assumed by differential privacy knows all records in the database except for one record, but is still unable to violate the privacy of the individual behind that record: the query result would be essentially indistinguishable (modulo e^{ε}) whether that individual's record was or was not in the data. The second breakthrough made by differential privacy is in formulating a general mechanism for adding noise to any continuous-valued query towards meeting that privacy measure. Another merit of differential privacy is that it is composable, in the sense that it can support multiple queries on the data.

3.1 A quick survey of differential privacy results

Many techniques have been proposed for applying differential privacy to specific data publishing and mining tasks. A survey by Dwork [19] provides a comprehensive review. For example, differential privacy has been applied to releasing query and click histograms from search logs [35, 47], recommender systems [58], publishing commuting patterns [55], publishing results of machine learning [7, 11, 43], clustering [28, 66], decision trees [29], mining frequent patterns [6], and aggregating distributed time-series [67].

Several recent works have focused on differentially private publishing of count queries. Hay et al. [37] propose an approach based on a hierarchy of intervals. Li et al. [52] propose a general framework that supports answering a given workload of count queries, and consider the problem of finding optimal strategies for a workload. Barak et al. [5] show how to publish a set of marginals of a contingency table while ensuring differential privacy by means of noise addition. Xiao et al. [88] propose to use the Haar wavelet for range count queries. They extend their wavelet approach to nominal attributes and multidimensional count queries. Ding et al. [16] also consider multidimensional count queries. They study the problem of publishing data cubes for a given fact table (microdata) while ensuring ε-differential privacy on one hand, and limiting the variance of the noise for better utility on the other hand. (A data cube is a set of aggregate counts for a projection of the basic table on subsets of its attribute set.) Xiao et al. [89] study the problem of differentially private histogram release based on an interactive differential privacy interface. They propose two multidimensional partitioning strategies including a baseline cell-based partitioning and an innovative kd-tree based partitioning.

4 PPDM and PPDP

There is a fundamental difference between the assumptions that underlie differential privacy and those that underlie syntactic privacy models. In fact, those two seemingly competing approaches are targeting two different playgrounds.

k-Anonymity and other syntactic notions of anonymity target PPDP. A typical scenario of PPDP is that in which a hospital wishes to release data about its patients for public scrutiny of

any type. The hospital possesses the data and is committed to the privacy of its patients. The goal is to *publish* the data in an anonymized manner without making any assumptions on the type of analysis and queries that will be executed on it. Once the data is published, it is available for any type of analysis.

Differential privacy, on the other hand, typically targets PPDM. In PPDM, as opposed to PPDP, the query that needs to be answered must be known prior to applying the privacy-preserving process. In the typical PPDM scenario, the data custodian maintains control of the data and does not publish it. Instead, the custodian responds to queries on the data, and ensures that the answers provided do not violate the privacy of the data subjects. In differential privacy this is typically achieved by adding noise to the data, and it is necessary to know the analysis to be performed in advance in order to calibrate the level of noise to the global sensitivity of the query and to the targeted differential privacy parameter ε [21]. While some differential privacy techniques (e.g., private histograms) are really intermediate analysis rather than a final data mining model, it is still necessary for the data custodian to know what analysis is intended to be performed.

In their criticism on syntactic models of privacy and defense of differential privacy, Narayanan and Shmatikov [63] state that PPDP is a bad idea and that only PPDM may provide sufficient privacy ("an interactive, query-based approach is generally superior from the privacy perspective to the 'release-and-forget' approach."). They acknowledge the impracticality of that conclusion by adding that "this can be a hard pill to swallow, because it requires designing a programming interface for queries, budgeting for server resources, performing regular audits, and so forth." Hence, while interactive approaches do have some advantages in the privacy vs. utility tradeoff, their inherent limitations are such that PPDP is likely here to stay.

The comments in [63] also miss the point that differential privacy does not necessarily imply an interactive approach. Noise and syntactic generalization have in fact been combined to support real-world data publishing [62]. The definition of differential privacy supports a query such as "return the dataset D", requiring that the returned data have noise added (as with some public use microdata sets) to ensure that the information related to any individual is sufficiently hidden. While differentially private data publishing has been shown to be possible [54, 8, 22, 36, 68, 12] (see Section 7), there has been little work to show that such an ε-differentially private dataset would be practical and useful.

Data publishing is a widespread practice (see, for example, public use microdata sets[1]); hence, it is important to develop appropriate techniques for PPDP. Fung et al. [30] argue that even if the data custodian knows in advance that data will be used for classification, it is not enough to just build and publish a classifier. First of all, even if the data custodian knows that the data will be used for classification, it may not know how the user may analyze the data. The user often has application-specific bias towards building the classifier. For example, some users prefer accuracy while others prefer interpretability, or some prefer recall while others prefer precision. In other cases, visualization or exploratory analysis of the data may guide the user toward the right approach to classification for their particular problem. Publishing the data provides the user a greater flexibility for data analysis. It should be noted that while data publishing techniques can be customized to provide better results for particular types of analysis [40, 65, 34, 51], data which is published towards a specific data mining goal can still be used for other data mining goals as

[1] http://www.census.gov/main/www/pums.html

well.

