UNIVERSITÀ DEGLI STUDI FIRENZE

FLORE
Repository istituzionale dell'Università degli Studi di Firenze

Relative privacy threats and learning from anonymized data

Questa è la Versione finale referata (Post print/Accepted manuscript) della seguente pubblicazione:

Original Citation:

Relative privacy threats and learning from anonymized data / Boreale M.; Corradi F.; Viscardi C.. - In: IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. - ISSN 1556-6013. - STAMPA. - 15:(2020), pp. 1379-1393. [10.1109/TIFS.2019.2937640]

Availability:

This version is available at: 2158/1176619 since: 2022-10-18T13:21:32Z

Published version:

DOI: 10.1109/TIFS.2019.2937640

Terms of use:

Open Access

La pubblicazione è resa disponibile sotto le norme e i termini della licenza di deposito, secondo quanto stabilito dalla Policy per l'accesso aperto dell'Università degli Studi di Firenze (https://www.sba.unifi.it/upload/policy-oa-2016-1.pdf)

Publisher copyright claim:

(Article begins on next page)

19 May 2024

Relative Privacy Threats and Learning From Anonymized Data

Michele Boreale, Fabio Corradi, and Cecilia Viscardi

***Abstract*— We consider group-based anonymization schemes, a popular approach to data publishing. This approach aims at protecting privacy of the individuals involved in a dataset, by releasing an *obfuscated* version of the original data, where the exact correspondence between individuals and attribute values is hidden. When publishing data about individuals, one must typically balance the *learner*'s utility against the risk posed by an *attacker*, potentially targeting individuals in the dataset. Accordingly, we propose a unified Bayesian model of group-based schemes and a related MCMC methodology to learn the population parameters from an anonymized table. This allows one to analyze the risk for any individual in the dataset to be linked to a specific sensitive value, when the attacker knows the individual's nonsensitive attributes, *beyond* what is implied for the general population. We call this *relative threat* analysis. Finally, we illustrate the results obtained with the proposed methodology on a real-world dataset.**

***Index Terms*— Privacy, anonymization, k-anonymity, MCMC methods.**

I. Introduction

WE CONSIDER a scenario where datasets containing personal microdata are released in anonymized form. The goal here is to enable the computation of general population characteristics with reasonable accuracy, at the same time preventing leakage of sensitive information about individuals *in the dataset*. The Database of Genotype and Phenotype [32], the U.K. Biobank [36] and the UCI Machine Learning repository [47] are well-known examples of repositories providing this type of datasets.

Anonymized datasets always have "personal identifiable information", such as names, SSNs and phone numbers, removed. At the same time, they include information derived from nonsensitive (say, gender, ZIP code, age, nationality) as well as sensitive (say, disease, income) attributes. Certain combinations of nonsensitive attributes, like ⟨gender, date of birth, ZIP code⟩, may be used to *uniquely* identify a significant fraction of the individuals in a population, thus forming so-called *quasi-identifiers*. For a given target individual, the *victim*, an attacker might easily obtain this piece of information (e.g. from personal web pages, social networks etc.), use it to identify him/her within a dataset and learn the corresponding sensitive attributes. This attack was famously demonstrated by L. Sweeney, who identified Massachusetts' Governor Weld medical record within the Group Insurance Commission (GIC) dataset [46]. Note that *identity disclosure*, that is the precise identification of an individual's record in a dataset, is not necessary to arrive at a privacy breach: depending on the dataset, an attacker might infer the victim's sensitive information, or even a few highly probable candidate values for it, without identity disclosure involved. This more general type of threat, *sensitive attribute disclosure*, is the one we focus on here.[1]

In an attempt to mitigate such threats for privacy, regulatory bodies mandate complex, often baroque syntactic constraints on the published data. As an example, here is an excerpt from the HIPAA *safe harbour* deidentification standard [48], which prescribes a list of 18 identifiers that should be removed or obfuscated, such as

> *all geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census: (1) the geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and (2) the initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000.*

There exists a large body of research, mainly in Computer Science, on syntactic methods. In particular, *group-based* anonymization techniques have been systematically investigated, starting with L. Sweeney's proposal of k-*anonymity* [46], followed by its variants, like ℓ*-diversity* [30] and *Anatomy* [49].In group-based methods, the anonymized - or obfuscated - version of a table is obtained by partitioning the set of records into groups, which are then processed to enforce certain properties. The rationale is that, even knowing that an individual belongs to a group of the anonymized table, it should not be possible for an attacker to link that individual to a specific sensitive value in the group. Two examples of group based anonymization are in Table I, adapted

Manuscript received January 18, 2019; revised May 2, 2019 and August 7, 2019; accepted August 15, 2019. This paper was presented at the Proceedings of SIS 2017 [5]. The associate editor coordinating the review of this article and approving it for publication was Prof. Xiaodong Lin. *(Corresponding author: Fabio Corradi.)*

The authors are with the Dipartimento di Statistica, Informatica, Applicazioni (DiSIA), Università di Firenze, Florence, Italy (e-mail: fabio.corradi@unifi.it).

Digital Object Identifier 10.1109/TIFS.2019.2937640

[1] Depending on the nature of the dataset, the mere *membership disclosure*, i.e. revealing that an individual is present in a dataset, may also be considered as a privacy breach: think of data about individuals who in the past have been involved in some form of felony. We will not discuss membership disclosure privacy breaches in this paper.

1556-6013 © 2019 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

IEEE Proof

IEEE Proof

TABLE I

A TABLE (TOP) ANONYMIZED ACCORDING TO 2-ANONYMITY VIA LOCAL RECODING (MIDDLE) AND ANATOMY (BOTTOM)

ID	Nat.	ZIP	Dis.
1	Malaysia	45501	Heart
2	Japan	45502	Flu
3	Japan	55503	Flu
4	Japan	55504	Stomach
5	China	66601	HIV
6	Japan	66601	Diabetes
7	India	77701	Flu
8	Malaysia	77701	Heart

a) Original table

ID	Nat.	ZIP	Dis.
1	{M, J}	4550*	Heart
2	{M, J}	4550*	Flu
3	Japan	5550*	Flu
4	Japan	5550*	Stomach
5	{C, J}	66601	HIV
6	{C, J}	66601	Diabetes
7	{I, M}	77701	Flu
8	{I, M}	77701	Heart

b) 2-anonymity via local recoding

GID	Nat.	ZIP	Dis.
1	Japan	45502	Heart
1	Malaysia	45501	Flu
2	Japan	55504	Flu
2	Japan	55503	Stomach
3	Japan	66601	HIV
3	China	66601	Diabetes
4	Malaysia	77701	Flu
4	India	77701	Heart

c) Anatomy

from [9]. The topmost, original table collects medical data from eight individuals; here *Disease* is considered as the only sensitive attribute. The central table is a 2-anonymous, 2-diverse table: within each group the nonsensitive attribute values have been generalized following group-specific rules (*local recoding*) so as to make them indistinguishable; moreover, each group features 2 distinct sensitive values. In general, each group in a k-anonymous table consists of at least k records, which are indistinguishable when projected on the nonsensitive attributes; ℓ-diversity additionally requires the presence in each group of at least ℓ distinct sensitive values, with approximately the same frequency. This is an example of *horizontal* scheme. Table I (c) is an example of application of the *Anatomy* scheme: within each group, the nonsensitive part of the rows are *vertically* and *randomly* permuted, thus breaking the link between sensitive and nonsensitive values. Again, the table is 2-diverse.

In recent years, the effectiveness of syntactic anonymization methods has been questioned, as offering weak guarantees against attackers with strong background knowledge – very precise contextual information about their victims. *Differential privacy* [18], which promises protection in the face of *arbitrary* background knowledge, while valuable in the release of summary statistics, still appears not of much use when it comes to data publishing (see the Related works paragraph). As a matter of fact, release of syntactically anonymized tables appears to be the most widespread data publishing practice, with quite effective tool support (see e.g. [37]).

In the present paper, discounting the risk posed by attackers with strong background knowledge, we pose the problem in relative terms: given that whatever is *learned about the general population* from an anonymized dataset represents legitimate and useful information ("smoke is associated with cancer"), one should prevent an attacker from drawing conclusions about specific individuals in the table ("almost certainly the target individual has cancer"): in other words, learning sensitive information for an individual in the dataset, *beyond* what is implied for the general population. To see what is at stake here, consider dataset (b) in Table I. Suppose that the attacker's victim is a Malaysian living at ZIP code 45501, and known to belong to the original table. The victim's record must therefore be in the first group of the anonymized table. The attacker may reason that, with the exception of the first group, a Japanese is never connected to Heart Disease; this hint can become a strong evidence in a larger, real-world table. Then the attacker can link with high probability the Malaysian victim in the first group to Heart Disease. In this attack, the attacker combines knowledge of the nonsensitive attributes of the victim (Malaysian, ZIP code 45501) with the group structure and the knowledge learned from the anonymized table.

We propose a unified probabilistic model to reason about such forms of leakage. In doing so, we clearly distinguish the position of the *learner* from that of the *attacker*: the resulting notion is called *relative privacy threat*. In our proposal, both the learner and the attacker activities are modeled as forms of Bayesian inference: the acquired knowledge is represented as a joint posterior probability distribution over the sensitive and nonsensitive values, given the anonymized table *and*, in the case of the attacker, knowledge of the victim's presence in the table. A comparison between these two distributions determines what we call relative privacy threat. Since posterior distributions are in general impossible to express analytically, we also put forward a MCMC method to practically estimate such posteriors. We also illustrate the results of applying our method to the Adult dataset from the UCI Machine Learning repository [47], a common benchmark in anonymization research.

A. Related Works

Sweeney's k-anonymity [46] is among the most popular proposals aiming at a systematic treatment of syntactic anonymization of microdata. The underlying idea is that every individual in the released dataset should be hidden in a "crowds of k". Over the years, k-anonymity has proven to provide weak guarantees against attackers who know much about their victims, that is have a strong background knowledge. For example, an attacker may know from sources other than the released data that his victim does *not* suffer from certain diseases, thus ruling out all possibilities but one in the victims's group. Additional constraints may be enforced in order to mitigate those attacks, like ℓ-diversity [30] and *t-closeness* [27]. Differential Privacy [18] promises protection in the face of arbitrary background knowledge. In its basic, interactive version, this means that, when querying a database via a differentially private mechanism, one will get approximately the same answers, whether the data of any specific individual is included or not in the database. This is typically achieved by injecting controlled levels of noise in the reported answer, e.g. Laplacian noise. Differential Privacy is very effective when applied to certain summary statistics, such as histograms. However, it raises a number of difficulties when applied to table publishing: in concrete cases, the level of noise necessary to guarantee an acceptable degree of privacy would destroy utility [12], [13], [44]. Moreover, due to correlation phenomena, it appears that Differential Privacy cannot in general be used to control evidence about the participation of individuals in a database [4], [26]. In fact, the no-free-lunch theorem of Kifer and Machanavajjhala [26] implies that it is impossible to guarantee both privacy *and* utility, without making assumptions about how the data have been generated (e.g., independence assumptions). Clifton and Tassa [10] critically review issues and criticisms involved in both syntactic methods and Differential Privacy, concluding that both have their place, in Privacy Preserving- Data Publishing and Data Mining, respectively. Both approaches have issues that call for further research. A few proposals involve blending the two approaches, with the goal to achieve both strong privacy guarantees and utility, see e.g. [28].

A major source of inspiration for our work has been Kifer's [25]. The main point of [25] is to demonstrate a pitfall of the *random worlds* model, where the attacker is assumed to assign equal probability to all cleartext tables compatible with the given anonymized one. Kifer shows that a Bayesian attacker willing to learn from the released table can draw sharper inferences than those possible in the random worlds model. In particular, Kifer shows that it is possible to extract from (anatomized) ℓ-diverse tables belief probabilities greater than $1/\ell$, by means of the so-called deFinetti attack. While pinpointing a deficiency of the random worlds model, it is questionable if this should be considered an attack, or just a legitimate learning strategy. Quoting [10] on the deFinetti attack:

> *The question is whether the inference of a general behavior of the population in order to draw belief probabilities on individuals in that population constitutes a breach of privacy (...). To answer this question positively for an attack on privacy, the success of the attack when launched against records that* are *part of the table should be significantly higher than its success against records that* are not *part of the table. We are not aware of such a comparison for the deFinetti attack.*

It is this very issue that we tackle in the present paper. Specifically, our main contribution here is to put forward a concept of relative privacy threat, as a means to assess the risks implied by publishing tables anonymized via group-based methods. To this end, we introduce: (a) a unified probabilistic model for group-based schemes; (b) rigorous characterizations of the learner and the attacker's inference, based on Bayesian reasoning; and, (c) a related MCMC method, which generalizes and systematizes that proposed in [25].

