



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)

Relative Privacy Threats and Learning From Anonymized Data

Michele Boreale, Fabio Corradi^{id}, 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

¹Depending 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.

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

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

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

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

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

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 \rightarrow posterior parameters \rightarrow 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

multiset of rows \rightarrow cleartext table \rightarrow obfuscated table.

Formally, fix $N \geq 1$. Given a multiset of N rows, $d = \{(s_1, r_1), \dots, (s_N, r_N)\}$, \mathcal{A} will first arrange d into a sequence of *groups*, $t = g_1, \dots, g_k$, the *cleartext table*. Each group in turn is a sequence of n_i rows, $g_i = (s_{i,1}, r_{i,1}), \dots, (s_{i,n_i}, r_{i,n_i})$, 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 $t^* = g_1^*, \dots, g_k^*$, where the obfuscation of each group g_i is a pair $g_i^* = (m_i, l_i)$. Here, each $m_i = s_{i,1}, \dots, s_{i,n_i}$ 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: $l_i \supseteq \{r_{i,1}, \dots, r_{i,n_i}\}$;
- for *vertical* schemes, the *multiset* of g_i 's nonsensitive values: $l_i = \{r_{i,1}, \dots, r_{i,n_i}\}$.

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 $4550^* = \{45500, 45501, \dots, 45509\}$ 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, $(s_1, r_1), (s_2, r_2), \dots, (s_1, s_N)$. 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

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:

$$\Pi \longrightarrow T \longrightarrow T^*. \quad (1)$$

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:

$$f(\pi, t, t^*) = f(\pi)f(t|\pi)f(t^*|t). \quad (2)$$

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 $\alpha = (\alpha_1, \dots, \alpha_{|\mathcal{S}|})$ and $\beta^s = (\beta_1^s, \dots, \beta_{|\mathcal{R}|}^s)$, respectively. In other words

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

The hyperparameters α and β may incorporate prior (background) knowledge on the population, if this is available. Otherwise, a uninformative 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

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

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 $\text{obf}(t)$. In the important special case $\text{obf}(t)$ acts as a deterministic function on tables, $f(t^*|t) = 1$ if and only if $\text{obf}(t) = t^*$, 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

473 to the density $f(\Pi = \pi | T^* = t^*)$. Formally, for each $(s, r) \in$
 474 $\mathcal{S} \times \mathcal{R}$:

$$475 \quad p_L(s, r | t^*) \triangleq E_{\pi \sim f(\pi | t^*)}[f(s, r | \pi)] = \int_{\mathcal{D}} f(s, r | \pi) f(\pi | t^*) d\pi. \quad (5)$$

477 Of course, we can condition p_L on any given r and obtain
 478 the conditional probability $p_L(s | r, t^*)$. Equivalently, we can
 479 compute

$$480 \quad p_L(s | r, t^*) \triangleq E_{\pi \sim f(\pi | t^*)}[f(s | r, \pi)] = \int_{\mathcal{D}} f(s | r, \pi) f(\pi | t^*) d\pi. \quad (6)$$

482 In particular, one can read off this distribution on a victim's
 483 nonsensitive attribute, say r_v , and obtain the corresponding
 484 distribution on \mathcal{S} .

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

493 Let us introduce in the above model a new random vari-
 494 able V , identifying the index of the victim within the clear-
 495 text table T . We posit that V is uniformly distributed on
 496 $\{1, \dots, N\}$, and independent from Π, T, T^* . Recalling that
 497 each row (S_j, R_j) is identified within T by a unique index
 498 j , we can define the attacker's probability distribution on \mathcal{S} ,
 499 after seeing t^* and r_v , as follows, where it is assumed that
 500 $f(R_V = r_v, t^*) > 0$, that is the observed victim's r_v is
 501 compatible with t^* :

$$502 \quad p_A(s | r_v, t^*) \triangleq f(S_V = s | R_V = r_v, t^*). \quad (7)$$

503 The following crucial lemma provides us with a characteri-
 504 zation of the above probability distribution that is only based
 505 on a selection of the marginals R_j given t^* . This will be the
 506 basis for actually computing $p_A(s | r_v, t^*)$. Note that, on the
 507 right-hand side, only those rows whose sensitive value - known
 508 from t^* - is s contribute to the summation. A proof of the
 509 lemma is reported in Appendix A.

510 *Lemma 1:* Let $T = (S_j, R_j)_{j \in 1 \dots N}$. Let s_j be the sensitive
 511 value in the j -th entry of t^* . Let r_v and t^* such that $f(R_V =$
 512 $r_v, t^*) > 0$. Then

$$513 \quad p_A(s | r_v, t^*) \propto \sum_{j: s_j = s} f(R_j = r_v | t^*). \quad (8)$$

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

518 *Example 1:* In order to illustrate the difference between
 519 the learner's and the attacker's inference, we reconsider the
 520 toy example in the Introduction. Let t^* be the 2-anonymous,
 521 2-diverse Table I(b). Assume the attacker's victim is
 522 the first individual of the original dataset, who is from
 523 Malaysia (= M) and lives in the ZIP code 45501 area, hence

TABLE III

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

	Heart	Flu	Stomach	HIV	Diabetes
$p_L(s r_v, t^*)$	0.343	0.317	0.113	0.114	0.113
$p_A(s r_v, t^*)$	0.580	0.420	0	0	0
$p_{RW}(s r_v, t^*)$	0.500	0.500	0	0	0

524 $r_v = (M, 45501)$. Table III shows the belief probabilities of
 525 the learner, $p_L(s | r_v, t^*)$, and of the attacker, $p_A(s | r_v, t^*)$, for
 526 the victim's disease s . We also include the random worlds
 527 model probabilities, $p_{RW}(s | r_v, t^*)$, which are just proportional
 528 to the frequency of each sensitive value within the victim's
 529 group. Note that the learner and the attacker distributions have
 530 the same mode, but the attacker is more confident about his
 531 prediction of the victim's disease. The random worlds model
 532 produces a multi-modal solution.

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

$$543 \quad p_L(s | r_v, t^*) \propto \int_{\mathcal{D}} f(\pi) f(s, r_v | \pi) \sum_{t: \mathcal{A}(t) = t^*} f(t | \pi) d\pi \quad (9)$$

$$544 \quad p_A(s | r_v, t^*) \propto \int_{\mathcal{D}} f(\pi) f(s_j, r_v | \pi) \sum_{t_{-j}: \mathcal{A}(t) = t^*} f(t | \pi) d\pi. \quad (10)$$

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

553 An alternative, more intuitive description of the inference
 554 process is as follows. The learner and the attacker first learn
 555 the parameters π given t^* , that is they evaluate $f(\pi_{Dis} | t^*)$,
 556 $f(\pi_{ZIP|S} | t^*)$ and $f(\pi_{Nat|S} | t^*)$, for all $s \in \mathcal{S}$. Due to the
 557 uncertainty on the ZIP code and/or Nationality, learning π
 558 takes the form of a mixture (this is akin to learning with
 559 soft evidence, see Corradi *et al.* [11]). After that, the learner,
 560 ignoring the victim is in the table, predicts the probability of
 561 r_v , $p_L(r_v | s, t^*)$, for all s , by using a mixture of Multinomial-
 562 Dirichlet. The attacker, on the other hand, while still basing
 563 his prediction $p_A(r_v | s, t^*)$ on the parameter learning outlined
 564 above, restricts his attention to the first two lines of t^* , thus
 565 realizing that $s \in \{\text{Heart, Flu}\}$. Then, by Bayes theorem,
 566 and adopting the relative frequencies of the diseases in t^* as
 567 an approximation of $f(s | t^*)$, the posterior probability of the
 568 diseases for the victim can be computed.