Mohammed et al. [60] too address the PPDP versus PPDM question. They provide additional arguments to support the necessity in publishing the data. First, the data custodian (e.g., a hospital, or a bank) often has neither the expertise nor the interest in performing data mining. Second, it is unrealistic to assume that the data custodian could attend to repeated requests of the user to produce different types of statistical information and fine-tune the data mining results for research purposes.

In conclusion, PPDP is an essential paradigm that coexists alongside PPDM. Differential privacy is viable for PPDM, but it is still an open question if it can practically support PPDP. The syntactic notions of privacy that we reviewed in Section 2 are viable solutions for PPDP.

5 Criticism of syntactic models of anonymity

In this section we describe some of the main criticisms of syntactic models, and explain why they are a challenge for further research rather than a justified cause to abandon the models.

5.1 The deFinetti attack

The random worlds model [4] is commonly used to reason about attackers. According to that model, all tables with specific quasi-identifier values that are consistent with the published anonymized table are equally likely, and the adversary uses that assumption in order to draw from the anonymized table belief probabilities regarding the linkage between quasi-identifier tuples in the table and sensitive values. Based on that assumption, it is argued in [56] that anonymized tables that are ℓ-diverse prevent inferring belief probabilities that are larger than $1/\ell$.

In [44], Kifer showed that it is possible to extract from ℓ-diverse tables belief probabilities greater than $1/\ell$ by means of the so-called deFinetti attack. That attack uses the anonymized table in order to learn a classifier that, given the quasi-identifier tuple of an individual in the underlying population, is able to predict the corresponding sensitive value with probability greater than the intended $1/\ell$ bound.

There are three arguments why that attack is not a solid argument to abandon syntactic privacy models in favor of differential privacy. First, a recent study of Cormode [13] found that the attack might not be useful. While Kifer showed that the effectiveness of the attack reduces with ℓ (since greater values of ℓ make the statistical learning process harder), and its computational complexity grows dramatically, Cormode found that even for small values of ℓ, the effectiveness of the attack diminishes substantially when the size of the ℓ-diverse blocks grows. Hence, ℓ-diverse k-anonymizations are immune to this attack for sufficiently large values of k.

Second, syntactic models of privacy are no more susceptible to the deFinetti attack than differential privacy. Indeed, the main finding in [13] is that the accuracy of inference of sensitive attributes, by means of the deFinetti attack, for differentially private data and ℓ-diverse data can be quite similar. His conclusion was that "rejecting all such (syntactic) anonymizations because the deFinetti attack exists is erroneous: by the same logic, we should also abandon differential privacy". His final conclusion, with which we concur, is that depending on the perceived threats, and the consequences of a successful attack, it may be appropriate to use de-identification, syntactic methods, differential privacy, or to withhold release entirely.

The third argument is more fundamental. The deFinetti attack relies on building a classifier based on the entire database. The question is whether the inference of a general behavior of the population in order to draw belief probabilities on individuals in that population constitutes a breach of privacy; differential privacy explicitly allows learning general behavior as long as it is not dependent on a single individual. To answer this question positively for an attack on privacy, the success of the attack when launched against records that *are* part of the table should be significantly higher than its success against records that *are not* part of the table. We are not aware of such a comparison for the deFinetti attack.

It is worth mentioning in this context the recent work by Last et al. [48]. They devised an algorithm that issues non-partition based anonymizations that are k-anonymous and ℓ-diverse. They then used those anonymizations in order to learn a classifier. They showed that the accuracy of that classifier over the original training data records was almost identical to its accuracy on new testing data records. Namely, even though data is published on Alice, Bob, and Carol, the "classifier attack" presents a similar level of risk for them, as well as for David, Elaine, and Frank who were not included in the original table that was used to generate the published anonymized data and, subsequently, to learn the classification model. Hence, such an "attack" cannot be regarded as a breach of privacy. It can only be regarded as a successful learning of the behavior of the general population, which is the raison d'être of any data publishing.

5.2 Minimality attacks

The minimality attack [85] exploits the knowledge of the anonymization algorithm in order to infer properties of the original data and, consequently, of individuals. An anonymized view of the original data induces a set of "possible worlds" for what the original data might have been. The knowledge of the anonymization algorithm and its decision making process enables, sometimes, eliminating some of the possible worlds, and thus increases the belief of the attacker in certain events to a level that is inconsistent with the desired privacy requirements.

Cormode et al. [14] conducted a detailed analysis of those attacks and showed that safeguards against such attacks can be found within the syntactic framework. They identified three criteria that render algorithms virtually immune to minimality attacks. The first one is the use of randomness – the anonymization algorithm must use randomness so that even an adversary who knows the algorithm cannot carry out the logical reasoning that underlies the minimality attack. (Note that randomness is not the same as noise addition; it simply requires that the method have random variability in the choice of how to perform syntactic anonymization. For example, an algorithm may face a choice of generalizing birthdate or address in order to achieve k-anonymity; making this decision randomly, as opposed to a deterministic optimality criterion, provides protection against minimality attacks. The generalization itself can still follow a generalization hierarchy rather than noise addition.) Second, algorithms that have a high degree of symmetry in their grouping choices are virtually invulnerable. And third, anonymization algorithms that do not jointly consider the quasi-identifiers and the sensitive attribute in determining the best way to generalize are immune to minimality attacks.

