On Privacy and Accuracy in Data Releases

Mário S. Alvim
Computer Science Department, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brasil

Natasha Fernandes
Department of Computing, Macquarie University, Sydney, Australia

Annabelle McIver
Department of Computing, Macquarie University, Sydney, Australia

Gabriel H. Nunes
Computer Science Department, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brasil

Abstract

In this paper we study the relationship between privacy and accuracy in the context of correlated datasets. We use a model of quantitative information flow to describe the the trade-off between privacy of individuals' data and and the utility of queries to that data by modelling the effectiveness of adversaries attempting to make inferences after a data release.

We show that, where correlations exist in datasets, it is not possible to implement optimal noise-adding mechanisms that give the best possible accuracy or the best possible privacy in all situations. Finally we illustrate the trade-off between accuracy and privacy for local and oblivious differentially private mechanisms in terms of inference attacks on medium-scale datasets.

2012 ACM Subject Classification Security and privacy

Keywords and phrases Privacy/utility trade-off, Quantitative Information Flow, inference attacks

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2020.1

Category Invited Paper

Funding *Mário S. Alvim*: Supported by CNPq, CAPES, and FAPEMIG.
Gabriel H. Nunes: Supported by CNPq, CAPES, and FAPEMIG.

1 Introduction

Consider the following stories about actual and potential privacy breaches.

(i) *In 2009 a lawsuit (Doe v. Netflix)* was filed against Netflix alleging that it had violated fair-trade laws and a federal privacy law. The complainants argued that Netflix's "anonymisation" still allowed individuals to be identified (and indeed plenty were) by combining movie preferences with other generally available data. And that Netflix should have known about these risks.
[Source: `https://www.wired.com/2009/12/netflix-privacy-lawsuit/`]

(ii) *Also in 2009, researchers in Natural Language Processing* showed that using powerful machine learning algorithms, authors of anonymously-penned documents could be identified with a high degree of accuracy.
[Source: On the Feasibility of Internet-Scale Author Identification [15]]

(iii) Finally *Privacy researchers at the University of Melbourne* noted that in any "anonymised datasets" records that can be characterised uniquely put the individuals who contributed thoese records at risk. "While uniqueness does not imply re-identification,

particular data that is known to be held by certain parties, does imply the opportunity for re-identification". [Source: `https://pursuit.unimelb.edu.au/articles/the-simple-process-of-re-identifying-patients-in-public-health-records`]

These three real examples share a common theme: the breaches, or the potential for breaches, are all related to "unintended inferences," meaning that the data published was deemed to be innocuous but turned out to have the potential for inferring information that individuals could claim as infringements to their privacy.

Modern approaches to privacy recognise that inferences are a problem and, whilst they likely cannot be eliminated, can be mitigated by providing "plausible deniability" for individuals. Differential privacy in particular does not rule out inferences, but rather provides guarantees couched in terms of *indistinguishability* – an individual can plausibly deny that their personal information contributed to some precise value v because the query to the data that is finally announced (so the argument goes) could have been produced just as plausibly with some other non-related value v'. This feature is formalised by equations of the form:

$$M(v)(Z) \leq M(v')(Z) \times e^{\epsilon d(v,v')} \ , \tag{1}$$

where M is a mechanism that delivers a noisy output, and v, v' are possible instances of input, one that could be related to an individual, and one that does not. The term $M(v)(Z)$ means "the probability that the output of M evaluated at input v is contained in the subset Z". Here $d(v, v')$ is a distance measure defined on inputs and $\epsilon \geq 0$ is a security parameter often described as a "privacy budget".

This powerful protection rests crucially on the particulars of the adversarial setting. However the practical application of differential privacy in machine learning settings (for example) has been shown to be problematic, with the ϵ parameter typically tuned to optimise accuracy (of some utility measure) without a good understanding of the ramifications for privacy [8]. Moreover the scope and variations of differential privacy mechanisms have been tabulated in other work [5] and what emerges is that there is often little clarification around the basic questions: how does some particular version of differential privacy (and its choice of ϵ) affect *both* the potential for unintended inferences to be made of the individuals contributing their data *and* the accuracy of the "useful" data that is released?

In this paper we aim to study these basic questions from the point of view of inferences. We use a model of *quantitative information flow* to analyse the extent to which inferences succeed once data has been released, and to estimate the accuracy of "useful" queries. From this viewpoint we are able to gauge direct risks to individuals when compared to the benefits of those wishing to use the data, as well as some understanding about the setting of the parameter ϵ in terms of the trade-off between privacy risks and accuracy of data releases.

Significant in this analysis is a contribution to the "no free lunch" style theorems of Kifer et al. [11]. We demonstrate that it is impossible to design differentially-private mechanisms that offer minimal risk benefits to all individuals in all situations at the same time as preserving a fixed accuracy threshold. And dually it is impossible to design differentially-private mechanisms that offer optimal accuracy in all scenarios whilst guaranteeing a fixed threshold for plausible deniability for all individuals.

The implication of these impossibility results is that the manner in which noisy outputs are created is crucial, and by making reasonable assumptions about adversaries, as argued by Kifer at al. [11], a better understanding of the relationship between inferences about privacy and accuracy can be obtained. Our final contribution is to compare experimentally the different risks to privacy versus accuracy in a large to medium-sized datasets when differential privacy is used as the noise-adding mechanism.

2 Informal example: does it matter how to randomise?

In this section we illustrate, using a simple scenario, the challenges faced by designers of privacy mechanisms in deciding what and how to randomise.

Suppose there are N participants invited to contribute to a survey. The survey question is of a personal nature and some respondents might be embarrassed if their true response came to be known. However the organisers of the survey only want to know the total number of (say) "yes" replies. Fig. (1) and Fig. (2) are two possible designs for collecting the responses and announcing a count of the total "yes" responses. Fig. (1) is the well-known "randomised response" protocol which was designed to give respondents plausible deniability so that even the data collectors do not know exactly whether the response they receive from any individual is the "true" response [20].

Fig. (2) follows the blueprint of traditional "oblivious" privacy mechanisms. An accurate tally of the true answers of the respondents is computed, and then some randomness is added. In this particular example we use the geometric distribution.

In both cases there is a corresponding differential privacy guarantee. In Fig. (1) the guarantee grants plausible deniability wrt. a $e^{\log 3}$ threshold for each respondent individually. For Fig. (2) the differential-privacy guarantee is not handed down (directly at least) to an individual, but rather gives a guarantee on the plausible deniability for the final output. In this case more work is required to determine the privacy risk to an individual, but it is relatively easy to provide a guarantee of accuracy: that's because the randomisation is added to the *useful* output, namely the true answer to the query and there are strong results which show that good accuracy can be guaranteed using the geometric distribution for this type of data release. In contrast for Fig. (1), as mentioned above, the randomness is added directly to the data that is considered to be private (by the owner of that data), and so in this scenario the survey organisers would want to know the affect on the accuracy of the final count, which also requires more work to compute.

Much of the theoretical analysis of privacy mechanisms is carried out within the context of an adversarial model. In the case of differential privacy that model is deliberately worst-case: namely it is assumed that the adversary knows *all the data* except that of a given individual. Within that setting though, it seems not straightforward to examine vulnerabilities in regard to "unintended inferences" or the potential for such inferences as described by (iii) above. Moreover Chatzikokolakis et al. [10] have shown that such adversarial models offer surprisingly weak guarantees of privacy operating against other reasonable adversarial settings.

In the case of trying to decide whether to use Fig. 1 or Fig. 2 if we use the weak differential privacy adversarial setting, a designer would be unable to determine how randomness might defend against an actual privacy breach i.e. an unintended inference. When focussing on randomness as a defence, relevant issues are not only that respondents have plausible deniability, but also what level of ϵ to use that balances the risk of an unintended inference with a reasonably accurate tally of "yes" respondents, and what might be the adversary's prior knowledge.