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

582 *When we cast this scenario in our framework, the individ-*
 583 *uals in the table play the role of the inhabitants (including*
 584 *the culprit), while r_v plays the role of the characteristic found*
 585 *on the crime scene, matching that of the suspect. In other*
 586 *words - perhaps ironically - our framework's victim plays here*
 587 *the role of the suspect S , while our attacker is essentially*
 588 *the investigator. Letting $\mathcal{S} = \{0, 1\}$ (innocent/guilty) and*
 589 $\mathcal{R} = \{0, 1\}$ (characteristic absent/present), the investigator's
 590 information is then summarized by an obfuscated horizontal
 591 table t^* of N rows with as many groups, where exactly one
 592 row, say the j -th, has $S_j = 1$ and $R_j^* = R_j = 1$ (the culprit),
 593 while for $i \neq j$, $S_i = 0$ and $R_i^* = *$ ($N - 1$ innocent
 594 inhabitants). Recalling that the variable V in our framework
 595 represents the suspect's index within the table, the probability
 596 that $I = C$ is

$$\Pr(V = j | R_V = 1, t^*) = \Pr(S_V = 1 | R_V = 1, t^*) \\ = p_A(s = 1 | r_v = 1, t^*).$$

597 Then applying (8), we find

$$p_A(s = 1 | r_v = 1, t^*) = \frac{f(R_j = 1 | t^*)}{f(R_j = 1 | t^*) + (N-1)f(R_{i \neq j} = 1 | t^*)} \\ = \frac{1}{1 + (N-1)f(R_{i \neq j} = 1 | t^*)}. \quad (11)$$

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

$$p_A(s = 1 | r_v = 1, t^*) = \frac{1}{1 + (N-1)p} \quad (12)$$

608 *which is the classical result known from the literature.*

609 *In DBS, the indicted exhibiting r_v is found after testing $1 \leq$*
 610 *$k < N$ individuals that do not exhibit r_v . This means the table*
 611 *t^* consists now of k rows $(s, r) = (0, 0)$ (the k innocent,*
 612 *tested inhabitants not exhibiting r_v), one row $(s, r) = (1, 1)$*
 613 *(the culprit) and $N - 1 - k$ rows $(s, r^*) = (0, *)$ (the $N - 1 - k$*
 614 *innocent, non-tested inhabitants). Accordingly, (11) becomes*
 615 *(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

$$p_A(s = 1 | r_v = 1, t^*) = \frac{1}{1 + (N - 1 - k)p}$$

619 *which again is the classical result known from the literature.*
 620 *Finally note that our methodology also covers the possibility*
 621 *to learn about the probability of the characteristic, $f(R_i =$*
 622 *$1 | t^*)$, but here we have only stressed how the attacker strategy*
 623 *solves the IP and DBS forensic problems. Uncertainty about*
 624 *population parameters and identification has been considered*
 625 *elsewhere by one of us [6].*

626 We now briefly discuss an extension of our framework to
 627 the more general case where the attacker has only partial
 628 information about his victim's nonsensitive attributes. For a
 629 typical application, think of a dataset where \mathcal{R} and \mathcal{S} are
 630 individuals' genetic profiles and diseases, respectively, with an
 631 adversary knowing only a partial DNA profile of his victim;
 632 e.g. only the alleles at a few loci. Formally, fix a nonempty
 633 set \mathcal{Y} and let $g : \mathcal{R} \rightarrow \mathcal{Y}$ be a (typically non-injective)
 634 function, modeling the attacker's observation of the victim's
 635 nonsensitive attribute. With the above introduced notation,
 636 consider the random variable $Y \triangleq g(R_V)$. It is natural to
 637 extend definition (7) as follows, where $g(r_v) = y_v \in \mathcal{Y}$ and
 638 $f(Y = y_v, t^*) > 0$:

$$p_A(s | y_v, t^*) \triangleq f(S_V = s | Y = y_v, t^*). \quad (14)$$

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

$$p_A(s | r_v, t^*) \propto \sum_{j : S_j = s} f(R_j \in g^{-1}(y_v) | t^*). \quad (15)$$

643 Also note that one has $f(R_j \in g^{-1}(y_v) | t^*) =$
 644 $\sum_{r \in g^{-1}(y_v)} f(R_j = r | t^*)$. An extension to the case of partial
 645 and noisy observations can be modeled similarly, by letting
 646 $Y = g(R_V, E)$, where E is a random variable representing
 647 an independent source of noise. We leave the details of this
 648 extension for future work. 649

IV. MEASURES OF PRIVACY THREAT AND UTILITY 650

651 We are now set to define the measures of *privacy threat* and
 652 *utility* we are after. We will do so from the point of view of
 653 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; 654
- (b) must decide whether to release t^* or not, weighing the privacy threats and the utility implied by this act. 655

656 The evaluator clearly distinguishes the position of the *learner*
 657 from that of the *attacker*. The learner is interested in learning
 658 from t^* the characteristics of the general population, via p_L .
 659 The attacker is interested in learning from t^* the sensitive
 660
 661

$$p_A(s = 1 | r_v = 1, t^*) = \frac{f(R_{k+1} = 1 | t^*)}{f(R_{k+1} = 1 | t^*) + kf(R_{i \in \{1, k\}} = 1 | t^*) + (N - 1 - k)f(R_{i > k+1} = 1 | t^*)} \quad (13)$$

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

674 a situation where, for some individual, p_A conveys
 675 much more information than that conveyed by p_L
 676 (learner's legitimate inference on general popula-
 677 tion), must be deemed as a privacy threat.