5.3 The curse of dimensionality

Aggarwal [1] showed that when the number of quasi-identifiers is large, most of the table entries have to be suppressed in order to achieve k-anonymity. Due to this so-called "curse of dimen-

sionality", applying k-anonymity on high-dimensional data would significantly degrade the data quality. This is an essential problem, but it may be addressed within the framework of syntactic privacy.

Mohammed et al. [60] suggested exploiting one of the limitations of the adversary: in real-life privacy attacks, it can be very difficult for an adversary to acquire complete background information on target individuals. Thus, it is reasonable to assume that the adversary's prior knowledge is bounded by at most L values of the quasi-identifiers, for some integer L that is smaller than the number of attributes in the dataset. Based on this assumption, they defined a new privacy model, called LKC-privacy, for anonymizing high-dimensional data. That privacy notion ensures that every combination of L quasi-identifier values is shared by at least K records, and the diversity of the sensitive value in each such group of records is no larger than $1/C$, for some specified parameters L, K, and C. In other words, LKC-privacy bounds the probability of a successful identity linkage to be at most $1/K$ and the probability of a successful attribute linkage to be at most $1/C$, provided that the adversary's prior knowledge does not exceed L quasi-identifiers. They then devised an anonymization algorithm for this privacy notion and tested it on real-life data. Their experiments showed that this privacy notion and the corresponding algorithm can effectively retain the essential information in anonymous data for data analysis.

In that context, it is important to understand that not all non-sensitive attributes should be automatically classified as quasi-identifiers. The data custodian should assess the chances of an adversary obtaining each of the attributes in the data schema. If the chance of an adversary getting hold of some attribute is small relative to the chance of acquiring the sensitive data by other means, then there is no need to consider such an attribute a quasi-identifier. Indeed, the standard assumption in PPDP is that the candidate table to be published includes multiple types of attributes. For example, Burnett et al. [9] define the following types: identifiers – attributes that uniquely identify an individual (e.g. `name`); quasi-identifiers – non-sensitive attributes like `zipcode`, `age`, or `gender`, that could be used in linkage attacks; non-identifiers – non-sensitive attributes that are not quasi-identifiers, in the sense that an adversary is unlikely to get hold of them; and sensitive attributes – personal attributes of private nature, such as `disease` or `income`.

Another important observation that mitigates the curse of dimensionality is that not all quasi-identifiers are needed for every data sharing purpose. The context in which the data is to be used (e.g., a medical research, a socioeconomic research, or a marketing research) may determine a different subset of the attributes that could be relevant in that context. Hence, instead of a single publication of the entire high-dimensional dataset, it is expected that the data will be published in several releases, where each release is an anonymization of a lower dimensional dataset which is a projection of the original dataset onto a subset of the attributes. In such cases, it is possible to combine information from several releases in order to breach privacy. Hence, the goal in this context is to protect the private information from adversaries who examine several releases of the underlying dataset. Algorithms for anonymizing datasets that are released in this manner were proposed in [72, 73, 80].

5.4 Composition attacks

Ganta et al. [31] described composition attacks on anonymized data. They considered settings in which multiple organizations independently release anonymized data about overlapping populations. An adversary who knows that some target individual occurs in the intersection of the

underlying populations of two (or more) such independent releases, may use that auxiliary knowledge in order to breach privacy. In their motivating example, Alice suffers from AIDS and she visits two different hospitals in her city. Those two hospitals independently issue anonymized releases about their patients. An adversary who knows that Alice appears in both releases may be able to infer that she suffers from AIDS, by intersecting the sets of sensitive values that may be linked to Alice by each of the two releases.

As the quality of data mining significantly improves if it is based on larger corpora of data, the best way to avoid such attacks is for data custodians to collaborate and issue one large release instead of separate smaller releases. For example, various hospitals in the same city or state can publish the anonymization of the unified data set that they jointly hold. Several studies have suggested secure multi-party protocols for computing anonymized views of databases that are distributed horizontally among several data holders, e.g., [41, 77, 90].

In addition to the improved utility from collaboratively published data, it also protects against privacy breaches. Malin describes a problem of trail re-identification [57]. He showed that multiple independent releases of data about an individual, coming from different sources, can result in a privacy violation even if each release independently satisfies the privacy constraint. He then proceeded to describe a solution to that attack, within the framework of syntactic anonymization.

6 Criticisms of differential privacy

It is important to be clear about the claims of differential privacy. Differential privacy bounds the impact an individual has on the outcome (data mining model, or published dataset.) The main premise is that if knowledge can be gained without an individual's data, then that individual's privacy is not violated – even if the knowledge can be used to learn private information about the individual. This means that certain types of background knowledge (e.g., how far an individual deviates from the mean) can be used with a differentially private result to learn specific values about the individual without violating differential privacy; the promise of differential privacy is (by design) not absolute secrecy. Many of the criticisms of both syntactic anonymity and differential privacy (such as some background knowledge attacks) presume any disclosure of information about an individual is a violation; Dwork showed in [18] that this cannot be achieved without entirely foregoing data utility. The definition of differential privacy sidesteps this issue by providing *relative* privacy; participating in the database should only slightly increase the risk of disclosure. (The additive noise δ in (ϵ, δ)-indistinguishability does not necessarily provide the same protection and must be used carefully, see [20].) That said, differential privacy is a strong notion of privacy – but it still suffers from a number of practical problems and limitations.