In the remaining sections we analyse the potential for inferences using a model of Quantitative Information Flow (QIF), and adversarial settings which include assumptions about an adversary's prior knowledge. We begin with a brief summary if QIF in the next section.

```
// Assume resp is an array of length N set to
// participants' responses, to a survey question.
i := 0;
count:= 0;
while (i<N) {
        coin:= 0 [1/2] 1;        // Random response
        count:= (count + coin
                [1/2]            // Randomly include or not
                count + resp[i]);
        i++;
}
Print count;     // Announce the approximate count
```

Assume that all variables cannot be observed by an adversary except for the final "Print" of the count. On each iteration, the participant i is randomly selected for inclusion in the count or not. In the case that the participant's true response $resp[i]$ is not included, a random response *coin* for that participant is delivered instead.
This mechanism $\mathcal{R}$ is able to guarantee the following differential-privacy constraint for each participant i providing plausible deniability for their true response:

$$\mathcal{R}(\texttt{resp[i]=0})(Z) \leq \mathcal{R}(\texttt{resp[i]=1})(Z)e^{\log 3} \ .$$

Figure 1 Randomised response with `N` participants.

```
// Assume resp is an array of length N set to
// participants' responses, to a survey question
// epsilon is a parameter for randomising the final tally.
i := 0;
tally:= 0;
while (i<N) {
        tally:= tally + resp[i];
        i++;
}
count:= Geom(tally,  epsilon);
Print count;     // Announce the approximate count
```

Assume that all variables cannot be observed by an adversary except for the final "Print" of the count. On each iteration, the participant i's response $resp[i]$ is included in the count. After the full count has been tallied, the result is randomised according to the (truncated) Geometric distribution. This mechanism $\mathcal{G}$ is able to guarantee the following differential-privacy constraint for the value of the tally, for $\epsilon = \log 3$:

$$\mathcal{G}(\texttt{tally} = k)(Z) \leq \mathcal{G}(\texttt{tally} = k{+}1)(Z)e^{\log 3} \ .$$

Figure 2 Oblivious response mechanism to announce the total for `N` participants.

3 Review of Quantitative Information Flow

The fundamental notion in the analysis of information flow is a *secret* which is a value that is unknown to an adversary, or at least about which there is uncertainty as to its precise value. The relation between secrets and privacy is that any information designated as "sensitive" means that it should be kept secret, meaning that its precise value should remain unknown, or uncertain from the point of view of the adversary. When a privacy mechanism releases information from a data set containing sensitive and insensitive information, of course some sensitive (or "secret") information is likely to be released as well. If the uncertainty about the secret's value is reduced sufficiently, then we might say that the privacy has been breached. This is the essence of an inference attack: if the uncertainty about a secret's value is reduced sufficiently then the adversary is able to predict the true values of secrets with high likelihood.

Quantitative Information Flow makes these intuitions mathematically precise. Given a range of possible secret values of (finite) type $\mathcal{X}$, we model a secret as a probability distribution of type $\mathbb{D}\mathcal{X}$, because it ascribes "probabilistic uncertainty" to the secret's exact value. Given $\pi\colon \mathbb{D}\mathcal{X}$, we write π_x for the probability that π assigns to $x\colon \mathcal{X}$, with the idea that the more likely it is that the real value is some specific x, then the closer π_x will be to 1. Normally the uniform distribution over $\mathcal{X}$ models a secret which could equally take any one of the possible values drawn from its type and we might say that, beyond the existence of the secret, nothing else is known. In any case, once we have a secret, we are interested in analysing whether a mechanism that uses it might leak some information about it. To do this we define a measure for uncertainty, and use it to compare the uncertainty of the secret before and after executing the algorithm. If we find that the two measurements differ then we can say that there has been an information leak.

The original QIF analyses of information leaks in computer systems [3, 4] used Shannon entropy [18] to measure uncertainty because it captures the idea that more uncertainty implies "more secrecy", and indeed the uniform distribution corresponds to maximum Shannon entropy (corresponding to maximum "Shannon uncertainty"). More recent treatments have shown that Shannon entropy is not the best way to measure uncertainty in security contexts because it does not model scenarios relevant to the goals of the adversary. In particular there are some circumstances where a Shannon analysis actually gives a more favourable assessment of security than is actually warranted if the adversary's motivation is taken into account [19].

Alvim et al. [2] proposed a more general notion of uncertainty based on "gain functions". This is the notion we will use. A *gain function* measures a secret's uncertainty according to how it affects an adversary's actions within a given scenario. We write $\mathcal{W}$ for a (usually finite) set of actions available to an adversary corresponding to an "attack scenario" where the adversary tries to infer something (e.g. some property) about the secret. For a given secret $x\colon \mathcal{X}$, an adversary's choice of $w\colon \mathcal{W}$ results in the adversary gaining something beneficial to his objective. This gain can vary depending on the adversary's choice (w) and the exact value of the secret (x). The more effective is the adversary's choice in how to act, the more he is able to overcome any uncertainty concerning the secret's value, and increase his gain.

▶ **Definition 1.** *Given a type $\mathcal{X}$ of secrets, a gain function $g\colon \mathcal{W}\times\mathcal{X} \to \mathbb{R}$ is a real-valued function such that $g(w, x)$ determines the gain to an adversary if he chooses w and the secret is x.*

A simple example of a gain function is bv, where $\mathcal{W} := \mathcal{X}$, and

$$\mathsf{bv}(x, x') \quad := \quad 1 \;\; \textit{if} \;\; x = x' \;\; \textit{else} \;\; 0\,. \tag{2}$$

For this scenario, the adversary's goal is to determine the exact value of the secret, so he receives a gain of 1 if he correctly guesses the value of a secret, and zero otherwise. Assuming that he knows the range of possible secrets $\mathcal{X}$, he therefore has $\mathcal{W}{:=}\mathcal{X}$ for his set of possible guesses.

Given a gain function we define the *vulnerability* of a secret in $\mathbb{D}\mathcal{X}$ relative to the scenario it describes: it is the maximum average gain to an adversary.

▶ **Definition 2.** *Let $g{:}\,\mathcal{W}{\times}\mathcal{X}{\to}\mathbb{R}$ be a gain function, and $\pi{:}\,\mathbb{D}\mathcal{X}$ be a secret. The* vulnerability *$V_g[\pi]$ of the secret wrt. g is:*

$$V_g[\pi] \quad := \quad \max_{w\in\mathcal{W}} \sum_{x\in\mathcal{X}} g(w,x){\times}\pi_x \ .$$

For a secret $\pi{:}\,\mathbb{D}\mathcal{X}$, the vulnerability wrt. bv is $V_{\mathsf{bv}}[\pi]{:=}\max_{x:\,\mathcal{X}}\pi_x$, i.e. the maximum probability assigned by π to possible values of x. The adversary's best strategy for optimising his gain would therefore be to choose the value x that corresponds to the maximum probability under π. This vulnerability V_{bv} is called *Bayes Vulnerability*.

A *mechanism* is an algorithm that inputs a secret and outputs some observable, which could be determined by the value of the secret. An example of a privacy mechanism could be an query to a database which outputs numbers of residents living in various regions or counties. We define $\mathcal{Y}$ to be the type for observables. The model of a mechanism now assigns a probability that $y{:}\,\mathcal{Y}$ can be observed given that the secret is x. Such observables could be sample timings in a timing analysis in cryptography, for example.

▶ **Definition 3.** *A* mechanism *is a stochastic channel*[1] *$C{:}\,\mathcal{X}{\times}\mathcal{Y}\to[0,1]$. The value C_{xy} is the probability that y is observed given that the secret is x.*

Given a (prior) secret $\pi{:}\,\mathbb{D}\mathcal{X}$ and mechanism C we write $\pi{\triangleright}C$ for the joint distribution *in $\mathcal{X}{\times}\mathcal{Y}$ defined*

$$(\pi{\triangleright}C)_{xy} \quad := \quad \pi_x{\times}C_{xy} \ .$$

For each $y{:}\,\mathcal{Y}$, the marginal probability *that y is observed is $p_y{:=}\sum_{x:\,\mathcal{X}}(\pi{\triangleright}C)_{xy}$. For each observable y, the corresponding* posterior probability *of the secret is the conditional $\pi|^y{:}\,\mathbb{D}\mathcal{X}$ defined $(\pi|^y)_x{:=}(\pi{\triangleright}C)_{xy}/p_y$.* [2]

Intuitively, given a prior secret $\pi\in\mathbb{D}\mathcal{X}$ and mechanism C, the entry $\pi_x{\times}C_{xy}$ of the joint distribution $\pi{\triangleright}C$ is the probability that the actual secret value is x and the observation is y. This joint distribution contains two pieces of information: the probability p_y of observing y and the corresponding posterior $\pi|^y$ which represents the adversary's updated view about the uncertainty of the secret's value. If the vulnerability of the posterior increases, then information about the secret has leaked and the adversary can use it to increase his gain by changing how he chooses to act. The adversary's average overall gain, taking the observations into account, is defined to be the average posterior vulnerability (i.e. the average gain of each posterior distribution, weighted according to their respective marginals):

$$V_g[\pi{\triangleright}C] := \sum_{y\in\mathcal{Y}} p_y{\times}V_g[\pi|^y] \ , \qquad \textit{where } p_y,\ \pi|^y \textit{ are defined at Def. 3.} \tag{3}$$

---

[1] Stochastic means that the rows sum to 1, i.e. $\sum_{y\in\mathcal{Y}} C_{xy}=1$.

[2] We assume for convenience that when we write p_y the terms C, π and y are understood from the context. Notation suited for formal calculation would need to incorporate C and π explicitly.

Now that we have Def. 2 and Def. 3 we can start to investigate whether the information leaked through observations $\mathcal{Y}$ actually have an impact in terms of whether it is useful to an adversary. It is easy to see that for any gain function g, prior π and mechanism C we have that $V_g[\pi] \leq V_g[\pi\rangle C]$. In fact the greater the difference between the prior and posterior vulnerability, the more the adversary is able to use the leaked information within the scenario defined by g.

We can also compare the information leak properties of mechanisms – if one mechanism C is more vulnerable than another D under all priors and gain functions, then we say that D is *more secure* than C.

▶ **Definition 4.** *Given C, D mechanisms we say that C is refined by D, or $C \sqsubseteq D$ if for all priors $\pi \in \mathbb{D}\mathcal{X}$ and gain functions $g\colon \mathcal{W}\times\mathcal{X} \to \mathbb{R}$ that $V_g[\pi\triangleright C] \geq V_g[\pi\triangleright D]$.*

Def. 4 is a very robust ordering as it applies to all scenarios. It also characterises post-processing of mechanisms. A post-processing step describes how outputs can be reassigned, or merged. If $C\colon \mathcal{X}\times\mathcal{Y} \to [0,1]$ that outputs values of type $\mathcal{Y}$, then we can "remap" them to outputs of type $\mathcal{Z}$ using a mechanism $D\colon \mathcal{Y}\times\mathcal{Z} \to [0,1]$. The result is therefore the matrix multiplication $C{\cdot}D$. Not surprisingly it can be shown that $C \sqsubseteq C{\cdot}D$, but perhaps surprisingly if $C \sqsubseteq D$ then there is some postprocessing mechanism E such that $C{\cdot}E = D$. We use the following facts about refinement [14].

F1 The maximal element in the refinement ordering is the *unit* channel $\mathbb{I}$ that releases no information at all. It satisfies $V_g[\pi\triangleright\mathbb{I}] = V_g[\pi]$, for all gain functions g.

F2 If $C \not\sqsubseteq D$ then there is some gain function such that $V_g[\pi\triangleright C] < V_g[\pi\triangleright D]$.

F3 QIF enables reasoning about correlated data as follows. Let $C\colon \mathcal{X}\times\mathcal{Y} \to [0,1]$, and let $\Pi \in \mathbb{D}(\mathcal{Z}\times\mathcal{X})$ be a joint distribution representing a correlation between $\mathcal{Z}$. We can define a channel $C^\star\colon \mathcal{Z}\times\mathcal{X}\times\mathcal{Y} \to [0,1]$ as $C^\star_{zxy} = C_{xy}$. Notice that $C^\star$ simply repeats rows of C; moreover $C^\star$ now allows us to investigate how much we can infer about $\mathcal{Z}$ from information flows about $\mathcal{X}$ through C.

F4 If $D \not\sqsubseteq C$ then there is some correlation $\Pi \in \mathbb{D}(\mathcal{Z}\times\mathcal{X})$ such that $V_{\mathsf{bv'}}[\pi\triangleright C] > V_{\mathsf{bv'}}[\pi\triangleright D]$, where bv' defines Bayes' Vulnerability for $\mathcal{Z}$.

3.1 Modelling inferences

An inference is commonly regarded as the ability to determine a property about an individual or a group of individuals using any available information. Thus the inferred property might not be obvious, which is why almost all data releases are vulnerable, to some extent, from inference attacks. In this section we show how to use a QIF model to study such vulnerabilities.

As we have noted above we can use a gain function to model an adversary trying to guess some value within a scenario defined prior knowledge $\pi \in \mathbb{D}\mathcal{X}$. We generalise the idea of Bayes vulnerability Def. 1, modelling an adversary trying to guess a specific value, to an adversary trying to guess a property (or set of values).

▶ **Definition 5.** *Given a state space $\mathcal{X}$ let $\mathcal{P}{:=}\{p_1, p_2 \ldots p_n\}$ be a partition of $\mathcal{X}$. Define gain function $\overline{\mathcal{P}} : \mathcal{P}\times\mathcal{X} \to \{0,1\}$:*

$$\overline{\mathcal{P}}(p,x) \;:=\; 1 \text{ iff } x \in p \text{ else } 0\ .$$

Suppose now that M is a mechanism operating within a scenario where the adversary's prior knowledge is π. We can determine the adversary's ability to use the information in the data release to infer a property defined by a partition $\mathcal{P}$: it is $V_{\overline{\mathcal{P}}}[\pi\triangleright M]$.

It turns out that we can use this idea to analyse the effectiveness of a mechanism in terms of both privacy *and* accuracy. To see how this works recall the mechanisms in Fig. 1 and Fig. 2. The *privacy aspect* is to prevent observers from inferring true responses, whereas the *accuracy* is to deliver a result from which observers can infer with a high level of confidence the true count. [3]

To analyse how well a participant's true response can be inferred from the data release we first set the state space $\mathcal{X}$ to be the set of total possible responses, and choose a prior distribution $\pi \in \mathbb{D}\mathcal{X}$. Later in our experiments §5 we describe how to choose a prior to capture reasonable prior knowledge of the adversary. Writing $\langle r_1, r_2\rangle$ for a scenario where participant 1's true response is $r_1 \in \{0,1\}$ and participant 2's true response is $r_2 \in \{0,1\}$ then $\mathcal{X}{:=}\{\langle 0,0\rangle, \langle 0,1\rangle, \langle 1,0\rangle, \langle 1,1\rangle\}$. Let $\mathcal{S}$ be the partition $\{\{\langle 0,0\rangle, \langle 0,1\rangle\}, \{\langle 1,0\rangle, \langle 1,1\rangle\}\}$ – notice that it contains two sets, with the first grouping the scenarios where participant 1's true response is 0, and the second where participant 1's true response is 1. Computing $V_{\overline{\mathcal{S}}}[\pi \triangleright M]$ gives the probability that the adversary can infer participant 1's true response, whatever it turns out to be. For M to defend strongly against unintended inferences about an individual participant we would like this to be as close to $V_{\overline{\mathcal{S}}}[\pi]$ as possible, because then the information delivered by the data release cannot be used by any adversary in his attack. In fact in our later experiments §5, we define privacy loss to be the ratio $V_{\overline{\mathcal{S}}}[\pi \triangleright M]/V_{\overline{\mathcal{S}}}[\pi]$.

On the other hand in both mechanisms the output count is randomised, even though delivering an accurate count is the purpose of the data release. We can therefore gauge the accuracy of the mechanism by calculating the probability with which the observer can infer the true count for the same prior π. In detail, let $\mathcal{U}$ be the partition $\{\{\langle 0,0\rangle\}, \{\langle 0,1\rangle, \langle 1,0\rangle\}, \{\langle 1,1\rangle\}\}$. Here there are three subsets – the first is the (only) case in which the true count is 0, the second contains two instances where the true count is 1 and the third is the unique case where the count is 2. Computing $V_{\overline{\mathcal{U}}}[\pi \triangleright M]$ therefore gives the probability that the observer (or adversary) can infer the true count. Clearly we would like this to be close to 1 for good accuracy.

We can see these ideas working out for our two examples above by constructing the information flow channels for each of them.

Consider first the channel associated with randomised response Fig. 1.

$$\texttt{randomresp}{:=} \quad \begin{array}{c} \\ \langle 0,0\rangle \\ \langle 1,0\rangle \\ \langle 0,1\rangle \\ \langle 1,1\rangle \end{array} \begin{array}{c} \begin{array}{ccc} \texttt{0} & \texttt{1} & \texttt{2} \end{array} \\ \begin{pmatrix} 0.56 & 0.38 & 0.06 \\ 0.19 & 0.62 & 0.19 \\ 0.19 & 0.62 & 0.19 \\ 0.06 & 0.38 & 0.56 \end{pmatrix} \end{array} \tag{4}$$

Each row of the matrix corresponds to the probabilities of observing the output count given by $0, 1$ or 2 when executed within a scenario defined by the participants' true responses. For example if both participants' responses are 0 (first row corresponding to $\langle 0,0\rangle$) the observer would see the output at 0 with probability 0.56 because 0 is output when both respondents' true values are counted and, when they are not, the random value that *is* counted is still zero. For each participant, his true value is counted with probability $1/2$, but also with probability $1/2$ he randomly picks to respond with zero anyway. Thus each participant contributes 0 to the final survey count with probability 0.75, and since participant responses are independent of each other the count reported will be 0 with probability $0.75{\times}0.75 \sim 0.56$.

[3] Thus in this scenario the adversary and observer are the same.

The other probabilities are computed similarly. Assuming then a uniform prior distribution over possible scenarios, we see that the probability of inferring participant 1's true response using the information in the data release is

$$V_{\overline{\mathcal{S}}}[\pi \triangleright \texttt{randomresp}] = 0.63 \ ,$$

which is only a little more than the prior $V_{\overline{\mathcal{S}}}[\pi] = 0.5$. On the other hand the accuracy of inferring the sum of the responses is *worse*:

$$V_{\overline{\mathcal{U}}}[\pi \triangleright \texttt{randomresp}] = 0.59 \ ,$$

signifying that the observer cannot have a great deal of confidence in inferring the true tally.

We can perform the same analysis on Fig. 2 for comparison. Here the channel matrix is:

$$\texttt{obliviousresp} := \begin{array}{c} \\ \langle 0,0\rangle \\ \langle 1,0\rangle \\ \langle 0,1\rangle \\ \langle 1,1\rangle \end{array} \begin{array}{c} \begin{array}{ccc} \texttt{0} & \texttt{1} & \texttt{2} \end{array} \\ \begin{pmatrix} 0.75 & 0.1667 & 0.0833 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.0833 & 0.1667 & 0.75 \end{pmatrix} \end{array} \tag{5}$$

Notice that the (truncated) geometric mechanism is used to compute the probabilities. In the first row, corresponding to scenario $\langle 0,0\rangle$, the true tally is 0 which is then reported accurately with probability 0.75. As above we an compute the probabilities of inferring participant 1's true response and the true tally as follows:

$$V_{\overline{\mathcal{S}}}[\pi \triangleright \texttt{obliviousresp}] = 0.67 \quad \textit{and} \quad V_{\overline{\mathcal{U}}}[\pi \triangleright \texttt{obliviousresp}] = 0.625 \ .$$

Notice that whilst the accuracy has improved from 0.59 to 0.625, this has come at a cost of increasing the probability of inferring participant 1's response from 0.63 to 0.67. The reason that the increase in accuracy incurs a decrease in privacy is because the two properties are related: if the ability to infer the participants' responses increases then the ability to infer an accurate tally which depends on those responses must also increase. The extent to which we can have accuracy of utility and privacy depends crucially on how the $\mathcal{S}$ and $\mathcal{U}$ are correlated in the prior.

In fact we can define a distribution $\Pi \in \mathbb{D}(\mathcal{S} \times \mathcal{U})$ which expresses the correlation between the privacy property defined by partition $\mathcal{S}$ and the useful property defined by partition $\mathcal{U}$:

$$\Pi_{su} := \pi(s \cap u) \ .$$

From here we can observe that $\mathcal{S}$ and $\mathcal{U}$ can be highly correlated for some of their partition subsets suggesting that a more accurate inference of one must lead to a more accurate inference of the other. For example if a tally of 0 can be accurately inferred then the probability of subsequently inferring that participant 1's true response is also 0 is $\Pi_{s_0,u_0}/\pi(u_0) = 1$, meaning that a 0 tally implies absolutely that participant 1's true value must be 0. On the other hand, the probability of inferring that his value is 0 if the tally is accurately reported as 1 is $\Pi_{s_0,u_1}/\pi(u_1) = 1/2$. Thus studying the impact of the mechanism in terms of the abstraction of the correlation will enable us to understand the trade off between privacy and accuracy, and the limitations on delivering highly accurate data releases in some scenarios.

3.2 Accuracy versus privacy

In this section we study the trade off between accuracy and privacy in terms of inferences. Rather than working with the raw data precisely we use an abstraction based on a correlation between $\mathcal{S}$ and $\mathcal{U}$ where, as described above, $\mathcal{S}$ is defined by a partition on raw data that specifies some privacy criterion, and $\mathcal{U}$ similarly is defined by a partition that specifies the useful data to be released as accurately as possible. We describe the correlation by a distribution $\mathbb{D}(\mathcal{S}\times\mathcal{U})$. For a stochastic channel in $\mathcal{X}\times\mathcal{Y}\to[0,1]$ representing a mechanism we use notation $\mathcal{X}\mapsto\mathcal{Y}$ to differentiate between the input type $\mathcal{X}$ (denoting the secret) and the output type $\mathcal{Y}$ denoting the observable.

Using these conventions, the next definition gives the effect of a mechanism in the ability for an adversary to infer the component $\mathcal{S}$.

▶ **Definition 6.** *Let $\Pi\colon\mathbb{D}(\mathcal{S}\times\mathcal{U})$ represent a correlation in $\mathcal{S}\times\mathcal{U}$, and let $M\colon(\mathcal{S}\times\mathcal{U})\mapsto\mathcal{Y}$ be a stochastic channel representing a data release. We say that M is susceptible to an inference leak for $\mathcal{S}$ if $V_{\overline{\mathcal{S}}}[\Pi] < V_{\overline{\mathcal{S}}}[\Pi\triangleright M]$.*

M is completely privacy preserving wrt. $\overline{\mathcal{S}}$ if and only if $V_{\overline{\mathcal{S}}}[\Pi] = V_{\overline{\mathcal{S}}}[\Pi\triangleright M]$. We measure the privacy loss by the ratio $V_{\overline{\mathcal{S}}}[\Pi\triangleright M]/V_{\overline{\mathcal{S}}}[\Pi]$.