678 Generally speaking, the evaluator should refrain from pub-
 679 lishing t^* if, for some individual, the *level* of relative pri-
 680 vacy threat exceeds a predefined threshold. Concerning the
 681 definition of the level of threat, the evaluator adopts the
 682 following Bayesian decision-theoretic point of view. Whatever
 683 distribution p is adopted to guess the victim's sensitive value,
 684 the attacker is faced with some utility function. Here, we con-
 685 sider a simple 0-1 utility function for the attacker, yielding 1 if
 686 the sensitive attribute is guessed correctly and 0 otherwise.
 687 The resulting attacker's expected utility is maximized by
 688 the Bayes act, i.e. by choosing $s = \operatorname{argmax}_{s' \in \mathcal{S}} p(s')$, and
 689 equals $p(s)$. The above discussion leads to the following
 690 definitions. Note that we consider threat measures both for
 691 individual rows and for the overall table. For each threatened
 692 row, the relative threat index **Ti** says how many times the
 693 probability of correctly guessing the secret is increased by
 694 the attacker's activity i.e. by exploiting the knowledge of
 695 the victim's presence in the table. At a global, table-wise
 696 level, the evaluator also considers the fraction **GT_A** of rows
 697 threatened by the attacker.

698 *Definition 1 (privacy threat):* We define the following pri-
 699 vacy threat measures.

- 700 • Let q be a full support distribution on \mathcal{S} and (s, r) be a
 701 row in t . We say (s, r) is *threatened under q* if $q(s) =$
 702 $\max_{s'} q(s')$, and that its *threat level under q* is $q(s)$.
- 703 • For a row (s, r) in t that is threatened by $p_A(\cdot|r, t^*)$, its
 704 *relative threat level* is

$$705 \quad \mathbf{Ti}(s, r, t, t^*) \triangleq \frac{p_A(s|r, t^*)}{p_L(s|r, t^*)}. \quad (16)$$

- 706 • Let $N_A(t, t^*)$ be the number of rows (s, r) in t threatened
 707 by $p_A(\cdot|r, t^*)$. The *global threat level* $\mathbf{GT}_A(t, t^*)$ is the
 708 fraction of rows that are threatened, that is

$$709 \quad \mathbf{GT}_A(t, t^*) \triangleq \frac{N_A(t, t^*)}{N}. \quad (17)$$

710 Similarly, we denote by $\mathbf{GT}_L(t, t^*)$ the fraction of rows
 711 (s, r) in t that are threatened under $p_L(\cdot|r, t^*)$.

- 712 • As a measure of how better the attacker performs than
 713 learner at a global level, we introduce *relative global*
 714 *threat*:

$$715 \quad \mathbf{RGT}_A(t, t^*) \triangleq \max\{0, \mathbf{GT}_A(t, t^*) - \mathbf{GT}_L(t, t^*)\}. \quad (18)$$

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

730 The evaluator is also interested in the potential utility
 731 conveyed by an anonymized table for a learner. Note that the
 732 learner's utility is distinct from the attacker's one. Indeed, the
 733 learner's interest is to make inferences that are as close as
 734 possible to the ones that could be done using the cleartext
 735 table. Accordingly, obfuscated tables that are *faithful* to the
 736 original table are the most useful. This leads us to compare two
 737 distributions on the population: the distribution learned from
 738 the anonymized table, p_L , and the *ideal* (**I**) distribution, p_I ,
 739 one can learn from the cleartext table t . The latter is formally
 740 defined as the expectation² of $f(s, r|\pi)$ under the posterior
 741 density $f(\pi|t)$. Explicitly, for each (s, r)

$$742 \quad p_I(s, r|t) \triangleq \int_{\mathcal{D}} f(s, r|\pi) f(\pi|t) d\pi. \quad (19)$$

743 Note that the posterior density $f(\pi|t)$ is in turn a Dirichlet
 744 density (see next section) and therefore a simple closed form
 745 of the above expression exists, based on the frequencies of
 746 the pairs (s, r) in t . In particular, recalling the α_s, β_r^s
 747 notation for the prior hyperparameters introduced in Section III, let
 748 $\alpha_0 = \sum_s \alpha_s$ and $\beta_0^s = \sum_r \beta_r^s$, and $\gamma_s(t)$ and $\delta_r^s(t)$ denote the
 749 frequency counts of s and (s, r) , respectively, in t . Then we
 750 have

$$751 \quad p_I(s, r|t) = \frac{\alpha_s + \gamma_s(t)}{\alpha_0 + N} \cdot \frac{\beta_r^s + \delta_r^s(t)}{\beta_0^s + \gamma_s(t)}. \quad (20)$$

752 The comparison between p_L and p_I can be based on some
 753 form of *distance* between distributions. One possibility is to
 754 rely on *total variation* (aka statistical) distance. Recall that,
 755 for discrete distributions q, q' defined on the same space \mathcal{X} ,
 756 the total variation distance is defined as

$$757 \quad \mathbf{TV}(q, q') \triangleq \sup_{A \subseteq \mathcal{X}} |q(A) - q'(A)| = \frac{1}{2} \sum_x |q(x) - q'(x)|.$$

758 Note that $\mathbf{TV}(q, q') \in [0, 1]$. Note that this is a quite
 759 conservative notion of diversity since it based on the event
 760 that shows the largest difference between distributions.

761 *Definition 2 (faithfulness):* The *relative faithfulness level* of
 762 t^* w.r.t. t is defined as

$$763 \quad \mathbf{RF}(t, t^*) \triangleq 1 - \mathbf{TV}(p_I(\cdot|t), p_L(\cdot|t^*)).$$

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

764 *Remark 4: In practice, the total variation of two high-*
 765 *dimensional distributions might be very hard to compute.*
 766 *Pragmatically, we note that for M large enough, $\mathbf{TV}(q, q') =$
 767 $\frac{1}{2} E_{x \sim q(x)} [|1 - \frac{q'(x)}{q(x)}|] \approx \frac{1}{2M} \sum_{i=1}^M |1 - \frac{q'(x_i)}{q(x_i)}|$, where the x_i are
 768 drawn i.i.d. according to $q(x)$. Then a proxy to total variation
 769 is the empirical total variation defined below, where (s_i, t_i) ,
 770 for $i = 1, \dots, M$, are generated i.i.d. according to $p_{\mathbf{I}}(\cdot, \cdot | t)$:*

$$771 \quad \mathbf{ETV}(t, t^*) \triangleq \frac{1}{2M} \sum_{i=1}^M \left| 1 - \frac{p_{\mathbf{L}}(s_i, r_i | t^*)}{p_{\mathbf{I}}(s_i, r_i | t)} \right|. \quad (21)$$

772
 773 *Remark 5 (ideal knowledge vs. attacker's knowledge):*
 774 *The following scenario is meant to further clarify the extra*
 775 *power afforded to the attacker, by the mere knowledge that*
 776 *his victim is in the table. Consider a trivial anonymization*
 777 *mechanism that simply releases the cleartext table, that is*
 778 $t^* = t$. As $p_{\mathbf{L}} = p_{\mathbf{I}}$ in this case, it would be tempting
 779 to conclude that the attacker cannot do better than the
 780 learner, hence there is no relative risk involved. However,
 781 this conclusion is wrong: for instance, $p_{\mathbf{I}}(\cdot | r_v, t)$ can fail to
 782 predict the victim's correct sensitive value if this value is
 783 rare, as we show below.