Very recently, partly inspired by differential privacy, a few authors have considered what might be called a relative or *differential* approach to assessing privacy threats, in conjunction with some notion of learning or inference from the anonymized data. Especially relevant to our work is *differential inference*, introduced in a recent paper by Kassem *et al.* [24]. These authors make a clear distinction between two different types of information that can be inferred from anonymized data: learning of "public" information, concerning the population, should be considered as legitimate; on the contrary, leakage of "private" information about individuals should be prevented. To make this distinction formal, given a dataset, they compare two probability distributions that can be machine-learned from two distinct training sets: one including and one excluding a target individual. An attack exists if there is a significant difference between the two distributions, measured e.g. in terms of Earth Moving Distance. While similar in spirit to ours, this approach is conceptually and technically different from what we do here. Indeed, in our case the attacker explicitly takes advantage of the extra piece of information concerning the presence of the victim in the dataset to attack the target individual, which leads to a more direct notion of privacy breach. Moreover, in [24] a Bayesian approach to inference is not clearly posed, so the obtained results lack a semantic foundation, and strongly depend on the adopted learning algorithm. Pyrgelis *et al.* [39] use Machine Learning for membership inference on aggregated location data, building a binary classifier that can be used to predict if a target user is part of the aggregate data or not. A similar goal is pursued in [35]. Again, a clear semantic foundation of these methods is lacking, and the obtained results can be validated only empirically. In a similar vein, [3] and [17] have proposed statistical techniques to detect privacy violations, but they only apply to differential privacy. Other works, such as [23] and [33], have just considered the problem of how to effectively learn from anonymized datasets, but not of how to characterize legitimate, as opposed to non-legitimate, inference.

On the side of the random worlds model, Chi-Wing Wong et al.'s work [9] shows how information on the population extracted from the anonymized table – in the authors' words, the *foreground* knowledge – can be leveraged by the attacker to violate the privacy of target individuals. The underlying reasoning, though, is based on the random worlds model, hence is conceptually and computationally very different from the Bayesian model adopted in the present paper. Bewong *et al.* [2] assess relative privacy threat for transactional data by a suitable extension of the notion of *t-closeness*, which is based on comparing the relative frequency of the victim's sensitive attribute in the whole table with that in the victim's group. Here the underlying assumption is that the attacker's prior knowledge about sensitive attributes matches the public knowledge, and that the observed sensitive attributes frequencies provide good

IEEE Proof

estimates both for the public knowledge and the attacker's belief. Our proposal yields more sophisticated estimates via a Bayesian inferential procedure. Moreover, in our scenario the assumption on the attacker's knowledge is relaxed requiring only the knowledge of the victim's presence in whatever group of the table.

A concept very different from the previously discussed proposals is Rubin's *multiple imputation* approach [43], by which only tables of *synthetic* data, generated sampling from a predictive distribution learned from the original table, are released. This avoids syntactic masking/obfuscation, whose analysis requires customized algorithms on the part of the learner, and leaves to the data producer the burden of synthesis. Note that this task can be nontrivial and raises a number of difficulties concerning the availability of auxiliary variables for non-sampled units, see [42]. In Rubin's view, synthetic data overcome all privacy concerns, in that no real individual's data is actually released. However, this position has been questioned, on the grounds that information about participants may leak through the chain: original table → posterior parameters → synthetic tables. In particular, Machanavajjhala *et al.* [31] study Differential Privacy of synthetic categorical data. They show that the release of such data can be made differentially private, at the cost of introducing very powerful priors. However, such priors can lead to a serious distortion in whatever is learned from the data, thus compromising utility. In fact, [50] argues that, in concrete cases, the required pseudo sample size hyperparameter could be larger than the size of the table. Experimental studies [7], [8] appear to confirm that such distorting priors are indeed necessary for released synthetic data to provide acceptable guarantees, in the sense of Differential Privacy. See [50] for a recent survey of results about synthetic data release and privacy. An outline of the model presented here, with no proofs of correctness, appeared in the conference paper [5].

B. Structure of the Paper

The rest of the paper is organized as follows. In Section II we propose a unified formal definition of vertical and horizontal schemes. In Section III we put forward a probabilistic model to reason about learner's and attacker's inference; the case of prior partial knowledge of the victim's attributes on the part of the attacker is also covered. Based on that, measures of (relative) privacy threats and utility are introduced in Section IV. In Section V, we study a MCMC algorithm to learn the population parameters posterior and the attacker's probability distribution from the anonymized data. In Section VI, we illustrate the results of an experiment conducted on a real-world dataset. A few concluding remarks and perspectives for future work are reported in Section VII. Some technical material has been confined to Appendix A.

II. Group Based Anonymization Schemes

A dataset consists of a collection of rows, where each row corresponds to an individual. Formally, let $\mathcal{R}$ and $\mathcal{S}$, ranged over by r and s respectively, be finite non-empty sets of *nonsensitive* and *sensitive* values, respectively. A *row* is a pair $(s, r) \in \mathcal{S} \times \mathcal{R}$. There might be more than one sensitive and nonsensitive characteristic, so s and r can be thought of as vectors.

A *group-based anonymization algorithm* $\mathcal{A}$ is an algorithm that takes a multiset of rows as input and yields an obfuscated table as output, according to the scheme

$$\text{multiset of rows} \longrightarrow \text{cleartext table} \longrightarrow \text{obfuscated table}.$$

Formally, fix $N \geq 1$. Given a multiset of N rows, $d = \{\!|(s_1, r_1), \ldots, (s_N, r_N)|\!\}$, $\mathcal{A}$ will first arrange d into a sequence of *groups*, $t = g_1, \ldots, g_k$, the *cleartext table*. Each group in turn is a sequence of n_i rows, $g_i = (s_{i,1}, r_{i,1}), \ldots, (s_{i,n_i}, r_{i,n_i})$, where n_i can vary from group to group. Note that both the number of groups, $k \geq 1$, and the number of rows in each group, n_i, depend in general on the original multiset d as well as on properties of the considered algorithm – such as ensuring k-anonymity and ℓ-diversity (see below). The obfuscated table is then obtained as a sequence $t^* = g_1^*, \ldots, g_k^*$, where the obfuscation of each group g_i is a pair $g_i^* = (m_i, l_i)$. Here, each $m_i = s_{i,1}, \ldots, s_{i,n_i}$ is the sequence of *sensitive* values occurring in g_i; each l_i, called *generalized nonsensitive value*, is one of the following:

- for *horizontal* schemes, a *superset* of g_i's nonsensitive values: $l_i \supseteq \{r_{i,1}, \ldots, r_{i,n_i}\}$;
- for *vertical* schemes, the *multiset* of g_i's nonsensitive values: $l_i = \{\!|r_{i,1}, \ldots, r_{i,n_i}|\!\}$.

Note that the generalized nonsensitive values in vertical schemes include all and only the values, with multiplicities, found in the corresponding original group. On the other hand, generalized nonsensitive values in horizontal schemes may include additional values, thus generating a superset. What values enter the superset depends on the adopted technique, e.g. micro-aggregation, generalization or suppression; in any case this makes the rows in each group indistinguishable when projected onto the nonsensitive attributes. For example, each of 45501, 45502 is generalized to the superset $4550{*} = \{45500, 45501, \ldots, 45509\}$ in the first group of Table I(b).

Sometimes it will be notationally convenient to ignore the group structure of t altogether, and regard the cleartext table t simply as a sequence of rows, $(s_1, r_1), (s_2, r_2), \ldots, (s_1, s_N)$. Each row (s_j, r_j) is then uniquely identified within the table t by its index $1 \leq j \leq N$.

An instance of horizontal schemes is k-*anonymity* [46]: in a k-anonymous table, each group consists of at least k$\geq$ 1 rows, where the different nonsensitive values appearing within each group have been generalized so as to make them indistinguishable. In the most general case, different occurrences of the same nonsensitive value might be generalized in different ways, depending on their position (index) within the table t: this is the case of *local recoding*. Alternatively, each occurrence of a nonsensitive value is generalized in the same way, independently of its position: this is the case of *global recoding*. Further conditions may be imposed on the resulting anonymized table, such as ℓ*-diversity*, requiring that at least $\ell \geq 1$ distinct values of the sensitive attribute appear in each group. Table I (center) shows an example of k= 2-anonymous and $\ell = 2$-diverse table: in each group the nonsensitive

IEEE Proof

TABLE II

SUMMARY OF NOTATION

Symbol	**Description**	**Symbol**	**Description**
A	attacker	β	$\pi_{R\|S}$ hyperparameters
α	π_S hyperparameters	δ	nonsensitive freq.
γ	sensitive freq.	g_i^*	obfuscated group i
g_i	group i	$\mathbf{GT}_A$	global threat level
ETV	emp. total variation	k	number of groups
I	evaluator (ideal)	k	min size of groups s
l_i	group i nonsens. values	L	learner
ℓ	min n. of sens. val.	m_i	group i sens. values
N	n. of rows in the table	π	parameters of R, S
$\pi_{R\|s}$	parameters of $R\|s$	π_S	parameters of S
R	nonsensitive r.v.	S	sensitive r.v.
t	clear text table	t^*	obfuscated table
Ti	rel. threat level	**TV**	total variation
RF	rel. faithfulness level	v	victim

values are indistinguishable and two different sensitive values (diseases) appear in each group.

An instance of vertical schemes is *Anatomy* [49]: within each group, the link between the sensitive and nonsensitive values is hidden by randomly permuting one of the two parts, for example the nonsensitive one. As a consequence, an anatomized table may be seen as consisting of *two* sub-tables: a sensitive and a nonsensitive one. Table I (c) shows an example of anatomized table: in the nonsensitive sub-table, the reference to the corresponding sensitive values is lost; only the multiset of nonsensitive values appears for each group.

Remark 1 (disjointness): Some anonymization schemes enforce the following disjointness property on the obfuscated table t^*:

> *Any two generalized nonsensitive values in t^* are disjoint: $i \neq j$ implies $l_i \cap l_j = \emptyset$.*

We need not assume this property in our treatment – although assuming it may be computationally useful in practice (see Section III).

For ease of reference, we provide a summary of the notation that will be used throughout the paper in Table II.

III. A Unified Probabilistic Model

We provide a unified probabilistic model for reasoning on group-based schemes. We first introduce the random variables of the model together with their joint density function. On top of these variables, we then define the probability distributions on $\mathcal{S} \times \mathcal{R}$ that formalize the *learner* and the *attacker* knowledge, given the obfuscated table.

A. Random Variables

The model consists of the following random variables.

- Π, taking values in the set of full support probability distributions $\mathcal{D}$ over $\mathcal{S} \times \mathcal{R}$, is the joint probability distribution of the sensitive and nonsensitive attributes in the population.
- $T = G_1, \dots, G_k$, taking values in the set of cleartext tables $\mathcal{T}$. Each group G_i is in turn a sequence of $n_i \geq 1$ consecutive rows in T, $G_i = (S_{i,1}, R_{i,1}), \dots, (S_{i,n_i}, R_{i,n_i})$. The number of groups k is not fixed, but depends on the anonymization scheme and the specific tuples composing T.
- $T^* = G_1^*, \dots, G_k^*$, taking values in the set of obfuscated tables $\mathcal{T}^*$.

We assume that the above three random variables form a Markov chain:

$$\Pi \longrightarrow T \longrightarrow T^*. \tag{1}$$

In other words, uncertainty on T is driven by Π, and T^* solely depends on the table T and the underlying obfuscation algorithm. As a result, $T^* \perp\!\!\!\perp \Pi \mid T$. Equivalently, the joint probability density function f of these variables can be factorized as follows, where π, t, t^* range over $\mathcal{D}$, $\mathcal{T}$ and $\mathcal{T}^*$, respectively:

$$f(\pi, t, t^*) = f(\pi) f(t|\pi) f(t^*|t). \tag{2}$$

Additionally, we shall assume the following:

- $\pi \in \mathcal{D}$ is encoded as a pair $\pi = (\pi_S, \pi_{R|S})$ where $\pi_{R|S} = \{\pi_{R|s} : s \in \mathcal{S}\}$. Here, π_S are the parameters of a full support categorical distribution over $\mathcal{S}$, and, for each $s \in \mathcal{S}$, $\pi_{R|s}$ are the parameters of a full support categorical distribution over $\mathcal{R}$. For each $(s, r) \in \mathcal{S} \times \mathcal{R}$

  $$f(s, r|\pi) = f(s|\pi) \cdot f(r|\pi_{R|s})$$

  We also posit that the π_S and the $\pi_{R|s}$'s are chosen independently, according to Dirichlet distributions of hyperparameters $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{|\mathcal{S}|})$ and $\boldsymbol{\beta}^s = (\beta_1^s, \dots, \beta_{|\mathcal{R}|}^s)$, respectively. In other words

  $$f(\pi) = \mathrm{Dir}(\pi_S \mid \boldsymbol{\alpha}) \cdot \prod_{s \in \mathcal{S}} \mathrm{Dir}(\pi_{R|s} \mid \boldsymbol{\beta}^s). \tag{3}$$

  The hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may incorporate prior (background) knowledge on the population, if this is available. Otherwise, a uniformative prior can be chosen setting $\alpha_i = \beta_j^s = 1$ for each i, s, j. When $r \in \mathcal{R}$ is a tuple of attributes, we shall assume conditional independence of those attributes given s, so that the joint probability of $r|s$ can be determined by factorization.
- The N individual rows composing the table t, say $(s_1, r_1), \dots, (s_N, r_N)$, are assumed to be drawn i.i.d. according to $f(\cdot|\pi)$. Equivalently

  $$f(t|\pi) = f(s_1, r_1|\pi) \cdots f(s_N, r_N|\pi). \tag{4}$$

Instances of the above model can be obtained by specifying an anonymization mechanism $\mathcal{A}$. In particular, the distribution $f(t^*|t)$ only depends on the obfuscation algorithm that is adopted, say $\mathrm{obf}(t)$. In the important special case $\mathrm{obf}(t)$ acts as a deterministic function on tables, $f(t^*|t) = 1$ if and only if $\mathrm{obf}(t) = t^*$, otherwise $f(t^*|t) = 0$.

B. Learner and Attacker Knowledge

We shall denote by p_{L} the probability distribution over $\mathcal{S} \times \mathcal{R}$ that can be learned given the anonymized table t^*. This distribution we take to be the average of $f(s, r|\pi)$ with respect

IEEE Proof

to the density $f(\Pi = \pi \,|T^* = t^*)$. Formally, for each $(s, r) \in \mathcal{S} \times \mathcal{R}$:

$$p_{\mathrm{L}}(s, r|t^*) \overset{\triangle}{=} E_{\pi \sim f(\pi|t^*)}[f(s, r|\pi)] = \int_{\mathcal{D}} f(s, r|\pi)\, f(\pi|t^*)\, \mathrm{d}\pi. \tag{5}$$

Of course, we can condition p_{L} on any given r and obtain the conditional probability $p_{\mathrm{L}}(s|r, t^*)$. Equivalently, we can compute

$$p_{\mathrm{L}}(s|r, t^*) \overset{\triangle}{=} E_{\pi \sim f(\pi|t^*)}[f(s|r, \pi)] = \int_{\mathcal{D}} f(s|r, \pi)\, f(\pi|t^*)\, \mathrm{d}\pi. \tag{6}$$

In particular, one can read off this distribution on a victim's nonsensitive attribute, say r_{v}, and obtain the corresponding distribution on $\mathcal{S}$.

We shall assume the attacker knows the values of $T^* = t^*$ and the nonsensitive value r_{v} of a target individual, the victim; *moreover the attacker knows the victim is an individual in the table*. Accordingly, in what follows we fix once and for all t^* and r_{v}: these are the values observed by the attacker. Given knowledge of a victim's nonsensitive attribute r_{v} *and* knowledge that the victim is actually in the table T, we can define the attacker's distribution on $\mathcal{S}$ as follows.

Let us introduce in the above model a new random variable V, identifying the index of the victim within the cleartext table T. We posit that V is uniformly distributed on $\{1, \ldots, N\}$, and independent from Π, T, T^*. Recalling that each row (S_j, R_j) is identified within T by a unique index j, we can define the attacker's probability distribution on $\mathcal{S}$, after seeing t^* and r_{v}, as follows, where it is assumed that $f(R_V = r_{\mathrm{v}}, t^*) > 0$, that is the observed victim's r_{v} is compatible with t^*:

$$p_{\mathrm{A}}(s|r_{\mathrm{v}}, t^*) \overset{\triangle}{=} f(S_V = s \mid R_V = r_{\mathrm{v}},\, t^*). \tag{7}$$

The following crucial lemma provides us with a characterization of the above probability distribution that is only based on a selection of the marginals R_j given t^*. This will be the basis for actually computing $p_{\mathrm{A}}(s|r_{\mathrm{v}}, t^*)$. Note that, on the right-hand side, only those rows whose sensitive value - known from t^* - is s contribute to the summation. A proof of the lemma is reported in Appendix A.

Lemma 1: Let $T = (S_j, R_j)_{j \in 1 \ldots N}$. Let s_j be the sensitive value in the j-th entry of t^*. Let r_{v} and t^* such that $f(R_V = r_{\mathrm{v}}, t^*) > 0$. Then

$$p_{\mathrm{A}}(s|r_{\mathrm{v}}, t^*) \propto \sum_{j\,:\,s_j = s} f(R_j = r_{\mathrm{v}} \,|\, t^*). \tag{8}$$

Note that the disjointness of generalized nonsensitive values of the groups can make the computation of (8) more efficient, restricting the summation on the right-hand side to a unique group.

Example 1: In order to illustrate the difference between the learner's and the attacker's inference, we reconsider the toy example in the Introduction. Let t^* be the 2-anonymous, 2-diverse Table I(b). Assume the attacker's victim is the first individual of the original dataset, who is from Malaysia($= M$) and lives in the ZIP code 45501 area, hence $r_{\mathrm{v}} = (M, 45501)$. Table III shows the belief probabilities of the learner, $p_{\mathrm{L}}(s|r_{\mathrm{v}}, t^*)$, and of the attacker, $p_{\mathrm{A}}(s|r_{\mathrm{v}}, t^*)$, for the victim's disease s. We also include the random worlds model probabilities, $p_{\mathrm{RW}}(s|r_{\mathrm{v}}, t^*)$, which are just proportional to the frequency of each sensitive value within the victim's group. Note that the learner and the attacker distributions have the same mode, but the attacker is more confident about his prediction of the victim's disease. The random worlds model produces a multi-modal solution.

TABLE III

POSTERIOR DISTRIBUTIONS OF DISEASES FOR A VICTIM WITH $r_{\mathrm{v}} = (M, 45501)$, FOR THE ANONYMIZED t^* IN TABLE I(B). NB: FIGURES AFFECTED BY ROUNDING ERRORS

	Heart	Flu	Stomach	HIV	Diabetes
$p_{\mathrm{L}}(s\|r_{\mathrm{v}}, t^*)$	0.343	0.317	0.113	0.114	0.113
$p_{\mathrm{A}}(s\|r_{\mathrm{v}}, t^*)$	0.580	0.420	0	0	0
$p_{\mathrm{RW}}(s\|r_{\mathrm{v}}, t^*)$	0.500	0.500	0	0	0

As to the computation of the probabilities in Table III, a routine application of the equations (2) – (8) shows that p_{L} and p_{A} reduce to the expressions (9) and (10) below, given in terms of the model's density (2). The crucial point here is that the adversary knows the group his victim is in, i.e. the first two lines of t^* in the example. Below, $s \in \mathcal{S}$; for $j = 1, 2$, s_j denotes the sensitive value of the j-th row, while t is a cleartext table, from which t_{-j} is obtained by removing (s_j, r_{v}). It is assumed that the obfuscation algorithm $\mathcal{A}$ is deterministic, so that $f(t^*|t) \in \{0, 1\}$.

$$p_{\mathrm{L}}(s|r_{\mathrm{v}}, t^*) \propto \int_{\mathcal{D}} f(\pi) f(s, r_{\mathrm{v}}|\pi) \sum_{t:\mathcal{A}(t)=t^*} f(t|\pi)\, \mathrm{d}\pi \tag{9}$$

$$p_{\mathrm{A}}(s_j|r_{\mathrm{v}}, t^*) \propto \int_{\mathcal{D}} f(\pi) f(s_j, r_{\mathrm{v}}|\pi) \sum_{t_{-j}:\mathcal{A}(t)=t^*} f(t|\pi)\, \mathrm{d}\pi. \tag{10}$$

Unfortunately, the analytic computation of the above integrals, even for the considered toy example, is a daunting task. For instance, the summation in (9) has as many terms as t^*-compatible tables t, that is 6.4×10^5 for Example 1 – although the resulting expression can be somewhat simplified using the independence assumption (4). Accordingly, the figures in Table III have been computed resorting to simulation techniques, see Section V.

An alternative, more intuitive description of the inference process is as follows. The learner and the attacker first learn the parameters π given t^*, that is they evaluate $f(\pi_{\mathrm{Dis}}|t^*)$, $f(\pi_{\mathrm{ZIP}|s}|t^*)$ and $f(\pi_{\mathrm{Nat}|s}|t^*)$, for all $s \in \mathcal{S}$. Due to the uncertainty on the ZIP code and/or Nationality, learning π takes the form of a mixture (this is akin to learning with soft evidence, see Corradi *et al.* [11]). After that, the learner, ignoring the victim is in the table, predicts the probability of r_{v}, $p_{\mathrm{L}}(r_{\mathrm{v}}|s, t^*)$, for all s, by using a mixture of Multinomial-Dirichlet. The attacker, on the other hand, while still basing his prediction $p_{\mathrm{A}}(r_{\mathrm{v}}|s, t^*)$ on the parameter learning outlined above, restricts his attention to the first two lines of t^*, thus realizing that $s \in \{\text{Heart, Flu}\}$. Then, by Bayes theorem, and adopting the relative frequencies of the diseases in t^* as an approximation of $f(s|t^*)$, the posterior probability of the diseases for the victim can be computed.

IEEE Proof

Remark 2 (attacker's inference and forensic identification): The attacker's inference is strongly reminiscent of two famous settings in forensic science: the Island Problem *(*IP*) and the* The Data Base Search Problem *(*DBS*), see e.g. [1], [14] and more recently [45]. In an island with N inhabitants a crime is committed; a characteristic of the criminal (e.g. a DNA trait) is found on the crime scene. It is known that the island's inhabitants posses this characteristic independently with probability p. It is assumed the existence of exactly one culprit C in the island. In* IP*, one island's inhabitant I, the suspect, is found to have the given characteristic, while the others are not tested. An investigator is interested in the probability that $I = C$.*

When we cast this scenario in our framework, the individuals in the table play the role of the inhabitants (including the culprit), while r_{v} plays the role of the characteristic found on the crime scene, matching that of the suspect. In other words - perhaps ironically - our framework's victim plays here the role of the suspect S, while our attacker is essentially the investigator. Letting $\mathcal{S} = \{0, 1\}$ (innocent/guilty) and $\mathcal{R} = \{0, 1\}$ (characteristic absent/present), the investigator's information is then summarized by an obfuscated horizontal table t^* of N rows with as many groups, where exactly one row, say the j-th, has $S_j = 1$ and $R^*_j = R_j = 1$ (the culprit), while for $i \neq j$, $S_i = 0$ and $R^*_i = *$ ($N - 1$ innocent inhabitants). Recalling that the variable V in our framework represents the suspect's index within the table, the probability that $I = C$ is

$$\begin{aligned}\Pr(V = j|R_V = 1, t^*) &= \Pr(S_V = 1|R_V = 1, t^*)\\ &= p_{\mathrm{A}}(s = 1|r_{\mathrm{v}} = 1, t^*).\end{aligned}$$

Then applying (8)*, we find*

$$\begin{aligned}p_{\mathrm{A}}(s = 1|r_{\mathrm{v}} = 1, t^*) &= \frac{f(R_j = 1|t^*)}{f(R_j = 1|t^*) + (N-1) f(R_{i\neq j} = 1|t^*)}\\ &= \frac{1}{1 + (N-1) f(R_{i\neq j} = 1|t^*)}. \qquad (11)\end{aligned}$$

By taking suitable prior hyperparameters, $f(R_{i\neq j} = 1|t^*)$ can be made arbitrarily close to p. For ease of comparison with the classical IP *and* DBS *settings, rather than relying on a learning procedure, we just assume here $f(R_i = 1|t^*) = p$ for $i \neq j$, so that* (11) *simplifies to*

$$p_{\mathrm{A}}(s = 1|r_{\mathrm{v}} = 1, t^*) = \frac{1}{1 + (N-1)p} \qquad (12)$$

which is the classical result known from the literature.
In DBS*, the indicted exhibiting r_{v} is found after testing $1 \leq k < N$ individuals that do not exhibit r_{v}. This means the table t^* consists now of k rows $(s, r) = (0, 0)$ (the k innocent, tested inhabitants not exhibiting r_{v}), one row $(s, r) = (1, 1)$ (the culprit) and $N - 1 - k$ rows $(s, r^*) = (0, *)$ (the $N - 1 - k$ innocent, non-tested inhabitants). Accordingly,* (11) *becomes (letting $j = k + 1$, and possibly after rearranging indices),* (13)*, as shown at the bottom of this page. Letting $f(R_i = 1|t^*) = p$ for $i > k + 1$, equation* (13) *becomes*

$$p_{\mathrm{A}}(s = 1|r_{\mathrm{v}} = 1, t^*) = \frac{1}{1 + (N-1-k)p}$$

which again is the classical result known from the literature. Finally note that our methodology also covers the possibility to learn about the probability of the characteristic, $f(R_i = 1|t^*)$, but here we have only stressed how the attacker strategy solves the IP *and* DBS *forensic problems. Uncertainty about population parameters and identification has been considered elsewhere by one of us [6].*

We now briefly discuss an extension of our framework to the more general case where the attacker has only partial information about his victim's nonsensitive attributes. For a typical application, think of a dataset where $\mathcal{R}$ and $\mathcal{S}$ are individuals' genetic profiles and diseases, respectively, with an adversary knowing only a partial DNA profile of his victim; e.g. only the alleles at a few loci. Formally, fix a nonempty set $\mathcal{Y}$ and let $g : \mathcal{R} \to \mathcal{Y}$ be a (typically non-injective) function, modeling the attacker's observation of the victim's nonsensitive attribute. With the above introduced notation, consider the random variable $Y \overset{\triangle}{=} g(R_V)$. It is natural to extend definition (7) as follows, where $g(r_{\mathrm{v}}) = y_{\mathrm{v}} \in \mathcal{Y}$ and $f(Y = y_{\mathrm{v}}, t^*) > 0$:

$$p_{\mathrm{A}}(s|y_{\mathrm{v}}, t^*) \overset{\triangle}{=} f(S_V = s \mid Y = y_{\mathrm{v}}, t^*). \qquad (14)$$

It is a simple matter to check that (8) becomes the following, where $g^{-1}(y) \subseteq \mathcal{R}$ denotes the counter-image of y according to g:

$$p_{\mathrm{A}}(s|r_{\mathrm{v}}, t^*) \propto \sum_{j\,:\,s_j = s} f(R_j \in g^{-1}(y_{\mathrm{v}}) \mid t^*). \qquad (15)$$

Also note that one has $f(R_j \in g^{-1}(y_{\mathrm{v}}) \mid t^*) = \sum_{r\in g^{-1}(y_{\mathrm{v}})} f(R_j = r \mid t^*)$. An extension to the case of partial and *noisy* observations can be modeled similarly, by letting $Y = g(R_V, E)$, where E is a random variable representing an independent source of noise. We leave the details of this extension for future work.

IV. Measures of Privacy Threat and Utility

We are now set to define the measures of *privacy threat* and *utility* we are after. We will do so from the point of view of a person or entity, the *evaluator*, who:

(a) has got a copy of the cleartext table t, and can build an obfuscated version t^* of it;
(b) must decide whether to release t^* or not, weighing the privacy threats and the utility implied by this act.

The evaluator clearly distinguishes the position of the *learner* from that of the *attacker*. The learner is interested in learning from t^* the characteristics of the general population, via p_L. The attacker is interested in learning from t^* the sensitive

$$p_{\mathrm{A}}(s = 1|r_{\mathrm{v}} = 1, t^*) = \frac{f(R_{k+1} = 1|t^*)}{f(R_{k+1} = 1|t^*) + k f(R_{i\in\{1,k\}} = 1|t^*) + (N-1-k) f(R_{i>k+1} = 1|t^*)} \qquad (13)$$

IEEE Proof

IEEE Proof

value of a target individual, the *victim*, via p_A. The last probability distribution is derived by exploiting the additional piece of information that the victim is an individual known to be in the original table, of whom the attacker gets to know the nonsensitive values. As pointed out in [34], information about the victim's nonsensitive attributes can be easily gathered from other sources such as personal blogs and social networks. These assumptions about the attacker's knowledge allow a comparison between the risks of a sensitive attribute disclosure for an individual *who is part of the table* and for individuals who are not. The evaluator adopts the following *relative*, or differential, point of view:

> a situation where, for some individual, p_A conveys much more information than that conveyed by p_L (learner's legitimate inference on general population), must be deemed as a privacy threat.

Generally speaking, the evaluator should refrain from publishing t^* if, for some individual, the *level* of relative privacy threat exceeds a predefined threshold. Concerning the definition of the level of threat, the evaluator adopts the following Bayesian decision-theoretic point of view. Whatever distribution p is adopted to guess the victim's sensitive value, the attacker is faced with some utility function. Here, we consider a simple 0-1 utility function for the attacker, yielding 1 if the sensitive attribute is guessed correctly and 0 otherwise. The resulting attacker's expected utility is maximized by the Bayes act, i.e. by choosing $s = \mathrm{argmax}_{s' \in \mathcal{S}}\, p(s')$, and equals $p(s)$. The above discussion leads to the following definitions. Note that we consider threat measures both for individual rows and for the overall table. For each threatened row, the relative threat index **Ti** says how many times the probability of correctly guessing the secret is increased by the attacker's activity i.e. by exploiting the knowledge of the victim's presence in the table. At a global, table-wise level, the evaluator also considers the fraction $\mathbf{GT}_A$ of rows threatened by the attacker.

Definition 1 (privacy threat): We define the following privacy threat measures.

- Let q be a full support distribution on $\mathcal{S}$ and (s,r) be a row in t. We say (s,r) is *threatened under* q if $q(s) = \max_{s'} q(s')$, and that its *threat level under* q is $q(s)$.
- For a row (s,r) in t that is threatened by $p_{\mathrm{A}}(\cdot|r,t^*)$, its *relative threat level* is
$$\mathbf{Ti}(s,r,t,t^*) \overset{\triangle}{=} \frac{p_{\mathrm{A}}(s|r,t^*)}{p_{\mathrm{L}}(s|r,t^*)}. \tag{16}$$
- Let $N_{\mathrm{A}}(t,t^*)$ be the number of rows (s,r) in t threatened by $p_{\mathrm{A}}(\cdot|r,t^*)$. The *global threat level* $\mathbf{GT}_A(t,t^*)$ is the fraction of rows that are threatened, that is
$$\mathbf{GT}_A(t,t^*) \overset{\triangle}{=} \frac{N_{\mathrm{A}}(t,t^*)}{N}. \tag{17}$$
Similarly, we denote by $\mathbf{GT}_{\mathrm{L}}(t,t^*)$ the fraction of rows (s,r) in t that are threatened under $p_{\mathrm{L}}(\cdot|r,t^*)$.
- As a measure of how better the attacker performs than learner at a global level, we introduce *relative global threat*:
$$\mathbf{RGT_A}(t,t^*) \overset{\triangle}{=} \max\{0, \mathbf{GT}_A(t,t^*) - \mathbf{GT}_{\mathrm{L}}(t,t^*)\}. \tag{18}$$

Remark 3 (setting a threshold for **Ti***): A difficult issue is how to set an acceptable threshold for the relative threat level* **Ti***. This is conceptually very similar to the question of how to set the level of ϵ in differential privacy: its proponents have always maintained that the setting of ϵ is a policy question, not a technical one. Much depends on the application at hand. For instance, when the US Census Bureau adopted differential privacy, this task was delegated to a committee (the Data Stewardship Executive Policy committee, DSEP); details on the operations of this committee can be found in [19, Sect.3.1]. We think that similar considerations apply when setting the threshold of* **Ti***. For instance, an evaluator might consider the distribution of the* **Ti** *values in the dataset (see Fig. 3a–3h in Section VI) and then choose a percentile as a cutoff.*

The evaluator is also interested in the potential *utility* conveyed by an anonymized table for a learner. Note that the learner's utility is distinct from the attacker's one. Indeed, the learner's interest is to make inferences that are as close as possible to the ones that could be done using the cleartext table. Accordingly, obfuscated tables that are *faithful* to the original table are the most useful. This leads us to compare two distributions on the population: the distribution learned from the anonymized table, p_{L}, and the *ideal* (I) distribution, p_{I}, one can learn from the cleartext table t. The latter is formally defined as the expectation[2] of $f(s,r|\pi)$ under the posterior density $f(\pi|t)$. Explicitly, for each (s,r)
$$p_{\mathrm{I}}(s,r|t) \overset{\triangle}{=} \int_{\mathcal{D}} f(s,r|\pi) f(\pi|t)\, \mathrm{d}\pi. \tag{19}$$
Note that the posterior density $f(\pi|t)$ is in turn a Dirichlet density (see next section) and therefore a simple closed form of the above expression exists, based on the frequencies of the pairs (s,r) in t. In particular, recalling the α_s, β_r^s notation for the prior hyperparameters introduced in Section III, let $\alpha_0 = \sum_s \alpha_s$ and $\beta_0^s = \sum_r \beta_r^s$, and $\gamma_s(t)$ and $\delta_r^s(t)$ denote the frequency counts of s and (s,r), respectively, in t. Then we have
$$p_{\mathrm{I}}(s,r|t) = \frac{\alpha_s + \gamma_s(t)}{\alpha_0 + N} \cdot \frac{\beta_r^s + \delta_r^s(t)}{\beta_0^s + \gamma_s(t)}. \tag{20}$$

The comparison between p_{L} and p_{I} can be based on some form of *distance* between distributions. One possibility is to rely on *total variation* (aka statistical) distance. Recall that, for discrete distributions q, q' defined on the same space $\mathcal{X}$, the total variation distance is defined as
$$\mathbf{TV}(q,q') \overset{\triangle}{=} \sup_{A \subseteq \mathcal{X}} |q(A) - q'(A)| = \frac{1}{2} \sum_x |q(x) - q'(x)|.$$
Note that $\mathbf{TV}(q,q') \in [0,1]$. Note that this is a quite conservative notion of diversity since it based on the event that shows the largest difference between distributions.

Definition 2 (faithfulness): The *relative faithfulness level of* t^* *w.r.t.* t is defined as
$$\mathbf{RF}(t,t^*) \overset{\triangle}{=} 1 - \mathbf{TV}\big(\, p_{\mathrm{I}}(\cdot\,|\,t)\,,\; p_{\mathrm{L}}(\cdot\,|\,t^*)\,\big).$$

[2] Another sensible choice would be taking $p_{\mathrm{I}}(s,r|\,t) = f(s,r|\,\pi_{\mathrm{MAP}})$, where $\pi_{\mathrm{MAP}} = \mathrm{argmax}_{\pi}\, f(\pi|t)$ is the maximum a posteriori distribution given t. This choice would lead to essentially the same results.

IEEE Proof

Remark 4: In practice, the total variation of two high-dimensional distributions might be very hard to compute. Pragmatically, we note that for M large enough, $\mathbf{TV}(q, q') = \frac{1}{2} E_{x\sim q(x)}[|1-\frac{q'(x)}{q(x)}|] \approx \frac{1}{2M}\sum_{i=1}^{M}|1-\frac{q'(x_i)}{q(x_i)}|$*, where the x_i are drawn i.i.d. according to $q(x)$. Then a proxy to total variation is the* empirical *total variation defined below, where (s_i, t_i), for $i = 1, \ldots, M$, are generated i.i.d. according to $p_{\mathrm{I}}(\cdot, \cdot \,|\, t)$:*

$$\mathbf{ETV}(t, t^*) \triangleq \frac{1}{2M}\sum_{i=1}^{M}\left|1 - \frac{p_{\mathrm{L}}(s_i, r_i \,|\, t^*)}{p_{\mathrm{I}}(s_i, r_i \,|\, t)}\right|. \tag{21}$$

Remark 5 (ideal knowledge vs. attacker's knowledge): The following scenario is meant to further clarify the extra power afforded to the attacker, by the mere knowledge that his victim is in the table. Consider a trivial anonymization mechanism that simply releases the cleartext table, that is $t^* = t$. As $p_{\mathrm{L}} = p_{\mathrm{I}}$ in this case, it would be tempting to conclude that the attacker cannot do better than the learner, hence there is no relative *risk involved. However, this conclusion is wrong: for instance, $p_{\mathrm{I}}(\cdot|r_{\mathrm{v}}, t)$ can fail to predict the vicitim's correct sensitive value if this value is rare, as we show below.*

For the sake of simplicity, consider the case where the observed victim's nonsensitive attribute r_{v} occurs just once in t in a row (s_0, r_{v}). Also assume a noninformative Dirichlet prior, that is, in the notation of Section III, set the hyperparameters to $\alpha_s = \beta_r^s = 1$ for each $s \in \mathcal{S}, r \in \mathcal{R}$. Then, simple calculations based on (20) *and the attacker's distribution characterization* (8)*, show the following. Here for each $s \in \mathcal{S}$, $\gamma_s = \gamma_s(t)$ denotes the frequency count of s in t, and c a suitable normalizing constant*:

$$p_{\mathrm{I}}(s|r_{\mathrm{v}}, t) = \begin{cases} \dfrac{1+\gamma_s}{|\mathcal{R}|+\gamma_s}c, & \textit{if } s \neq s_0 \\ \dfrac{2(1+\gamma_{s_0})}{|\mathcal{R}|+\gamma_{s_0}}c, & \textit{if } s = s_0 \end{cases}$$

$$p_{\mathrm{A}}(s|r_{\mathrm{v}}, t^*) = \begin{cases} 0, & \textit{if } s \neq s_0 \\ 1, & \textit{if } s = s_0. \end{cases} \tag{22}$$

As far as the target individual $(s_0, r_{\mathrm{v}}) \in t$ is concerned, we see that while p_{A} predicts s_0 with certainty, predictions based on $p_{\mathrm{L}} = p_{\mathrm{I}}$ will be blatantly wrong, if there are values $s \neq s_0$ that occur very frequently in t, while s_0 is rare, and N is large compared to $|\mathcal{R}|$. To make an extreme numeric case, consider $|\mathcal{S}| = 2$, $|\mathcal{R}| = 1000$ and $\gamma_{s_0} = 1$ in a table t of $N =$ 10^6 *rows: plugging these values in* (22) *yields* $p_{\mathrm{L}}(s_0|r_{\mathrm{v}}, t^*) =$ $p_{\mathrm{I}}(s_0|r_{\mathrm{v}}, t) \approx 0.004$*, hence a relative threat for (s_0, r_{v}) of* $1/p_{\mathrm{L}}(s_0|r_{\mathrm{v}}, t^*) \approx 250$.

V. Learning From the Obfuscated Table by MCMC

Estimating the privacy threat and faithfulness measures defined in the previous section, for specific tables t and t^*, implies being able to compute the distributions (5), (6) and (8). Unfortunately, these distributions, unlike (19), are not available in closed form, since $f(\Pi = \pi \,|\, T^* = t^*) = f(\pi|t^*)$ cannot be derived analytically. Indeed, in order to do so, one should integrate $f(\pi, t|t^*)$ with respect to the density $f(t|t^*)$, which appears not to be feasible.

To circumvent this difficulty, we will introduce a *Gibbs sampler*, defining a Markov chain $(X_i)_{i\geq 0}$, with $X_i = (\Pi_i, T_i)$, converging to the density

$$\begin{aligned} &f(\Pi = \pi, T = t|t^*) \\ &\quad = f\big(\Pi = \pi,\ S_1 = s_1, R_1 = r_1, \ldots, S_N = s_N, R_N = r_N \,|\, t^*\big) \end{aligned}$$

(note that the sensitive values s_j in T are in fact fixed and known, given t^*). General results (see e.g. [41]) ensure that, if $\Pi_0, \Pi_1, \ldots$ are the samples drawn from the Π-marginal of such a chain, then for each $(s, r) \in \mathcal{S} \times \mathcal{R}$

$$\overset{1}{-}M\sum_{\ell=0}^{M} f(s, r|\Pi_\ell) \rightarrow \int_{\mathcal{D}} f(s, r|\pi) f(\pi|t^*)\mathrm{d}\pi \ = p_{\mathrm{L}}(s, r|t^*) \tag{23}$$

$$\overset{1}{-}M\sum_{\ell=0}^{M} f(s|r, \Pi_\ell) \rightarrow \int_{\mathcal{D}} f(s|r, \pi) f(\pi|t^*)\mathrm{d}\pi \ = p_{\mathrm{L}}(s|r, t^*) \tag{24}$$

almost surely as $M \longrightarrow +\infty$. Therefore, by selecting an appropriately large M, one can build approximations of $p_{\mathrm{L}}(s, r|t^*)$ and $p_{\mathrm{L}}(s|r, t^*)$ using the arithmetical means on the left-hand side of (23) and (24), respectively. Moreover, for each index $1 \leq j \leq N$, using samples drawn from the R_j-marginals of the same chain, one can build an estimate of $f(R_j = r_j \,|\, t^*)$. Consequently, using (8) (resp. (15), in the case of partial observation) one can estimate $p_{\mathrm{A}}(s|r_{\mathrm{v}}, t^*)$ (resp. $p_{\mathrm{A}}(s|y_{\mathrm{v}}, t^*)$) for any given r_{v} (resp. y_{v}).

In the rest of the section, we will first introduce the MCMC for this problem and then show its convergence. We will then discuss details of the sampling procedures for each of the two possible schemes, horizontal and vertical.

A. Definition and Convergence of the Gibbs Sampler

Simply stated, our problem is sampling from the marginals of the following target density function, where $t^* = g_1^*, \ldots, g_k^*$ and $t = g_1, \ldots, g_k$ (note that the number of groups k is known and fixed, given t^*).

$$f(\pi, t|t^*). \tag{25}$$

Note that the r_j's of interest, for $1 \leq j \leq N$, are the elements of the groups g_i's, for $1 \leq i \leq k$. The Gibbs scheme allows for some freedom as to the blocking of variables. Here we consider $k + 1$ blocks, coinciding with π and $g_1, \ldots, g_k$. This is natural as, in the considered schemes, $(R_i, S_i) \perp\!\!\!\perp (R_j, S_j)|\pi, t^*$ for (R_i, S_i) and (R_j, S_j) occurring in distinct groups. Formally, let $x^0 = \pi^0, t^0$ (with $t^0 = g_1^0, \ldots, g_k^0$) denote any initial state satisfying $f(\pi^0, t^0|t^*) > 0$. Given a state at step h, $x^h = \pi^h, t^h$ ($t^h = g_1^h, \ldots, g_k^h$), one lets $x^{h+1} \triangleq \pi^{h+1}, t^{h+1}$, where $t^{h+1} = g_1^{h+1}, \ldots, g_k^{h+1}$ and

$$\pi^{h+1} \quad \text{is drawn from} \quad f(\pi|t^h, t^*) \tag{26}$$

$$\begin{aligned} &g_i^{h+1} \quad \text{is drawn from} \\ &\quad f(g|\pi^{h+1}, g_1^{h+1}, \ldots, g_{i-1}^{h+1}, g_{i+1}^h, \ldots, g_k^h, t^*) \\ &\quad (1 \leq i \leq k). \end{aligned} \tag{27}$$

IEEE Proof

Running this chain presupposes we know how to sample from the *full conditional* distributions on the right-hand side of (26) and (27). In particular, there are several possible approaches to sample from g. In this subsection we provide a general discussion about convergence, postponing the details of sampling from the full conditionals to the next subsection.

Let us denote by $t_{-i} \overset{\triangle}{=} g_1, \ldots, g_{i-1}, g_{i+1}, \ldots, g_k$ the table obtained by removing the i-th group g_i from t. The following relations for the full conditionals of interest can be readily checked, relying on the conditional independencies of the model (2) and (4) (we presuppose that in each case the conditioning event has nonzero probability)

$$f(\pi|t, t^*) = f(\pi|t) \tag{28}$$

$$f(g|\pi, t_{-i}, t^*) \propto f(g|\pi) f(t^*|g, t_{-i}) \quad (1 \le i \le k). \tag{29}$$

As we shall see, each of the above two relations enables sampling from the densities on the left-hand side. Indeed, (28) is a posterior Dirichlet distribution, from which effective sampling can be easily performed (see next subsection). A straightforward implementation of (29) in a Acceptance-Rejection (AR) sampling perspective is as follows: draw g according to $f(g|\pi)$ and accept it with probability $f(t^*|g, t_{-i}) = f(t^*|t)$. Here, $f(t^*|t)$ is just the probability that the obfuscation algorithm returns t^* as output when given $t = g, t_{-i}$ as input. Actually, to make sampling from the RHS of (29) effective, further assumptions will be introduced (see next subsection). Note that, since the sensitive values are fixed in t and known from the given t^*, sampling g in (29) is actually equivalent to sampling the *nonsensitive* values of the group.

In addition to (29), to simplify our discussion about convergence, we shall henceforth assume that, for each group index $1 \le i \le k$, the set of instances of the i-th group that are compatible with t^* does *not* depend on the rest of the table, t_{-i}. That is, we assume that for each i $(1 \le i \le k)$:

$$\{g : f(t^*|g, t_{-i}) > 0\} = \{g : f(t^*|g, t'_{-i}) > 0\} \; \forall t_{-i} \text{ and } t'_{-i} \overset{\triangle}{=} \mathcal{G}_i. \tag{30}$$

For instance, (30) holds true if the anonymization algorithm ensures t^* is independent from t_{i-1} given a i-th group g: $t^* \perp\!\!\!\perp t_{-i} \mid g$.

Let $x = (\pi, g_1, \ldots, g_k)$ denote a generic state of this Markov chain. Under the assumption (30), the *support* of the target density $f(x|t^*)$ is the product space

$$\mathcal{X} \overset{\triangle}{=} \mathcal{D} \times \mathcal{G}_1 \times \cdots \times \mathcal{G}_k. \tag{31}$$

By this, we mean that $\{x : f(x|t^*) > 0\} = \mathcal{X}$. This is a consequence of: (a) the fact that Dirichlet only considers full support distributions; and (b) equation (29), taking into account the assumption (30). Let $X_0, X_1, \ldots$ denote the Markov chain defined by the sampler over $\mathcal{X}$ and denote by $\kappa(\cdot|\cdot)$ its conditional kernel density over $\mathcal{X}$. Slightly abusing notation, let us still indicate by $f(\cdot|t^*)$ the probability distribution over $\mathcal{X}$ induced by the density $f(x|t^*)$. Convergence in distribution follows from the following proposition, which is an instance of general results – see e.g. the discussion following Corollary 1 of [41].

Proposition 1 (convergence): Assume (30). For each (measurable) set $A \subseteq \mathcal{X}$ such that $f(A|t^*) > 0$ and each $x^0 \in \mathcal{X}$, we have $\kappa(X^1 \in A|X^0 = x^0) > 0$. As a consequence, the Markov chain $(X_i)_{i \ge 0}$ is irreducible and aperiodic, and its stationary density is $f(x|t^*)$ in (25).

B. Sampling From the Full Conditionals

Let us consider (28) first. It is a standard fact that the posterior of the Dirichlet distribution $f(\pi|t)$, given the N i.i.d. observations t drawn from the categorical distribution $f(\cdot|\pi)$, is still a Dirichlet, where the hyperparameters have been updated as follows. Denote by $\boldsymbol{\gamma}(t) = (\gamma_1, \ldots, \gamma_{|\mathcal{S}|})$ the vector of the frequency counts γ_i of each s_i in t. Similarly, given s, denote by $\boldsymbol{\delta}^s(t) = (\delta^s_1, \ldots, \delta^s_{|\mathcal{R}|})$ the vector of the frequency counts δ_i of the pairs (r_i, s), for each r_i, in t. Then, for each $\pi = (\pi_S, \pi_{R|S})$, we have

$$f(\pi|t) = \mathrm{Dir}(\pi_S \mid \boldsymbol{\alpha} + \boldsymbol{\gamma}(t)) \cdot \prod_{s \in \mathcal{S}} \mathrm{Dir}(\pi_{R|s} \mid \boldsymbol{\beta}^s + \boldsymbol{\delta}^s(t)). \tag{32}$$

Let us now discuss (29). In what follows, for the sake of notation we shall write a generic i-th group as $g_i = (s_1, r_1), \ldots, (s_n, r_n)$ (thus avoiding double subscripts), and let $g^*_i = (m_i, l_i)$ denote the corresponding obfuscated group in t^*. As already observed, given an obfuscated i-th group $g^*_i = (l_i, m_i)$, when sampling a i-th group g from (29), one actually needs to generate only the nonsensitive values of g, which are constrained by l_i, as the sensitive ones are already fixed by the sequence m_i. In what follows, to make sampling from (29) effective, will shall work under the following assumptions, which are stronger than (30).

(a) Deterministic obfuscation function: for each t and t^*, $f(t^*|t)$ is either 0 or 1.

(b) For each $1 \le i \le k$, letting $g^*_i = (l_i, m_i)$, with $m_i = s_1, \ldots, s_n$, the i-th obfuscated group in t^*, the following holds true:

Horizontal schemes

$$\mathcal{G}_i = \{g = (s_1, r_1), \ldots, (s_n, r_n) : r_\ell \in l_i \quad \text{for } 1 \le \ell \le n\} \tag{33}$$

Vertical schemes

$$\mathcal{G}_i = \{g = (s_1, r_{i_1}), \ldots, (s_n, r_{i_n}) : \text{for } r_{i_1}, \ldots, r_{i_n} \text{ a permutation of } l_i\}. \tag{34}$$

Assumption (a) is realistic in practice. In horizontal schemes, assumption (b) makes the considered sets $\mathcal{G}_i$'s possibly larger than the real ones, that is $l_i \supset \{r_1, \ldots, r_n\}$. This happens, for instance, if in certain groups the ZIP code is constrained to just, say, two values, while the generalized code "5013*" allows for all values in the set $\{50130, \ldots, 50139\}$. We will not attempt here a formal analysis of this assumption. In some cases, such as in schemes based on global recoding, this assumption is realistic. Otherwise, we only note that the support $\mathcal{X}$ of the resulting Markov chain may be (slightly) larger than the one that would be obtained not assuming (33) or (34). Heuristically, this leads one to sampling from a more dispersed density than the target one. At least, the resulting distributions can be taken to represent a lower bound of what the attacker can actually learn.