Next, as described above, we can also define the accuracy of inferring the utility.

▶ **Definition 7.** *Let $\Pi\colon\mathbb{D}(\mathcal{S}\times\mathcal{U})$ represent a correlation in $\mathcal{S}\times\mathcal{U}$, and let $M\colon(\mathcal{S}\times\mathcal{U})\mapsto\mathcal{Y}$ be a stochastic channel representing a data release. M's accuracy for $\mathcal{U}$ is the probability that $\mathcal{U}$ can be inferred, i.e. $V_{\overline{\mathcal{U}}}[\Pi\triangleright M]$.*

M is completely accurate for $\mathcal{U}$ if and only if $V_{\overline{\mathcal{U}}}[\Pi\triangleright M] = 1$.

Now we have these definitions, we can provide a simple proof of the well-known "no free lunch" theorem of Kifer [11], which states that it is not possible to have a single mechanism that is arbitrarily accurate and arbitrarily private for all possible correlated secrets.

▶ **Theorem 8.** *There exists no mechanism M which guarantees* both *arbitrary levels of accuracy* and *privacy for all datasets $\Pi\colon\mathbb{D}(\mathcal{S}\times\mathcal{U})$.*

Proof. Let M be a mechanism acting on a (correlated) data set. Pick $\mathcal{S}=\mathcal{U}$, and partition $\mathcal{P}$ on both $\mathcal{S},\mathcal{U}$ so that the private and the useful property are entirely correlated. For any $\epsilon>0$ we might hope that $V_{\overline{\mathcal{P}}}[\Pi\triangleright M]\geq 1-\epsilon$ for good accuracy, whilst at the same time $V_{\overline{\mathcal{P}}}[\Pi\triangleright M]\leq V_{\overline{\mathcal{P}}}[\Pi]+\epsilon$ for good privacy. However if both constraints hold simultaneously then we deduce that $2\epsilon\geq 1-V_{\overline{\mathcal{P}}}[\Pi]$ for all Π, something that cannot be guaranteed. ◀

Thm. 8 applies to all privacy mechanisms, including differential privacy of course. In the next section we investigate two common implementations for differential privacy – one which uses the randomisation in a way to obtain good accuracy, and the other to obtain good privacy. We investigate the trade-off between accuracy and privacy in both cases.

4 Differential privacy

Alvim et al. [1] show that when a mechanism is modelled as a channel the above definition (1) is equivalent to comparing rows of channels relating to x, x'. In particular (1) applied to a channel M says that M is ϵ-differentially private for x, x' if

$$M_{xy}\leq e^{\epsilon}M_{x'y}\quad\textit{and}\quad M_{x'y}\leq e^{\epsilon}M_{xy} \tag{6}$$

for all $y\in\mathcal{Y}$. Notice that (6) compares two channel rows corresponding to secret values x, x'.

When we think of modelling mechanisms in this way, in the context of correlated data, we see that there are two ways in which we can add the noise. The first is to apply it directly to the $\mathcal{S}$ component, in the style of random response sometimes called "local differential privacy". The second is to the $\mathcal{U}$ component, i.e. the accurate result of the query; this is sometimes referred to as an "oblivious (differentially private) mechanism". As we saw in Fig. (1) which is an example of the first kind, and Fig. (2), an example of the second kind, both have an impact on accuracy and privacy. We define these two approaches to differential privacy more generally and examine their respective trade-offs.

▶ **Definition 9.** *A mechanism $C \in (\mathcal{S}\times\mathcal{U}) \to\!\!\triangleright \mathcal{Y}$ is $\mathcal{U}$-insensitive if there is some $L \in \mathcal{S} \to\!\!\triangleright \mathcal{Y}$ such that $C_{suy} = L_{sy}$ (i.e. $C = L^\star$). (Compare [F3] above.) Dually, C is $\mathcal{S}$-insensitive if there is some $R \in \mathcal{U} \to\!\!\triangleright \mathcal{Y}$ such that $C_{suy} = R_{uy}$ (i.e. $C = R^\star$).*

Such insensitive mechanisms as described in Def. 9 add noise either to the $\mathcal{S}/\mathcal{U}$-component respectively, so that the information flow on the other i.e. $\mathcal{U}/\mathcal{S}$-component is derived from the original correlation between $\mathcal{U}$ and $\mathcal{S}$. In general the use of randomisation can often be tailored to achieve particular probabilistic effects. The use of the Laplace distribution for example in oblivious mechanisms means that good utility can be maintained on $\mathcal{U}$, and the impact on privacy therefore can be investigated through the original correlation. We explore the trade-off between accuracy and privacy in differentially-private mechanisms next.

4.1 Utility-focussed privacy

An *utility-focussed mechanism* for differential privacy adds noise to the result of a query, thereby creating opportunities to tune the randomness for accuracy of that query. The differentially-private guarantee for utility-focussed mechanisms entails a property of indistinguishability only on similar query *results*. Any indistinguishability relating to privacy of any sensitive data is dependent on how it correlates with those query results.

With our model for correlated data we can decompose an oblivious mechanism into one which acts directly on $\mathcal{U}$ independently of $\mathcal{S}$, effectively treating it as the secret, and then reinstalling the correlation with $\mathcal{S}$.

▶ **Definition 10.** *$M \in (\mathcal{S}\times\mathcal{U})\to\!\!\triangleright\mathcal{Y}$ is a* utility-focussed *ϵ-private mechanism if it is ϵ-private on $\mathcal{U}$, and $\mathcal{S}$-insensitive.*

The well-known oblivious mechanisms for differential privacy are examples of utility-focused mechanisms. Whilst the randomisation can be tuned to optimise the accuracy of the query this mode of adding randomness does not tell us about ability of an adversary that has knowledge of the correlation to infer the secret component. Thm. 8 tells us we need to consider that correlation to determine the risk. [4]

A weaker property though is to consider whether there is a distinguished mechanism that protects better than all other (oblivious) mechanisms wrt. inference attacks. We write $\overline{\mathcal{S}}$ ($\overline{\mathcal{U}}$) for the gain function derived from the partitioning $\mathcal{S}$ ($\mathcal{U}$) into its singleton sets.

▶ **Definition 11.** *An ϵ-differentially private mechanism M for datasets $\mathbb{D}(\mathcal{S}\times\mathcal{U})$ is optimal wrt. inference attacks on $\mathcal{S}$, if $V_{\overline{\mathcal{S}}}[\Pi \triangleright M] \leq V_{\overline{\mathcal{S}}}[\Pi \triangleright M']$ for all $\Pi \in \mathbb{D}(\mathcal{S}\times\mathcal{U})$ and all ϵ-differentially private mechanisms M'.*

[4] This problem of some individuals being "outliers" and thus potentially identified even in the release of aggregate data is well known and a worst-case mitigation of this situation is the use of sensitivity analysis in for example Laplacian mechanisms [6]. The result is to make any individual's contribution to a noisy announcements of an aggregate statistic minuscule.

Unfortunately there is no mechanism that is optimal wrt. inference attacks except for the trivial mechanisms that release no information at all.

▶ **Theorem 12.** *There is no non-trivial utility-focused mechanism amongst all ϵ-private mechanisms on $\mathcal{S}\times\mathcal{U}$ that is universally optimal wrt. inference attacks on $\mathcal{S}$.*

Proof. (Sketch) Let M be such a non-trivial ϵ-private mechanism. We show that it cannot be universally optimal wrt. inference attacks on $\mathcal{S}$. By Def. 10 this means that M can be decomposed with ϵ-private component $U \in \mathcal{U} \twoheadrightarrow \mathcal{Y}$ such that $U^\star = M$. Moreover if M is not the trivial mechanism, then there exists some mechanism G such that $U \sqsubset G$ (strict refinement, F1), implying that $G \not\sqsubseteq U$. In this case, by (F4) we have immediately that there is some $\Pi \in \mathbb{D}(\mathcal{S}\times\mathcal{U})$ such that $V_{\overline{\mathcal{S}}}[\Pi \triangleright U^\star] > V_{\overline{\mathcal{S}}}[\Pi \triangleright G^\star]$, as required. ◀

Thm. 12 tells us that if privacy is of utmost concern, the use of a utility-focussed mechanism is really about preserving accuracy of the result of the query, and gives no optimal guarantees regarding whether the actual sensitive data is vulnerable to an inference attack under some other mechanism. We see for example with Fig 2 that whilst this does deliver the most accurate result it is not the "most private" as measured by vulnerability to inferences, amongst non-trivial mechanisms.

4.2 Secrecy-focussed privacy

There are many alternative ways to add randomness in privacy mechanisms rather than to the result of a query. We can model this idea as follows.

▶ **Definition 13.** *$M \in (\mathcal{S}\times\mathcal{U})\twoheadrightarrow\mathcal{Y}$ is a* secrecy-focussed *ϵ-private mechanism if it is ϵ-private on $\mathcal{S}$, and is $\mathcal{U}$-insensitive.*

Locally differential private mechanisms are somewhat in this style and therefore are, in some sense, dual to utility-focussed mechanisms because they first produce a noisy version of the data through a noise-adding mechanism applied to $\mathcal{S}$, rather than to the result of a query. Subsequent queries are then applied to the noisy version of the data, in the style of randomised response Fig.1. Next, we define optimality wrt. utility, and show that secrecy-focussed mechanisms cannot be universally optimal wrt. accuracy of utility.

▶ **Definition 14.** *An ϵ-differentially private mechanism $M \in \mathbb{D}(\mathcal{S}\times\mathcal{U})$ is optimal wrt. accuracy on $\mathcal{U}$, if $V_{\overline{\mathcal{U}}}[\Pi \triangleright M] \geq V_{\overline{\mathcal{U}}}[\Pi \triangleright M']$ for all $\Pi \in \mathbb{D}(\mathcal{S}\times\mathcal{U})$ and all ϵ-differentially private mechanisms $M' \in \mathbb{D}(\mathcal{S}\times\mathcal{U})$.*

▶ **Theorem 15.** *Let $|\mathcal{S}| \geq 3$. There is no secrecy-focussed ϵ-differentially private mechanisms on $\mathbb{D}(\mathcal{S}\times\mathcal{U})$ amongst all ϵ-differentially private mechanisms which is optimal wrt. accuracy on $\mathcal{U}$.*

Proof. (Sketch.) Let M be such a secrecy-focussed mechanism, and let L be such that $L^\star = M$ as per Def. 10. The proof is dual to that of Thm. 12 once we show that there is some ϵ-differentially private L' over $\mathcal{S}$ such that $L \not\sqsubseteq L'$. But this follows from an analysis of optimal ϵ-private mechanisms [10]. ◀

Thm. 15 is interesting because it says that mechanisms that add randomness to enable a direct guarantee to the privacy component, cannot be optimised universally for all utility measures. For example, mechanisms that are designed to be locally-differentially private use post-processing to compute the utility on the noisy data. Post-processing, however

is just refinement. Indeed it can be shown, for example, that Fig.1 is a refinement of a *secrecy-focussed* ϵ-private mechanism but more work is needed to explain the observation that locally private mechanisms often exhibit poor accuracy on medium-sized datasets. We end this section by showing that any refinements of secrecy-focussed mechanisms cannot provide universal accuracy.

▶ **Corollary 16.** *If $M \sqsubseteq M'$ and M is secrecy-focussed as in Thm. 15, then M' is not optimal wrt. accuracy on $\mathcal{U}$.*

Proof. Follows since $V_{\overline{\mathcal{U}}}[\Pi \triangleright M] \geq V_{\overline{\mathcal{U}}}[\Pi \triangleright M']$. ◀

5 Experiments

In this section we present some experiments illustrating the main points presented in previous sections. We consider scenarios in which both a data analyst and an adversary can observe the output of a mechanism that reports a (possibly randomised) count of the number of rows in a dataset that satisfy some property. However, whereas the data analyst wants to infer the real value of the counting query performed on the dataset, the adversary wants only to infer the value of a sensitive attribute in a row just added to the dataset. More precisely, we consider experimental scenarios under the following conditions.

1. There is a dataset D of interest, consisting in a multiset of *rows*, each of which is a tuple defined on a set $\mathcal{A}$ of *attributes*. Each attribute $a{\in}\mathcal{A}$ has domain $\mathit{domain}(a)$, and we denote by $\mathit{rows}(\mathcal{A})$ the set of all possible rows that can be formed from attribute set $\mathcal{A}$, and by $x[a]$ the value assumed by attribute $a{\in}\mathcal{A}$ on a row $x{\in}\mathit{rows}(\mathcal{A})$.
2. A new row $x^\star{\in}\mathit{rows}(\mathcal{A})$ will be added to D (due to, e.g., data from a new individual), yielding an extended dataset denoted (with a slight abuse of notation) by $D{\cup}x^\star$.
3. A *data analyst* wants to learn the result of a counting query $\mathtt{count}_q$ that returns the number of rows in the extended dataset $D{\cup}x^\star$ satisfying a given condition q on a *useful attribute* $a_u{\in}\mathcal{A}$ (e.g., how many rows have attribute *gender* set to *female*). The value $\mathtt{count}_q(D{\cup}x^\star)$ is the *real count* for the counting query performed on the extended dataset.
4. An *adversary* has full knowledge of all rows in the dataset D, but is unsure about the contents of the newly added row $x^\star$. His goal is to learn the value $x^\star[a_s]$ of a *sensitive attribute* $a_s{\in}\mathcal{A}$ for this new row.
5. Both the data analyst and the adversary learn the count on the extended dataset $D{\cup}x^\star$ via a *query mechanism* $M_{\mathtt{count}_q}$ that returns a (possibly randomised) version of the real count $\mathtt{count}_q(D{\cup}x^\star)$. We call the output $M_{\mathtt{count}_q}(D{\cup}x^\star)$ the *reported count*, by the mechanism, for the counting query performed on the extended dataset.
6. Both the data analyst and the adversary know the value of the real count $\mathtt{count}_q(D)$ on the original dataset D, and both have as prior knowledge a distribution $\pi^\star{:}\mathbb{D}(\mathit{rows}(\mathcal{A}))$ on all values that the new added row $x^\star$ can assume (e.g., the adversary and the data analyst may be the same entity). From that, they can derive, in the usual way, distributions on the value $x^\star[a_s]$ of the sensitive value of the newly added individual and on the real count $\mathtt{count}_q(D{\cup}x^\star)$ for the query.

To properly formalize the privacy loss and utility of such scenarios, we introduce the following notation. Given a scenario Γ as described above, let Pr^Γ denote the corresponding joint probability distribution –depending on the coin tosses of the distribution $\pi^\star$ on the values for the new row $x^\star$ and on the query mechanism $M_{\mathtt{count}_q}$ employed–, s.t. $Pr^\Gamma(x^\star{=}x, x^\star[a_s]{=}s, \mathtt{count}_q(D{\cup}x^\star){=}u, M_{\mathtt{count}_q}(D{\cup}x^\star){=}u')$ is the probability that in scenario Γ: (i) the

new added row $x^\star$ assumes value $x \in rows(\mathcal{A})$; (ii) the sensitive value $x^\star[a_s]$ of the new added row $x^\star$ assumes value $s \in domain(a_s)$; (iii) the real count of query $\texttt{count}_q$ performed on the extended dataset $D \cup x^\star$ assumes value $u \in \mathbb{N}$; and (iv) the reported count of query $\texttt{count}_q$ produced by the mechanism $M_{\texttt{count}_q}$, w.r.t. the extended dataset $D \cup x^\star$, assumes value $u' \in \mathbb{N}$.

We then define the *privacy loss* of a scenario as the multiplicative Bayes leakage of the new row's sensitive value $x^\star[a_s]$ given the reported count $M_{\texttt{count}_q}(D \cup x^\star)$ on the extended dataset $D \cup x^\star$. (Compare Def. 6.) Intuitively, privacy loss reflects by how much knowledge of the reported count increases the adversary's chance of correctly guessing the secret value in one try. Formally:

$$\textit{privacy-loss}(\Gamma) \stackrel{\text{def}}{=} \frac{\sum_{u' \in \mathbb{N}} \max_{s \in domain(a_s)} Pr^\Gamma(x^\star[a_s]{=}s, M_{\texttt{count}_q}(D \cup x^\star){=}u')}{\max_{s \in domain(a_s)} Pr^\Gamma(x^\star[a_s]{=}s)} \ . \tag{7}$$

On the other hand, we define the *utility* of a scenario as the posterior Bayes vulnerability of the real count $\texttt{count}_q(D \cup x^\star)$ of query $\texttt{count}_q$ performed on the extended dataset $D \cup x^\star$ given the reported count $M_{\texttt{count}_q}(D \cup x^\star)$ on the extended dataset $D \cup x^\star$. (Compare Def. 7.) Formally:

$$\textit{utility}(\Gamma) \stackrel{\text{def}}{=} \sum_{u' \in \mathbb{N}} \max_{u \in \mathbb{N}} Pr^\Gamma(\texttt{count}_q(D \cup x^\star){=}u, M_{\texttt{count}_q}(D \cup x^\star){=}u') \ . \tag{8}$$

Equations (7) and (8) depend on the joint probability distribution Pr^Γ induced by the scenario, which itself depends on the coin tosses of the query mechanism $M_{\texttt{count}_q}$ employed. Hence, to fully flesh out these definitions we need to determine how the mechanism $M_{\texttt{count}_q}$ works. We consider two deferentially-private mechanisms adding noise in different ways.

- An *oblivious mechanism* $M^{obv}_{\texttt{count}_q}$ that first applies counting query $\texttt{count}_q$ to the input dataset to produce a real count u, and then applies a (differentially-private) *oblivious randomisation function* $R^{obv}{:}\mathbb{N}{\to}\mathbb{D}(\mathbb{N})$ to u in order to produce a reported count u' as the output of the mechanism. This is similar to a utility-focussed mechanism.
- A *local mechanism* $M^{loc}_{\texttt{count}_q}$ that first applies a (differentially-private) *local randomisation function* $R^{loc}{:}domain(a_u){\to}\mathbb{D}(domain(a_u))$ independently to each row's useful value to produce a randomised dataset D', and only then applies the counting query $\texttt{count}_q$ to D' in order to produce a reported count $u'{=}\texttt{count}_q(D')$ as the output of the mechanism. This is in the spirit of a privacy-focussed mechanism.

In our experiments we used the dataset released by ProPublica [5] corresponding to two years worth of data from the COMPAS tool (Correctional Offender Management Profiling for Alternative Sanctions), which is one of the most popular algorithmic tools used in the United States criminal justice system for pretrial and sentencing evaluation of the risk of bad behaviour for criminal defendants. We focused on the tool's assessment scores for "Risk of Failure to Appear", and eliminated from the dataset all but the most recent record for any given person, as well as all records with invalid entries for attributes `marital_status` and `score_text`. This treatment has left us with a database containing 11,710 unique records.

We then considered two scenarios. In both, the adversary's goal is to learn the newly added row's value for sensitive attribute `custody_status`, which can be 0-"pretrial defendant", 1-"residential program", 2-"probation", 3-"parole", 4-"jail inmate", or 5-"prison inmate". However, in scenario A the data analyst's query of interest is `"select count *`

[5] `https://github.com/propublica/compas-analysis`

`from D where custody_status=0"`, whose result is highly correlated with the secret information, whereas in scenario B the query of interest is `"select count * from D where marital_status=0"` (the range for `marital_status` is 0-"single", 1-"significant other", 2-"married", 3-"separeted", 4-"divorced", or 5-"widowed"), whose result is highly independent from the secret information. In both scenarios data analyst and adversary assume the new row $x^\star$ added to dataset D follows a distribution $\pi^\star{:}\mathbb{D}(\mathit{rows}(\mathcal{A}))$ s.t. the probability of each $x{\in}\mathit{rows}(\mathcal{A})$ is the value's frequency in the dataset D. Moreover, for fairness in comparison, we instantiate both the oblivious randomisation function R^{obv} and the local randomisation function R^{loc} as the truncated geometric mechanism with the same value of ϵ. [6]

Table 1 presents the results of our experiments for various values of ϵ. In terms of inferences, we want the privacy loss to be low to offer good protection for individuals. This means, in the scale of Bayes vulnerability leakage, that the value should be close to 1 because then we can argue that in the studied scenario the information flow does not increase the adversary's prior knowledge. [7] For accuracy however, to offer good utility, we want it to be as high as possible. In the scale of Bayes vulnerability a value close to 1 represents high certainty that the true value of the query can be accurately inferred.

As we can notice, in both scenarios described above and for each value of ϵ the local mechanism is consistently more private, but less useful, than its oblivious counterpart. The experiments also illustrate the well known fact from the literature that local mechanisms tend not to provide good utility in small datasets like the one we consider here: utility remains at its theoretical minimum for relatively high values of ϵ ($\approx \ln 200$ in Scenario A and $\approx \ln 10^3$ in Scenario B). Finally, note that in Scenario A, where real count and secret are highly correlated, it is hard to achieve a satisfactory trade-off between utility and privacy, as these values are always in opposition to each other. On the other hand, since in Scenario B the real count is practically independent of the secret, it is possible to maintain privacy at a minimum level even for very high values of utility.

6 Related work

Our work builds upon the "no free lunch" theorem of Kifer and Machanavajjhala [11], who provided the first analysis of the limitations of differential privacy in the presence of correlations between secrets. Their work also uses inference attack models to demonstrate the effect of correlations on possible inferences resulting in unexpected privacy breaches. Our work complements theirs by utilising the QIF framework to model the effect of inference attacks through the lens of information flow. The same authors proposed Pufferfish, a framework providing a more nuanced approach to privacy that depends on the idiosyncracies of particular datasets [12]. Inspired by Pufferfish, He et al. [7] introduced the Blowfish framework to allow a more tailored approach to privacy policies which depends on known (public) correlations in the dataset.

Other authors have observed that differential privacy does not protect known correlations in datasets. Zhu et al. [21] proposed strengthening privacy mechanisms for correlated datasets based on a modified measure of the sensitivity of the dataset. Liu et al. [13] demonstrate an inference attack against a differentially private dataset by exploiting known correlations in

[6] The *truncated geometric mechanism* with parameter $\epsilon{>}0$, domain $\{0,1,2,\ldots,m\}$, and co-domain $\{0,1,2,\ldots,n\}$ is defined as $G(j|i){=}{}^{1}\!/_{(1+\alpha)}\cdot\alpha^{i}$, if $j{=}0$, or $G(j|i){=}{}^{(1-\alpha)}\!/_{(1+\alpha)}\cdot\alpha^{|i-j|}$, if $0{<}j{<}n$, or $G(j|i){=}{}^{1}\!/_{(1+\alpha)}\cdot\alpha^{|i-n|}$, if $j{=}n$, for every $0{\leq}i{\leq}m$ and $0{\leq}j{\leq}n$, where $\alpha{=}e^{-\epsilon}$, and each value $G(j|i)$ represents the probability that integer i is remapped to integer j.

[7] In a leakage measure we do not tabulate the actual probability of inference, but rather than increase in inference compared to the prior.

Table 1 Results of privacy loss and utility on Scenarios A (with highly correlated counting query and secret) and B (with practically independent counting query and secret) for the COMPAS dataset A value close to 1 for privacy means the data release does not pose a risk to an individual; a value close to 1 for utility means that the data release is accurate.

	Scenario A				**Scenario B**			
	Oblivious mechanism		**Local mechanism**		**Oblivious mechanism**		**Local mechanism**	
ϵ	**Priv. loss**	**Util.**	**Priv. loss**	**Util.**	**Priv. loss**	**Util.**	**Priv. loss**	**Util.**
0 (theoretical minimum)	1.0000	0.7984	1.0000	0.7984	1.0000	0.7704	1.0000	0.7704
ln **3**	1.0000	0.7984	1.0000	0.7984	1.0000	0.7704	1.0000	0.7704
ln **5**	1.0452	0.8333	1.0000	0.7984	1.0000	0.8333	1.0000	0.7704
ln **10**	1.1402	0.9091	1.0000	0.7984	1.0000	0.9091	1.0000	0.7704
ln **100**	1.2418	0.9901	1.0000	0.7984	1.0000	0.9901	1.0000	0.7704
ln **200**	1.2480	0.9950	1.0001	0.7984	1.0000	0.9950	1.0000	0.7704
ln **500**	1.2517	0.9980	1.0035	0.8011	1.0000	0.9980	1.0000	0.7704
ln $\mathbf{10^3}$	1.2529	0.9990	1.0234	0.8169	1.0000	0.9990	1.0000	0.7704
ln $\mathbf{10^5}$	1.2542	1.0000	1.2481	0.9952	1.0000	1.0000	1.0000	0.9330
∞ (theoretical maximum)	1.2607	1.0000	1.2607	1.0000	1.2607	1.0000	1.2607	1.0000

the data. Works in this area focus on known correlations – particularly correlations within the dataset – and do not address the more general problem of unknown correlations which may be known only to an adversary.

Other works on inference attacks on private data consider alternate privacy metrics to measure privacy loss. Salamatian et al. [17] use traditional information theoretic measures such as entropy and mutual information to produce an alternate privacy framework to differential privacy which is focussed on protecting against inference attacks. Most recently, Jayaraman et al. [9] propose new privacy metrics to evaluate the risk of inference attacks.

Finally, there is considerable interest in understanding inference attacks on machine learning models which employ differential privacy to protect their training data. Rahman et al. [16] empirically evaluate the success of membership inference attacks against machine learning models trained with different privacy parameters. Most recently, Jayaraman et al. [9] empirically evaluate the risk of inference attacks on differentially private machine learned models using their own privacy metrics. Works in this area typically use ad hoc methods to evaluate privacy leakage, whereas our QIF framework permits rigorous analysis based on an operational inference attack model.

7 Conclusion

In this paper we have studied the relationship between accuracy and privacy in data releases from the perspective of inferences. We have shown, using a channel model for quantitative information flow, how to capture reasonable assumptions about prior knowledge to compare accuracy of a query result versus the ability for the adversary to infer additional information about individuals in the database. We have demonstrated how correlations in databases of moderate size pose challenges for protecting privacy of individuals.

In future work we hope to use this kind of analysis to expose potential vulnerabilities in databases by considering the threats posed by inferences in proposed data releases.

References

1 Mário S. Alvim, Konstantinos Chatzikokolakis, Pierpaolo Degano, and Catuscia Palamidessi. Differential privacy versus quantitative information flow. *CoRR*, abs/1012.4250, 2010. `arXiv:1012.4250`.

2 Mário S. Alvim, Kostas Chatzikokolakis, Catuscia Palamidessi, and Geoffrey Smith. Measuring information leakage using generalized gain functions. In *Proc. 25th IEEE Computer Security Foundations Symposium (CSF 2012)*, pages 265–279, June 2012.

3 David Clark, Sebastian Hunt, and Pasquale Malacaria. Quantitative analysis of the leakage of confidential data. *Electr. Notes Theor. Comput. Sci.*, 59(3):238–251, 2001.

4 David Clark, Sebastian Hunt, and Pasquale Malacaria. Quantified interference for a while language. *Electr. Notes Theor. Comput. Sci.*, 112:149–166, 2005.

5 Damien Desfontaines and Balázs Pejó. Sok: Differential privacies. *Proceedings on Privacy Enhancing Technologies*, 2020(2):288–313, 2020.

6 Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006. `doi:10.1007/11681878_14`.

7 Xi He, Ashwin Machanavajjhala, and Bolin Ding. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pages 1447–1458, 2014.

8 Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 18951912, USA, 2019. USENIX Association.

9 Bargav Jayaraman, Lingxiao Wang, David Evans, and Quanquan Gu. Revisiting membership inference under realistic assumptions. *arXiv preprint*, 2020. `arXiv:2005.10881`.

10 C. Palamidessi K. Chatzikokolakis, N. Fernandes. Comparing systems: Max-case refinement orders and application to differential privacy. In *Proc. CSF*. IEEE Press, 2019.

11 Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, SIGMOD Õ11, page 193Ð204, New York, NY, USA, 2011. Association for Computing Machinery. `doi:10.1145/1989323.1989345`.

12 Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.*, 39(1), January 2014. `doi:10.1145/2514689`.

13 Changchang Liu, Supriyo Chakraborty, and Prateek Mittal. Dependence makes you vulnberable: Differential privacy under dependent tuples. In *NDSS*, volume 16, pages 21–24, 2016.

14 Alvim M, K. Chatzikokolakis, A.K. McIver, C.C. Morgan, G. Smith, and C. Palamidessi. *The Science of Quantitative Information Flow*. Information Security and Cryptography. Springer, 2020. To appear.

15 Arvind Narayanan, Hristo S. Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. On the feasibility of internet-scale author identification. In *IEEE Symposium on Security and Privacy, SP 2012, 21-23 May 2012, San Francisco, California, USA*, pages 300–314. IEEE Computer Society, 2012. `doi:10.1109/SP.2012.46`.

16 Md Atiqur Rahman, Tanzila Rahman, Robert Laganière, Noman Mohammed, and Yang Wang. Membership inference attack against differentially private deep learning model. *Trans. Data Priv.*, 11(1):61–79, 2018.

17 Salman Salamatian, Amy Zhang, Flavio du Pin Calmon, Sandilya Bhamidipati, Nadia Fawaz, Branislav Kveton, Pedro Oliveira, and Nina Taft. Managing your private and public data: Bringing down inference attacks against your privacy. *IEEE Journal of Selected Topics in Signal Processing*, 9(7):1240–1255, 2015.

18 C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

19 Geoffrey Smith. On the foundations of quantitative information flow. In Luca de Alfaro, editor, *Proc. 12th International Conference on Foundations of Software Science and Computational Structures (FoSSaCS '09)*, volume 5504 of *Lecture Notes in Computer Science*, pages 288–302, 2009.

20 S.L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60:63Ð69, 1965.

21 Tianqing Zhu, Ping Xiong, Gang Li, and Wanlei Zhou. Correlated differential privacy: Hiding information in non-iid data set. *IEEE Transactions on Information Forensics and Security*, 10(2):229–242, 2014.