784 For the sake of simplicity, consider the case where the
 785 observed victim's nonsensitive attribute r_v occurs just once in t
 786 in a row (s_0, r_v) . Also assume a noninformative Dirichlet prior,
 787 that is, in the notation of Section III, set the hyperparameters
 788 to $\alpha_s = \beta_r^s = 1$ for each $s \in \mathcal{S}, r \in \mathcal{R}$. Then, simple
 789 calculations based on (20) and the attacker's distribution
 790 characterization (8), show the following. Here for each $s \in \mathcal{S}$,
 791 $\gamma_s = \gamma_s(t)$ denotes the frequency count of s in t , and c a
 792 suitable normalizing constant:

$$793 \quad p_{\mathbf{I}}(s | r_v, t) = \begin{cases} \frac{1 + \gamma_s}{|\mathcal{R}| + \gamma_s} c, & \text{if } s \neq s_0 \\ \frac{2(1 + \gamma_{s_0})}{|\mathcal{R}| + \gamma_{s_0}} c, & \text{if } s = s_0 \end{cases}$$

$$794 \quad p_{\mathbf{A}}(s | r_v, t^*) = \begin{cases} 0, & \text{if } s \neq s_0 \\ 1, & \text{if } s = s_0. \end{cases} \quad (22)$$

795 As far as the target individual $(s_0, r_v) \in t$ is concerned, we
 796 see that while $p_{\mathbf{A}}$ predicts s_0 with certainty, predictions based
 797 on $p_{\mathbf{L}} = p_{\mathbf{I}}$ will be blatantly wrong, if there are values $s \neq s_0$
 798 that occur very frequently in t , while s_0 is rare, and N is large
 799 compared to $|\mathcal{R}|$. To make an extreme numeric case, consider
 800 $|\mathcal{S}| = 2$, $|\mathcal{R}| = 1000$ and $\gamma_{s_0} = 1$ in a table t of $N =$
 801 10^6 rows: plugging these values in (22) yields $p_{\mathbf{L}}(s_0 | r_v, t^*) =$
 802 $p_{\mathbf{I}}(s_0 | r_v, t) \approx 0.004$, hence a relative threat for (s_0, r_v) of
 803 $1/p_{\mathbf{L}}(s_0 | r_v, t^*) \approx 250$.

804 V. LEARNING FROM THE OBFUSCATED TABLE BY MCMC

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

To circumvent this difficulty, we will introduce a *Gibbs sam-*
 813 *pler*, defining a Markov chain $(X_i)_{i \geq 0}$, with $X_i = (\Pi_i, T_i)$,
 814 converging to the density
 815

$$816 \quad f(\Pi = \pi, T = t | t^*)$$

$$817 \quad = f(\Pi = \pi, S_1 = s_1, R_1 = r_1, \dots, S_N = s_N, R_N = r_N | t^*)$$

(note that the sensitive values s_j in T are in fact fixed and
 818 known, given t^*). General results (see e.g. [41]) ensure that,
 819 if Π_0, Π_1, \dots are the samples drawn from the Π -marginal of
 820 such a chain, then for each $(s, r) \in \mathcal{S} \times \mathcal{R}$
 821

$$822 \quad \frac{1}{M} \sum_{\ell=0}^M f(s, r | \Pi_{\ell}) \rightarrow \int_{\mathcal{D}} f(s, r | \pi) f(\pi | t^*) d\pi = p_{\mathbf{L}}(s, r | t^*)$$

$$823 \quad (23)$$

$$824 \quad \frac{1}{M} \sum_{\ell=0}^M f(s | r, \Pi_{\ell}) \rightarrow \int_{\mathcal{D}} f(s | r, \pi) f(\pi | t^*) d\pi = p_{\mathbf{L}}(s | r, t^*)$$

$$825 \quad (24)$$

826 almost surely as $M \rightarrow +\infty$. Therefore, by selecting
 827 an appropriately large M , one can build approximations of
 828 $p_{\mathbf{L}}(s, r | t^*)$ and $p_{\mathbf{L}}(s | r, t^*)$ using the arithmetical means on
 829 the left-hand side of (23) and (24), respectively. Moreover,
 830 for each index $1 \leq j \leq N$, using samples drawn from the
 831 R_j -marginals of the same chain, one can build an estimate of
 832 $f(R_j = r_j | t^*)$. Consequently, using (8) (resp. (15), in the case
 833 of partial observation) one can estimate $p_{\mathbf{A}}(s | r_v, t^*)$ (resp.
 834 $p_{\mathbf{A}}(s | y_v, t^*)$) for any given r_v (resp. y_v).

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

839 A. Definition and Convergence of the Gibbs Sampler

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

$$844 \quad f(\pi, t | t^*). \quad (25)$$

845 Note that the r_j 's of interest, for $1 \leq j \leq N$, are the elements
 846 of the groups g_i 's, for $1 \leq i \leq k$. The Gibbs scheme allows
 847 for some freedom as to the blocking of variables. Here we
 848 consider $k + 1$ blocks, coinciding with π and g_1, \dots, g_k .
 849 This is natural as, in the considered schemes, $(R_i, S_i) \perp\!\!\!\perp$
 850 $(R_j, S_j) | \pi, t^*$ for (R_i, S_i) and (R_j, S_j) occurring in distinct
 851 groups. Formally, let $x^0 = \pi^0, t^0$ (with $t^0 = g_1^0, \dots, g_k^0$)
 852 denote any initial state satisfying $f(\pi^0, t^0 | t^*) > 0$. Given
 853 a state at step h , $x^h = \pi^h, t^h$ ($t^h = g_1^h, \dots, g_k^h$), one lets
 854 $x^{h+1} \triangleq \pi^{h+1}, t^{h+1}$, where $t^{h+1} = g_1^{h+1}, \dots, g_k^{h+1}$ and

$$855 \quad \pi^{h+1} \text{ is drawn from } f(\pi | t^h, t^*) \quad (26)$$

$$856 \quad g_i^{h+1} \text{ is drawn from}$$

$$857 \quad f(g | \pi^{h+1}, g_1^{h+1}, \dots, g_{i-1}^{h+1}, g_{i+1}^h, \dots, g_k^h, t^*)$$

$$858 \quad (1 \leq i \leq k). \quad (27)$$

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 $t_{-i} \triangleq g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_k$ 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)

$$f(\pi|t, t^*) = f(\pi|t) \quad (28)$$

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

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 $1 \leq i \leq k$, 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 ($1 \leq i \leq k$):

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

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

Let $x = (\pi, g_1, \dots, g_k)$ 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

$$\mathcal{X} \triangleq \mathcal{D} \times \mathcal{G}_1 \times \dots \times \mathcal{G}_k. \quad (31)$$

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, \dots 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 $(X_i)_{i \geq 0}$ 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 $\boldsymbol{\gamma}(t) = (\gamma_1, \dots, \gamma_{|\mathcal{S}|})$ the vector of the frequency counts γ_i of each s_i in t . Similarly, given s , denote by $\boldsymbol{\delta}^s(t) = (\delta_{r_1}^s, \dots, \delta_{r_{|\mathcal{R}|}}^s)$ the vector of the frequency counts δ_i^s of the pairs (r_i, s) , for each r_i , in t . Then, for each $\pi = (\pi_S, \pi_{R|S})$, we have

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

Let us now discuss (29). In what follows, for the sake of notation we shall write a generic i -th group as $g_i = (s_1, r_1), \dots, (s_n, r_n)$ (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 $1 \leq i \leq k$, letting $g_i^* = (l_i, m_i)$, with $m_i = s_1, \dots, s_n$, the i -th obfuscated group in t^* , the following holds true:

Horizontal schemes

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

Vertical schemes

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

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 $l_i \supset \{r_1, \dots, r_n\}$. 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 $\{50130, \dots, 50139\}$. 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.

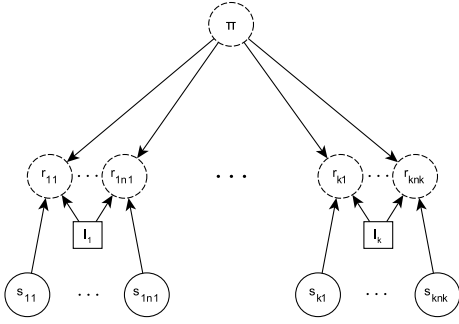


Fig. 1. Sampling from $f(g|\pi, t_{-i}, t^*)$ ($g \in \mathcal{G}_i$) for horizontal schemes, across all the groups.

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

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

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

973 a) *Horizontal schemes*: In order to generate $g =$
 974 $(r_1, s_1), \dots, (r_n, s_n) \in \mathcal{G}_i$, for each $\ell = 1, \dots, n$, we draw
 975 $r_\ell \in l_i$ with probability $\propto f(r_\ell|s_\ell, \pi)$. Explicitly, (29) now
 976 becomes

$$977 f(g|\pi, t_{-i}, t^*) = \begin{cases} 0, & \text{if } g \notin \mathcal{G}_i \\ \prod_{\ell=1}^n \frac{f(r_\ell|s_\ell, \pi)}{\sum_{r \in l_i} f(r|s_\ell, \pi)}, & \text{if } g \in \mathcal{G}_i \end{cases} \quad (36)$$

978 thus satisfying (35). Note that this is equivalent to sam-
 979 pling each row independently. The sampling process of
 980 $f(g|\pi, t_{-i}, t^*)$ for horizontal schemes across all the groups
 981 of the table is illustrated graphically in Fig. 1.

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

988 We discuss here a more straightforward sampling procedure,
 989 based on generating $g_i \in \mathcal{G}_i$ in a single shot. We adopt a
 990 *single-iteration Metropolis within Gibbs* scheme. Essentially,
 991 this consists in running a Metropolis method that targets the
 992 distribution $\propto f(g|\pi)$ with support \mathcal{G}_i , for one iteration.
 993 Specifically, let us write the current value of the i -th group in
 994 the Gibbs Markov chain as g_i^h . Following Casella and Robert
 995 [40, Ch.10], this step consists in drawing $g \in \mathcal{G}_i$ according to
 996 a proposal distribution $J(g|g_i^h)$ and accepting it, that is letting
 997 $g_i^{h+1} = g$, with probability

$$998 \epsilon \triangleq \min \left\{ 1, \frac{f(g|\pi)J(g_i^h|g)}{f(g_i^h|\pi)J(g|g_i^h)} \right\} \quad (37)$$

999 while keeping $g_i^{h+1} = g_i^h$ with probability $1 - \epsilon$. The
 1000 resulting MCMC method is still theoretically sound: see Casella

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

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

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, \dots, r_n is the
sequence of sensitive values in the proposed g :

$$1008 \epsilon = \min \left\{ 1, \frac{\prod_{\ell=1}^n f(r_\ell|s_\ell, \pi)}{\prod_{\ell=1}^n f(r_\ell^h|s_\ell, \pi)} \right\}.$$

1009 VI. EXPERIMENTS

1010 We have put a proof-of-concept implementation³ of our
 1011 methodology at work on a subset of the Adult dataset extracted
 1012 by Barry Becker from the 1994 US Census database and
 1013 available from the UCI machine learning repository [47]. This
 1014 is a common benchmark for experiments on anonymization
 1015 [38]. In particular, we have focused on the subset of 5692 rows
 1016 also considered by the authors of [38], with the following
 1017 categorical attributes: *sex, age, race, marital status, education,*
 1018 *native country, workclass, salary class, occupation,* with *occu-*
 1019 *vation* (14 values) considered as the only sensitive attribute.
 1020 We will discuss implementation and results details separately
 1021 for vertical and horizontal schemes. We will then briefly
 1022 discuss convergence issues of the employed MCMC method.

1023 A. Horizontal Schemes: k -Anonymity

1024 Using the ARX anonymization tool [37] we obtained two
 1025 different k -anonymous versions of the considered dataset,
 1026 enjoying respectively k -anonymity and ℓ -diversity⁴ for $k =$
 1027 $\ell = 4$ and $k = \ell = 6$. The average size of the groups
 1028 was respectively of 38 rows ($k = \ell = 4$) and of 355 rows
 1029 ($k = \ell = 6$).

1030 The results we have obtained are summarized in Table IV.
 1031 For reference, we include the following information in the last
 1032 two lines: *baseline accuracy*, the fraction of rows correctly
 1033 classified using the empirical distribution obtained from the
 1034 frequencies of the sensitive values in the anonymized table
 1035 – i.e., the fraction of the most frequent sensitive value; and

³Python code and data available from the authors.

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

1036 *ideal accuracy*, the fraction of tuples threatened under p_L .
 1037 As a further element of comparison, we also consider an
 1038 attacker whose reasoning is based on the random worlds
 1039 models, and include in the table \mathbf{GT}_{RW} , the fraction of rows
 1040 correctly classified assuming all tables compatible with t^*
 1041 equally likely. Like in [25], we compute \mathbf{ABS}_A and \mathbf{ABS}_{RW} ,
 1042 the *absolute error* under the distribution derived under p_A and
 1043 under the random worlds distribution p_{RW} , respectively. \mathbf{ABS}
 1044 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
 1045 be either of $p_A(\cdot)$ or $p_{RW}(\cdot)$. Note that, since the considered
 1046 anonymized tables do not enjoy disjointness between groups
 1047 (see Remark 1), also in the random worlds perspective the
 1048 probability of each sensitive attribute may well be $\geq 1/\ell$.
 1049 In our experiments, when $\ell = 4$ the attacker outperforms
 1050 random worlds classification, while when a more powerful
 1051 obfuscation is adopted the two results are quite similar.

1052 The remaining rows in Table IV consider the privacy threats
 1053 and faithfulness measures introduced in Section IV. As a
 1054 general comment, small variations of ℓ and/or k do not produce
 1055 dramatic changes. The faithfulness level is stable, but does not
 1056 reach a satisfactory level. The attacker is anyway in a position
 1057 to correctly classify the sensitive attribute of individuals in the
 1058 table $\approx 2.3 - 2.5\%$ more often than the learner. We found the
 1059 maximum value of \mathbf{Ti}_A for the threatened rows is about 13.8,
 1060 meaning the attacker can be up to ≈ 14 times more confident
 1061 than the learner about the guessed value.

1062 A more informative summary of our analysis is provided by
 1063 the scatter plots and histograms of Figure 2. The scatter plots
 1064 are obtained from the threat levels under p_L and under p_A .
 1065 The number of rows (s, r) in which $p_A(s|r, t^*) \geq p_L(s|r, t^*)$
 1066 roughly equals those in which $p_A(s|r, t^*) \leq p_L(s|r, t^*)$,
 1067 although globally the attacker has a slight advantage in terms
 1068 of number of threatened rows. In Figure 2 we also report the
 1069 empirical distribution $\log_2 \mathbf{Ti}_A$ for tuples threatened under p_A
 1070 and under p_L . We also have evidence of positive skewness,
 1071 as shown by the value of γ (the third standardized moments
 1072 of the empirical distributions). Recalling that $\log_2 \mathbf{Ti}_A = 1$
 1073 means $p_A(s|r, t^*) = 2p_L(s|r, t^*)$, the histograms show that
 1074 $p_A(s|r, t^*)$ is often more than twice $p_L(s|r, t^*)$ leading to a
 1075 $\log_2 \mathbf{Ti}_A \geq 1$. In particular, when $k = \ell = 4$, $\log_2 \mathbf{Ti}_A$
 1076 is at least 1 for $\approx 6\%$ of the individuals threatened under p_A ,
 1077 meaning $\approx 0.6\%$ of the whole table. Conversely, $\log_2 \mathbf{Ti}_A$
 1078 is close to 0 for most of the rows in which $p_A(s|r, t^*) \leq$
 1079 $p_L(s|r, t^*)$.

1080 B. Vertical Schemes: Anatomy

1081 Using a freely available anonymization tool [22], we have
 1082 obtained two anatomized versions of the considered dataset,
 1083 with groups of size $\ell = 4$ and $\ell = 6$, respectively. The
 1084 resulting tables also enjoy ℓ -diversity. The results we have
 1085 obtained are summarized in Table V. Concerning the random
 1086 worlds approach, we note the following. Anatomy partitions
 1087 the tables in groups all of size ℓ . Therefore, although disjoint-
 1088 ness is not satisfied, just as in the horizontal case, the sensitive
 1089 attribute frequencies equal $1/\ell$ in each group. This implies
 1090 that the probability of a sensitive value depends on how many
 1091 groups contain the victim's nonsensitive attributes and on

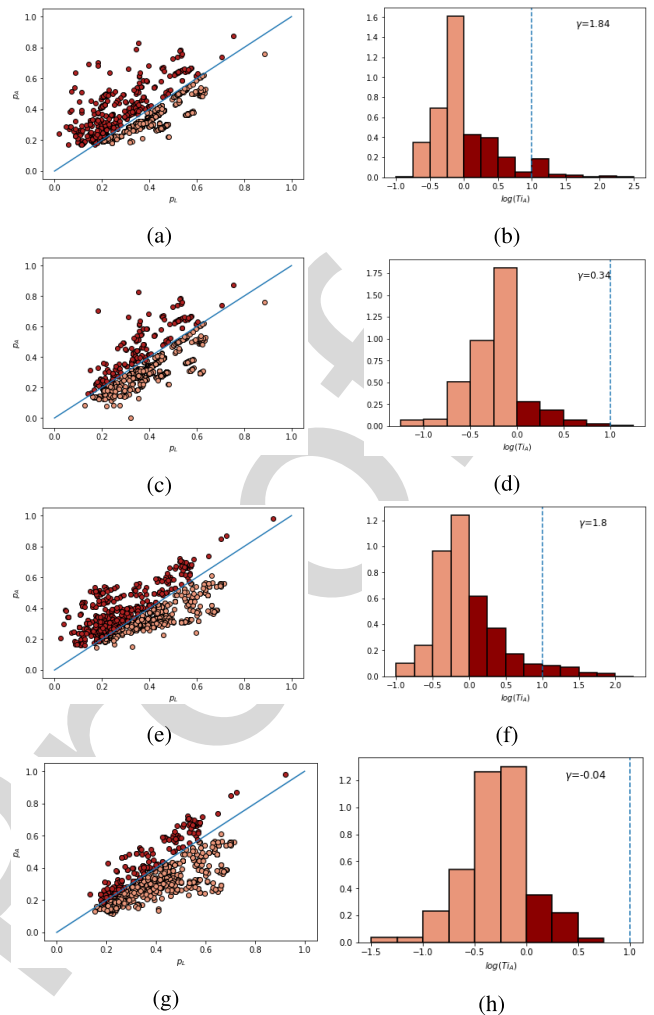


Fig. 2. Results for k -anonymity. Top ($\ell = k = 6$): scatter plots of p_L vs p_A for tuples threatened under p_A (a), and under p_L (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_A$ for these two cases. Bottom: same for $\ell = k = 4$. The skewness value (γ) represents the third standardized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

1092 their frequencies in each group, leading often to multimodal
 1093 distributions. We assume that a guess may be obtained ran-
 1094 domly choosing between the equally likely sensitive attributes.
 1095 Accordingly, the fractions of threatened rows, \mathbf{GT}_{RW} , are
 1096 averaged over 500 different sampling. Here, it is apparent that
 1097 the our attacker is able to classify better than the random
 1098 worlds scenario. We note that, as ℓ increases from 4 to 6,
 1099 the fraction of rows threatened under the distributions derived
 1100 by the learner (\mathbf{GT}_L) and by the attacker (\mathbf{GT}_A) decreases
 1101 significantly. Moreover, as ℓ grows both the relative threat
 1102 \mathbf{RGT}_A and the faithfulness level \mathbf{RF} decrease, which implies
 1103 a trade-off between privacy and the utility conveyed by the
 1104 table.

1105 Again, for a more informative summary of our analysis,
 1106 we look at scatter plots and histograms, displayed in Figure 3,
 1107 where we compare p_A and p_L on threatened rows. It is
 1108 apparent here that the attacker is more confident than the
 1109 learner in the majority of the cases, even when focusing on
 1110 the rows threatened under p_L . This is in contrast with the
 1111 horizontal case, where the attacker exhibits smaller threat

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	

1112 levels on the rows threatened under p_L (Figure 2, (d) and (h)).
 1113 As far as the histograms are concerned, an even greater
 1114 skewness than the horizontal case is evident here. In particular,
 1115 the attacker can be up to ≈ 287 times more confident than
 1116 the learner, being the maximum \mathbf{Ti}_A about 286.19. Moreover,
 1117 when $\ell = 4$, the individuals with $\log_2 \mathbf{Ti}_A \geq 1$ are $\approx 26\%$
 1118 of the rows threatened under p_A ($\approx 8\%$ of the whole table). This
 1119 means that there are 483 individuals in the dataset for which
 1120 the threat level under p_A is at least twice as much the threat
 1121 level under p_L .

1122 C. Discussion

1123 Comparing the horizontal and the vertical cases for the
 1124 considered dataset, the following considerations are in order.

- 1125 • In the horizontal case, we have a situation of low faith-
 1126 fulness and low privacy threat, irrespective of the value
 1127 of k and ℓ . Indeed, in both cases the average group size
 1128 is well above k , and this has a negative effect on the
 1129 inference capabilities of both the learner and the attacker.
 1130 The slight numerical differences observed between the
 1131 cases $k = \ell = 4$ and $k = \ell = 6$ are basically an artifact
 1132 of the anonymization tool. Yet, in relative terms, one can
 1133 observe a significant increase in the number of tuples
 1134 threatened by the attacker, over the learner.
- 1135 • In the vertical case, one obtains a greater faithfulness
 1136 at the price of a greater privacy threat. This difference
 1137 from the horizontal case is partly explained by the smaller
 1138 group size, which now coincides with ℓ . Now moving
 1139 from $\ell = 4$ to $\ell = 6$ has a tangible negative impact
 1140 on the inference capabilities of both the learner and the
 1141 attacker. In relative terms, one can observe an even more
 1142 marked increase of the number of tuples threatened by
 1143 the attacker, over the learner.

1144 The above considerations partly depend on both the original
 1145 dataset and the details of the employed anonymization tool.

1146 D. Assessing MCMC Convergence

1147 For each of the considered anonymized datasets, we ran a
 1148 MCMC as introduced in Section V for $M = 100,000$ runs.
 1149 The convergence of each chain to the stationary distribu-
 1150 tion was assessed via a methodology based on comparing
 1151 sub-sequences of the sample sequences with one another. More
 1152 precisely, as for the population parameters distribution (32),
 1153 we used the method proposed by Geweke [21]. The Geweke

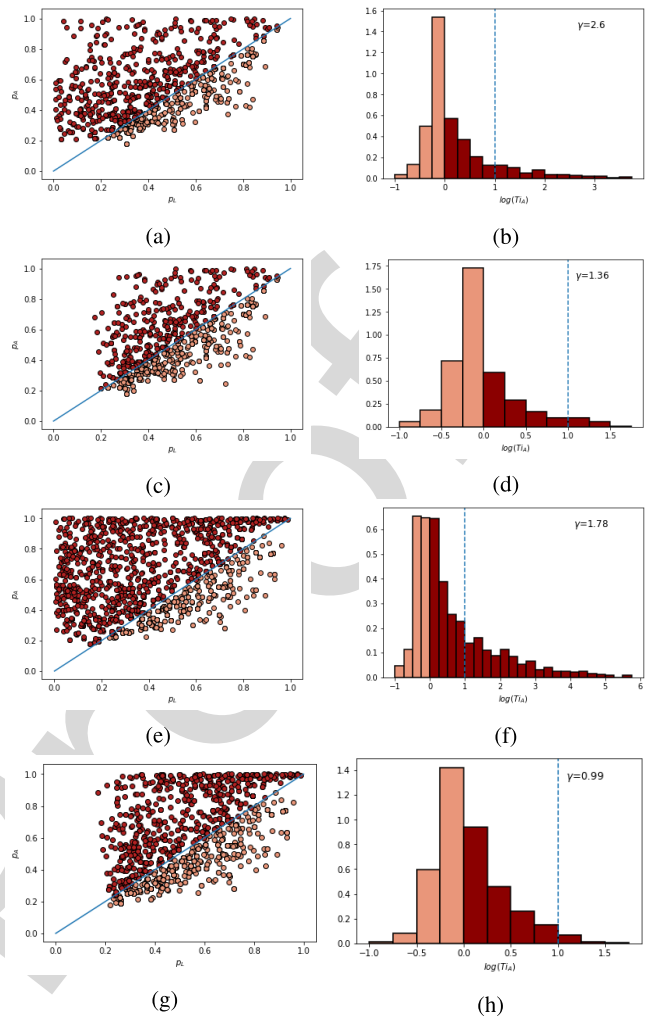


Fig. 3. Results for Anatomy. Top ($\ell = 6$): scatter plots of p_L vs p_A for tuples threatened under p_A (a), and under p_L (c); (b) and (d) are the histograms of $\log_2 \mathbf{Ti}_A$ for these two cases. Bottom: same for $\ell = 4$. The skewness value (γ) represents the third standardized moment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

1154 proposal is based on an adapted two-samples test on the means
 1155 in sub-sequences of the chain.

1156 After a burn-in of 50,000 iterations, we compared the last
 1157 25,000 samples against 5 blocks of 5,000 consecutive sam-
 1158 ples each, taken starting from the 50,000-th iteration. We found
 1159 that all the distributions $\pi_{R|S}$ produced a test statistic within
 1160 two standard deviations from zero, thus providing evidence of
 1161 convergence.

1162 As for the distribution of the cleartext table, $f(t|\pi, t^*)$, we
 1163 used a test specifically designed for categorical distributions
 1164 by Deonovich and Smith, called Weiß procedure [15]. The
 1165 approach is based on a χ^2 test adjusted for the autocorrelation
 1166 induced by the chain. The test is based on partitioning the
 1167 whole sample sequence into sub-sequences, and then testing
 1168 the homogeneity between the empirical distribution of each
 1169 sub-sequence and the empirical distribution of the whole
 1170 chain. After a burn-in of 50,000 observations, we compared
 1171 5 sub-sequences of 10,000 consecutive samples each. For the
 1172 vertical scheme, we assessed the convergence for each row of
 1173 the table, thereby demonstrating the stationary of $f(t|\pi, t^*)$.

1174 For the horizontal scheme, some of the rows did not exhibit
 1175 evidence of convergence. However, we found that, starting
 1176 with several independent chains, very similar results in terms
 1177 of the proposed assessment measures were obtained.

1178 In the vertical case, within the Metropolis step both the pure
 1179 random permutation and the swap group generation strategies
 1180 (Section V-B) were experimented. The obtained results are
 1181 consistent; however, the pure random permutation strategy
 1182 shows a much higher rate of rejection, suggesting that the
 1183 swap strategy should be preferred.

1184 VII. CONCLUSION

1185 We have put forward a notion of relative privacy threat that
 1186 applies to group-based anonymization schemes. Our proposal
 1187 is based on a rigorous characterization of the learner's and
 1188 of the attacker's inference, in a unified Bayesian model of
 1189 group-based schemes. A related MCMC algorithm for posterior
 1190 parameters estimation has also been introduced. Experiments
 1191 conducted on the well-known Adult dataset [47] have been
 1192 illustrated.

1193 Our analysis emphasizes the risks posed by the mere fact
 1194 that an attacker can look up a released anonymized table.
 1195 This prompts an obvious alternative: release the parameters
 1196 of the posterior distribution learned from the cleartext table
 1197 (p_I , in our notation). This may not always be possible, or be
 1198 a good idea, for several reasons. First, certain organizations
 1199 must release datasets as part of their mission, e.g. census
 1200 bureaus. Second, especially in the case of high-dimensional
 1201 data, the computation of the posterior is feasible only assum-
 1202 ing suitable conditional independencies, whereby potentially
 1203 important correlations are lost; see [10] and references therein.
 1204 Third, parameters release itself is not exempt from risks for
 1205 privacy. In particular, although differentially private release of
 1206 the parameters is possible [16], it seems that quite strong
 1207 priors are necessary to obtain acceptable guarantees; see
 1208 [50, Ch.6] and references therein. In conclusion, further
 1209 research is called for in order to understand under what
 1210 circumstances data and/or parameters release can be done
 1211 safely.

1212 APPENDIX A 1213 PROOF OF LEMMA 1

1214 We first characterize the probability $f(V = j | R_V = r_v, t^*)$,
 1215 for an arbitrary $j \in \{1, \dots, N\}$. Bayes theorem yields

$$\begin{aligned}
 1216 f(V = j | R_V = r_v, t^*) &\propto f(R_V = r_v | V = j, t^*) f(V = j | t^*) \\
 1217 &= f(R_j = r_v | V = j, t^*) f(V = j | t^*) \\
 1218 &\propto f(R_j = r_v | V = j, t^*) \quad (38) \\
 1219 &= f(R_j = r_v | t^*) \quad (39)
 \end{aligned}$$

1220 where (38) follows from $f(V = j | t^*) = f(V = j) = 1/N$
 1221 (independence of V), and (39) follows because, as easily
 1222 checked, for any fixed j , independence of R_j and V is
 1223 preserved by conditioning on t^* . Now we have, for every $s \in S$

$$\begin{aligned}
 1224 p_A(s | r_v, t^*) &\quad (40) \\
 1225 &= f(S_V = s | R_V = r_v, t^*) \\
 1226 &= \sum_j f(S_V = s, V = j | R_V = r_v, t^*)
 \end{aligned}$$

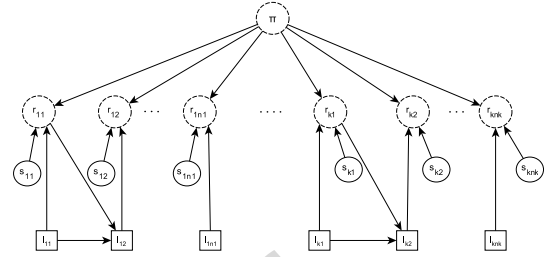


Fig. 4. Sampling from $\theta(g|\pi, t^*)$ for vertical schemes.

$$= \sum_j f(S_V = s | V = j, R_V = r_v, t^*) f(V = j | R_V = r_v, t^*) \quad 1227$$

$$= \sum_j f(S_j = s | V = j, R_j = r_v, t^*) f(V = j | R_V = r_v, t^*) \quad 1228$$

$$= \sum_{j: s_j=s} f(S_j = s | V = j, R_j = r_v, t^*) f(V = j | R_V = r_v, t^*) \quad 1229 \quad (41) \quad 1230$$

$$= \sum_{j: s_j=s} f(V = j | R_V = r_v, t^*) \quad (42) \quad 1231$$

$$\propto \sum_{j: s_j=s} f(R_j = r_v | t^*), \quad (43) \quad 1232$$

1233 where (41) and (42) follow from the fact that, for $s_j \neq s$,
 1234 $f(S_j = s, t^*) = 0$, while for $s_j = s$ obviously $f(S_j = s | V =$
 1235 $j, R_j = r_v, t^*) = 1$. Finally, (43) follows from (39).
 1236

1237 Note that in (43) each term on the RHS actually is the joint
 1238 probability $f(R_j = r_v, S_j = s | t^*)$, being $s_j = s$ embedded in
 1239 the range of the summation.

1239 APPENDIX B

1240 AN ALTERNATIVE GROUP SAMPLING METHOD FOR 1241 VERTICAL SCHEMES

1242 We consider the following method for sampling $g \in \mathcal{G}_i$.
 1243 Draw n values $r_{i\ell}$, $\ell = 1, \dots, n$, as follows:

1. draw r_{i1} from l_i according to a distribution $\propto f(r | s_1, \pi)$; 1244
2. draw r_{i2} from $l_i \setminus \{r_{i1}\}$ according to a distribution \propto 1245
 $f(r | s_2, \pi)$; 1246
- ...
- n. draw r_{in} from $l_i \setminus \{r_{i1}, \dots, r_{i(n-1)}\}$ according to a distrib- 1248
 ution $\propto f(r | s_n, \pi)$. 1249

1250 For a multiset l' , let $\sigma(l' | s_\ell, \pi) \triangleq \sum_{r \text{ in } l'} f(r | s_\ell, \pi)$ denote
 1251 the probability of extracting some element appearing in l'
 1252 (disregarding multiplicities) according to $f(\cdot | s_\ell, \pi)$. Using this
 1253 notation, the probability of returning exactly the sequence
 1254 r_{i1}, \dots, r_{in} , hence $g = (s_1, r_{i1}), \dots, (s_n, r_{in}) \in \mathcal{G}_i$, as a result
 1255 of the above n drawings, can be written as

$$\begin{aligned}
 1256 \theta(g|\pi, t^*) &\triangleq \frac{f(r_{i1} | s_1, \pi)}{\sigma(l_i | s_1, \pi)} \cdot \frac{f(r_{i2} | s_2, \pi)}{\sigma(l_i \setminus \{r_{i1}\} | s_2, \pi)} \cdots \frac{f(r_{in} | s_n, \pi)}{f(r_{in} | s_n, \pi)} \\
 1257 &= \frac{\prod_{\ell=1}^n f(r_{i\ell} | s_\ell, \pi)}{\nu(g|\pi)}
 \end{aligned}$$

1258 where we denote by $\nu(g|\pi)$ the denominator of the expression
 1259 on the RHS of \triangleq above. The sampling process of $\theta(g|\pi, t^*)$
 1260 for vertical schemes across all the groups of the table is
 1261 illustrated in Fig. 4. We note that $\theta(g|\pi, t^*)$ is dependent on
 1262 the chosen ordering of the sensitive values s_1, \dots, s_n , which

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/ICSS*, