Managing your Private and Public Data: Bringing down Inference Attacks against your Privacy

Salman Salamatian*, Amy Zhang†, Flavio du Pin Calmon*, Sandilya Bhamidipati‡,
Nadia Fawaz‡, Branislav Kveton§, Pedro Oliveira¶, Nina Taft‖

***Abstract*—We propose a practical methodology to protect a user's private data, when he wishes to publicly release data that is correlated with his private data, to get some utility. Our approach relies on a general statistical inference framework that captures the privacy threat under inference attacks, given utility constraints. Under this framework, data is distorted before it is released, according to a probabilistic privacy mapping. This mapping is obtained by solving a convex optimization problem, which minimizes information leakage under a distortion constraint. We address practical challenges encountered when applying this theoretical framework to real world data. On one hand, the design of optimal privacy mappings requires knowledge of the prior distribution linking private data and data to be released, which is often unavailable in practice. On the other hand, the optimization may become untractable when data assumes values in large size alphabets, or is high dimensional. Our work makes three major contributions. First, we provide bounds on the impact of a mismatched prior on the privacy-utility tradeoff. Second, we show how to reduce the optimization size by introducing a quantization step, and how to generate privacy mappings under quantization. Third, we evaluate our method on two datasets, including a new dataset that we collected, showing correlations between political convictions and TV viewing habits. We demonstrate that good privacy properties can be achieved with limited distortion so as not to undermine the original purpose of the publicly released data, e.g. recommendations.**

I. Introduction

One of the central problems of managing privacy in the Internet is that of managing both users' public and private data simultaneously. Many users are willing to release *some* data about themselves to a service provider, such as their movie watching history [2]; they do so because such data enable useful services and is often not considered sensitive or private. However users also have other data they consider private, such as income level, political affiliation, or medical conditions. These private attributes can often be inferred from the data the user considers public, by using inference algorithms that have been trained using machine learning techniques. We use the term *inference attack* to refer to the use of an inference algorithm that infers something about a user they may consider private (called a *private attribute*) from their public data. These threats can be seen as hidden threats to users' privacy since most users are unaware of the correlations between their private and public data, and of the ability of machine learning techniques to learn these correlations and exploit them in inference algorithms. For example, most users may not realize that their political views can be inferred from their movie watching history [1], [3]. Indeed, the research community has illustrated numerous scenarios in which personal information can be inferred from the data trails people create. In addition to political views, inference has been used to learn age [4], sexual orientation [5], gender [1], [4] and drug use [5]. The emergence of the Internet of Things (IoT) has generated much fascination along with much concern over rapidly expanding privacy risks, including those related to the inference of users' personal behavior in great detail from data collected by IoT devices over time [6]. This personal information is used rampantly in online services for the purposes of personalized recommendations, and targeted ads. However, when users consider a particular characteristic private, they may wish to prevent inference algorithms from inferring such attributes.

In this work, we focus on a method which allows a user to release her public data, while preventing against inference attacks that may infer her private data from the public information. Our solution relies on a privacy mapping, which informs a user on how to distort her public data, before releasing it, such that no inference attack can successfully infer her private attribute. We aim to provide protection against any inference algorithm that may be used by an attacker. At the same time, the distortion should be bounded so that the personalization service such as a recommendation, provided based on the distorted data, continues to be relevant to the user.

In this paper we adopt the privacy framework presented in [7]. This general framework considers the privacy threat incurred by a user when an adversary attempts to infer the user's private information from the user's public (released) data. The privacy leakage is measured in terms of an inference cost gain that the adversary has by observing the released data. The goal of the framework is to inform a user on how to randomly distort their data before releasing it. We call this a *privacy mapping* from the original user data to the distorted data. The framework finds this mapping while confining the amount of distortion according to utility constraints that ensure the data remains useful for the personalized service. The authors in [7] formulate the problem of determining this mapping for a general inference cost function as a convex

* MIT, Cambridge, MA, {salmansa,flavio}@mit.edu
† SET Media, San Francisco, CA, amy@set.tv
‡ Technicolor, Los Altos, CA, {sandilya.bhamidipati,nadia.fawaz}@technicolor.com
§ Adobe Research, San Jose, CA, kveton@adobe.com
¶ Disqus, San Francisco, CA, cpdomina@gmail.com
‖ Google, Mountain View, CA, ninataft@google.com

This work was done while all eight authors were with Technicolor, Palo Alto, CA, including the internships of S. Salamatian, F. Calmon and A. Zhang at Technicolor. Parts of this work appeared in the proceedings of IEEE GlobalSIP 2013 [1].

program. Without significant loss of generality, [7] argues that the privacy loss can be measured in terms of mutual information, which leads to an optimization similar to the one found in rate-distortion theory. This formulation, albeit general and theoretically sound, faces a number of practical challenges when applied to actual data available within web services, in particular data that assumes value from large alphabets. The first challenge is that this method relies on knowing a joint distribution between the private and public data, called the *prior*. Often the true prior distribution is not available and instead only a limited set of samples of the private and public data can be observed. This leads to the *mismatched prior* problem. We seek to provide a meaningful distortion and bring privacy even in the face of a mismatched prior. Our first contribution centers around this. Starting with the set of observable data samples, we find an estimate of the prior, based on which the privacy mapping is derived. We develop bounds on the impact on the privacy-utility tradeoff of the mismatched prior. More precisely, we show that the private information leakage increases log-linearly with the $\mathcal{L}_1$-distance between our estimate and the prior; that the distortion rate increases linearly with the $\mathcal{L}_1$-distance between our estimate and the prior; and that the $\mathcal{L}_1$-distance between our estimate and the prior decreases as the sample size increases.

The second challenge, that occurs when the size of the underlying alphabet of the user data is very large, e.g. due to a large number of features representing the data, is that of an intractable optimization. We introduce a quantization preprocessing step that limits the dimensionality of the problem, and also reduces the impact of the mismatch. More precisely, we first quantize the original data. We then determine how to distort the data in the space defined by the quantization representative points. The privacy mapping is computed on the representative points, using a convex solver that minimizes privacy leakage subject to a distortion constraint. The advantage of our quantization scheme is that it is computationally efficient - we reduce the number of optimized variables from being quadratic in the size of the underlying alphabet to being quadratic in the number of representative points, and thus make the optimization independent from the number of observable data samples. For some real world examples, this can lead to orders of magnitude reduction in the optimization size. We also show that any additional distortion introduced by quantization increases linearly with the maximum distance between a sample datapoint and the closest representative point. This quantization step, our second contribution, provides a fundamental extension to the original method [7] which cannot easily be applied in practice when the data is too high dimensional. The problem of the optimization size was also studied in [8], where the authors were interested in scaling up the optimization in [7] using linear programming techniques. Our method is complimentary to theirs and can be used regardless of the optimization algorithm used.

Our third area of contribution centers around evaluations. In [7] the authors only proposed and reasoned about their framework but did not evaluate it. Here, we evaluate our methods on two datasets, one well-known dataset used by the machine learning community, and one new dataset that we collected ourselves. This latter dataset is one that contains users TV show ratings and their political affiliation. The existence of correlations between political affiliation and opinions about TV has been shown in [9]. In our study, we consider users opinions about TV shows as the public be data to be released, and a user's political affiliation as information to be kept private. In our solution for producing a mapping, we consider different types of distortions: in some cases we use *erasure-distortions* in which an element of a user's public data is removed, while in other cases we use *exchange-distortions* in which specific elements in a public profile are altered.

Our evaluations demonstrate multiple things. First, even when we do not have a fully specified prior distribution on the public and private data, we show that we can still provide privacy in this difficult environment at the extra cost of a small amount of additional distortion in the public data. Second, we illustrate that our quantization approach works well, namely that it is possible to provide good privacy even when quantization is needed to reduce the dimensionality of the data. Third, we show that in our Politics-and-TV dataset, perfect privacy can be achieved with a 15% distortion of the original public data. In practice less than 15% distortion could provide sufficient privacy. We also illustrate examples of specific distortions (changes to particular public data profiles) and show these are intuitively reasonable, yet not trivial.

This paper differs from our prior work as follows. In [7], no evaluations were carried out, nor did it address the problems in applying this theoretical framework to real world datasets (i.e. incomplete prior data, and high dimensionality), both of which we do here. In our short paper [1], we presented the idea of quantization, but only evaluated it on a limited dataset. In this current paper, we blend quantization with the mismatched prior and carry out broader evaluations that include additional datasets, as well as illustrating how reducing mutual information reduces the success rate of the attacker. We also include all related mathematical proofs herein.

The paper is organized as follows. In Section II, we review related works and explain differences with previous works. In Section III, we formally define the problem. In Section IV, we provide bounds on the impact of a mismatched prior on the privacy-distortion tradeoff. In Section V, we propose a quantization step to reduce the optimization size. Our datasets are described in Section VI, the results of our evaluations are provided in Section VII, and we conclude in Section VIII.

II. Related Work

A. Privacy

The prevalent notion of privacy in the research community is differential privacy [10], [11]. Traditionally, differential privacy considers a centralized statistical database privacy setting, in which a database contains private data from multiple users, and an untrusted analyst asks a query (aggregate function or population quantity) over the entries of the database. An ϵ-differentially private mechanism produces a randomized answer to the query, such that the distribution of the randomized answer to the query does not vary more than a factor e^{ϵ} if one entry of the database varies. This guarantees that it is

difficult to distinguish "neighboring" databases based solely on the observation of the output. The centralized statistical database setting traditionally considered by differential privacy differs from the privacy setting considered in this paper as follows. First, we consider a local privacy setting focused on an individual privacy-conscious user, in which the entity collecting data from the user– the service provider– is not trusted. The user data is not aggregated in a database before it is randomized, but it stays locally at the user where it will be randomized according to a privacy mapping. Local privacy dates back to randomized response in surveys [12]. Second, in our setting, privacy is ensured at the individual user level. More precisely, the quantity that is randomized prior to the release to the service provider is not an aggregate quantity over multiple users. It is an individual quantity of an individual privacy-conscious user, such as the user's very own movie ratings. The private quantity that is protected is another piece of individual data of this privacy conscious user, for example his political views. Neither the ratings nor political views are aggregate quantities over multiple users, or answers to queries over multiple user data. Third, in our setting, recommendations are provided to the privacy-conscious user based on his randomized rating vector, using a recommendation engine that has been a priori trained by the service provider based on original data from non-privacy conscious users. This is in contrast to [13], which considers the problem of training a privacy-preserving recommendation engine whose modeling parameters satisfy differential privacy with respect to the data entries of privacy-conscious users on which the engine is trained.

Differential privacy does not take into account the distribution of the entries of the database, which makes the formulation mathematically tractable and simplifies the implementation of differentially private systems. Moreover, differential privacy is robust against arbitrary side information from the attacker (also called background knowledge or auxiliary information), which is a property that our scheme cannot guarantee as such, even though our recent works promises great progress on defining the privacy-utility trade-off under side information. However, differential privacy does not quantify the amount of information that is leaked from the system. Furthermore, when inputs are correlated, guaranteeing differential privacy does not necessarily guarantee a bounded information leakage. As shown in [7], for certain input distributions, an adversary might able to infer with arbitrarily high precision the entries of the input database from a differentially private answer to a query on this database.

Other general and flexible frameworks similar to differential privacy exist such as the Pufferfish framework [14]. In this framework, a pair of mutually exclusive statements are output, such that the adversary does not know which, if either of the two statements is true. This framework does not take try to minimize distortion of the data, and ignores utility preservation. In our paper, we focus on the privacy-utility trade-off. We also assume that the adversary has knowledge of the data generation process, and knows the same prior distribution as the system which designs the privacy mapping. The Pufferfish framework can accommodate any assumption about the adversary's knowledge of the prior distribution, but also requires that the system designing the privacy mapping knows what the adversary's belief on the prior distribution is, which is not knowledge we can assume.

Another existing trend in the privacy research community is to apply information-theoretic tools to quantify and design privacy-preserving mechanisms [7], [15]–[19]. Information theory provides a natural framework to measure the amount of private information that an adversary can learn by observing a given user's public data. This was first noted by Reed [16], and has since appeared in different forms in the information theory and privacy literature. One line of work, adopted in [17], [18], provides asymptotic and fundamental limits for an adversary's average equivocation of the private data as the number of data samples grows arbitrarily large and characterize rate-distortion-equivocation regions.

Non-asymptotic approaches to information-theoretic privacy were discussed, for example, in [7], [15], [19]. In [15], information-theoretic metrics were directly applied to design privacy mechanisms without considering distortion constraints. Afterwards, [19] presented a formulation for designing privacy mechanisms similar to the ones found in rate-distortion theory. More recently, [7] introduced a general framework for privacy against statistical inference that takes into account distortion constraints for the user's public data. This framework was first applied to real data in [1], where the quantization method was also proposed. Built on these ideas, a system implementing privacy mappings for TV shows was presented in [20]. Similarities and differences between this privacy-utility framework and the information bottleneck method were studied in [21], and a sparse optimization algorithm to solve efficiently the privacy problem was proposed in [8].

Information-theoretic approaches have also been used to quantify the information flow in security systems (e.g. [22] and the references therein). In this case, different information-theoretic metrics are used to quantify the change of an attackers belief on the input of a system given an observation of the output. These approaches, such as the one used in [22], also take into account possible prior mismatches and extra knowledge that an attacker might have. Even though in this paper we also use information-theoretic metrics to quantify the change in the attacker's belief, our results are fundamentally different in what they seek to accomplish. Our main goal is not to simply quantify the adversarial threat, but create a practical framework that allows the design of privacy-preserving mechanisms that also maintain a certain level of utility of the data. Therefore, we simultaneously consider the utility of the data and the variation of the adversary's belief, instead of focusing solely on the information flow.

Our focus on inference attacks that occur in recommendations, differs from other privacy problems with recommendation systems. For example, in [2], the focus is on de-anonymization of user records through linkage attacks: by linking records in two databases, the authors recover the identity of some users in an anonymized database. Linkage attacks are a different kind of threat that we do not address in this paper, and thus our work is orthogonal to this one. Our focus is on preventing inference attacks on private attributes

of an individual user, rather than on recovering the identity of a user from anonymized records through linkage attacks. K-anonymity was introduced by [23] and was intended to hide the identity of a user by making his record indistinguishable from the records of $K-1$ other users. Our setting is different because we are attempting to prevent private attributes of a user from being inferred, while simultaneously allowing the user to disclose some other data.

B. Quantization

Data quantization [24] are methods that reduce the size of datasets. In summary, all of these methods select k representative examples from the set of n examples, where $k \ll n$. The difference between the methods is in their objectives. One of the most popular methods is k-means clustering, which minimizes the mean squared error between the examples and their closest representative example [24]. Another popular metric is to minimize is the maximum distance between the example and its closest representative example. Online k-center clustering [25] and cover trees [26] find nearly optimal solutions to this problem.

III. Problem Statement and Background

Notations: We denote by $\mathrm{Simplex}$ the probability simplex defined by $\sum_x p(x) = 1, \quad p(x) \geq 0 \ \forall x$. Let $A \in \mathcal{A}$ and $B \in \mathcal{B}$ be random vectors taking values in the finite alphabets $\mathcal{A}$ and $\mathcal{B}$ respectively. The joint probability distribution of the elements of A and B is denoted $p_{A,B} : \mathcal{A} \times \mathcal{B} \to [0,1]$. The marginal distribution of vector A is defined by $p_A(a) = \sum_{b \in \mathcal{B}} p_{A,B}(a,b) \ \ \forall a \in \mathcal{A}$, while the conditional distribution of A given B is given by $p_{A|B}(a|b) = \frac{p_{A,B}(a,b)}{p_B(b)}$.

The entropy $H(A) = -\sum_{a \in \mathcal{A}} p_A(a) \log(p_A(a))$ of a random vector A depends only on the distribution p_A, while the mutual information $I(A;B) = \sum_{a \in \mathcal{A}, b \in \mathcal{B}} p_{A,B}(a,b) \log\left(\frac{p_{A,B}(a,b)}{p_A(a) p_B(b)}\right)$ of vectors A and B depends only on the joint distribution $p_{A,B}$, since the marginals p_A and p_B can be obtained from $p_{A,B}$.

In this section, we define the threat model, and describe the privacy-accuracy framework we consider. Then, we point out two challenges encountered when applying this framework in practice, and outline our approaches to address them. These approaches are treated in more details in Sections IV and V.

A. Threat Model

Setting: We consider the setting described in [7], where a user has two types of data: some data that he would like to remain private, e.g. his income level, his political views, etc., and some data that he is willing to release publicly and from which he will derive some utility, for example the release of his media preferences to a service provider would allow the user to receive content recommendations.

We denote by $A \in \mathcal{A}$ the vector of personal attributes that the user wants to keep private, and by $B \in \mathcal{B}$ the vector of data he is willing to make public. We assume that the user private attributes A are linked to his data B by the joint probability distribution $p_{A,B}$. Thus, an adversary who would observe B could infer some information about A from B. To reduce this inference threat, instead of releasing B, the user will release a *distorted version* of B, denoted $\hat{B} \in \hat{\mathcal{B}}$, generated according to a conditional probabilistic mapping $p_{\hat{B}|B}$, called the *privacy mapping*. Note that the set $\hat{\mathcal{B}}$ may differ from the set $\mathcal{B}$. It should be pointed out that both the prior distribution $p_{A,B}$ and the privacy mapping $p_{\hat{B}|B}$ are assumed to be known to the adversary. **Privacy Threat**: To formalize the privacy threat model, we assume the standard statistical inference threat model in [7]. The attack considered is that of statistical inference of the private attribute A from the observation B. The inference attack, defined below, is statistical because it makes use of the prior distribution $p_{A,B}$ to infer private attribute A from the observation B.

Definition 1 (Inference attack)**.** *An* inference attack *on $A \in \mathcal{A}$ given the observation $B = b \in \mathcal{B}$ takes as input the distribution $p_{A,B}$, and the observation $B = b$, and outputs a probability distribution $q^* : \mathcal{A} \to [0,1]$ from the set $\mathcal{P}(\mathcal{A})$ of distributions with support $\mathcal{A}$, as the solution to the minimization*

$$q^* = \arg\min_q E_{A|B}[C(A,q)|B=b], \tag{1}$$

for some cost function $C(A,q)$.

In other words, the inference attack takes as inputs the observed $B = b$, and outputs a belief distribution q_A on A given this observation by minimizing its inference cost. Note that this definition is more general than the one where the inference attacks only outputs an estimate $\hat{A}$. The minimization of an expected cost in (1) is a standard approach in statistical inference [27, Chapter 8].

We now define the privacy leakage as follows. Prior to observing $\hat{B}$, the adversary would choose a distribution q on A as the solution of the minimization

$$c_0^* = \min_q E_A[C(A,q)].$$

After observing $\hat{B}$, the adversary would update his belief q such that it minimizes

$$c_{\hat{b}}^* = \min_q E_{A|\hat{B}}[C(A,q)|\hat{B}=\hat{b}].$$

The average cost gain by the adversary after observing $\hat{B}$ is the difference

$$\Delta C = c_0^* - E_{\hat{B}}[c_{\hat{b}}^*].$$

Definition 2 (Privacy leakage)**.** *The* privacy leakage *on $A \in \mathcal{A}$ from the observation of $B \in \mathcal{B}$ is given by ΔC.*

The privacy leakage quantifies how much an adversary gains in term of inference of the private attributes A thanks to the observation of $\hat{B}$. The goal of the privacy mapping will be to minimize this gain. In the particular case of *perfect privacy* $\Delta C = 0$, the released data $\hat{B}$ does not provide any information that is helpful for the inference of A, and the inference cannot outperform an uninformed guess. This general framework does not assume a particular inference algorithm.

If an adversary uses the log-loss[1] cost function $C(A, q) = -\log(q_A)$, it can easily be shown [7] that

$$\Delta C = I(A; \hat{B}). \tag{2}$$

Hence, the privacy leakage is captured by the mutual information between the private attributes A and the publicly released data $\hat{B}$. It should be noted that in the case of perfect privacy ($I(A; \hat{B}) = 0$), the privacy mapping $p_{\hat{B}|B}$ renders the released data $\hat{B}$ statistically independent from the private data A.

A natural question is how the privacy leakage modeled by the mutual information $I(A; \hat{B})$ can be related to the probability of success of an inference algorithm used by the adversary. In [28], using Fano's Inequality [29, Theorem 2.11.1], the probability of error $\mathbb{P}(\tilde{A} \neq A)$ of any inference algorithm that infers $\tilde{A}$ based on the observation $\hat{B}$ is lower-bounded by

$$\mathbb{P}(\tilde{A} \neq A) \geq \frac{H(A) - I(A; \hat{B}) - 1}{\log |\mathcal{A}|} \tag{3}$$

From (3), it is clear that making $I(A; \hat{B})$ small enough, more precisely in the order of $I(A; \hat{B}) < \epsilon H(A)$, increases the bound on the probability of error. Also note that this bound is maximized for $\frac{H(A)-1}{\log(\mathcal{A})}$, and cannot be made arbitrarily large. Indeed, even if $\hat{B}$ and A are statistically independent ($I(A; \hat{B}) = 0$), an adversary can use its knowledge of p_A to make an uniformed guess on the value of A: by inferring $\tilde{A} = \arg\max_{a \in \mathcal{A}} p_A(a)$ as the most probable value in $\mathcal{A}$, he will be correct with probability $p_A(a)$.

It should be mentioned that, although we model the privacy threat using the average cost gain ΔC in this paper, Calmon and Fawaz [7] also proposed a worst-case model $\Delta C^* = c_0^* - \min_{\hat{b} \in \mathcal{B}} c_{\hat{b}}^*$, where the privacy threat is measured in terms of the most informative output, i.e. the output that gives the largest gain in cost. We would like to point out that in the case of perfect privacy under the log-loss, the average threat model $\Delta C = 0$ and the worst-case threat model $\Delta C^* = 0$ are equivalent. Thus conclusions drawn on distortion to achieve perfect privacy under the average threat model also hold for the worst-case model. In general, the worst-case threat is an upperbound on the average threat, and its analysis and application are the object of some of our ongoing work.

Distortion Constraint: The privacy mapping $p_{\hat{B}|B}$ should be designed in such a way that it renders any statistical inference of A based on the observation of $\hat{B}$ harder, yet, at the same time, preserves some utility to the released data $\hat{B}$, by limiting the distortion generated by the mapping. This can be modeled by a constraint $\Delta \geq 0$ on the average distortion:

$$E_{B,\hat{B}}[d(B, \hat{B})] \leq \Delta, \tag{4}$$

for some distortion function $d : \mathcal{B} \times \hat{\mathcal{B}} \to \mathbb{R}^+$. Any distortion function can be used, such as the Hamming distance if B and $\hat{B}$ are binary vectors, or the l_2-norm if B and $\hat{B}$ are real vectors, or even more complex functions, possibly non-symmetric, modeling the variation in utility that a user would derive from the release of $\hat{B}$ instead of B. The latter could, for example, represent the difference in the quality of content recommended to the user based on his distorted media preferences $\hat{B}$ instead of his true preferences B.

[1]For a justification of the relevance and generality of the log-loss cost, we refer the reader to [7, Section IV.A] and to [21].

Algorithm 1 Privacy mapping design.

Input: prior $p_{A,B}$

solve the problem for $p_{\hat{B}|B}$:

$$\begin{aligned} \underset{p_{\hat{B}|B}}{\text{minimize}} \quad & J(p_{A,B}, p_{\hat{B}|B}) \\ \text{subject to} \quad & \mathbb{E}_{p_{B,\hat{B}}}\Big[d(B, \hat{B})\Big] \leq \Delta \\ & p_{\hat{B}|B} \in \text{Simplex} \end{aligned}$$

Output: mapping $p_{\hat{B}|B}$

B. Privacy-Accuracy Framework

In this section, we describe how the privacy mapping is designed to address the inference privacy threat, under a constraint on the distortion.

The mutual information $I(A; \hat{B})$ is a function of the joint distribution $p_{A,\hat{B}}$, which in turn depends on both the prior distribution $p_{A,B}$ and the privacy mapping $p_{\hat{B}|B}$. Indeed, $A \to B \to \hat{B}$ form a Markov chain, thus

$$\begin{aligned} p_{A,\hat{B}}(a, \hat{b}) &= \sum_{b \in \mathcal{B}} p_{\hat{B}|B}(\hat{b}|b) p_{A,B}(a, b), \\ p_{\hat{B}}(\hat{b}) &= \sum_{b \in \mathcal{B}} p_{\hat{B}|B}(\hat{b}|b) p_B(b), \end{aligned} \tag{5}$$

and using Eq. (5) in the definition of $I(A; \hat{B})$, we can write

$$I(A; \hat{B}) = \sum_{a,b,\hat{b}} p_{A,B}(a, b) p_{\hat{B}|B}(\hat{b}|b) \log \frac{\sum_{b"} p(\hat{b}|b") p(b"|a)}{\sum_{a',b'} p(\hat{b}|b') p(a', b')}. \tag{6}$$

To stress the dependency of the privacy leakage on the prior distribution and the privacy mapping, we will denote

$$I(A; \hat{B}) = J(p_{A,B}, p_{\hat{B}|B}).$$

Similarly, the average distortion $E_{B,\hat{B}}[d(B, \hat{B})]$ is a function of the joint distribution $p_{B,\hat{B}}$, which in turn depends both on the prior distribution $p_{A,B}$, through the marginal p_B, and on the privacy mapping $p_{\hat{B}|B}$. Consequently, given a prior distribution $p_{A,B}$, the privacy mapping $p_{\hat{B}|B}$ minimizing the privacy leakage subject to a distortion constraint is obtained as the solution to the optimization

$$\begin{aligned} \underset{p_{\hat{B}|B}}{\text{minimize}} \quad & J(p_{A,B}, p_{\hat{B}|B}) \\ \text{subject to} \quad & E_{B,\hat{B}}[d(B, \hat{B})] \leq \Delta \\ & p_{\hat{B}|B} \in \text{Simplex}, \end{aligned} \tag{7}$$

which is summarized in Algorithm 1. It was shown in [7] that this problem is convex, and can thus be efficiently solved using standard algorithms. Note that this problem bears some resemblance with a modified rate-distortion problem.

In the case of discrete data (numerical or categorical), the privacy mapping is obtained as the solution of optimization (7)

over probabilistic distributions, which has a finite number of variables $|\hat{\mathcal{B}}|\times|\mathcal{B}|$. The case of continuous data, and the case of mixed continuous data and discrete (categorical or numerical) data, can be handled by solving the optimization over a class of parametric distributions, whose parameters are optimized in the design of the privacy mapping, or by discretizing the alphabets. For instance, in [21], [28], the optimization is solved for Gaussian priors and Gaussian mappings.

C. Practical Challenges

In this section, we describe two practical challenges encountered when applying the theoretical privacy-accuracy framework described in Section III-B.

Consider the setting where data assumes values from a large alphabet, then two practical challenges arise:

Mismatched prior: Finding the privacy mapping as the solution to the convex optimization in Algorithm 1 relies on the fundamental assumption that the prior distribution $p_{A,B}$ that links private attributes A and data B is known and can be fed as an input to the algorithm. In practice, the true prior distribution may not be known, but may rather be estimated from a set of sample data that can be observed, for example from a set of users who do not have privacy concerns and publicly release both their attributes A and their original data B. The prior estimated based on this set of samples from non-private users is then used to design the privacy-preserving mechanism that will be applied to new users, who are concerned about their privacy. If data assumes values from a large alphabet, then obtaining an accurate estimate of the prior distribution requires a large amount of samples, which may not be available in practice. Thus, in practice, there may exist a mismatch between the estimated prior and the true prior, due for example to a small number of observable samples, or to the incompleteness of the observable data. In Section IV, we characterize the actual privacy-accuracy tradeoff that results from first running Algorithm 1 with a mismatched prior as input, and then using the so-obtained privacy mapping, instead of the mapping that would have been obtained under the knowledge of the true prior.

Large number of optimization variables: Designing the privacy mapping $p_{\hat{B}|B}$ requires characterizing the value of $p_{\hat{B}|B}(\hat{b}|b)$ for all possible pairs $(b,\hat{b})\in\mathcal{B}\times\hat{\mathcal{B}}$, i.e. solving the convex optimization problem over $|\mathcal{B}||\hat{\mathcal{B}}|$ variables. When $\hat{\mathcal{B}}=\mathcal{B}$, and the size of the alphabet $|\mathcal{B}|$ is large, solving the convex optimization over $|\mathcal{B}|^2$ variables may be intractable.

Now assuming that, although data takes values from a large alphabet, it actually lies in a low dimensional space, then leveraging the structural properties of the data can help addressing both practical challenges mentioned above. Indeed, by leveraging these structural properties, a quantization pre-processing step reduces the size of the alphabet of the data, which helps addressing the challenges as follows:

1) Mismatched prior: If quantization is used to map a point B to a representative example C in a pre-processing step, then the optimization for the design of the privacy takes as input the prior distribution $p_{A,C}$ rather than the prior distribution pA,B. The size of the alphabet of C is smaller than that of B. Thus, obtaining an accurate estimate of the prior $p(A,C)$ requires less samples than estimating the prior $p(A,B)$. Therefore, the mismatch effect is attenuated.

2) Number of optimization variables: If the quantization pre-processing step is used, the optimization for the design of the privacy mapping is solved over the probabilistic mappings $p_{\hat{C}|C}$ over the representative points C, rather than over the mappings $p_{\hat{B}|B}$. The number of optimization variables becomes quadratic in the number of representative points, rather than quadratic in the size of the alphabet of B. This results in a reduction in the number of variables since the size of the alphabet of C is smaller than that of B.

In Section V, we introduce a pre-processing step based on quantization. We show that this method does not affect the privacy levels that can be achieved, but comes at the expense of a limited amount of additional distortion, that we characterize.

IV. Privacy under a mismatched prior

Suppose that we do not have perfect knowledge of the true prior distribution $p_{A,B}$ but that we have its estimate $q_{A,B}$. Let the $\|p_{A,B}-q_{A,B}\|_1$ represent the mismatch between the true prior $p_{A,B}$ and the estimate $q_{A,B}$. Let $p^*_{\hat{B}|B}$ denote the optimal privacy mapping obtained when $p_{A,B}$ is fed as an input to the optimization (7), and let $q^*_{\hat{B}|B}$ denote the solution obtained when feeding the mismatched distribution $q_{A,B}$ as an input to the optimization (7). Then, if $q_{A,B}$ is a good estimate of $p_{A,B}$ (low mismatch), then $q^*_{\hat{B}|B}$ should be close to $p^*_{\hat{B}|B}$. In particular, we distinguish between two desirable properties:

- **Consistency:** As the true prior is $p_{A,B}$, the *actual* privacy leakage when using privacy mappings $q^*_{\hat{B}|B}$ is given by $J(p_{A,B},q^*_{\hat{B}|B})$, and not by the quantity $J(q_{A,B},q^*_{\hat{B}|B})$ that is optimized when the estimate $q_{A,B}$ is fed as an input to the optimization. By consistency, we mean that the privacy mappings $q^*_{\hat{B}|B}$ obtained using the estimate $q_{A,B}$ should have a good performance, both in terms of actual privacy leakage and distortion, when used under the true prior $p_{A,B}$. Theorem 1 expresses the difference in privacy leakage $|J(p_{A,B},q^*_{\hat{B}|B})-J(q_{A,B},q^*_{\hat{B}|B})|$ in terms of the mismatch $\|p_{A,B}-q_{A,B}\|_1$.
- **Near-Optimality:** For near-optimality, we wish that the performance of the privacy mappings $q^*_{\hat{B}|B}$ be close to that of the optimal mappings $p^*_{\hat{B}|B}$. Theorem 2 expresses the difference in privacy leakage $|J(q_{A,B},q^*_{\hat{B}|B})-J(p_{A,B},p^*_{\hat{B}|B})|$ in terms of the mismatch $\|p_{A,B}-q_{A,B}\|_1$.

Theorem 1 (Consistency)**.** *Let $q^*_{\hat{B}|B}$ be a solution to the optimization problem* (7) *with $q_{A,B}$ as input. Then:*

$$\left|J(p_{A,B},q^*_{\hat{B}|B})-J(q_{A,B},q^*_{\hat{B}|B})\right| \leq 3\,\|p_{A,B}-q_{A,B}\|_1\log\frac{|\mathcal{A}|\,|\mathcal{B}|}{\|p_{A,B}-q_{A,B}\|_1}$$

$$\mathbb{E}_{p_{\hat{B},B}}\Big[d(\hat{B},B)\Big]\leq\Delta+d_{\max}\,\|p_{A,B}-q_{A,B}\|_1$$

where $d_{\max}=\max_{\hat{b},b}d(\hat{b},b)$ is the maximum distance in the feature space and $\mathbb{E}_{p_{\hat{B},B}}$ is the expectation over $p_{\hat{B},B}(\hat{b},b)=$

$\sum_a p_{A,B}(a,b)q^*_{\hat{B}|B}(\hat{b}|b)$.

Theorem 1 can be interpreted as a consistency result. Indeed, the optimized privacy leakage $J(q_{A,B}, q^*_{\hat{B}|B})$ and the actual leakage $J(p_{A,B}, q^*_{\hat{B}|B})$ are close if the priors are close. Note, however, that there is no mention of the true optimal leakage $J(p_{A,B}, p^*_{\hat{B}|B})$. Theorem 2 bounds the difference between the optimum with the true distribution $J(p_{A,B}, p^*_{\hat{B}|B})$ and the optimum with the estimate distribution $J(q_{A,B}, q^*_{\hat{B}|B})$.

Theorem 2 (Near-optimality)**.** *Let $q^*_{\hat{B}|B}$ and $p^*_{\hat{B}|B}$ be the solutions of the optimization problem* (7) *respectively with $q_{A,B}$ and $p_{A,B}$ as inputs and distortion constraint Δ. Then,*

$$|J(p_{A,B}, p^*_{\hat{B}|B}) - J(q_{A,B}, q^*_{\hat{B}|B})| \leq 7\|p_{A,B} - q_{A,B}\|_1 \frac{d_{max}}{d_{min}} \log \frac{|\mathcal{A}||\mathcal{B}|}{\|p_{A,B} - q_{A,B}\|_1} \quad (8)$$

with d_{max} defined as in Thm. 1, and d_{min} the smallest non-zero value of the distortion, i.e.*, $d_{min} = \min_{b,\hat{b},s.t.,d(b,\hat{b})>0} d(b,\hat{b})$.*

Theorem 1 and Theorem 2 can be combined using the triangle inequality to give a bound on the difference between the actual leakage when having $q^*_{\hat{B}|B}$,*i.e.*, $J(p_{A,B}, q^*_{\hat{B}|B})$ and the optimal leakage $J(p_{A,B}, p^*_{\hat{B}|B})$. The proof of these theorems are inspired by existing results in Information Theory regarding uniform continuity of information theoretic measures such as [30], [31], and methods for the proof of Theorem 2 can be found in [32]. The results can be tighten by using a tighter version of Lemma 1 in the appendix, as in [33, Problem 3.10], but the order of the error stays the same.

This set of results allows us to construct mappings $q^*_{\hat{B}|B}$ that have close to optimal performance, even though the mapping is not perfectly known. The error grows in the order of $O(-\|p_{A,B} - q_{A,B}\|_1 \log \|p_{A,B} - q_{A,B}\|_1)$ with the mismatch. Note that only this distance is necessary to compute the bounds, and not the true prior itself. In Prop. 1 below, we provide a bound on the probability of $\|p_{A,B} - q_{A,B}\|_1$ being large, when $q_{A,B}$ is simply the empirical distribution obtained from counting on n samples.

Proposition 1. *Let $q_{A,B} = \frac{\#\{a_i=a, b_i=b\}}{n}$ be the empirical distribution of $p_{A,B}$, where n is the total number of samples, and $\#\{a_i = a, b_i = b\}$ is the number of examples where $A = a$ and $B = b$. Then,*

$$\mathbb{P}(\|q_{A,B} - p_{A,B}\|_1 \geq \varepsilon) \leq (n+1)^{|\mathcal{A}||\mathcal{B}|} 2^{-2n\epsilon^2}$$

The proof of Prop. 1 follows from Sanov's theorem [29, Thm 12.4.1] and Pinsker's Inequality [33, Problem 3.18].

Therefore, as the sample size n increases, the probability of having a poor empirical estimator of the true distribution in terms of $\mathcal{L}_1$-norm decreases with rate $(n+1)^{|\mathcal{A}||\mathcal{B}|}2^{-2n\epsilon^2}$. This proposition allows us to formulate corollaries of the following form, here by combining it with Theorem 1:

Corollary 1. *Le $q_{A,B}$ be the empirical distribution over n samples, and let $0 < \epsilon \leq \frac{1}{2}$. Then,*

$$\left|J(p_{A,B}, p^*_{\hat{B}|B}) - J(q_{A,B}, p^*_{\hat{B}|B})\right| \leq 3\epsilon \log \frac{|\mathcal{A}||\mathcal{B}|}{\epsilon} \quad (9)$$

$$\mathbb{E}_{p_{\hat{B},B}}\left[d(\hat{B}, B)\right] \leq \Delta + d_{\max}\epsilon \quad (10)$$

with probability $(n+1)^{|\mathcal{A}||\mathcal{B}|}2^{-2n\epsilon^2}$

This corollary shows the impact on the privacy-accuracy tradeoff of the number of samples available to estimate the distribution and the size of the alphabets.

V. Optimization Size Reduction by Quantization

In real-world datasets, the alphabet $\mathcal{B}$ is often large. In particular, the number of symbols in the alphabet $\mathcal{B}$ observed in the available dataset may be $\theta(n)$, linear in the number of samples n in the dataset. Suppose that $\hat{\mathcal{B}} = \mathcal{B}$. Then the number of optimized variables in Problem (7) is $\theta(n^2)$. Note that the distortion constraint is linear in $p_{\hat{B}|B}(\hat{b} \mid b)$, but the objective function is neither linear nor quadratic. As a result, Optimization (7) cannot be solved using fast linear or quadratic programming solvers. In general, the problem is hard to solve when the size of alphabet $\mathcal{B}$ exceeds a few hundreds symbols.

However, in many problems of interest, data lies on a low-dimensional manifold. For instance, in recommender systems, the ratings of a user can be viewed as a low-dimensional vector in the so-called latent space, whose length is the number of latent factors [34]. In such cases, quantization is guaranteed to reduce the dimensionality of the problem. In particular, let the data lie in a compact d-dimensional latent space where d is small. Then based on a standard sphere packing argument [35], this space can be covered by k representative points such the maximum distance of any point from the closest representative point is $\theta(k^{-1/d})$. In other words, to guarantee that the maximum distance is δ, $\theta((1/\delta)^d)$ representative points are necessary. Note that this quantity is independent of the number of data samples n.

We leverage this observation to propose an approach to reduce the number of optimization variables. Our method comprises three steps. First, a quantization [24] step maps the symbols in alphabet $\mathcal{B}$ to $|\mathcal{C}|$ representative examples in a smaller alphabet $\mathcal{C}$. Second, we learn a privacy-preserving mapping $q_{\hat{C}|C}$ on the new alphabet, where $\hat{\mathcal{C}} = \mathcal{C}$. Third, the symbols in $\mathcal{B}$ are mapped to the representative examples $\hat{\mathcal{C}}$ based on the learned mapping $q_{\hat{C}|C}$. Our approach is summarized in Algorithm 2 and Diagram 1.

Our solution has several notable properties. To begin with, the privacy-preserving mapping $q_{\hat{C}|C}$ is learned on the reduced alphabet $\mathcal{C}$. Thus, we need to solve the convex optimization (7) for only $|\mathcal{C}||\hat{\mathcal{C}}|$ variables instead of $|\mathcal{B}||\hat{\mathcal{B}}|$. In practice, $|\mathcal{C}| \ll |\mathcal{B}|$ and this results in major computational savings. Second, quantization and privacy-preserving optimization are done separately. Therefore, any quantization method can be easily combined with our approach. In particular, we can minimize the quantization error in the quantization step, and then our privacy mechanism guarantees the optimal mapping in terms of additional distortion. It should be noted that the distance used in the quantization phase and the distortion

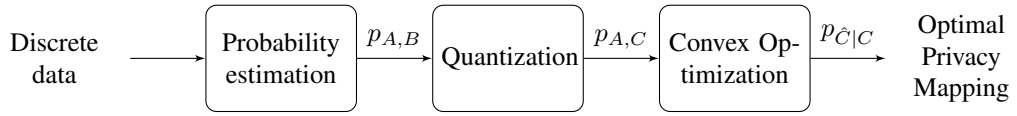


Fig. 1: The quantization approach for large alphabets

function in the constraint of the privacy mapping optimization need not be the same. In the case where they differ, the end-to-end distortion can be obtained by first computing the value of the distortion function for the representative points resulting from quantization, and then adding this value to the distortion generated by the privacy mapping. Finally, quantization obviously yields a suboptimal privacy-accuracy tradeoff, since the quantization step is an additional source of distortion. However, in Theorem 3, we quantify how quantization affects the privacy-accuracy tradeoff, and show that the levels of privacy that can be achieved are not affected, but come at the expense of a bounded amount of distortion.

In the rest of this section, we analyze Algorithm 2, which essentially solves the following variant of problem (7):

$$\begin{aligned} \underset{p_{\hat{C}|C}}{\text{minimize}} \quad & J(q_{A,C}, p_{\hat{C}|C}) \\ \text{subject to:} \quad & \mathbb{E}_{p_{C,\hat{C}}}\left[d(C,\hat{C})\right] \leq \Delta \\ & p_{\hat{C}|C} \in \text{Simplex}; \end{aligned} \tag{11}$$

where alphabets $\mathcal{B}$ and $\hat{\mathcal{B}}$ are substituted for alphabets $\mathcal{C}$ and $\hat{\mathcal{C}}$, and the joint distribution over A and C is defined as

$$q_{A,C}(a,c) = \sum_{b \sim c} p_{A,B}(a,b), \tag{12}$$

where $b \sim c$ means that the symbol b is in the cluster represented by center c. The above equation aggregates the probability mass of all symbols in the cluster in its center. The symbols in $\mathcal{B}$ are mapped to $\hat{\mathcal{C}}$ according to

$$p_{\hat{C}|B}(\hat{c} \mid b) = q_{\hat{C}|C}(\hat{c} \mid \psi(b)), \tag{13}$$

where $\psi : B \to C$ is a function that maps a symbol in $\mathcal{B}$ to a cluster center in $\mathcal{C}$. Note that the probability distributions that are associated with optimization (20) are marked by q. We now prove our main claim.

Theorem 3. *Let $q_{\hat{C}|C}$ be a solution to problem* (20) *and $p_{\hat{C}|B}$ be the corresponding mapping from $\mathcal{B}$ (Equation 13). Moreover, let $\mathcal{C}$ be an alphabet such that $\max_{b \in \mathcal{B}} \min_{c \in \mathcal{C}} d(b,c) \leq r$. Then the privacy leakage $J(p_{A,B}, p_{\hat{C}|B})$ of the mapping $p_{\hat{C}|B}$ is equal to the value of the objective function of (20):*

$$J(p_{A,B}, p_{\hat{C}|B}) = J(q_{A,C}, q_{\hat{C}|C}),$$

and its total distortion rate is no more than r larger than the target Δ:

$$\mathbb{E}_{p_{B,\hat{C}}}\left[d(B,\hat{C})\right] \leq \Delta + r.$$

Proof: The information-leakage equality can be proved as follows. First, both $J(p_{A,B}, q_{\hat{C}|B})$ and $J(q_{A,C}, q_{\hat{C}|C})$ can be rewritten as

$$J(p_{A,B}, q_{\hat{C}|B}) = H(p_A) + H(p_{\hat{C}}) - H(p_{A,\hat{C}}) \tag{14}$$

$$J(q_{A,C}, q_{\hat{C}|C}) = H(q_A) + H(q_{\hat{C}}) - H(q_{A,\hat{C}}), \tag{15}$$

where

$$p(a,\hat{c}) = \sum_b q(\hat{c}|\psi(b)) p(a,b) \tag{16}$$

$$q(a,\hat{c}) = \sum_c q(\hat{c}|c) q(a,c). \tag{17}$$

Second, note that

$$\begin{aligned} p(a,\hat{c}) &= \sum_b q(\hat{c}|\psi(b)) p(a,b) \\ &= \sum_c q(\hat{c}|c) \sum_{b \sim c} p(a,b) \\ &= \sum_c q(\hat{c}|c) q(a,c) \\ &= q(a,\hat{c}). \end{aligned} \tag{18}$$

The two distributions are identical, thus $H(p_{A,\hat{C}}) = H(q_{A,\hat{C}})$. An analogous result holds for the entropies of the marginals. As a result, the privacy leakage of the mapping $q_{\hat{C}|B}$ on $\mathcal{B}$ is equal to the privacy leakage of the mapping $q_{\hat{C}|C}$ on $\mathcal{C}$.

The distortion inequality is proved as follows. (13) implies

$$\begin{aligned} q_{B,\hat{C}}(b,\hat{c}) &= \sum_a q_{\hat{C}|B}(\hat{c}|b) p_{A,B}(a,b) \\ &= \sum_a q_{\hat{C}|C}(\hat{c}|\psi(b)) p_{A,B}(a,b). \end{aligned} \tag{19}$$

Based on this equality, we can bound the distortion as

$$\begin{aligned} \mathbb{E}_{q_{B,\hat{C}}}\left[d(B,\hat{C})\right] &= \sum_{b,\hat{c}} q(b,\hat{c}) d(b,\hat{c}) \\ &= \sum_{a,b,\hat{c}} q(\hat{c}|\psi(b)) p(a,b) d(b,\hat{c}) \\ &= \sum_{a,c,\hat{c}} q(\hat{c}|c) \sum_{b \sim c} p(a,b) d(b,\hat{c}) \\ &\leq \sum_{a,c,\hat{c}} q(\hat{c}|c) \sum_{b \sim c} p(a,b) [d(b,c) + d(c,\hat{c})] \\ &= \sum_{a,c,\hat{c}} q(\hat{c}|c) \underbrace{\sum_{b \sim c} p(a,b)}_{q(a,c)} d(c,\hat{c}) + \\ &\quad \sum_{a,c} \underbrace{\sum_{\hat{c}} q(\hat{c}|c)}_{1} \sum_{b \sim c} p(a,b) d(b,\psi(b)) \\ &\leq \mathbb{E}_{q_{C,\hat{C}}}\left[d(C,\hat{C})\right] + r \sum_{a,b} p(a,b) \\ &\leq \Delta + r. \end{aligned}$$

Algorithm 2 Quantized privacy mapping design.

Input: prior $p_{A,B}$
for all $(a, c) \in (\mathcal{A}, \mathcal{C})$ **do**
$q_{A,C}(a, c) \leftarrow \sum_{b \sim c} p_{A,B}(a, b)$
end for
solve the convex optimization problem over $p_{\hat{C}|C}$:

$$\begin{aligned} \underset{p_{\hat{C}|C}}{\text{minimize}} \quad & J(q_{A,C}, p_{\hat{C}|C}) \\ \text{subject to} \quad & \mathbb{E}_{p_{C,\hat{C}}}\left[d(C, \hat{C})\right] \leq \Delta \\ & p_{\hat{C}|C} \in \text{Simplex}; \end{aligned}$$

return optimal solution $q_{\hat{C}|C}$
for all $(b, \hat{c}) \in (\mathcal{B}, \hat{\mathcal{C}})$ **do**
$p_{\hat{C}|B}(\hat{c}|b) \leftarrow q_{\hat{C}|C}(\hat{c}|\psi(c))$
end for
Output: mapping $p_{\hat{C}|B}$

■

Theorem 3 states that the information leakage of the mapping $p_{\hat{C}|B}$ is the same as that of the optimized mapping $q_{\hat{C}|C}$. So we optimize the quantity of interest $J(p_{A,B}, p_{\hat{C}|B})$ in a time which is independent of the size of the input alphabet $\mathcal{B}$. The total distortion increases due to quantization, linearly with the maximum distance r between any example b and its closest representative example $\psi(b)$.

The maximum distance r can be minimized by existing quantization techniques, e.g. online k-center clustering [25] and cover trees [26]. Both methods quantize data nearly optimally. In particular, if the minimum quantization error by $|\mathcal{C}|$ examples is r^*, then the maximum error produced by these methods is $8r^*$. Note that finding $|\mathcal{C}|$ examples that minimize the quantization error is NP hard.

VI. Datasets

In order to evaluate our framework, we apply it to two datasets. The first one, the *Census* data is a well-known publicly available dataset used by the machine-learning community. The other one, called *Politics-and-TV*, is a dataset on political convictions and TV preferences, that we collected by conducting a survey, as explained in Section VI-B. In this paper, datasets (politics and TV, census) are used to estimate the prior distribution $p_{A,B}$. Once this estimate is available, we study how to design a privacy mapping that can be applied to the data of an individual privacy-conscious user. We use the *Census* dataset to illustrate the basic performance of Algorithm 1. We evaluate Algorithm 2 on the *Politics-and-TV* datasets. The *Politics-and-TV* data has a high-dimensional alphabet and thus allows us to evaluate how quantization influences our ability to provide privacy. We present the optimal privacy-accuracy curve for each case, and give some insights on the privacy mappings.

A. Census Dataset

The *Census* dataset is a well studied dataset in the Machine Learning community ([36] and references therein). Based on the 1994 Census, the dataset is a sample of the United States population, and contains both categoric and numerical features. More precisely, for each entry in the dataset, there are features such as age, workclass, education, gender, and native country, as well as income category (smaller or larger than 50k per year). For our purposes, we consider the information to be released publicly as the education, gender, and age, while the income category is the private information to be protected. It is noteworthy to know that about 76% of the people in the dataset have an income smaller than 50k.

Our privacy mechanism in this case uses erasures. Erasure policies are ones in which we advise a user how to modify their public profile before it is released, by erasing 1, 2 or 3 pieces of information, in order to make it hard to infer income category. The suggestion is tailored to each individual.

The joint probability distribution $p_{A,B}$ is estimated over the available data. Because of the discrete nature of the data, the low dimension of the feature space considered, and the large number of available observations (about 50,000 entries), the joint distribution can be estimated easily with very high confidence. In this case, there is essentially no prior mismatch.

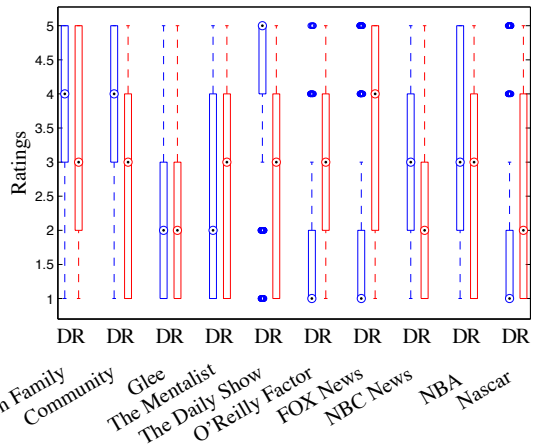


Fig. 2: Box plots of ratings for 12 TV shows by Democrats (D) and Republicans (R)

B. Politics and TV Dataset

The *Politics-and-TV* dataset gathers data on political convictions and TV preferences of viewers in the USA in Fall 2012. The collection of such data was motivated by large scale surveys such as [37], [38], which illustrated that the audiences for a number of TV shows can be distinctly characterized. Opinion polls have also published articles in the press with lists of top-10 or 20 TV shows that are most indicative of political affiliation. For example, *The Colbert Report* is predominantly watched by Democrats, whereas *Fox News* and *Swamp Loggers* are primarily watched by Republicans. We thus started from the premise that it is possible to use public information about a user's TV preferences, such as the list and ratings of TV shows he watches, to infer some private information, namely political convictions. It should be noted that fewer than 1% of Facebook users disclose their

political views in their public profile, which seems to indicate that political convictions are deemed private information. We describe hereafter the data collection process, and our dataset.

Data Collection: We designed a survey that users take voluntarily. In our survey, users were first asked to provide demographic information (gender, age group, state they live in) as well as their political views (Democrat, Republican). Then users were asked to complete a sequence of 6 panels, each panel presenting the user with 6-8 TV shows of a certain genre, namely Sitcoms, Reality Shows, TV series, Talk Shows, News, and Sports, for a total of 50 TV shows. Users were asked to rate only those TV shows that they watched on a scale from 1 to 5— the usual star rating system. After providing their ratings, users were shown, for each genre, how their ratings compared with the average ratings given by Democrats and Republicans. In our privacy policy, users were informed that no private information that can be used to identify an individual was stored— we did not store cookies, nor IP addresses, etc. Thus the data collected is by consenting users.

We ran our survey in two phases. In phase 1 (October 2012), we ran it on Mechanical Turk requesting only US-based workers. An initial experiment revealed that 80% of users completing the survey were Democrats. To diminish this bias, we reran the survey in two batches. For the first batch, we limited the user pool to Democrats only, and in the second batch we limited it to Republicans only. This mechanism helped although it still did not produce equal numbers of Democrats and Republicans. In total, we obtained 854 surveys, with 518 Democrats and 336 Republicans. In phase 2 (November 2012), we launched our survey on the public web at www.PoliticsandMedia.org. We drove traffic to the survey website by running advertising campaigns on MyLikes.com and Google AdWords, shortly before the U.S. 2012 presidential election. From this, we obtained another 364 completed surveys, with 226 Democrats and 138 Republicans. We conducted this survey in two places (Mechanical Turk and the Web) to create more diversity of users in our survey. An advantage of the Mechanical Turk approach is that users are incentivized to properly complete the survey. We threw out surveys which were clearly never finished, e.g. no ratings, and the numbers above reflect the final retained surveys.

Dataset: The dataset contains entries for 1,218 users, broken into 744 Democrats, and 474 Republicans. For each user, the entry is a vector $[\text{age}, \text{gender}, \text{state}, \text{politics}, r_1, \ldots ...r_{50}]$ where $r_i \in \{0, 1, \ldots, 5\}$ is the user's star rating for show i if the user rated the show, and 0 otherwise. The 5 most watched TV shows are *The Daily Show with Jon Stewart, The Colbert Report, NFL, The Big Bang Theory,* and *Family Guy*. In the sequel, we will consider two versions of the rating vector: the 5-star rating vector $R \in \{0, 1, \ldots, 5\}^{50}$, and the binarized rating vector $B \in \{0, 1\}^{50}$. The binarized rating b_i of show i is obtained by setting $b_i = 1$ if the original rating $r_i >= 4$ (the user likes the show), and $b_i = 0$ otherwise.

VII. Results

A. Baseline Convex optimization and mismatched priors on Census

We demonstrate here a direct application of the convex optimization approach Algorithm-1 described earlier on the Census dataset. This can be seen as a simple application as we do not need to apply a quantization step. For this dataset, we will use the *erasure-distortion* approach meaning that our proposed distortion to an individual's public data (age, education, and gender) may be to remove a subset of features. In this way, we distort without lying, and our distortion metric is the number of erasures.

Formally, let $B(u) = (b_1, b_2, b_3, a)$ be the features of user u, where $b_1 \in \{male, female\}$, $b_2 \in \{$*young, adult, old*$\}$ and $b_3 \in \{$*high-school, college degree, master degree, doctorate*$\}$. The feature a is the private attribute defined as $a \in \{high, low\}$ where high/low refers to an income above/below 50K$ respectively. In this case the output alphabet $\hat{\mathcal{B}}$ after the privacy mapping is larger than the input alphabet $\mathcal{B}$ as each feature can be replaced by an erasure. Because of the mapping restriction $p_{\hat{b}|b}$ can have non zero values if b and $\hat{b}$ differ only in positions where $\hat{b}$ has an erasure. We define the distortion metric $d(\hat{b}, b)$ as the number of erasures in $\hat{b}$, when b and $\hat{b}$ match in non-erasure positions and $d(\hat{b}, b) = \infty$ otherwise.

We have tested the algorithm for different distortion constraint values and obtained the black privacy-distortion curve shown in Fig. 4. The y-axis captures the privacy leakage measured by the mutual information. The x-axis quantifies the distortion in terms of average number of erasures. Without any distortion (0 erasures), the privacy leakage, or mutual information, is 0.15 bits. If, on average, we erase one of the three features in these user profiles, then the privacy leakage drops to roughly 0.025 bits. This can be interpreted as requiring an adversary to ask many more questions in order to learn the private information. Perfect privacy (mutual information is zero) is obtained when the expected erasures is 1.5 features (out of three). This confirms that gender, age and education are related to one's income.

To illustrate the impact of mismatched priors in Fig. 4, we have also generated estimates $q_{A,B}$ of the true prior $p_{A,B}$ by using only a subset of the data, resp. 1%, 10%, 50% and 80% of the available data. For each prior estimate, we can generate mappings $q^*_{\hat{B}|B}$. In Fig. 4, the dashed curves correspond to the actual leakage $J(p_{A,B}, q^*_{\hat{B}|B})$. Except the 1% samples estimate, the privacy-utility curves that we obtain are very close to the optimal, and increasingly close as we improve the quality of the estimate by taking more and more samples. This shows the stability of the optimization problem against small variations of the prior input, and demonstrates how nearly optimal mappings can be found even though the estimated prior is not perfect. On the other hand, if the estimated prior is too poor, it is hard to give any guarantees on the nearly-optimality, as it can be seen from the 1% dashed curve. Yet, the privacy leakage is still decreasing as a function of the available distortion, which is the desired behavior.

		Original features			Private mapping		
1	< 50k	male	young	College	male	-	-
2	N	male	adult	College	-	adult	College
3	< 50k	male	young	HS	male	-	HS
4	N	male	adult	HS	male	-	HS
5	< 50k	female	young	HS	-	-	-
6	< 50k	female	young	College	-	-	-
7	> 50k	male	adult	Masters	male	-	-
8	< 50k	female	adult	College	-	adult	College

Fig. 3: Most probable mapping for the Top 8 categories in the Census Dataset. Initially some set of attributes may be highly correlated with income (denoted by $< 50k$ and $> 50k$), or be more neutral (denoted by N). HS stands for High School degree.

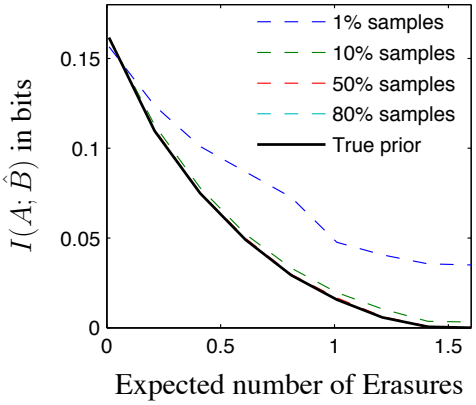


Fig. 4: Census data: Privacy Distortion curve, and impact of mismatched priors

To provide more insights on the privacy mapping, we have represented some specific cases of mappings in Fig. 3. We see that different inputs features are mapped to the same output, *e.g.*, rows 1 and 7 are both mapped to 'male' with two erasures. By observing the latter, the adversary cannot determine whether the original features were those from row 1 or those from row 7— this creates confusion on the income.

B. Mismatched prior and quantization on Politics-and-TV

We demonstrate here a more realistic privacy preservation application over the Politics-and-TV dataset described earlier.

Consider the setting where a user wishes to release his TV show ratings $R \in \{0, 1, \ldots, 5\}^{50}$ (or $B \in \{0, 1\}^{50}$), in the hope of getting good recommendations, but is concerned about them leaking information about his political affiliation $A \in \{\text{Democrat}, \text{Republican}\}$. Note that although we focus on the case where the private data is a single variable representing political affiliation, the privacy-accuracy framework [7] can handle protecting a set of private variables, e.g. we could protect any subset of a user's three attributes [age,gender, politics]. The rating vector R (resp B) lives in a large alphabet of size 6^{50} (resp. 2^{50})[2]. Solving Algorithm 1 over 6^{100} variables would be intractable, and justifies resorting to quantization. In this section, we first describe the privacy threat on political affiliation from the release of TV show ratings, then we characterize the privacy-accuracy trade-off under quantization. We illustrate the success of our privacy approach by showing how an inference algorithm degrades down to an uninformed guess at perfect privacy. Finally, we compare the quality of recommendations based on the actual user ratings versus the privatized ratings.

Privacy threat: The threat comes from the underlying existence of TV shows that are highly correlated with political affiliation, e.g. *The Daily Show* is predominantly liked by Democrats, while *Fox News* is preferred by Republicans. Fig. 2 shows boxplots of ratings for 12 shows—two shows from each genre in the dataset— by Democrats and Republicans.

[2]The number of survey samples is small relative to the size of the alphabet. Estimating the prior $p_{A,R}$ from the dataset may lead to a mismatched prior.

Those shows for which there is little overlap in the opinions of Republicans and Democrats clearly demonstrate high correlation between political affiliation and opinion of those shows. Such shows have high discriminative power that inference algorithms can exploit. There exists a broad variety of shows in terms of their discriminative power - some are very much so, while others exhibit low correlation. Users who rate highly shows such as *The O'Reilly Factor*, or *The Daily Show*, may be facing a stronger threat than those who only watch and rate shows with little discriminating power. Broadly speaking, across our 50 shows, we found that roughly one third of them have strong correlation with political affiliation.

To understand the inherent threat in this dataset, we quantify the potential privacy leakage using mutual information $I(A; R)$. To provide an illustrative example, we consider a reduced set of our data for which we can compute the mutual information. We consider the top 5 most seen TV shows, and use the binarized version of the rating vector with ratings in $\{0, 1\}^5$. For this case, we observe that the mutual information between the observed features and the political orientation is already at 0.191 bits. An adversary, with this information on hand (the 5 tuple of binarized ratings), could use a maximum a posteriori (MAP) detector and guess the political affiliation of somebody with an accuracy of 71%. Hence, the privacy threat is real. Note that because mutual information is a non-decreasing function, as we add additional shows, the threat either stays the same or increases.

Privacy-accuracy trade-off: We now apply our quantization approach and investigate its impact on the privacy-distortion tradeoff. We first consider the full dataset (all 50 shows) with the binarized version of the ratings. For this scenario we use an *exchange-distortion* in which we exchange on TV show for another. We use Algorithm 2 that first quantizes the data using a clustering algorithm (K-means with a Hamming distance) into 25 clusters; then we apply the convex optimization on the quantized points. The resulting tradeoff is depicted by the blue curve in Fig. 5. The curve shows that the quantization step alone introduces an average Hamming distortion of about 12% (leftmost point on x-axis) per rating, or 6.1 over all 50 shows, and results in a mutual information of 1.6 bits on the

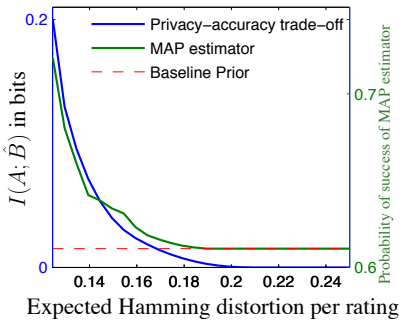


Fig. 5: Politics & TV data: Privacy-accuracy trade-off on binarized ratings after quantization.

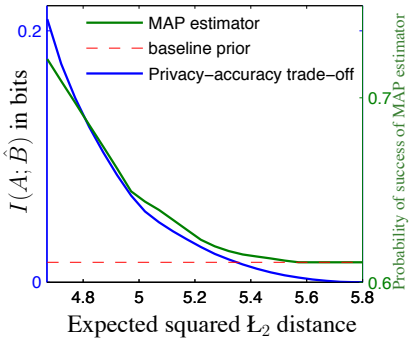


Fig. 6: Politics & TV data: Privacy-accuracy trade-off on actual ratings after quantization.

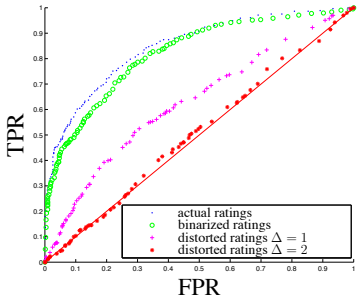


Fig. 7: Politics & TV data: ROC curve of a logistic regression classifier for the political views based on TV show ratings

representative points (cluster centers). As this is still high, we are motivated to apply further distortion. Fig. 5 shows that using the optimal privacy mapping resulting from convex optimization, we can steadily decrease the privacy threat with increasing distortion. Not only is our privacy-distortion curve properly behaved, but small increases in Hamming distance bring the privacy leakage down quickly. Moreover, we can achieve perfect privacy ($I = 0$) at the cost of an additional 6% in average Hamming distortion (beyond the clustering distortion). Perfect privacy is achieved at an overall Hamming Distortion of less than 10 out of 50; put alternatively, perfect privacy is obtainable if on average we change just less than 20% of a user's rating data before it is released.

We next consider the same tradeoff using the version of our data with the actual ratings. We use K-means clustering with squared L2 distance, and the results are given by the blue curve in Fig 6. We cannot calculate the original mutual information because we do not know the distribution of actual ratings (the number of unique rating vectors is too large compared to the size of our data set), but the mutual information after quantization is 0.201. The quantization step introduces an average squared L2 of 4.66. Using the actual ratings requires slightly higher distortion to reach perfect privacy than with binarized ratings. In this case, we are able to achieve perfect privacy with an extra squared L2 distortion of about 1.2.

In Table I, some privacy mappings are represented. Recall that mappings go from clusters, to clusters, therefore we chose to characterize each cluster by the top three TV shows for people from that cluster. The results are intuitive, and we see for example that some clusters associated with politically neutral shows (such as the Family Guy, NFL, Dexter cluster) are mostly mapped to themselves. On the other hand, some cluster are associated with political shows (For example the third cluster in Table. I contains Daily Show and Colbert Report), and are therefore mapped to other clusters. Because of the distorition constraint, it is mapped with highest probability to a cluster which shares some neutral TV shows (For example the thrid cluster is mapped to a neutral cluster NFL, Dexter, Family Guy which shares Family Guy with the initial cluster).

Inference defeat: The previous plots show the reduction in privacy leakage that is achieved by our distortion. Another key performance metric is to examine how much the accuracy of a Democrat/Republican classifier is reduced when distorted ratings are used instead of the non-distorted ones. We consider the example of a logistic regression classifier to infer political affiliation (similar to the one used in [4] to infer gender from movie ratings). We used 10-fold cross validation on our full dataset, considered both cases of actual and binarized ratings, and a distortion that achieves perfect privacy ($I = 0$). After perturbing the ratings to reach $I = 0$, any inference algorithm cannot perform better than an uninformed guess. In Fig. 7 we plot the false positive rate, the number of Democrats falsely classified as Republicans, against the true positive rate, the number of Republicans who are correctly classified. With a distortion bound of $\Delta = 1$, we see that we can significantly reduce the classifier's performance but not yet reach perfect privacy; however with $\Delta = 2$ the classifier is reduced to nothing more than an uninformed classifier. This demonstrates that our approach can indeed successfully render inference attempts useless. Finally, note that logistic regression also performs almost equally well with binarized and actual ratings, which means merely perturbing existing ratings is not enough. The adversary can ignore the actual rating values, consider only binarized ratings, and classify almost equally well on whether or not a user rated a show. Therefore, we must add and/or delete ratings to protect privacy.

In addition to the ROC curve, we also depict the evolution of the probability of success of a MAP estimator, as we increase distortion, by the green curves in Fig. 5 and Fig. 6. More precisely, the MAP estimator tries to infer the Political view from the distorted ratings. We see that the probability of success of this estimator decreases as we introduce more and more distortion, starting around 72% success rate, both in the binarized case in Fig. 5, and in the full ratings case in Fig. 6, and attains the baseline 61% of an uninformed random guess, which precisely corresponds to the ratio of democrats in the dataset. Indeed, an uninformed random guess that always outputs Democrat (most probable political views given the prior distribution) as the result of its inference, regardless of the ratings, always yields a 61% success rate.

Recommendation quality: As a final performance metric, we consider the impact of our distortion on the recommendations

	Initial Cluster			Privacy Mapping		
N	Family Guy	NFL	Dexter	Family Guy	NFL	Dexter
R	FOX News	NFL	O'Reilly Factor	Daily Show	NFL	Colbert Report
D	Daily Show	Colbert Report	Family Guy	NFL	Dexter	Family Guy
D	Daily Show	Colbert Report	Dexter	Daily Show	Colbert Report	Dexter
N	Modern Family	The Big bang theory	CNN	Modern Family	The Big bang theory	CNN

TABLE I: Some privacy mappings from clusters to clusters. Each row is a cluster by the 3 most seen TV shows for people within that cluster. Initially, some cluster may be highly correlated with a political affiliation (denoted by D and R), or may be more neutral (denoted by N) in the sense that the distribution of democrats and republicans in the cluster is close to the base distribution in the dataset.

that would be produced by a recommender system based on matrix factorization (MF) [34].Using 5-fold cross validation, we split the rating dataset into a training set containing 80% of the data, representing non-privacy conscious users on which the MF recommendation system is trained using alternating least squares [34]; and a test set containing 20% of the data, representing privacy-conscious users on which the recommendation system is tested both on the original ratings and on the distorted ratings. Our goal is to compare the relevance of recommendations when the privacy-conscious user does not distort his ratings versus the case where he distorts his ratings using the privacy mapping. The random splitting into training and test sets is done 5 times as shown by the columns of Table II. The MF recommendation system works by trying to predict the ratings of TV shows that have not been rated by the user, from the ratings of TV shows that have been rated by the user. In practice, in the testing phase, we first randomly set aside a subsest $\mathcal{I}$ of 10% of the ratings in the test set, and try to predict them using the recommendation system. The prediction is done twice: once based on the original ratings, and once based on the distorted ratings. The quality of the recommendation is measured by the Root Mean Square Error (RMSE) in rating prediction, defined by $RMSE = \sqrt{\sum_{i\in\mathcal{I}}(\hat{r}_i - r_i)^2}$, where r_i denotes the true rating for TV show i, whereas $\hat{r}_i$ denotes the predicted rating. In Table II, RMSE1 is computed for ratings predicted based on original ratings, whereas RMSE2 computed for ratings predicted based on distorted ratings. We can see that the degradation in the RMSE of rating prediction, due to the use of the privacy mapping to distort ratings, is small. Recently, in [20], even better results were obtained by coupling quantization with dimensionality reduction, prior to distortion.

VIII. Conclusion

Privacy attacks are receiving more and more attention, both from a theoretical perspective, and from a practical point of view. The amount of information shared everyday, and the recent improvements in inference models have brought in the attention of all, the urge for effective yet private systems. This fundamental contradiction is the core of the privacy problem. In this paper, we show a practical approach to privacy that has roots in a strong theoretical framework. We show that is possible to have private systems by adding a layer of privacy, without changing the way the data is processed afterwards, or its purpose. Using techniques from different fields, such as information theory, convex optimization, estimation, and quantization, we address some challenges introduced by the diversity and complexity of real world data. Namely we show that a mismatched prior estimation does not hurt too much in terms of privacy-distortion tradeoff. Moreover, we propose a generic methodology to deal with large data through quantization. We show that distortion grows linearly in the quantization error, and that the privacy leakage remains controlled.

TABLE II: Rating prediction RMSE

Set	1	2	3	4	5
RMSE1	1.2506	1.1820	1.2461	1.2155	1.2101
RMSE2	1.6972	1.6763	1.6215	1.7248	1.8036

Appendix A
Proof of Theorem 1

The following lemma [29], which bounds the difference in the entropies of two distributions, will be useful in the proof of the Theorems.

Lemma 1 ([29, Thm 17.3.3]). *Let p and q be distributions with the same support $\mathcal{X}$ such that $\|p-q\|_1 \leq \frac{1}{2}$. Then:*

$$|H(p) - H(q)| \leq \|p-q\|_1 \log \frac{|\mathcal{X}|}{\|p-q\|_1}.$$

Proof of Theorem 1: The first inequality can be proved in four steps. Initially, we note that the objective function can be rewritten as

$$J(p_{A,B}, p_{\hat{B}|B}) = H(p_A) + H(p_{\hat{B}}) - H(p_{A,\hat{B}}). \tag{20}$$

Therefore, the difference between the objective functions with respect to $p_{A,B}$ and $q_{A,B}$ is bounded as:

$$\begin{aligned} &\left|J(p_{A,B}, p_{\hat{B}|B}) - J(q_{A,B}, p_{\hat{B}|B})\right| \\ &\quad\leq |H(p_A) - H(q_A)| + \\ &\quad\quad |H(p_{\hat{B}}) - H(q_{\hat{B}})| + \\ &\quad\quad |H(p_{A,\hat{B}}) - H(q_{A,\hat{B}})|. \end{aligned} \tag{21}$$

The bound in Lemma 1 can be used to bound each of the

terms in Equation (21). For instance:

$$\begin{aligned}\left\|p_{A,\hat{B}} - q_{A,\hat{B}}\right\|_1 &= \sum_{a,\hat{b}} \left|\sum_b p(\hat{b}|b)[p(a,b) - q(a,b)]\right| \\ &\leq \sum_{a,b,\hat{b}} p(\hat{b}|b)\,|p(a,b) - q(a,b)| \\ &= \sum_{a,b} \underbrace{\sum_{\hat{b}} p(\hat{b}|b)}_{1}\,|p(a,b) - q(a,b)| \\ &= \|p_{A,B} - q_{A,B}\|_1 \end{aligned} \tag{22}$$

and therefore:

$$\begin{aligned}&|H(p_{A,\hat{B}}) - H(q_{A,\hat{B}})| \\ &\quad\leq \|p_{A,B} - q_{A,B}\|_1 \log \frac{|\mathcal{A}|\,|\mathcal{B}|}{\|p_{A,B} - q_{A,B}\|_1}.\end{aligned} \tag{23}$$

Similarly, it can be shown that:

$$\begin{aligned}&|H(p_A) - H(q_A)| \\ &\quad\leq \|p_{A,B} - q_{A,B}\|_1 \log \frac{|\mathcal{A}|}{\|p_{A,B} - q_{A,B}\|_1}\end{aligned} \tag{24}$$

$$\begin{aligned}&|H(p_{\hat{B}}) - H(q_{\hat{B}})| \\ &\quad\leq \|p_{A,B} - q_{A,B}\|_1 \log \frac{|\mathcal{B}|}{\|p_{A,B} - q_{A,B}\|_1}.\end{aligned} \tag{25}$$

Finally, the three upper bounds can be substituted into Equation (21), which yields:

$$\begin{aligned}&\left|J(p_{A,B}, p_{\hat{B}|B}) - J(q_{A,B}, p_{\hat{B}|B})\right| \\ &\quad\leq 3\,\|p_{A,B} - q_{A,B}\|_1 \log \frac{|\mathcal{A}|\,|\mathcal{B}|}{\|p_{A,B} - q_{A,B}\|_1}.\end{aligned} \tag{26}$$

Our first claim is proved by substituting $p^*_{\hat{B}|B}$ for $p_{\hat{B}|B}$ in the above equation.

The proof of our second claim is based on the inequality:

$$\begin{aligned}&\left|\mathbb{E}_{p_{\hat{B},B}}\left[d(\hat{B},B)\right] - \mathbb{E}_{q_{\hat{B},B}}\left[d(\hat{B},B)\right]\right| \\ &= \left|\sum_{a,b,\hat{b}} p(\hat{b}|b)[p(a,b) - q(a,b)]d(b,\hat{b})\right| \\ &\leq \sum_{a,b,\hat{b}} p(\hat{b}|b)d(b,\hat{b})\,|p(a,b) - q(a,b)| \\ &\leq d_{\max} \sum_{a,b} \underbrace{\sum_{\hat{b}} p(\hat{b}|b)}_{1}\,|p(a,b) - q(a,b)| \\ &= d_{\max}\,\|p_{A,B} - q_{A,B}\|_1\,.\end{aligned} \tag{27}$$

Based on this observation, it follows that:

$$\begin{aligned}\mathbb{E}_{p_{\hat{B},B}}\left[d(\hat{B},B)\right] &\leq \mathbb{E}_{q_{\hat{B},B}}\left[d(\hat{B},B)\right] + \\ &\qquad d_{\max}\,\|p_{A,B} - q_{A,B}\|_1 \\ &\leq \Delta + d_{\max}\,\|p_{A,B} - q_{A,B}\|_1\,.\end{aligned} \tag{28}$$

The last step is due to the constraint $\mathbb{E}_{q_{\hat{B},B}}\left[d(\hat{B},B)\right] \leq \Delta$ that is enforced in our problem (7). ■

APPENDIX B
PROOF OF THEOREM 2

First, let us introduce some useful notation. Consider the optimization problem 7, and denote by $R(p_{A,B}, \Delta)$ the optimal privacy leakage for input $p_{A,B}$ and distortion constraint Δ. We also denote by $\mathcal{A}(\Delta)$ the set of feasible mappings, *i.e.*, $\mathcal{A}(\Delta) = \left\{p_{\hat{B}|B} : \mathbb{E}_{B,\hat{B}}[d(B,\hat{B})] \leq \Delta\right\}$. The following lemma is useful in the proof of Thm. 2, and allows us to construct distributions that are close in a $\mathcal{L}_1$ sense but have specific expected distortions.

Lemma 2. *Let q be a distribution over $\mathcal{X}$ such that $\mathbb{E}_q[f] = \delta$, with f a non-negative function. For any $\delta > 0$, there exist a distribution p over the same support, such that $\mathbb{E}_p[f] = 0$ and $\|q - p\|_1 \leq \frac{2\delta}{f_{min}}$, where $f_{min} = \min_{x, f(x)>0} f(x)$ is the smallest non-zero value of f.*

Proof: We do the proof by construction. Consider p such that for all $x \in \mathcal{X}$ with $f(x) > 0$, let $p(x) = 0$. For all other $x \in \mathcal{X}$, set $p(x) = q(x) + \frac{\sum_{x\in\mathcal{X}, f(x)>0} q(x)}{|x\in\mathcal{X}: d(x)>0|}$, where the second term corresponds to adding uniformly the missing mass so that $\sum_x p(x) = 1$. We have:

$$\|p - q\|_1 \leq \sum_{x\in\mathcal{X}, f(x)>0} |p(x) - q(x)| \tag{29}$$

$$+ \sum_{x\in\mathcal{X}, f(x)=0} |p(x) - q(x)| \tag{30}$$

$$= 2 \sum_{x\in\mathcal{X}, f(x)>0} q(x) \tag{31}$$

Next, we have that:

$$\delta = \mathbb{E}_q[f] = \sum_{x\in\mathcal{X}} f(x)q(x) \tag{32}$$

$$\geq f_{\min} \sum_{x\in\mathcal{X}, f(x)>0} q(x) \tag{33}$$

$$\geq \frac{f_{\min}}{2}\|p - q\|_1 \tag{34}$$

where (34) folows from (31). Noticing that $\mathbb{E}_p[f] = 0$ gives the desired result. ■

Proof of Theorem 2: Recall that we denote by $R(p_{A,B}, \Delta)$ the result of the optimization problem (7) with input $p_{A,B}$ and distortion constraint Δ, and that we use $\mathcal{A}(\Delta)$ to denote the feasible region of this optimization problem. We use $\epsilon = \|p - q\|_1$. Our goal is to bound $|R(p_{A,B}, \Delta) - R(q_{A,B}, \Delta)|$. We have:

$$R(p_{A,B}, \Delta + \epsilon d_{\max}) \leq J(p_{A,B}, q^*_{\hat{B}|B}) \tag{35}$$

$$\begin{aligned}&\leq J(q_{A,B}, q^*_{\hat{B}|B}) + |J(p_{A,B}, q^*_{\hat{B}|B}) - J(q_{A,B}, q^*_{\hat{B}|B})| \\ &= R(q_{A,B}, \Delta) + |J(p_{A,B}, q^*_{\hat{B}|B}) - J(q_{A,B}, q^*_{\hat{B}|B})|\end{aligned} \tag{36}$$

where (35) follows from the distortion inequality of Thm. 1 which means that $q^*_{\hat{B}|B}$ is in the feasible set $\mathcal{A}(\Delta + \epsilon d_{\max})$. Adding $R(p_{A,B}, \Delta)$ on both sides of (36), and rearranging

terms, we obtain:

$$\begin{aligned} &R(p_{A,B},\Delta) - R(q_{A,B},\Delta) \\ &\leq |J(p_{A,B}, q^*_{\hat{B}|B}) - J(q_{A,B}, q^*_{\hat{B}|B})| \\ &\quad + R(p_{A,B},\Delta) - R(p_{A,B},\Delta + \epsilon d_{\max}) \end{aligned} \tag{37}$$

Notice that the first term of (37) can be bounded using Thm. 1. The second term corresponds to the difference in the solution of the optimization problem when we have expanded the feasible set by allowing an additional distortion $\epsilon d_{\max}$. We have the following cases:

- $p^*_{\hat{B}|B}$ was not on the border of the feasible set $\mathcal{A}(\Delta)$. Then, as the problem is convex, $p^*_{\hat{B}|B}$ is also a minimizing distribution of the optimization problem with expanded feasible set $\mathcal{A}(\Delta + \epsilon d_{\max})$. Therefore, $R(p_{A,B},\Delta) - R(p_{A,B},\Delta + \epsilon d_{\max}) = 0$.
- $p^*_{\hat{B}|B}$ is on the border of the feasible set $\mathcal{A}(\Delta)$. First, notice that $R(p,\Delta)$ is convex in Δ. This can be seen as $\mathbb{E}_{p_{\hat{B},B}}[d(\hat{B},B)]$ is linear and that the mutual information $J(p_{A,B}, p_{\hat{B}|B})$ is convex in $p_{\hat{B}|B}$. Therefore, if we let Δ_1 and Δ_2 be two distortion value, and let p^*_1 and p^*_2 be the respective minimizing distributions, then it is the case that for $p_\alpha = \alpha p^*_1 + (1-\alpha)p^*_2$, with $0 \leq \alpha \leq 1$, we have:

$$\begin{aligned} R(p_\alpha,\Delta) &\leq J(p_{A,B}, p_\alpha) && (38)\\ &\leq \alpha J(p_{A,B}, p^*_1) + (1-\alpha) J(p_{A,B}, p^*_2) && (39)\\ &= \alpha R(p_{A,B},\Delta_1) + (1-\alpha) R(p_{A,B},\Delta_2) && (40) \end{aligned}$$

  As the function $R(p,\Delta)$ is convex and non-increasing with respect to Δ, its steepest descent is at zero, that is :

$$\begin{aligned} &R(p_{A,B},\Delta) - R(p_{A,B},\Delta + \epsilon d_{\max}) \\ &\quad \leq R(p_{A,B},0) - R(p_{A,B},\epsilon d_{\max}) \end{aligned} \tag{41}$$

  Then, by Lemma 2 with $f = d(\hat{B},B)$, and $\delta = \epsilon d_{\max}$, there is a $\tilde{p}_{\hat{B}|B} \in \mathcal{A}(0)$, such that the distance between $\tilde{p}_{\hat{B}|B}$ and the minimizing distribution of the optimization problem with expanded feasible set $\mathcal{A}(\epsilon d_{\max})$ is at most $\epsilon \frac{2d_{\max}}{d_{\min}}$. If $\epsilon \leq \frac{d_{\min}}{4d_{\max}}$, we can use Lemma 1 and equations similar to those in (22) to obtain:

$$\begin{aligned} &R(p_{A,B},\Delta) - R(p_{A,B},\Delta + \epsilon d_{\max}) \\ &\quad \leq 4\epsilon \frac{d_{\max}}{d_{\min}} \log \frac{d_{\min}|\mathcal{A}||\mathcal{B}|}{\epsilon d_{\max}} && (42)\\ &\quad \leq 4\epsilon \frac{d_{\max}}{d_{\min}} \log \frac{|\mathcal{A}||\mathcal{B}|}{\epsilon} && (43) \end{aligned}$$

Using (43) in (37) gives the desired bound.■

REFERENCES

[1] S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, P. Oliveira, and N. Taft, "How to hide the elephant- or the donkey- in the room: Practical privacy against statistical inference for large data," in *IEEE GlobalSIP*, 2013.

[2] A. Narayanan and V. Schmatikov, "Robust de-anonymization of large sparce datasets," in *IEEE Symposium on Security and Privacy*, 2008.

[3] S. Bhagat, U. Weinsberg, S. Ioannidis, and N. Taft, "Recommending with an agenda: Active learning of private attributes using matrix factorization," in *ACM Conference on Recommender Systems*, 2014.

[4] U. Weinsberg and S. Bhagat and S. Ioannidis and N. Taft, " BlurMe: Inferring and Obfuscating User Gender Based on Ratings," in *ACM Conference on Recommender Systems (RecSys)*, September 2012.

[5] M. Kosinski, D. Stillwell, and G. T., "Private traits and attributes are predictable from digital records of human behavior," in *PNAS*, 2013.

[6] "FTC staff report on the workshop "internet of things: Privacy and security in a connected world"," 2015. [Online]. Available: http://www.ftc.gov/reports/federal-trade-commission-staff-report-november-2013-workshop-entitled-internet-things

[7] F. Calmon and N. Fawaz, "Privacy against statistical inference," in *Allerton Conference on Communication, Control, and Computing*, 2012.

[8] S. Salamatian, N. Fawaz, B. Kveton, and N. Taft, "SPPM: Sparse Privacy Preserving Mappings," *Uncertainty in Artificial Intelilgence UAI*, 2014.

[9] J. Fetto, "Top tv shows for reaching key voters," http://www.experian.com/blogs/marketing-forward/2012/08/28/top-tv-shows-for-reaching-key-voters/, 2012.

[10] C. Dwork, F. Mcsherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *TCC*, 2006.

[11] C. Dwork, "Differential privacy," in *Automata, Languages and Programming*. Springer, 2006, vol. 4052, pp. 1–12.

[12] S. L. Warner, "Randomized response: A survey technique for eliminating evasive answer bias," *Journal of American Statistical Association*, 1965.

[13] F. McSherry and I. Mironov, "Differentially private recommender systems: building privacy into the net," in *ACM SIGKDD*, 2009.

[14] D. Kifer and A. Machanavajjhala, "A rigorous and customizable framework for privacy," in *ACM PODS*, 2012, pp. 77–88.

[15] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *ACM PODS*, 2003.

[16] I. S. Reed, "Information Theory and Privacy in Data Banks," in *Proc. of national computer conference and exposition*. ACM, 1973.

[17] H. Yamamoto, "A source coding problem for sources with additional outputs to keep secret from the receiver of wiretappers," *IEEE Trans. Inf. Theory*, vol. 29, no. 6, 1983.

[18] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoff in databases: An information-theoretic approach," *IEEE Trans. Inf. Forensics Security*, 2013.

[19] D. Rebollo-Monedero, J. Forné, and J. Domingo-Ferrer, "From t-closeness-like privacy to postrandomization via information theory," *IEEE Transactions on Knowledge and Data Engineering*, 2010.

[20] A. Zhang, S. Bhamidipati, N. Fawaz, and B. Kveton, "Priview: Media consumption and recommendation meet privacy against inference attacks," *IEEE Web 2.0 Security and Privacy Workshop (W2SP)*, 2014.

[21] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," *ITW*, 2014.

[22] S. Hamadou, V. Sassone, and C. Palamidessi, "Reconciling belief and vulnerability in information flow," in *2010 IEEE Symposium on Security and Privacy (SP)*, 2010, pp. 79–92.

[23] L. Sweeney, "k-anonymity: a model for protecting privacy," in *JUFKS*, 2002.

[24] R. Gray and D. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998.

[25] M. Charikar, C. Chekuri, T. Feder, and R. Motwani, "Incremental clustering and dynamic information retrieval," in *STOC*, 1997.

[26] A. Beygelzimer, S. Kakade, and J. Langford, "Cover trees for nearest neighbor," in *ICML*, 2006.

[27] M. H. DeGroot, "Decision problems," in *Optimal Statistical Decisions*, 2005, pp. 119–154.

[28] N. F. Ali Makhdoumi, "Privacy-Utility Tradeoff under Statistical Uncertainty," in *Allerton Conference on Communication, Control, and Computing*, 2013.

[29] T. M. Cover and J. A. Thomas, *Elements of information theory*. New York, NY, USA: Wiley-Interscience, 1991.

[30] Z. Zhang, "Estimating Mutual Information Via Kolmogorov Distance," *Information Theory, IEEE Transactions on*, vol. 53, no. 9.

[31] K. M. R. Audenaert, "A Sharp Fannes-type Inequality for the von Neumann Entropy," 2006.

[32] H. Palaiyanur and A. Sahai, "On the uniform continuity of the rate-distortion function," *ISIT 2008*, pp. 1–5, jan 2008.

[33] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Cambridge University Press, 2011.

[34] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, Aug 2009.

[35] J. Conway and N. Sloane, *Sphere Packings, Lattices and Groups*. New York, NY: Springer, 1998.

[36] "Census income data set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Census+Income

[37] "What Your Favorite TV Shows And Networks Say About Your Politics," http://www.buzzfeed.com/rubycramer/what-your-favorite-tv-shows-say-about-your-politic, 2012.
[38] "Simmons Consumer Segmentations: PublicPersonas," http://www.experian.com/simmons-research/simmons-consumer-research.html, 2012.

Salman Salamatian Salman Salamatian is a PhD candidate in Electrical Engineering and Computer Science at MIT. Previously he obtained his Bachelor, in 2012, and Masters, in 2014, in the School of Computer and Communication Sciences at EPFL, Switzerland, where he also was a member of the Laboratory of Information Theory (LTHI) in 2014/2015. His research interests spans topics from Information Theory and its application to network problems, to estimation and statistical learning theory, security, and privacy. Email: salmansa@mit.edu

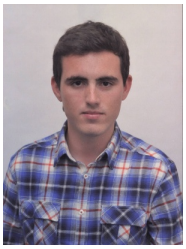
Amy Zhang Amy Zhang is currently a machine learning engineer at SET Media working on video classification. She received a B.S. in Mathematics in 2011 and a B.S. and M.Eng. in Electrical Engineering and Computer Science in 2011 and 2012 from MIT. Her interests are in deep learning, graphical models, and inference. Email: amy@set.tv

Flavio du Pin Calmon Flavio du Pin Calmon is a PhD candidate in Electrical Engineering and Computer Science (with a minor in Mathematics) at MIT, and a member of the Network Coding and Reliable Communications Group at the Research Laboratory of Electronics (RLE). He will join Harvards School of Engineering and Applied Sciences as an Assistant Professor of Electrical Engineering in July/2017. Before coming to MIT, he received an M.Sc. in Electrical Engineering from the Universidade Estadual de Campinas, Brazil, and a B.Sc. in Communications Engineering from the Universidade de Brasilia, Brazil. His research interests include information theory, statistical learning theory, estimation theory, security and privacy. Email: flavio@mit.edu

Sandilya Bhamidipati Sandilya Bhamidipati is currently a Systems Architect at Technicolor. He leads engineering efforts in Content Discovery, Recommendation Systems, User Analytics and User Privacy. His areas of interest include applied machine learning and data mining for large scale systems. Sandilya has a Masters in Computer Science from Rutgers University in 2010 and a Bachelor's in Computer Science and Engineering from JNTU, Hyderabad in 2006. Email: sandilya.bhamidipati@technicolor.com

Nadia Fawaz Nadia Fawaz is a senior researcher at Technicolor research center in Los Altos, CA. Her current research interests include data privacy and personalization. Her work leverages techniques from information theory, random matrix theory, statistics and privacy theory, and aims at bridging theory and practice. From 2009 to 2011, she was a postdoctoral researcher in the Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (MIT), Cambridge, MA. She received her Ph.D. degree in 2008 and her Diplôme dingnieur (M.Sc.) in 2005 both in electrical engineering, from cole Nationale Supérieure des Tlcommunications de Paris and EURECOM, France. She is a Member of IEEE and of ACM. Email: nadia.fawaz@technicolor.com

Branislav Kveton Branislav Kveton is a machine learning scientist at Adobe Research in San Jose. He proposes, analyzes, and applies algorithms that learn incrementally, run in real time, and converge to near optimal solutions as the number of training examples increases. Most of his recent work is focused on online learning of structured problems, such as graphs, submodularity, matroids, polymatroids, and reinforcement learning. He was at Technicolor's Research Center from 2011 to 2014, and at Intel Research from 2006 to 2011. Before 2006, he was a graduate student in the Intelligent Systems Program at the University of Pittsburgh. His advisor was Milos Hauskrecht. Email: kveton@adobe.com



Pedro Oliveira Pedro Oliveira is a Data engineer at Disqus in San Francisco, CA. From 2011 to 2013, he was a software engineer at Technicolor research center in Palo Alto, CA. He received a M.Sc. in Informatics Engineering in 2009 from University of Coimbra, Portugal. His interests are in machine learning and natural language processing. Email: cpdomina@gmail.com

Nina Taft Nina Taft is currently a senior staff Research Scientist at Google. Her interests span privacy, intrusion detection, networking and machine learning. She has previously worked at Technicolor Research Los Altos, Intel Research Labs Berkeley, Sprint Labs (Burlingame, CA) and SRI (Menlo Park, CA). She received her PhD from University of California Berkeley in 1994, and a B.S.E from the University of Pennsylvania. She is a Senior Member of the IEEE. Email: ninataft@google.com