An Ad Omnia Approach to Defining and Achieving Private Data Analysis

Cynthia Dwork

Microsoft Research
dwork@microsoft.com

Abstract. We briefly survey several privacy compromises in published datasets, some historical and some on paper. An inspection of these suggests that the problem lies with the nature of the privacy-motivated promises in question. These are typically syntactic, rather than semantic. They are also *ad hoc*, with insufficient argument that fulfilling these syntactic and *ad hoc* conditions yields anything like what most people would regard as privacy. We examine two comprehensive, or *ad omnia*, guarantees for privacy in statistical databases discussed in the literature, note that one is unachievable, and describe implementations of the other.

In this note we survey a body of work, developed over the past five years, addressing the problem known variously as statistical disclosure control, inference control, privacy-preserving datamining, and private data analysis. Our principal motivating scenario is a *statistical database*. A statistic is a quantity computed from a sample. Suppose a trusted and trustworthy curator gathers sensitive information from a large number of respondents (the sample), with the goal of learning (and releasing to the public) statistical facts about the underlying population. The problem is to release statistical information without compromising the privacy of the individual respondents. There are two settings: in the *non-interactive* setting the curator computes and publishes some statistics, and the data are not used further. Privacy concerns may affect the precise answers released by the curator, or even the set of statistics released. Note that since the data will never be used again the curator can destroy the data (and himself) once the statistics have been published.

In the *interactive* setting the curator sits between the users and the database. Queries posed by the users, and/or the responses to these queries, may be modified by the curator in order to protect the privacy of the respondents. The data cannot be destroyed, and the curator must remain present throughout the lifetime of the database.

There is a rich literature on this problem, principally from the satistics community [11, 15, 24, 25, 26, 34, 36, 23, 35] (see also the literature on controlled release of tabular data, contingency tables, and cell suppression), and from such diverse branches of computer science as algorithms, database theory, and cryptography [1, 10, 22, 28], [3, 4, 21, 29, 30, 37, 43], [7, 9, 12, 13, 14, 19, 8, 20]; see also the survey [2] for a summary of the field prior to 1989.

Clearly, if we are not interested in utility, then privacy can be trivially achieved: the curator can be silent, or can release only random noise. Throughout the discussion we will implicitly assume the statistical database has some non-trivial utility, and we will focus on the definition of privacy.

When defining privacy, or any other security goal, it is important to specify both what it means to compromise the goal and what power and other resources are available to the adversary. In the current context we refer to any information available to the adversary from sources other than the statistical database as *auxiliary information.* An attack that uses one database as auxiliary information to compromise privacy in a different database is frequently called a *linkage attack*. This type of attack is at the heart of the vast literature on hiding small cell counts in tabular data ("cell suppression").

1 Some Linkage Attacks

1.1 The Netflix Prize

Netflix recommends movies to its subscribers, and has offered a $1,000,000 prize for a 10% improvement in its recommendation system (we are not concerned here with how this is measured). To this end, Netflix has also published a training data set. According to the Netflix Prize rules webpage, "The training data set consists of more than 100 million ratings from over 480 thousand randomly-chosen, anonymous customers on nearly 18 thousand movie titles" and "The ratings are on a scale from 1 to 5 (integral) stars. *To protect customer privacy, all personal information identifying individual customers has been removed and all customer ids have been replaced by randomly-assigned ids.* The date of each rating and the title and year of release for each movie are provided" (emphasis added).

Netflix data are not the only movie ratings available on the web. There is also the International Movie Database (IMDb) site, where individuals may register for an account and rate movies. The users need not choose to be anonymous. Publicly visible material includes the user's movie ratings and comments, together with the dates of the ratings.

Narayanan and Shmatikov [32] cleverly used the IMDb in a linkage attack on the anonymization of the Netflix training data set. They found, "with 8 movie ratings (of which we allow 2 to be completely wrong) and dates that may have a 3-day error, 96% of Netflix subscribers whose records have been released can be uniquely identified in the dataset" and "for 89%, 2 ratings and dates are enough to reduce the set of plausible records to 8 out of almost 500,000, which can then be inspected by a human for further deanonymization." In other words, the removal of all "personal information" did not provide privacy to the users in the Netflix training data set. Indeed, Narayanan and Shmatikov were able to identify a particular user, about whom they drew several unsavory conclusions. Note that Narayanan and Shmatikov may have been correct in their conclusions or they may have been incorrect, but *either way this user is harmed*.

1.2 *k*-anonymization and Sequelae

The most famous linkage attack was obtained by Sweeney [40], who identified the medical records of the governor of Massachusetts by linking voter registration records to "anonymized" Massachusetts Group Insurance Commission (GIC) medical encounter data, which retained the birthdate, sex, and zip code of the patient. Sweeney proposed an antidote: k-anonymity [38, 39, 41, 42]. Roughly speaking, this is a syntactic condition requiring that every "quasi-identifier" (essentially, combination of non-sensitive attributes) must appear at least k times in the published database, if it occurs at all. This can be achieved by coarsening attribute categories, for example, replacing 5-digit zipcodes by their 3-digit prefixes. There are many problems with k-anonymity (computational complexity and the fact that the choice of category coarsenings may reveal information about the database, to name two), but the biggest problem is that it simply does not provide strong privacy; a lot of information my still be leaked about respondents/individuals in the database. Machanavajjhala, Gehrke, and Kifer [30] discuss this problem, and respond by proposing a new criterion for the published database: ℓ-diversity. However, Xiao and Tao [43] note that multiple ℓ-diverse data releases completely compromise privacy. They propose a different syntactic condition: m-invariance.

The literature does not contain any direct attack on m-invariance (although, see Section 2.1 for general difficulties). However it is clear that something is going wrong: the "privacy" promises are syntactic conditions on the released datasets, but there is insufficient argument that the syntactic conditions have the correct semantic implications.

1.3 Anonymization of Social Networks

In a social network graph, nodes correspond to users (or e-mail accounts, or telephone numbers, etc), and edges have various social semantics (friendship, frequent communications, phone conversations, and so on). Companies that hold such graphs are frequently asked to release an anonymized version, in which node names are replaced by random strings, for study by social scientists. The intuition is that the anonymized graph reveals only the structure, not the potentially sensitive information of who is socially connected to whom. In [5] it is shown that anonymization does not protect this information at all; indeed it is vulnerable both to active and passive attacks. Again, anonymization is just an *ad hoc* syntactic condition, and has no privacy semantics.

2 On Defining Privacy for Statistical Databases

One source of difficulty in defining privacy for statistical databases is that the line between "inside" and "outside" is slightly blurred. In contrast, when Alice and her geographically remote colleague Bob converse, Alice and Bob are the "insiders," everyone else is an "outsider," and privacy can be obtained by any cryptosystem that is semantically secure against a passive eavesdropper.

Let us review this notion. Informally, semantic security says that the ciphertext (encryption of the message to be transmitted) reveals no information about the plaintext (the message). This was formalized by Goldwasser and Micali [27] along the following lines. The ability of the adversary, having access to both the ciphertext and any auxiliary information, to learn (anything about) the plaintext is compared to the ability of a party having access *only* to the auxiliary information (and not the ciphertext), to learn anything about the plaintext[1]. Clearly, if this difference is very, very tiny, then in a rigorous sense the ciphertext leaks (almost) no information about the plaintext.

The formalization of semantic security along these lines is one of the pillars of modern cryptography. It is therefore natural to ask whether a similar property can be achieved for statistical databases. However, unlike the eavesdropper on a conversation, the statistical database attacker is also a user, that is, a legitimate consumer of the information provided by the statistical database, so this attacker is both a little bit of an insider (not to mention that she may also be a respondent in the database), as well as an outsider, to whom certain fine-grained information should not be leaked.

2.1 Semantic Security for Statistical Databases?

In 1977 Tor Dalenius articulated an *ad omnia* privacy goal for statistical databases: anything that can be learned about a respondent from the statistical database should be learnable without access to the database. Happily, this formalizes to semantic security (although Dalenius' goal predated the Goldwasser and Micali definition by five years). Unhappily, however, it cannot be achieved, both for small and big reasons. It is instructive to examine these in depth.

Many papers in the literature attempt to formalize Dalenius' goal (in some cases unknowingly) by requiring that the adversary's prior and posterior views about an individual (*i.e.*, before and after having access to the statistical database) shouldn't be "too different," or that access to the statistical database shouldn't change the adversary's views about any individual "too much." Of course, this is clearly silly, if the statistical database teaches us anything at all. For example, suppose the adversary's (incorrect) prior view is that everyone has 2 left feet. Access to the statistical database teaches that almost everyone has one left foot and one right foot. The adversary now has a very different view of whether or not any given respondent has two left feet. Even when used correctly, in a way that is decidedly not silly, this prior/posterior approach suffers from definitional awkwardness [21, 19, 8].

At a more fundamental level, a simple hybrid argument shows that it is impossible to achieve cryptographically small levels of "tiny" difference between an adversary's ability to learn something about a respondent given access to the database, and the ability of someone without access to the database to learn something about a respondent. Intuitively, this is because the user/adversary is

[1] The definition in [27] deals with probabilistic polynomial time bounded parties. This is not central to the current work so we do not emphasize it in the discussion.

supposed to learn unpredictable and non-trivial facts about the data set (this is where we assume some degree of utility of the database), which translates to learning more than cryptographically tiny amounts about an individual. However, it may make sense to relax the definition of "tiny." Unfortunately, even this relaxed notion of semantic security for statistical databases cannot be achieved.

The final nail in the coffin of hope for Dalenius' goal is a formalization of the following difficulty. Suppose we have a statistical database that teaches average heights of population subgroups, and suppose further that it is infeasible to learn this information (perhaps for financial reasons) any other way (say, by conducting a new study). Consider the auxiliary information "Terry Gross is two inches shorter than the average Lithuanian woman." Access to the statistical database teaches Terry Gross' height. In contrast, someone without access to the database, knowing only the auxiliary information, learns much less about Terry Gross' height.

A rigorous impossibility result generalizes and formalizes this argument, extending to essentially any notion of privacy compromise. The heart of the attack uses extracted randomness from the statistical database as a one-time pad for conveying the privacy compromise to the adversary/user [16].

This brings us to an important observation: Terry Gross did not have to be a member of the database for the attack described above to be prosecuted against her. This suggests a new notion of privacy: minimize the increased risk to an individual incurred by joining (or leaving) the database. That is, we move from comparing an adversary's prior and posterior views of an individual to comparing the risk to an individual when included in, versus when not included in, the database. This new notion is called *differential privacy*.

Remark 1. It might be remarked that the counterexample of Terry Gross' height is contrived, and so it is not clear what it, or the general impossibility result in [16], mean. Of course, it is conceivable that counterexamples exist that would not appear contrived. More significantly, the result tells us that it is impossible to construct a privacy mechanism that both preserves utility and provably satisfies at least one natural formalization of Dalenius' goal. But proofs are important: they let us know exactly what guarantees are made, and they can be verified by non-experts. For these reasons it is extremely important to find *ad omnia* privacy goals and implementations that provably ensure satisfaction of these goals.

2.2 Differential Privacy

In the sequel, the randomized function $\mathcal{K}$ is the algorithm applied by the curator when releasing information. So the input is the data set, and the output is the released information, or *transcript*. We do not need to distinguish between the interactive and non-interactive settings.

Think of a database as a set of rows. We say databases D_1 and D_2 *differ in at most one element* if one is a subset of the other and the larger database contains just one additional row.

Definition 1. *A randomized function* $\mathcal{K}$ *gives* ϵ-differential privacy *if for all data sets* D_1 *and* D_2 *differing on at most one element, and all* $S \subseteq Range(\mathcal{K})$*,*

$$\Pr[\mathcal{K}(D_1) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D_2) \in S], \tag{1}$$

where the probability space in each case is over the coin flips of the mechanism $\mathcal{K}$*.*

A mechanism $\mathcal{K}$ satisfying this definition addresses all concerns that any participant might have about the leakage of her personal information: even if the participant removed her data from the data set, no outputs (and thus consequences of outputs) would become significantly more or less likely. For example, if the database were to be consulted by an insurance provider before deciding whether or not to insure a given individual, then the presence or absence of that individual's data in the database will not significantly affect her chance of receiving coverage.

Differential privacy is therefore an *ad omnia* guarantee. It is also a very strong guarantee, since it is a statistical property about the behavior of the mechanism and therefore is independent of the computational power and auxiliary information available to the adversary/user.

Differential privacy is not an absolute guarantee of privacy. As we have seen, any statistical database with any non-trivial utility can compromise privacy. However, in a society that has decided that the benefits of certain databases outweigh the costs, differential privacy ensures that only a limited amount of additional risk is incurred by participating in the (socially beneficial) databases.

Remark 2.
1. The parameter ϵ is public. The choice of ϵ is essentially a social question and is beyond the scope of this paper. That said, we tend to think of ϵ as, say, 0.01, 0.1, or in some cases, $\ln 2$ or $\ln 3$. If the probability that some bad event will occur is very small, it might be tolerable to increase it by such factors as 2 or 3, while if the probability is already felt to be close to unacceptable, then an increase of $e^{0.01} \approx 1.01$ might be tolerable, while an increase of e, or even only $e^{0.1}$, would be intolerable.
2. Definition 1 extends to group privacy as well (and to the case in which an individual contributes more than a single row to the database). A collection of c participants might be concerned that their collective data might leak information, even when a single participant's does not. Using this definition, we can bound the dilation of any probability by at most $\exp(\epsilon c)$, which may be tolerable for small c. Of course, the point of the statistical database is to disclose aggregate information about large groups (while simultaneously protecting individuals), so we should expect privacy bounds to disintegrate with increasing group size.

3 Achieving Differential Privacy in Statistical Databases

We now describe an interactive mechanism, $\mathcal{K}$, due to Dwork, McSherry, Nissim, and Smith [20]. A *query* is a function mapping databases to (vectors of) real numbers. For example, the query "Count P" counts the number of rows in the database having property P.

When the query is a function f, and the database is X, the *true answer* is the value $f(X)$. The $\mathcal{K}$ mechanism adds appropriately chosen random noise to the true answer to produce what we call the *response*. The idea of preserving privacy by responding with a noisy version of the true answer is not new, but this approach is delicate. For example, if the noise is symmetric about the origin and the same question is asked many times, the responses may be averaged, cancelling out the noise[2]. We must take such factors into account.

Definition 2. *For $f : \mathcal{D} \to \mathbb{R}^d$, the* sensitivity *of f is*

$$\Delta f = \max_{D_1, D_2} \|f(D_1) - f(D_2)\|_1 \tag{2}$$

for all D_1, D_2 differing in at most one element.

In particular, when $d = 1$ the sensitivity of f is the maximum difference in the values that the function f may take on a pair of databases that differ in only one element. For now, let us focus on the case $d = 1$.

For many types of queries Δf will be quite small. In particular, the simple counting queries discussed above ("How many rows have property P?") have $\Delta f = 1$. Our techniques work best – ie, introduce the least noise – when Δf is small. Note that sensitivity is a property of the function alone, and is independent of the database. The sensitivity essentially captures how great a difference (between the value of f on two databases differing in a single element) must be hidden by the additive noise generated by the curator.

On query function f the privacy mechanism $\mathcal{K}$ computes $f(X)$ and adds noise with a scaled symmetric exponential distribution with standard deviation $\sqrt{2}\Delta f/\epsilon$. In this distribution, denoted $\mathrm{Lap}(\Delta f/\epsilon)$, the mass at x is proportional to $\exp(-|x|(\epsilon/\Delta f))$.[3] Decreasing ϵ, a publicly known parameter, flattens out this curve, yielding larger expected noise magnitude. When ϵ is fixed, functions f with high sensitivity yield flatter curves, again yielding higher expected noise magnitudes.

The proof that $\mathcal{K}$ yields ϵ-differential privacy on the single query function f is straightforward. Consider any subset $S \subseteq Range(\mathcal{K})$, and let D_1, D_2 be any pair of databases differing in at most one element. When the database is D_1, the probability mass at any $r \in Range(\mathcal{K})$ is proportional to $\exp(-|f(D_1) - r|(\epsilon/\Delta f))$, and similarly when the database is D_2. Applying the triangle inequality in the

[2] We do not recommend having the curator record queries and their responses so that if a query is issued more than once the response can be replayed. One reason is that if the query language is sufficiently rich, then semantic equivalence of two syntactically different queries is undecidable. Even if the query language is not so rich, the devastating attacks demonstrated by Dinur and Nissim [14] pose completely random and unrelated queries.

[3] The probability density function of $\mathrm{Lap}(b)$ is $p(x|b) = \frac{1}{2b}\exp(-\frac{|x|}{b})$, and the variance is $2b^2$.

exponent we get a ratio of at most $\exp(-|f(D_1) - f(D_2)|(\epsilon/\Delta f))$. By definition of sensitivity, $|f(D_1) - f(D_2)| \leq \Delta f$, and so the ratio is bounded by $\exp(-\epsilon)$, yielding ϵ-differential privacy.

It is easy to see that for any (adaptively chosen) query sequence $f_1, \ldots, f_d$, ϵ-differential privacy can be achieved by running $\mathcal{K}$ with noise distribution $\mathrm{Lap}(\sum_i \Delta f_i/\epsilon)$ on *each* query. In other words, the quality of each answer deteriorates with the sum of the sensitivities of the queries. Interestingly, it is sometimes possible to do better than this. Roughly speaking, what matters is the maximum possible value of $\Delta = ||(f_1(D_1), f_2(D_1), \ldots, f_d(D_1)) - (f_1(D_2), f_2(D_2), \ldots, f_d(D_2))||_1$. The precise formulation of the statement requires some care, due to the potentially adaptive choice of queries. For a full treatment see [20]. We state the theorem here for the non-adaptive case, viewing the (fixed) sequence of queries $f_1, f_2, \ldots, f_d$ as a single d-ary query f and recalling Definition 2 for the case of arbitrary d.

Theorem 1. *For $f : \mathcal{D} \to \mathrm{I\!R}^d$, the mechanism $\mathcal{K}_f$ that adds independently generated noise with distribution* $\mathrm{Lap}(\Delta f/\epsilon)$ *to each of the d output terms enjoys ϵ-differential privacy.*

Among the many applications of Theorem 1, of particular interest is the class of *histogram* queries. A histogram query is an arbitrary partitioning of the domain of database rows into disjoint "cells," and the true answer is the set of counts describing, for each cell, the number of database rows in this cell. Although a histogram query with d cells may be viewed as d individual counting queries, the addition or removal of a single database row can affect the entire d-tuple of counts in at most one location (the count corresponding to the cell to (from) which the row is added (deleted); moreover, the count of this cell is affected by at most 1, so by Definition 2, every histogram query has sensitivity 1.

4 Utility of $\mathcal{K}$ and some Limitations

The mechanism $\mathcal{K}$ described above has excellent accuracy for insensitive queries. In particular, the noise needed to ensure differential privacy depends only on the sensitivity of the function and on the parameter ϵ. Both are independent of the database and the number of rows it contains. Thus, if the database is very large, the errors for many questions introduced by the differential privacy mechanism is relatively quite small.

We can think of $\mathcal{K}$ as a differential privacy-preserving interface between the analyst and the data. This suggests a line of research: finding algorithms that require few, insensentitive, queries for standard datamining tasks. As an example, see [8], which shows how to compute singular value decompositions, find the ID3 decision tree, carry out k-means clusterings, learn association rules, and learn anything learnable in the statistical queries learning model using only a relatively small number of counting queries. See also the more recent work on contingency tables (and OLAP cubes) [6].

It is also possible to combine techniques of secure function evaluation with the techniques described above, permitting a collection of data holders to cooperatively simulate $\mathcal{K}$; see [17] for details.

Recent Extensions Sensitivity of a function f is a *global* notion: the worst case, over *all* pairs of databases differing in a single element, of the change in the value of f. Even for a function with high sensitivity, it may be the case that "frequently" – that is, for "many" databases or "much" of the time – the function is locally insensitive. That is, much of the time, adding or deleting a single database row may have little effect on the value of the function, even if the worst case difference is large.

Given any database D, we would like to generate noise according to the local sensitivity of f at D. Local sensitivity is itself a legitimate query ("What is the local sensitivity of the database with respect to the function f?"). If, for a fixed f, the local sensitivity varies wildly with the database, then to ensure differential privacy the local sensitivity must not be revealed too precisely. On the other hand, if the curator simply adds noise to $f(D)$ according to the local sensitivity of f at D, then a user may ask the query f several times in an attempt to guage the local sensitivity, which we have just argued cannot necessarily be safely learned with great accuracy. To prevent this, we need a way of *smoothing* the change in magnitude of noise used so that on locally insensitive instances that are sufficiently far from highly sensitive ones the noise is small. This is the subject of recent work of Nissim, Raskhodnikova, and Smith [33].

In some tasks, the addition of noise makes no sense. For example, the function f might map databases to strings, strategies, or trees. McSherry and Talwar address the problem of optimizing the output of such a function while preserving ϵ-differential privacy [31]. Assume the curator holds a database X and the goal is to produce an object y. In a nutshell, their *exponential mechanism* works as follows. There is assumed to be a *utility function u(X,y)* that measures the quality of an output y, given that the database is X. For example, if the database holds the valuations that individuals assign a digital good during an auction, $u(X, y)$ might be the revenue, with these valuations, when the price is set to y. The McSherry-Talwar mechanism outputs y with probability proportional to $\exp(u(X, y)\epsilon)$ and ensures ϵ-differential privacy. Capturing the intuition, first suggested by Jason Hartline, that privacy seems to correspond to truthfulness, the McSherry and Talwar mechanism yields approximately-truthful auctions with nearly optimal selling price. Roughly speaking, this says that a participant cannot dramatically reduce the price he pays by lying about his valuation. Interestingly, McSherry and Talwar note that one can use the simple composition of differential privacy, much as was indicated in Remark 2 above for obtaining privacy for groups of c individuals, to obtain auctions in which no cooperating group of c agents can significantly increase their utility by submitting bids other than their true valuations.

Limitations As we have seen, the magnitude of the noise generated by $\mathcal{K}$ increases with the number of questions. A line of research initiated by Dinur and Nissim indicates that this increase is inherent [14]. They showed that if the database is a vector x of n bits and the curator provides relatively accurate (within $o(\sqrt{n})$) answers to $n \log^2 n$ random subset sum queries, then by using linear programming the adversary can reconstruct a database x' agreeing with x in all but $o(n)$ entries, ie, satisfying $support(x - x') \in o(n)$. We call this *blatant non-privacy*. This result was later strengthened by Yekhanin, who showed that if the attacker asks the n Fourier queries (with entries ± 1; the true answer to query vector y is the inner product $\langle x, y \rangle$) and the noise is always $o(\sqrt{n})$, then the system is blatantly non-private [44].

Additional strengthenings of these results were obtained by Dwork, Mscherry, and Talwar [18]. They considered the case in which the curator can sometimes answer completely arbitrarily. When the queries are vectors of standard normals and again the true answer is the inner product of the database and the query vector, they found a sharp threshold $\rho^* \approx 0.239$ so that if the curator replies completely arbitrarily on a $\rho < \rho*$ fraction of the queries, but is confined to $o(\sqrt{n})$ error on the remaining queries, then again the system is blatantly non-private even against only $O(n)$ queries. Similar, but slightly less strong results are obtained for ± 1 query vectors.

These are not just interesting mathematical exercises. While at first blush simplistic, the Dinur-Nissim setting is in fact sufficiently rich to capture many natural questions. For example, the rows of the database may be quite complex, but the adversary/user may know enough information about an individual in the database to uniquely identify his row. In this case the goal is to prevent any single *additional* bit of information to be learned from the database. (In fact, careful use of hash functions can handle the "row-naming problem" even if the adversary does not know enough to uniquely identify individuals at the time of the attack, possibly at the cost of a modest increase in the number of queries.) Thus we can imagine a scenario in which an adversary reconstructs a close approximation to the database, where each row is identified with a set of hash values, and a "secret bit" is learned for many rows. At a later time the adversary may learn enough about an individual in the database to deduce sufficiently many of the hash values of her record to identify the row corresponding to the individual, and so obtain her "secret bit." Thus, naming a set of rows to specify a query is not just a theoretical possibility, and the assumption of only a single sensitive attribute per user still yields meaningful results.

Research statisticians like to "look at the data." Indeed, conversations with experts in this field frequently involve pleas for a "noisy table" that will permit highly accurate answers to be derived for computations that are not specified at the outset. For these people the implications of the Dinur-Nissim results are particularly significant: no "noisy table" can provide very accurate answers to too many questions; otherwise the table could be used to simulate the interactive mechanism, and a Dinur-Nissim style attack could be mounted against the table. Even worse, while in the interactive setting the noise can be adapted to the queries, in the non-interactive setting the curator does not have this freedom to aid in protecting privacy.

5 Conclusions and Open Questions

We have surveyed a body of work addressing the problem known variously as statistical disclosure control, privacy-preserving datamining, and private data analysis. The concept of ϵ-differential privacy was motivated and defined, and a specific technique for achieving ϵ-differential privacy was described. This last involves calibrating the noise added to the true answers according to the sensitivity of the query sequence and to a publicly chosen parameter ϵ.

Of course, statistical databases are a very small part of the overall problem of defining and ensuring privacy. How can we sensibly address privacy in settings in which the boundary between "inside" and "outside" is completely porous, for example, in outsourcing of confidential data for processing, bug reporting, and managing cookies? What is the right notion of privacy in a social network (and what are the questions of interest in the study of such networks)?

We believe the notion of differential privacy may be helpful in approaching these problems.

References

[1] J. O. Achugbue and F. Y. Chin, The Effectiveness of Output Modification by Rounding for Protection of Statistical Databases, *INFOR 17*(3), pp. 209–218, 1979.

[2] N. R. Adam and J. C. Wortmann, Security-Control Methods for Statistical Databases: A Comparative Study, *ACM Computing Surveys 21*(4), pp. 515–556, 1989.

[3] D. Agrawal and C. C. Aggarwal, On the design and Quantification of Privacy Preserving Data Mining Algorithms, *Proceedings of the 20th Symposium on Principles of Database Systems*, pp. 247-255, 2001.

[4] R. Agrawal and R. Srikant, Privacy-Preserving Data Mining, *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 439–450, 2000.

[5] L. Backstrom, C. Dwork, J. Kleinberg, Wherefore Art Thou r3579x?: Anonymized Social Networks, Hidden Patterns, and Structural Steganography, *Proceedings of the 16th International World Wide Web Conference*, pp. 181–190, 2007.

[6] B. Barak, K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, K. Talwar, Privacy, Accuracy, and Consistency Too: A Holistic Solution to Contingency Table Release, *Proceedings of the 26th Symposium on Principles of Database Systems*, pp. 273–282, 2007.

[7] L. L. Beck, A Security Mechanism for Statistical Databases, *ACM TODS 5*(3), pp. 316–338, 1980.

[8] A. Blum, C. Dwork, F. McSherry, and K. Nissim, Practical Privacy: The SuLQ framework. *Proceedings of the 24th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, June 2005.

[9] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, Toward Privacy in Public Databases, *Proceedings of the 2nd Theory of Cryptography Conference*, 2005.

[10] F. Y. Chin and G. Ozsoyoglu, *Auditing and infrence control in statistical databases*, IEEE Trans. Softw. Eng., SE-8(6):113–139, April 1982.
[11] T. Dalenius, Towards a Methodology for Statistical Disclosure Control. *Statistik Tidskrift 15*, pp. 429–222, 1977.
[12] D. E. Denning, Secure Statistical Databases with Random Sample Queries, *ACM Transactions on Database Systems 5*(3), pp. 291–315, 1980.
[13] D. Denning, P. Denning, and M. Schwartz, The Tracker: A Threat to Statistical Database Security, *ACM Transactions on Database Systems 4*(1), pp. 76–96, 1979.
[14] I. Dinur and K. Nissim, Revealing Information While Preserving Privacy, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 202-210, 2003.
[15] G. Duncan, Confidentiality and statistical disclosure limitation. In N. Smelser & P. Baltes (Eds.), International Encyclopedia of the Social and Behavioral Sciences. New York: Elsevier. 2001
[16] C. Dwork, Differential Privacy, *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)(2)*, pp. 1–12, 2006.
[17] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, Our Data, Ourselves: Privacy Via Distributed Noise Generation, *Proceedings of EUROCRYPT 2006*, pp. 486–503, 2006.
[18] C. Dwork, F. McSherry, and K. Talwar, The Price of Privacy and the Limits of LP Decoding, *Proceedings of the 39th ACM Symposium on Theory of Computing*, pp. 85–94, 2007.
[19] C. Dwork and K. Nissim, Privacy-Preserving Datamining on Vertically Partitioned Databases, *Proceedings of CRYPTO 2004* , vol. 3152 of LNCS volume 3152, pp. 528]-544, 2004.
[20] C. Dwork, F. McSherry, K. Nissim, and A. Smith, Calibrating Noise to Sensitivity in Private Data Analysis, *Proceedings of the 3rd Theory of Cryptography Conference*, pp. 265–284, 2006.
[21] A. V. Evfimievski, J. Gehrke and R. Srikant, Limiting Privacy Breaches in Privacy Preserving Data Mining, *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 211-222, 2003.
[22] D. Dobkin, A. Jones and R. Lipton, Secure Databases: Protection Against User Influence, *ACM TODS 4*(1), pp. 97–106, 1979.
[23] I. Fellegi, On the question of statistical confidentiality, *Journal of the American Statistical Association 67*, pp. 7-18, 1972.
[24] S. Fienberg, Confidentiality and Data Protection Through Disclosure Limitation: Evolving Principles and Technical Advances, IAOS Conference on Statistics, Development and Human Rights September, 2000, available at `http://www.statistik.admin.ch/about/international/fienberg_final_paper.doc`
[25] S. Fienberg, U. Makov, and R. Steele, Disclosure Limitation and Related Methods for Categorical Data, *Journal of Official Statistics*, *14*, pp. 485–502, 1998.
[26] L. Franconi and G. Merola, Implementing Statistical Disclosure Control for Aggregated Data Released Via Remote Access, Working Paper No. 30, United Nations Statistical Commission and European Commission, joint ECE/EUROSTAT work session on statistical data confidentiality, April, 2003, available at `http://www.unece.org/stats/documents/2003/04/confidentiality/wp.30.e.pdf`
[27] S. Goldwasser and S. Micali, Probabilistic Encryption, *J. Comput. Syst. Sci. 28*(2), pp. 270–299, 1984
[28] D. Gusfield, A Graph Theoretic Approach to Statistical Data Security, *SIAM J. Comput. 17*(3), pp. 552–571, 1988.

[29] E. Lefons, A. Silvestri, and F. Tangorra, *An analytic approach to statistical databases*, In 9th Int. Conf. Very Large Data Bases, pages 260– 274. Morgan Kaufmann, Oct-Nov 1983.

[30] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam: l-Diversity: Privacy Beyond k-Anonymity, Proceedings of the 22nd International Conference on Data Engineering (ICDE'06), page 24, 2006.

[31] F. McSherry and K. Talwar, Mechanism Design via Differential Privacy, *Proceedings of the 48th Annual Symposium on Foundations of Computer Science*, 2007.

[32] A. Narayanan, V. Shmatikov, How to Break Anonymity of the Netflix Prize Dataset, available at `http://www.cs.utexas.edu/~shmat/shmat_netflix-prelim.pdf`

[33] K. Nissim, S. Raskhodnikova, and A. Smith, Smooth Sensitivity and Sampling in Private Data Analysis, *Proceedings of the 39th ACM Symposium on Theory of Computing*, pp. 75–84, 2007.

[34] T.E. Raghunathan, J.P. Reiter, and D.B. Rubin, Multiple Imputation for Statistical Disclosure Limitation, *Journal of Official Statistics 19*(1), pp. 1 – 16, 2003.

[35] S. Reiss, Practical Data Swapping: The First Steps, *ACM Transactions on Database Systems 9*(1), pp. 20–37, 1984.

[36] D.B. Rubin, Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics 9*(2), pp. 461 – 469, 1993.

[37] A. Shoshani, Statistical databases: Characteristics, problems and some solutions, *Proceedings of the 8th International Conference on Very Large Data Bases (VLDB'82)*, pages 208–222, 1982.

[38] P. Samarati and L. Sweeney, Protecting Privacy when Disclosing Information: k-Anonymity and its Enforcement Through Generalization and Specialization, Technical Report SRI-CSL-98-04, SRI Intl., 1998.

[39] P. Samarati and L. Sweeney, Generalizing Data to Provide Anonymity when Disclosing Information (Abstract), *Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, p. 188, 1998.

[40] L. Sweeney, Weaving Technology and Policy Together to Maintain Confidentiality, *J Law Med Ethics, 1997. 25*(2-3), pp. 98-110, 1997.

[41] L. Sweeney, k-anonymity: A Model for Protecting Privacy, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), pp. 557–570, 2002.

[42] L. Sweeney, Achieving k-Anonymity Privacy Protection Using Generalization and Suppression, *International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10*(5), pp. 571–588, 2002.

[43] X. Xiao ad Y. Tao, *M-invariance: towards privacy preserving re-publication of dynamic datasets*, In SIGMOD 2007, pp. 689-700.

[44] S. Yekhanin, *private communication*, 2006.