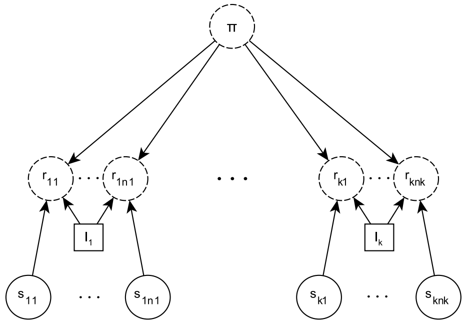


Fig. 1. Sampling from $f(g|\pi, t_{-i}, t^*)$ ($g \in \mathcal{G}_i$) for horizontal schemes, across all the groups.

Under assumptions (a) and (b) above, for each $1 \leq i \leq k$, it holds that $g \in \mathcal{G}_i$ if and only if $f(t^*|g, t_{-i}) = 1$. Therefore sampling according to the right-hand side of (29) reduces to the following:

$$\text{draw } g \in \mathcal{G}_i \text{ with probability } \propto f(g|\pi) \quad (1 \leq i \leq k). \quad (35)$$

We discuss now how to implement (35) effectively. This will achieve sampling from the full conditionals (29) without resorting to a presumably inefficient AR method. We deal with the two cases, horizontal and vertical, separately.

a) Horizontal schemes: In order to generate $g = (r_1, s_1), \ldots, (r_n, s_n) \in \mathcal{G}_i$, for each $\ell = 1, .., n$, we draw $r_\ell \in l_i$ with probability $\propto f(r_\ell|s_\ell, \pi)$. Explicitly, (29) now becomes

$$f(g|\pi, t_{-i}, t^*) = \begin{cases} 0, & \text{if } g \notin \mathcal{G}_i \\ \displaystyle\prod_{\ell=1}^{n} \frac{f(r_\ell|s_\ell, \pi)}{\sum_{r \in l_i} f(r|s_\ell, \pi)}, & \text{if } g \in \mathcal{G}_i \end{cases} \quad (36)$$

thus satisfying (35). Note that this is equivalent to sampling each row independently. The sampling process of $f(g|\pi, t_{-i}, t^*)$ for horizontal schemes across all the groups of the table is illustrated graphically in Fig. 1.

b) Vertical schemes: Let $l_i = \{\!| r_1, \ldots, r_n |\!\}$. We have that $g \in \mathcal{G}_i$ if and only if $g = (s_1, r_{i_1}), \ldots, (s_n, r_{i_n})$, for some permutation $(r_{i_\ell})_{1 \leq \ell \leq n}$ of $r_1, \ldots, r_n$. Here, sampling the nonsensitive values of g row by row would involve to gradually reduce the sample space. A sampling procedure along these lines is possible, but nontrivial, see Appendix B.

We discuss here a more straightforward sampling procedure, based on generating $g_i \in \mathcal{G}_i$ in a single shot. We adopt a *single-iteration Metropolis within Gibbs* scheme. Essentially, this consists in running a Metropolis method that targets the distribution $\propto f(g|\pi)$ with support $\mathcal{G}_i$, for one iteration. Specifically, let us write the current value of the i-th group in the Gibbs Markov chain as g_i^h. Following Casella and Robert [40, Ch.10], this step consists in drawing $g \in \mathcal{G}_i$ according to a proposal distribution $J(g|g_i^h)$ and accepting it, that is letting $g_i^{h+1} = g$, with probability

$$\epsilon \overset{\triangle}{=} \min\left\{1, \frac{f(g|\pi)J(g_i^h|g)}{f(g_i^h|\pi)J(g|g_i^h)}\right\} \quad (37)$$

while keeping $g_i^{h+1} = g_i^h$ with probability $1 - \epsilon$. The resulting MCMC method is still theoretically sound: see Casella and Robert [40, Ch.10.3.3]. As to the proposal distribution $J(g|g_i^h)$, a possibility is generating $g \in \mathcal{G}_i$ via a pure random permutation of the n nonsensitive values in l_i; or just to swap the nonsensitive values of two randomly chosen positions in g_i^h. In both cases, the proposal is symmetric, and (37) simplifies accordingly as follows, where $r_1, \ldots, r_n$ is the sequence of sensitive values in the poposed g:

$$\epsilon = \min\left\{1, \frac{\prod_{\ell=1}^{n} f(r_\ell|s_\ell, \pi)}{\prod_{\ell=1}^{n} f(r_\ell^h|s_\ell, \pi)}\right\}.$$

TABLE IV

SUMMARY OF THREAT AND FAITHFULNESS MEASURES FOR ANONYMIZATION ACCORDING TO K-ANONYMITY AND ℓ- DIVERSITY

		Group size and diversity	
		k = ℓ = 4	k = ℓ = 6
Global threat level under p_A	$\mathbf{GT}_A$	0.2930	0.2994
Global threat level under p_L	$\mathbf{GT}_\text{L}$	0.2681	0.2756
Global threat level under p_RW	$\mathbf{GT}_\text{RW}$	0.2131	0.2890
Relative global threat	$\mathbf{RGT_A}$	0.0249	0.0232
Empirical relative faithfulness level	**RF**	0.3106	0.3011
Absolute error under p_A	$\mathbf{ABS}_\text{A}$	9795.58	9699.09
Absolute error under p_RW	$\mathbf{ABS}_\text{RW}$	9980.35	9451.53
Baseline accuracy	0.1656		
Ideal accuracy	0.3534		

VI. EXPERIMENTS

We have put a proof-of-concept implementation[3] of our methodology at work on a subset of the Adult dataset extracted by Barry Becker from the 1994 US Census database and available from the UCI machine learning repository [47]. This is a common benchmark for experiments on anonymization [38]. In particular, we have focused on the subset of 5692 rows also considered by the authors of [38], with the following categorical attributes: *sex*, *age*, *race*, *marital status*, *education*, *native country*, *workclass*, *salary class*, *occupation*, with *occupation* (14 values) considered as the only sensitive attribute. We will discuss implementation and results details separately for vertical and horizontal schemes. We will then briefly discuss convergence issues of the employed MCMC method.

A. Horizontal Schemes: k-Anonymity

Using the ARX anonymization tool [37] we obtained two different k-anonymous versions of the considered dataset, enjoying respectively k-anonymity and ℓ-diversity[4] for k = ℓ = 4 and k = ℓ = 6. The average size of the groups was respectively of 38 rows (k = ℓ = 4) and of 355 rows (k = ℓ = 6).

The results we have obtained are summarized in Table IV. For reference, we include the following information in the last two lines: *baseline accuracy*, the fraction of rows correctly classified using the empirical distribution obtained from the frequencies of the sensitive values in the anonymized table – i.e., the fraction of the most frequent sensitive value; and

[3] Python code and data available from the authors.

[4] Recall that ℓ-diversity requires at least ℓ distinct values of the sensitive attribute in each group.

IEEE Proof

ideal accuracy, the fraction of tuples threatened under p_{I}. As a further element of comparison, we also consider an attacker whose reasoning is based on the random worlds models, and include in the table $\mathbf{GT}_{\mathrm{RW}}$, the fraction of rows correctly classified assuming all tables compatible with t^* equally likely. Like in [25], we compute $\mathbf{ABS}_{\mathrm{A}}$ and $\mathbf{ABS}_{\mathrm{RW}}$, the *absolute error* under the distribution derived under p_{A} and under the random worlds distribution p_{RW}, respectively. $\mathbf{ABS}$ is defined as $\sum_{i=1}^{N}\sum_{s\in\mathcal{S}}|\mathbf{1}_{\{s_i=s\}} - p(s|r_i,t^*)|$, where $p(\cdot)$ might be either of $p_{\mathrm{A}}(\cdot)$ or $p_{\mathrm{RW}}(\cdot)$. Note that, since the considered anonymized tables do not enjoy disjointness between groups (see Remark 1), also in the random worlds perspective the probability of each sensitive attribute may well be $\geq 1/\ell$. In our experiments, when $\ell = 4$ the attacker outperforms random worlds classification, while when a more powerful obfuscation is adopted the two results are quite similar.

The remaining rows in Table IV consider the privacy threats and faithfulness measures introduced in Section IV. As a general comment, small variations of ℓ and/or k do not produce dramatic changes. The faithfulness level is stable, but does not reach a satisfactory level. The attacker is anyway in a position to correctly classify the sensitive attribute of individuals in the table $\approx 2.3 - 2.5\%$ more often than the learner. We found the maximum value of $\mathbf{Ti}_{\mathrm{A}}$ for the threatened rows is about 13.8, meaning the attacker can be up to ≈14 times more confident than the learner about the guessed value.

A more informative summary of our analysis is provided by the scatter plots and histograms of Figure 2. The scatter plots are obtained from the threat levels under p_{L} and under p_{A}. The number of rows (s, r) in which $p_{\mathrm{A}}(s|r, t^*) \geq p_{\mathrm{L}}(s|r, t^*)$ roughly equals those in which $p_{\mathrm{A}}(s|r, t^*) \leq p_{\mathrm{L}}(s|r, t^*)$, although globally the attacker has a slight advantage in terms of number of threatened rows. In Figure 2 we also report the empirical distribution $\log_2 \mathbf{Ti}_{\mathrm{A}}$ for tuples threatened under p_{A} and under p_{L}. We also have evidence of positive skewness, as shown by the value of γ (the third standardized moments of the empirical distributions). Recalling that $\log_2 \mathbf{Ti}_{\mathrm{A}} = 1$ means $p_{\mathrm{A}}(s|r, t^*) = 2p_{\mathrm{L}}(s|r, t^*)$, the histograms show that $p_{\mathrm{A}}(s|r, t^*)$ is often more than twice $p_{\mathrm{L}}(s|r, t^*)$ leading to a $\log_2 \mathbf{Ti}_{\mathrm{A}} \geq 1$. In particular, when $k = \ell = 4$, $\log_2 \mathbf{Ti}_{\mathrm{A}}$ is at least 1 for ≈ 6% of the individuals threatened under p_{A}, meaning ≈ 0.6% of the whole table. Conversely, $\log_2 \mathbf{Ti}_{\mathrm{A}}$ is close to 0 for most of the rows in which $p_{\mathrm{A}}(s|r, t^*) \leq p_{\mathrm{L}}(s|r, t^*)$.

B. Vertical Schemes: Anatomy

Using a freely available anonymization tool [22], we have obtained two anatomized versions of the considered dataset, with groups of size $\ell = 4$ and $\ell = 6$, respectively. The resulting tables also enjoy ℓ-diversity. The results we have obtained are summarized in Table V. Concerning the random worlds approach, we note the following. Anatomy partitions the tables in groups all of size ℓ. Therefore, although disjointness is not satisfied, just as in the horizontal case, the sensitive attribute frequencies equal $1/\ell$ in each group. This implies that the probability of a sensitive value depends on how many groups contain the victim's nonsensitive attributes and on their frequencies in each group, leading often to multimodal distributions. We assume that a guess may be obtained randomly choosing between the equally likely sensitive attributes. Accordingly, the fractions of threatened rows, $\mathbf{GT}_{\mathrm{RW}}$, are averaged over 500 different sampling. Here, it is apparent that the our attacker is able to classify better than the random worlds scenario. We note that, as ℓ increases from 4 to 6, the fraction of rows threatened under the distributions derived by the learner ($\mathbf{GT}_{\mathrm{L}}$) and by the attacker ($\mathbf{GT}_{A}$) decreases significantly. Moreover, as ℓ grows both the relative threat $\mathbf{RGT}_{\mathbf{A}}$ and the faithfulness level $\mathbf{RF}$ decrease, which implies a trade-off between privacy and the utility conveyed by the table.

Again, for a more informative summary of our analysis, we look at scatter plots and histograms, displayed in Figure 3, where we compare p_{A} and p_{L} on threatened rows. It is apparent here that the attacker is more confident than the learner in the majority of the cases, even when focusing on the rows threatened under p_{L}. This is in contrast with the horizontal case, where the attacker exhibits smaller threat

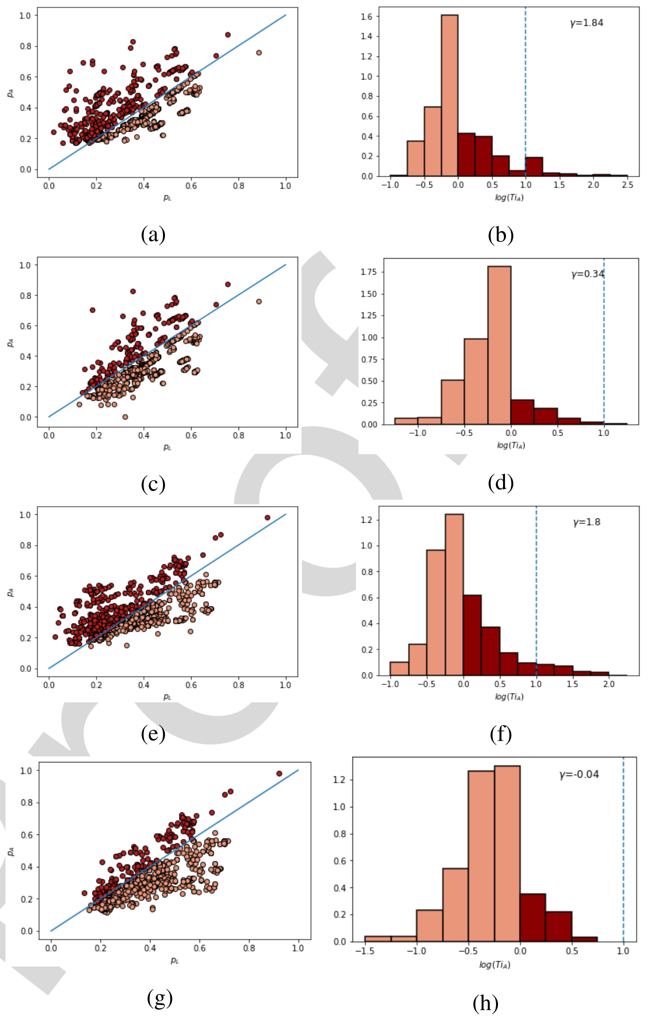


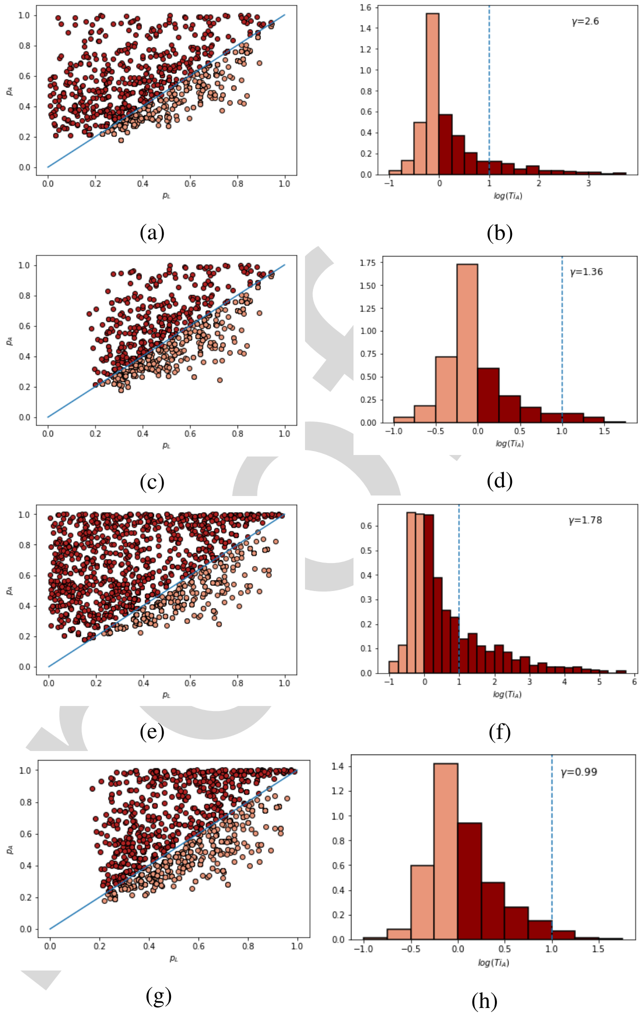
Fig. 2. Results for k-anonymity. Top (ℓ = k = 6): scatter plots of p_{L} vs p_{A} for tuples threatened under p_{A} (a), and under p_{L} (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_{\mathrm{A}}$ for these two cases. Bottom: same for ℓ = k = 4. The skewness value (γ) represents the third standardized moment of the empirical distribution. Dark red areas show where the attacker performs better than the learner.

IEEE Proof

TABLE V

SUMMARY OF THREAT AND FAITHFULNESS MEASURES FOR ANONYMIZATION ACCORDING TO ANATOMY

		Group size and diversity	
		$\ell = 4$	$\ell = 6$
Global threat level under p_A	$\mathbf{GT}_A$	0.3273	0.2396
Global threat level under p_L	$\mathbf{GT}_L$	0.2653	0.2136
Global threat level under p_{RW}	$\mathbf{GT}_{RW}$	0.1669	0.1689
Relative global threat	$\mathbf{RGT}_A$	0.0620	0.0260
Empirical relative faithfulness level	$\mathbf{RF}$	0.6493	0.5341
Absolute error under p_A	$\mathbf{ABS}_A$	8391.66	9276.25
Absolute error under p_{RW}	$\mathbf{ABS}_{RW}$	9471.94	9889.07
Baseline accuracy		0.1656	
Ideal accuracy		0.3534	



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 3. Results for Anatomy. Top ($\ell = 6$): scatter plots of p_L vs p_A for tuples threatened under p_A (a), and under p_L (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_A$ for these two cases. Bottom: same for $\ell = 4$. The skewness value (γ) represents the third standardized moment of the empirical distribution. Dark red areas show where the attacker performs better than the learner.

levels on the rows threatened under p_L (Figure 2, (d) and (h)). As far as the histograms are concerned, an even greater skewness than the horizontal case is evident here. In particular, the attacker can be up to ≈ 287 times more confident then the learner, being the maximum $\mathbf{Ti}_A$ about 286.19. Moreover, when $\ell = 4$, the individuals with $\log_2 \mathbf{Ti}_A \geq 1$ are $\approx 26\%$ of the rows threatened under p_A ($\approx 8\%$ of the whole table). This means that there are 483 individuals in the dataset for which the threat level under p_A is at least twice as much the threat level under p_L.

C. Discussion

Comparing the horizontal and the vertical cases for the considered dataset, the following considerations are in order.

- In the horizontal case, we have a situation of low faithfulness and low privacy threat, irrespective of the value of k and ℓ. Indeed, in both cases the average group size is well above k, and this has a negative effect on the inference capabilities of both the learner and the attacker. The slight numerical differences observed between the cases $k = \ell = 4$ and $k = \ell = 6$ are basically an artifact of the anonymization tool. Yet, in relative terms, one can observe a significant increase in the number of tuples threatened by the attacker, over the learner.
- In the vertical case, one obtains a greater faithfulness at the price of a greater privacy threat. This difference from the horizontal case is partly explained by the smaller group size, which now coincides with ℓ. Now moving from $\ell = 4$ to $\ell = 6$ has a tangible negative impact on the inference capabilities of both the learner and the attacker. In relative terms, one can observe an even more marked increase of the number of tuples threatened by the attacker, over the learner.

The above considerations partly depend on both the original dataset and the details of the employed anonymization tool.

D. Assessing MCMC *Convergence*

For each of the considered anonymized datasets, we ran a MCMC as introduced in Section V for $M = 100,000$ runs. The convergence of each chain to the stationary distribution was assessed via a methodology based on comparing sub-sequences of the sample sequences with one another. More precisely, as for the population parameters distribution (32), we used the method proposed by Geweke [21]. The Geweke proposal is based on an adapted two-samples test on the means in sub-sequences of the chain.

After a burn-in of 50,000 iterations, we compared the last 25,000 samples against 5 blocks of of 5,000 consecutive samples each, taken starting from the 50,000-th iteration. We found that all the distributions $\pi_{R|S}$ produced a test statistic within two standard deviations from zero, thus providing evidence of convergence.

As for the distribution of the cleartext table, $f(t|\pi, t^*)$, we used a test specifically designed for categorical distributions by Deonovich and Smith, called Weiß procedure [15]. The approach is based on a χ^2 test adjusted for the autocorrelation induced by the chain. The test is based on partitioning the whole sample sequence into sub-sequences, and then testing the homogeneity between the empirical distribution of each sub-sequence and the empirical distribution of the whole chain. After a burn-in of 50,000 observations, we compared 5 sub-sequences of 10,000 consecutive samples each. For the vertical scheme, we assessed the convergence for each row of the table, thereby demonstrating the stationary of $f(t|\pi, t^*)$.

For the horizontal scheme, some of the rows did not exhibit evidence of convergence. However, we found that, starting with several independent chains, very similar results in terms of the proposed assessment measures were obtained.

In the vertical case, within the Metropolis step both the pure random permutation and the swap group generation strategies (Section V-B) were experimented. The obtained results are consistent; however, the pure random permutation strategy shows a much higher rate of rejection, suggesting that the swap strategy should be preferred.

VII. Conclusion

We have put forward a notion of relative privacy threat that applies to group-based anonymization schemes. Our proposal is based on a rigorous characterization of the learner's and of the attacker's inference, in a unified Bayesian model of group-based schemes. A related MCMC algorithm for posterior parameters estimation has also been introduced. Experiments conducted on the well-known Adult dataset [47] have been illustrated.

Our analysis emphasizes the risks posed by the mere fact that an attacker can look up a released anonymized table. This prompts an obvious alternative: release the parameters of the posterior distribution learned from the cleartext table (p_{I}, in our notation). This may not always be possible, or be a good idea, for several reasons. First, certain organizations must release datasets as part of their mission, e.g. census bureaus. Second, especially in the case of high-dimensional data, the computation of the posterior is feasible only assuming suitable conditional independencies, whereby potentially important correlations are lost; see [10] and references therein. Third, parameters release itself is not exempt from risks for privacy. In particular, although differentially private release of the parameters is possible [16], it seems that quite strong priors are necessary to obtain acceptable guarantees; see [50, Ch.6] and references therein. In conclusion, further research is called for in order to understand under what circumstances data and/or parameters release can be done safely.

Appendix A
Proof of Lemma 1

We first characterize the probability $f(V=j|R_V=r_{\mathrm{v}},t^*)$, for an arbitrary $j\in\{1,\ldots,N\}$. Bayes theorem yields

$$\begin{aligned}f(V=j|R_V=r_{\mathrm{v}},t^*) &\propto f(R_V=r_{\mathrm{v}}|V=j,t^*)f(V=j|t^*)\\ &= f(R_j=r_{\mathrm{v}}|V=j,t^*)f(V=j|t^*)\\ &\propto f(R_j=r_{\mathrm{v}}|V=j,t^*) &(38)\\ &= f(R_j=r_{\mathrm{v}}|t^*) &(39)\end{aligned}$$

where (38) follows from $f(V=j|t^*)=f(V=j)=1/N$ (independence of V), and (39) follows because, as easily checked, for any fixed j, independence of R_j and V is preserved by conditioning on t^*. Now we have, for every $s\in\mathcal{S}$

$$\begin{aligned}&p_{\mathrm{A}}(s|r_{\mathrm{v}},t^*) &(40)\\ &=f(S_V=s\mid R_V=r_{\mathrm{v}},\ t^*)\\ &=\sum_j f(S_V=s,V=j|R_V=r_{\mathrm{v}},t^*)\\ &=\sum_j f(S_V=s|V=j,R_V=r_{\mathrm{v}},t^*)f(V=j|R_V=r_{\mathrm{v}},t^*)\\ &=\sum_j f(S_j=s|V=j,R_j=r_{\mathrm{v}},t^*)f(V=j|R_V=r_{\mathrm{v}},t^*)\\ &=\sum_{j\,:\,s_j=s} f(S_j=s|V=j,R_j=r_{\mathrm{v}},t^*)f(V=j|R_V=r_{\mathrm{v}},t^*)\\ & &(41)\\ &=\sum_{j\,:\,s_j=s} f(V=j|R_V=r_{\mathrm{v}},t^*) &(42)\\ &\propto\sum_{j\,:\,s_j=s} f(R_j=r_{\mathrm{v}}|t^*). &(43)\end{aligned}$$

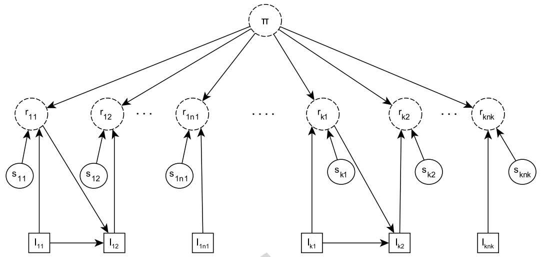


Fig. 4. Sampling from $\theta(g|\pi,t^*)$ for vertical schemes.

where (41) and (42) follow from the fact that, for $s_j\neq s$, $f(S_j=s,t^*)=0$, while for $s_j=s$ obviously $f(S_j=s|V=j,R_j=r_{\mathrm{v}},t^*)=1$. Finally, (43) follows from (39).