6.1 Computing global sensitivity

Computing a realistic bound on the global sensitivity of multidimensional queries requires a very complex analysis of the domain of all possible tuples in the multidimensional space. For example, assessing the global sensitivity of queries that relate height and weight, based only on the ranges of each of those attributes, without taking into consideration their correlation, may give unreasonably high values; specifically, even though a typical range of heights includes the height 2 meters, and a typical range of weights includes the weight 3 kilograms, it would be devastating to add noise for calculating the body mass index for protecting against the possibility that the database includes

a person with height 2 meters and weight 3 kilograms. Unrealistic sensitivity values give excessive noise, resulting in little utility from a differentially privacy result.

While specific types of queries may be amenable to specific techniques that do not pose these issues (e.g., the previously mentioned histogram queries), in general computing a global sensitivity that both guarantees privacy and provides usable levels of noise is a difficult task.

6.2 Non-compact uncertainty

Another problem with the applicability of differential privacy is the inherent uncertainty in the answer. In disciplines such as biostatistics or biomedical research, it is imperative to have known bounds on the value of the original data [25]. This is the case with syntactic anonymization models, in which data is generalized according to accepted generalization rules. This is not the case with perturbation models in which the correlation between the original and perturbed data is probabilistic. Because of those reasons, syntactic privacy models, such as k-anonymity, are still perceived by practitioners as sufficient for mitigating risk in the real world while maximizing utility, and real life applications still utilize them for sanitizing data (see [24, 25]).

In addition to the inherent uncertainty in the answer, the quality of results obtained from a differentially private mechanism can vary greatly. Many of the positive results have been obtained using histogram-style queries on Boolean data. However, the differentially private mechanism of adding Laplacian noise can significantly alter the answer. An example is provided in [71]: a differentially private query for the mean income of a single U.S. county, with $\varepsilon = 0.25$ (resp. $\varepsilon = 1.0$), deviates from the true value by \$10,000 or less only 3% (resp. 12%) of the time! This can be extremely misleading, given that the true value is \$16,708. (This is a real-world example of a query with high-income outliers that cause high global sensitivity. In methods of syntactic anonymity, such outliers may only have local effect on records that were grouped with them in the same anonymity block.)

Wasserman and Zhou [82] show similar results for the differentially private histogram method of [21]; substantial error arises with smaller sample sizes. They also formally analyzed such accuracy variation for the exponential mechanism of [59]. They showed that the accuracy is linked to the rate at which the empirical distribution concentrates around the true distribution.

6.3 How to set ε?

The U.S. HIPAA safe harbor rules [26] specify legally acceptable syntactic anonymizations. Certain types of identifying information must be removed, and dates and locations have to be generalized to some extent: locations must be generalized into geographic units that have at least 20,000 residents; date of birth must be rounded up to the year of birth only (unless the age is 90 years or more, in which case wider ranges are required). A simple "back of the envelope" calculation yields the level k of anonymity that those rules induce. In differential privacy, on the other hand, little has been done to address the practically essential question of how to set the privacy parameter ε. While the definition of differential privacy clearly addresses the issue of identification (if it is hard to determine whether an individual is in the database, it is certainly hard to identify that individual's record in the database), the way in which ε affects the ability to identify an individual is not as clear. The parameter ε in ε-differential privacy is not a measure of privacy in the normal sense: it bounds the impact an individual has on the result, not what is disclosed about an individual. Queries that specifically ask information about an individual,

e.g. "Is Bob in the database", are an exception. In such queries, ε directly relates to disclosure of information on that particular individual. However, for queries that ask more general properties of the data, the impact of ε on identifying an individual is less clear. As shown in [49], for a given setting of ε, the confidence an adversary can have that a specific individual is in the database can change depending on the query, values in the data, and even on values not in the data. While ε-differential privacy does adjust for changes in values in both the data and values outside the dataset (for example, both are incorporated in the calculation of the query's global sensitivity for queries that are based on a numeric function of the data values), this is not a direct measure of what is revealed about an individual.

This may not be an insurmountable problem; a *differential identifiability* approach that has much in common with differential privacy is given in [50]. In differential identifiability, the adversary model is essentially the same as in differential privacy. The key distinction is that the parametrization of the noise to be added is based on the posterior confidence an adversary, knowing the value of ε, can have about the inclusion of any specific individual in the database. This mechanism allows calibrating the added noise to enforce identifiability requirements such as those derived from the HIPAA safe harbor rules. Having said that, there are limitations and assumptions in the adversary model, such as the assumption of a uniform prior adversary belief in the presence of individuals in the database, that demand further research.

6.4 Independence assumption

Many differential privacy mechanisms make some hidden assumptions that are not necessary in syntactic models. One such assumption is that individuals are independent. The problem becomes quite apparent with relational learning, where values of one individual can influence what is learned about another. When one individual can influence another, what does it mean to calculate the sensitivity, or impact that one individual may have on the query's result? Suppose, for example, that we want to predict election results in a differentially private manner. While removing one individual from the dataset would seem to change only one vote, the effect on the prediction made by a relational learner may be significantly larger, depending on the social role of that individual. In the real world, if a leader of an organization decides to change the candidate whom he supports, many members of the organization may consequently also change their vote. In the same sense, a relational learning function that predicts the outcome of a vote may have sensitivity much greater than one, as removing an influential individual would lead to changes in the predicted vote for neighboring individuals. Such dependencies between individuals can thus cause nearly unbounded (and very difficult to calculate) changes in a query outcome.

The dependence of the applicability of differential privacy on assumptions about the data was already shown by Kifer and Machanavajjhala [45]: they proved a no-free-lunch theorem by which differential privacy cannot provide privacy and utility without making assumptions about the data. They argue that the privacy of an individual is preserved when it is possible to limit the inference of an attacker about the *participation* of the individual in the data generating process. This is different from limiting the inference about the *presence of its tuple* in the database. For example, Bob's participation in a social network may cause edges to form between pairs of his friends, so that it affects more than just the tuple labeled as "Bob". The definition of evidence of participation, in turn, depends on how the data are generated; this is how assumptions enter the picture. Kifer and Machanavajjhala believe that under any reasonable formalization of evidence

of participation, such evidence can be encapsulated by exactly one tuple only when all tuples are independent, and that this independence assumption is a good rule of thumb when considering the applicability of differential privacy.

Achieving meaningful results from differential privacy may require assumptions on the model for data generation (e.g., all tuples are independent, though not necessarily generated from the same distribution) [45], new ways of defining what constitutes information about a single individual [76], or even entirely new privacy definitions [32, 46].

Syntactic models avoid this problem since in such models all individuals are anonymized, and the method of anonymization is independent of the social relations between the individuals.

Immunity to background knowledge One of the main claims of differential privacy is that it is immune to attacks based on the adversary's background knowledge. In some cases this claim is not as strong as it might appear. An example is given in [71]: given relative background knowledge (an individual earns \$5M more than the U.S. average), a differentially private query for the needed information (U.S. average income) can return quite accurate results – essentially violating the privacy of the rich individual. Hence, some background knowledge may allow an adversary to learn information on one individual from a differentially private answer that is computed from the values of other individuals.

7 Summary and conclusions

This study examined two types of privacy models: syntactic models of anonymity and differential privacy. Those two approaches are sometimes perceived as competing approaches, and that one can be used instead of the other. The first point that we made in this study is that the above conception is wrong. We explained that the syntactic models are designed for privacy-preserving data *publishing* (PPDP) while differential privacy is typically applicable for privacy-preserving data *mining* (PPDM). Hence, one approach cannot replace the other, and they both have a place alongside the other.

Next, we discussed criticisms of syntactic anonymization models (the deFinetti attack, minimality attacks, the curse of dimensionality, and composition attacks) and explained why none is a show stopper. Then, we proceeded to point out problems (or issues that need to be resolved) with the differential privacy approach. We explained the genuine difficulty in computing global sensitivity of queries, especially multidimensional ones; we discussed the problem with the utility of differentially private query answers due to the inherent uncertainty and the fact that the errors may be significant with high probability; we raised the question of how to set the differential privacy parameter ε and how to relate it to the probability of identification; and we highlighted some of the hidden assumptions that underlie differential privacy.

Our conclusion is that while differential privacy is a valuable weapon in the fight to both maintain privacy and foster use of data, it is not the universal answer. It provides a way to deal with a previously unanswered question in PPDM: how to ensure that the model developed does not inherently violate privacy of the individuals in the training data? While there are still issues related to both privacy and utility to be resolved, as pointed out in Section 6, the basic concept is a strong one.

At the same time, PPDP remains a pertinent and essential notion. While privacy advocates may not like it, societal practice (and laws such as HIPAA and those mandated by [23]) recognize that the right to privacy must be balanced against the public good. Syntactic models substantially

reduce privacy risk compared to a release of actual data values, and provide guarantees on the correctness (or range of correctness) of analysis of the anonymized data. In many cases, this is preferable to many noise addition techniques (particularly the Laplacian noise mechanism for differential privacy), as the latter still allow the possibility that the result obtained is very far from the true value, and thus extremely misleading.

It should be clarified that the two paradigms are not necessarily exclusive: recent work by Li, Qardaji, and Su suggests a link [54]. By first randomly selecting a subset of the data, and then applying k-anonymization, they show that the resulting syntactic anonymization can be made consistent with (ε, δ)-differential privacy. A key point is that the k-anonymization algorithm must introduce some random variability in the anonymization *process* (as recommended by [14], see Section 5.2). In particular, the generalization function must be developed using an ε-differentially private mechanism. They do require a slight relaxation of the background knowledge available to the attacker as discussed in Section 3 (see more details in [54]). A research challenge for PPDP is privacy definitions with adversary models that capture issues such as data correlation and inherently control potential real-world problems such as the deFinetti and minimality attacks.

A recent work by Cormode et al. [15], published after the completion of this work, provides empirical evidence of the need to pursue the study of both approaches. They advocate the notion of empirical privacy as a measurement tool; it represents the precision with which the sensitive values of individuals can be inferred from the released data. They also consider an empirical approach to measuring utility, based on a workload of queries that can essentially be used to describe the distribution of data and serve as the building blocks of more complex data analysis. They consider k-anonymity, ℓ-diversity, t-closeness and differential privacy. Although the theoretical guarantees of these models are different and cannot be directly compared, the notions of empirical privacy and empirical utility apply to all of them. Their findings are that the difference between the various models is less dramatic than previously thought. Differential privacy often provides the best empirical privacy for a fixed (empirical) utility level, but for more accurate answers it might be preferable to adopt a syntactic anonymity model like ℓ-diversity. Hence, the selection of a suitable privacy model should be made based on the use scenario.

In conclusion, in both paradigms, the issues raised should be viewed as opportunities for future research, rather than a call for abandoning one approach or the other. Advances in both paradigms are needed to ensure that the future provides reasonable protections on privacy as well as supporting legitimate learning from the ever-increasing data about us.

Acknowledgement. The authors thank Ehud Gudes and Christine Task for stimulating discussions on the topics of this study. This work was partially supported by MURI award FA9550-08-1-0265 from the Air Force Office of Scientific Research, and by the National Science Foundation under Grant No. 1012208.

References

[1] C. C. Aggarwal. On k-anonymity and the curse of dimensionality. In *VLDB*, pages 901–909, 2005.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D. Thomas, and A. Zhu. Achieving anonymity via clustering. In *ACM-SIGMOD Symposium on Principles of Database Systems (PODS)*, pages 153–162, 2006.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *ACM-SIGMOD Conference on Management of Data*, pages 439–450, May 2000.

[4] F. Bacchus, A. J. Grove, D. Koller, and J. Y. Halpern. From statistics to beliefs. In *AAAI*, pages 602–608, 1992.

[5] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *PODS*, pages 273–282, 2007.

[6] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. Discovering frequent patterns in sensitive data. In *KDD*, pages 503–512, 2010.

[7] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *STOC*, pages 609–618, 2008.

[8] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, pages 609–618, 2008.

[9] L. Burnett, K. Barlow-Stewart, A. Proos, and H. Aizenberg. The GeneTrustee: a universal identification system that ensures privacy and confidentiality for human genetic databases. *Journal of Law and Medicine*, 10(4):506–513, 2003.

[10] J. Cao and P. Karras. Publishing microdata with a robust privacy guarantee. *PVLDB*, 5:1388–1399, 2012.

[11] K. Chaudhuri and C. Monteleoni. Privacy-preserving logistic regression. In *NIPS*, pages 289–296, 2008.

[12] R. Chen, N. Mohammed, B. C. M. Fung, B. C. Desai, and L. Xiong. Publishing set-valued data via differential privacy. *PVLDB*, 4(11):1087–1098, 2011.

[13] G. Cormode. Personal privacy vs population privacy: learning to attack anonymization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1253–1261, 2011.

[14] G. Cormode, N. Li, T. Li, and D. Srivastava. Minimizing minimality and maximizing utility: Analyzing method-based attacks on anonymized data. *PVLDB*, 3:1045–1056, 2010.

[15] G. Cormode, C. M. Procopiuc, E. Shen, D. Srivastava, and T. Yu. Empirical privacy and empirical utility of anonymized data. In *ICDE Workshop on Privacy-Preserving Data Publication and Analysis (PRIVDB)*, 2013.

[16] B. Ding, M. Winslett, J. Han, and Z. Li. Differentially private data cubes: optimizing noise sources and consistency. In *SIGMOD Conference*, pages 217–228, 2011.

[17] D. Dobkin, A. K. Jones, and R. J. Lipton. Secure databases: Protection against user influence. *ACM Trans. Database Syst.*, 4(1):97–106, Mar. 1979.

[18] C. Dwork. Differential privacy. In *ICALP (2)*, pages 1–12, 2006.

[19] C. Dwork. Differential privacy: A survey of results. In *TAMC*, pages 1–19, 2008.

[20] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology (EUROCRYPT 2006)*, page 486503, Saint Petersburg, Russia, May 2006.

[21] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, pages 265–284, 2006.

[22] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. P. Vadhan. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 381–390, 2009.

[23] 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. *Official Journal of the European Communities*, No I.(281):31–50, Oct. 24 1995.

[24] K. E. Emam and F. Dankar. Protecting privacy using k-anonymity. *Journal of the American Medical Informatics Association*, 15:627–637, 2008.

[25] K. E. Emam, F. K. Dankar, R. Issa, E. Jonker, D. Amyot, E. Cogo, J.-P. Corriveau, M. Walker, S. Chowdhury, R. Vaillancourt, T. Roffey, and J. Bottomley. A globally optimal k-anonymity method for the de-identification of health information. *Journal of the American Medical Informatics Association*, 16:670–682, 2009.

[26] Standard for privacy of individually identifiable health information. *Federal Register*, Special Edition:768–769, Oct. 1 2007. 45 CFR 164.514(b)(2).

[27] Federal Committee on Statistical Methodology. Statistical policy working paper 22 (second version, 2005): Report on statistical disclosure limitation methodology. Technical report, Statistical and Science Policy, Office of Information and Regulatory Affairs, Office of Management and Budget, Dec. 2005. http://www.fcsm.gov/working-papers/spwp22.html.

[28] D. Feldman, A. Fiat, H. Kaplan, and K. Nissim. Private coresets. In *STOC*, pages 361–370, 2009.

[29] A. Friedman and A. Schuster. Data mining with differential privacy. In *KDD*, pages 493–502, 2010.

[30] B. C. M. Fung, K. Wang, and P. S. Yu. Anonymizing classification data for privacy preservation. *IEEE Trans. on Knowl. and Data Eng.*, 19(5):711–725, 2007.

[31] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, pages 265–273, 2008.

[32] J. Gehrke, E. Lui, and R. Pass. Towards privacy for social networks: A zero-knowledge based definition of privacy. In *Theory of Cryptography Conference*, pages 432–449, 2011.

[33] A. Gionis, A. Mazza, and T. Tassa. k-anonymization revisited. In *International Conference on Data Engineering (ICDE)*, pages 744–753, 2008.

[34] J. Goldberger and T. Tassa. Efficient anonymizations with enhanced utility. *Transactions on Data Privacy*, 3(2):149–175, 2010.

[35] M. Götz, A. Machanavajjhala, G. Wang, X. Xiao, and J. Gehrke. Publishing search logs - a comparative study of privacy guarantees. *IEEE Trans. Knowl. Data Eng.*, 24:520–532, 2012.

[36] M. Hardt, K. Ligett, and F. McSherry. A simple and practical algorithm for differentially private data release. *CoRR*, abs/1012.4763, 2010.

[37] M. Hay, V. Rastogi, G. Miklau, and D. Suciu. Boosting the accuracy of differentially private histograms through consistency. *PVLDB*, 3:1021–1032, 2010.

[38] A. Hundepool, J. Domingo-Ferrer, L. Franconi, S. Giessing, E. S. Nordholt, K. Spicer, and P.-P. de Wolf. *Statistical Disclosure Control*. Wiley, 2012.

[39] Interagency Confidentiality and Data Access Group: An Interest Group of the Federal Committee on Statistical Methodology. Checklist on disclosure potential of proposed data releases. Technical report, Statistical Policy Office, Office of Information and Regulatory Affairs, Office of Management and Budget, July 1999. http://www.fcsm.gov/committees/cdac/.

[40] V. Iyengar. Transforming data to satisfy privacy constraints. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 279–288, 2002.

[41] P. Jurczyk and L. Xiong. Distributed anonymizations: Achieving privacy for both data subjects and data providers. In *Data and Applications Security*, pages 191–207, 2009.

[42] H. Kargupta, S. Datta, Q. Wang, and K. Sivakumar. Random data perturbation techniques and privacy preserving data mining. *Knowledge and Information Systems*, 7(4):387–414, May 2005.

[43] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith. What can we learn privately? In *FOCS*, pages 531–540, 2008.

[44] D. Kifer. Attacks on privacy and definetti's theorem. In *SIGMOD Conference*, pages 127–138, 2009.

[45] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *SIGMOD*, pages 193–204, 2011.

[46] D. Kifer and A. Machanavajjhala. A rigorous and customizable framework for privacy. In *PODS*, pages 77–88, Scottsdale, Arizona, May 21-23 2012.

[47] A. Korolova, K. Kenthapadi, N. Mishra, and A. Ntoulas. Releasing search queries and clicks privately. In *WWW*, pages 171–180, 2009.

[48] M. Last, T. Tassa, and A. Zhmudyak. Improving accuracy of classification models induced from anonymized datasets. *To appear in Information Sciences*.

[49] J. Lee and C. Clifton. How much is enough? choosing ϵ for differential privacy. In *The 14th Information Security Conference (ISC 2011)*, pages 325–340, Xi'an, China, Oct. 26-29 2011.

[50] J. Lee and C. Clifton. Differential identifiability. In *The 19th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1041–1049, Beijing, China, Aug. 12-16 2012.

[51] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. Workload-aware anonymization techniques for large-scale datasets. *ACM Trans. Database Syst*, 33(3):1–47, 2008.

[52] C. Li, M. Hay, V. Rastogi, G. Miklau, and A. McGregor. Optimizing linear counting queries under differential privacy. In *PODS*, pages 123–134, 2010.

[53] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and ℓ-diversity. In *Proceedings of IEEE International Conference on Data Engineering (ICDE) 2007*, pages 106–115, 2007.

[54] N. Li, W. Qardaji, and D. Su. On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy. In *7th ACM Symposium on Information, Computer and Communications Security (ASIACCS'2012)*, pages 32–33, Seoul, Korea, May 2-4 2012.

[55] A. Machanavajjhala, D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber. Privacy: Theory meets practice on the map. In *ICDE*, pages 277–286, 2008.

[56] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. ℓ-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3, 2007.

[57] B. Malin. *Trail Re-Identification and Unlinkability in Distributed Databases*. PhD thesis, Carnegie Mellon University, May 2006.

[58] F. McSherry and I. Mironov. Differentially private recommender systems: Building privacy into the netflix prize contenders. In *KDD*, pages 627–636, 2009.

[59] F. McSherry and K. Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103, Providence, Rhode Island, Oct. 21-23 2007.

[60] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. kwong Lee. Anonymizing healthcare data: a case study on the blood transfusion service. In *KDD*, pages 1285–1294, 2009.

[61] R. A. Moore, Jr. Analysis of the kim-winkler algorithm for masking microdata files – how much masking is necessary and sufficient? conjectures for the development of a controllable algorithm. Statistical Research Division Report Series RR 96-05, U.S. Bureau of the Census, Washington, DC., 1996.

[62] R. A. Moore, Jr. Controlled data-swapping techniques for masking public use microdata sets. Statistical Research Division Report Series RR 96-04, U.S. Bureau of the Census, Washington, DC., 1996.

[63] A. Narayanan and V. Shmatikov. Myths and fallacies of personally identifiable information. *Comm. of the ACM*, 53(6):24–26, 2010.

[64] National Center for Biotechnology Information. dbGaP: Genotypes and Phenotypes. http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/about.html.

[65] M. E. Nergiz and C. Clifton. Thoughts on k-anonymization. *Data and Knowledge Engineering*, 63(3):622–645, Dec. 2007.

[66] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, pages 75–84, 2007.

[67] V. Rastogi and S. Nath. Differentially private aggregation of distributed time-series with transformation and encryption. In *SIGMOD Conference*, pages 735–746, 2010.

[68] A. Roth and T. Roughgarden. Interactive privacy via the median mechanism. In *STOC*, pages 765–774, 2010.

[69] P. Samarati. Protecting respondents' identities in microdata release. *IEEE Trans. on Knowl. and Data Eng.*, 13(6):1010–1027, 2001.

[70] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information (abstract). In *PODS '98: Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, page 188, 1998.

[71] R. Sarathy and K. Muralidhar. Evaluating laplace noise addition to satisfy differential privacy for numeric data. *Transactions on Data Privacy*, 4(1):1–17, 2011.

[72] E. Shmueli and T. Tassa. Privacy by diversity in sequential releases of databases. *Submitted*.

[73] E. Shmueli, T. Tassa, R. Wasserstein, B. Shapira, and L. Rokach. Limiting disclosure of sensitive data in sequential releases of databases. *Information Sciences*, 191:98–127, 2012.

[74] L. Sweeney. Uniqueness of simple demographics in the U.S. population. *Laboratory for international Data Privacy (LIDAP-WP4)*, 2000.

[75] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.

[76] C. Task and C. Clifton. A guide to differential privacy theory in social network analysis. In *The 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, Istanbul, Turkey, Aug. 26-29 2012.

[77] T. Tassa and E. Gudes. Secure distributed computation of anonymized views of shared databases. *Transactions on Database Systems*, 37(2):11, 2012.

[78] T. Tassa, A. Mazza, and A. Gionis. k-Concealment: An alternative model of k-type anonymity. *Transactions on Data Privacy*, 5(1):189–222, 2012.

[79] T. Truta, A. Campan, and P. Meyer. Generating microdata with p-sensitive k-anonymity property. In *SDM*, pages 124–141, 2007.

[80] K. Wang and B. Fung. Anonymizing sequential release. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*, pages 414–423, 2006.

[81] L. Wang, D. Wijesekera, and S. Jajodia. Cardinality-based inference control in data cubes. *Journal of Computer Security*, 12(5):655–692, 2005.

[82] L. Wasserman and S. Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, Mar. 2010.

[83] L. Willenborg and T. D. Waal. *Elements of Statistical Disclosure Control*, volume 155 of *Lecture Notes in Statistics*. Springer Verlag, New York, NY, 2001.

[84] R. Wong, J. Li, A. Fu, and K. Wang. (α, k)-anonymity: an enhanced k-anonymity model for privacy preserving data publishing. In *In ACM SIGKDD*, pages 754–759, 2006.

[85] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing. In *VLDB*, pages 543–554, 2007.

[86] W. K. Wong, N. Mamoulis, and D. W.-L. Cheung. Non-homogeneous generalization in privacy preserving data publishing. In *SIGMOD Conference*, pages 747–758, 2010.

[87] X. Xiao and Y. Tao. Anatomy: Simple and effective privacy preservation. In *Proceedings of 32nd International Conference on Very Large Data Bases (VLDB 2006)*, pages 139–150, Seoul, Korea, Sept. 12-15 2006.

[88] X. Xiao, G. Wang, and J. Gehrke. Differential privacy via wavelet transforms. In *ICDE*, pages 225–236, 2010.

[89] Y. Xiao, L. Xiong, and C. Yuan. Differentially private data release through multidimensional partitioning. In *Secure Data Management*, pages 150–168, 2010.

[90] S. Zhong, Z. Yang, and R. Wright. Privacy-enhancing k-anonymization of customer data. In *PODS*, pages 139–147, 2005.