Note that in (43) each term on the RHS actually is the joint probability $f(R_j=r_{\mathrm{v}},S_j=s|t^*)$, being $s_j=s$ embedded in the range of the summation.

Appendix B
An Alternative Group Sampling Method for Vertical Schemes

We consider the following method for sampling $g\in\mathcal{G}_i$. Draw n values r_{i_ℓ}, $\ell=1,\ldots,n$, as follows:

1. draw r_{i_1} from l_i according to a distribution $\propto f(r|s_1,\pi)$;
2. draw r_{i_2} from $l_i\setminus\{\!|\, r_{i_1}\,|\!\}$ according to a distribution $\propto f(r|s_2,\pi)$;

   …

n. draw r_{i_n} from $l_i\setminus\{\!|\, r_{i_1},\ldots,r_{i_{n-1}}\,|\!\}$ according to a distribution $\propto f(r|s_n,\pi)$.

For a multiset l', let $\sigma(l'|s_\ell,\pi)\overset{\triangle}{=}\sum_{r\text{ in }l'}f(r|s_\ell,\pi)$ denote the probability of extracting some element appearing in l' (disregarding multiplicities) according to $f(\cdot|s_\ell,\pi)$. Using this notation, the probability of returning exactly the sequence $r_{i_1},\ldots,r_{i_n}$, hence $g=(s_1,r_{i_1}),\ldots,(s_n,r_{i_n})\in\mathcal{G}_i$, as a result of the above n drawings, can be written as

$$\begin{aligned}\theta(g|\pi,t^*)&\overset{\triangle}{=}\frac{f(r_{i_1}|s_1,\pi)}{\sigma(l_i|s_1,\pi)}\cdot\frac{f(r_{i_2}|s_2,\pi)}{\sigma(l_i\setminus\{\!|\, r_{i_1}\,|\!\}|s_2,\pi)}\cdots\frac{f(r_{i_n}|s_n,\pi)}{f(r_{i_n}|s_n,\pi)}\\ &=\frac{\prod_{\ell=1}^n f(r_{i_\ell}|s_\ell,\pi)}{\nu(g|\pi)}\end{aligned}$$

where we denote by $\nu(g|\pi)$ the denominator of the expression on the RHS of $\overset{\triangle}{=}$ above. The sampling process of $\theta(g|\pi,t^*)$ for vertical schemes across all the groups of the table is illustrated in Fig. 4. We note that $\theta(g|\pi,t^*)$ is dependent on the chosen ordering of the sensitive values $s_1,\ldots,s_n$, which

IEEE Proof

may invalidate condition (35). A possible solution could be to sweep the order of sampling according to the Random Sweep Gibbs sampler scheme originally proposed by [20] and further developed by [29].

References

[1] D. J. Balding and P. Donnelly, "Inference in forensic identification," *J. Roy. Stat. Soc. A, Statist. Soc.*, vol. 158, no. 1, pp. 21–53, 1995.

[2] M. Bewong, J. Liu, L. Liu, J. Li, and K. K. R. Choo, "A relative privacy model for effective privacy preservation in transactional data," in *Proc. IEEE Trustcom/BigDataSE/ICESS*, Aug. 2017, pp. 394–401.

[3] B. Bichsel, T. Gehr, D. Drachsler-Cohen, P. Tsankov, and T. Vechev, "DP-finder: Finding differential privacy violations by sampling and optimization," in *Proc. ACM CCS*, 2018, pp. 508–524.

[4] M. Boreale and M. Paolini, "Worst- and average-case privacy breaches in randomization mechanisms," in *Theoretical Computer Science*, vol. 597, Berlin, Germany: Springer, 2015, pp. 40–61.

[5] M. Boreale and F. Corradi, "Relative privacy risks and learning from anonymized data," in *Proc. SIS*, A. Petrucci, F. Racioppi, and R. Verde, Eds. Firenze Univ. Press, 2017, pp. 199–204.

[6] D. Cavallini and F. Corradi, "Forensic identification of relatives of individuals included in a database of DNA profiles," *Biometrika*, vol. 93, pp. 525–536, Sep. 2006.

[7] A.-S. Charest, "How can we analyze differentially-private synthetic datasets? " *J. Privacy Confidentiality*, vol. 2, no. 2, 2011. AQ:4

[8] A.-S. Charest, "Empirical evaluation of statistical inference from differentially-private contingency tables," in *Proc. Int. Conf. Privacy Stat. Databases (PSD)*. Berlin, Germany: Springer-Verlag, 2012, pp. 257–272.

[9] R. C.-W. Wong, A. W. C. Fu, K. Wang, P. S. Yu, and J. Pei, "Can the utility of anonymized data be used for privacy breaches?" *ACM Trans. Knowl. Discovery Data*, vol. 5, no. 3, pp. 16-1–16-24, 2011.

[10] C. Clifton and T. Tassa, "On syntactic anonymity and differential privacy," in *Proc. Trans. Data Privacy*, vol. 6, 2013, pp. 161–183.

[11] F. Corradi, V. Pinchi, S. Garatti, and I. Barsanti, "Probabilistic classification of age by third molar development: The use of soft evidence," *J. Forensic Sci.*, vol. 58, no. 1, pp. 51–59, 2013.

[12] F. K. Dankar and K. El Emam, "The application of differential privacy to health data," in *Proc. Joint EDBT/ICDT Workshops (EDBT-ICDT)*, 2012, pp158–166.

[13] F. K. Dankar and K. El Emam, "Practicing differential privacy in health care: A review," in *Trans. Data Privacy*, vol. 6, no. 1, pp. 35–67, 2013.

[14] A. P. Dawid, "The island problem: Coherent use of identification evidence," in *Aspects of Uncertainty: A Tribute to D. V. Lindley*, P. R. Freeman and A. F. M. Smith, Eds. Hoboken, NJ, USA: Wiley, 1994, pp. 159–170.

[15] B. E. Deonovic and B. J. Smith, "Convergence diagnostics for MCMC draws of a categorical variable," 2017, *arXiv:1706.04919*. [Online]. Available: https://arxiv.org/abs/1706.04919

[16] C. Dimitrakakis, B. Nelson, A. Mitrokotsa, and B. I. P. Rubinstein, "Robust and Private Bayesian Inference," in *Proc. 25th Int. Conf. Algorithmic Learn. Theory (ALT)*, Bled, Slovenia, Oct. 2014, pp. 291–305.

[17] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer, "Detecting violations of differential privacy," in *Proc. ACM CCS*, 2018, pp. 475–489.

[18] C. Dwork, "Differential privacy," in *Proc. 33rd Int. Conf. Automata, Lang. Program.* Berlin, Germany: Springer-Verlag, 2006, pp. 1–12.

[19] S. L. Garfinkel, J. M. Abowd, and S. Powazek, "Issues encountered deploying differential privacy," in *Proc. ACM Workshop Privacy Electron. Soc.*, 2018, pp. 133–137.

[20] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, no. 6, pp. 721–741, Nov. 1984.

[21] J. Geweke, "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in *Bayesian Statistics*. Oxford, U.K.: Oxford Univ. Press, 1992, pp. 169–193.

[22] Q. Gong. (2014). *Anatomize, GitHub*. [Online]. Available: https://github.com/qiyuangong/Anatomize

[23] A. Inan, M. Kantarcioglu, and E. Bertino, "Using anonymized data for classification," in *Proc. IEEE Int. Conf. Data Eng. (ICDE)*, Washington, DC, USA, Mar./Apr. 2009, pp. 429–440.

[24] A. Kassem, G. Acs, C. Castelluccia, and C. Palamidessi, "Differential inference testing a practical approach to evaluate anonymized data," INRIA, Res. Rep., 2018, pp. 1–21. AQ:5

[25] D. Kifer, "Attacks on privacy and deFinetti's theorem," in *Proc. SIGMOD Conf.*, 2006, pp. 127–138.

[26] D. Kifer and A. Machanavajjhala, "No free lunch in data privacy," in *Proc. SIGMOD*, 2011, pp. 193–204.

[27] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and ℓ-diversity," in *Proc. ICDE*, 2007, pp. 106–115. doi: 10.1109/ICDE.2007.367856.

[28] N. Li, W. Qardaji, and D. Su, "On sampling, anonymization, and differential privacy: Or, k-anonymization meets differential privacy," in *Proc. 7th ACM Symp. Inf., Comput. Commun. Secur. (ASIACCS)*, 2012, pp. 32–33. AQ:6

[29] J. S. Liu, "Markov chain Monte Carlo and related topics," Dept. Statist., Stanford Univ., Stanford, CA, USA, Tech. Rep., 1995.

[30] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, "ℓ-diversity: Privacy beyond k-anonymity," in *Proc. ICDE*, 2006, p. 24.

[31] A. Machanavajjhala, D. Kifer, J. Abowd, J. Gehrke, and L. Vilhube, "Privacy: Theory meets practice on the map," in *Proc. IEEE 24th Int. Conf. Data Eng.*, Apr. 2008, pp. 277–286. doi: 10.1109/ICDE.2008.4497436.

[32] M. D. Mailman *et al.*, "The NCBI dbGaP database of genotypes and phenotypes," *Nature Genet.*, vol. 39, no. 10, pp. 1181–1186, 2007. doi: 10.1038/ng1007-1181.

[33] K. Mancuhan and C. Clifton, "Statistical learning theory approach for data classification with ℓ-diversity," 2016, *arXiv:1610.05815*. [Online]. Available: https://arxiv.org/abs/1610.05815

[34] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large datasets (how to break anonymity of the Netflix prize dataset)," Univ. Texas Austin, Austin, TX, USA, Tech. Rep., 2008.

[35] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Toronto, ON, Canada, 2018, pp. 634–646. doi: 10.1145/3243734.3243855.

[36] W. Ollier, T. Sprosen, and T. Peakman, "UK Biobank: From concept to reality," *Future Med.*, vol. 6, no. 6, pp. 639–646, 2005.

[37] F. Prasser and F. Kohlmayer, "Putting statistical disclosure control into practice: The ARX data anonymization tool," in *Medical Data Privacy Handbook*, A. Gkoulalas-Divanis and G. Loukides, Eds. Cham, Switzerland: Springer, Nov. 2015.

[38] F. Prasser, F. Kohlmayer, and K. A. Kuhn, "A benchmark of globally-optimal anonymization methods for biomedical data," in *Proc. 27th IEEE Int. Symp. Comput.-Based Med. Syst.*, New York, NY, USA, May 2014, pp. 66–71.

[39] A. Pyrgelis, C. Troncoso, and E De Cristofaro, "Knock knock, who's there? Membership inference on aggregate location data," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, Feb. 2018, pp. 18–21. doi: 10.14722/ndss.2018.23183.

[40] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. 2nd ed. Springer, 2004. AQ:7

[41] G. O. Roberts and A. F. M. Smith, "Simple conditions for the convergence of the Gibbs sampler and Metropolis–Hastings algorithms," *Stochastic Processes Appl.*, vol. 49, pp. 207–216, Feb. 1994.

[42] T. E. Raghunathan, J. P. Rubin, and D. B. Reiter, "Multiple imputation for statistical disclosure limitation," *J. Off. Statist.*, vol. 19, no. 1, pp. 1–16, 2003.

[43] D. B. Rubin, "Statistical disclosure limitation," *J. Off. Statist.*, vol. 9, no. 2, pp. 461–468, 1993.

[44] R. Sarathy and K. Muralidhar, "Evaluating Laplace noise addition to satisfy differential privacy for numeric data," *Trans. Data Privacy*, vol. 4, no. 1, pp. 1–17, 2011.

[45] K. Slooten and R. Meester, "Forensic identification: The island problem and its generalisations," 2017. *arXiv:1201.4647*. [Online]. Available: https://arxiv.org/abs/1201.4647

[46] L. Sweeney, "K-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 10, no. 5, pp. 557–570, 2002.

[47] *UCI Machine Learning Repository, Adult Dataset*. (1996). [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Adult

[48] U.S. Office for Civil Rights. (Nov. 26, 2012). *Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance With the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule*. [Online]. Available: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf

[49] X. Xiao and Y. Tao, "Anatomy: Simple and effective privacy preservation," in *Proc. VLDB*, 2006, pp. 139–150.

[50] S. Zheng, "The differential privacy of Bayesian inference," B.S. thesis, Harvard College, Cambridge, MA, USA, 2015. [Online]. Available: https://dash.harvard.edu/handle/1/14398533

IEEE Proof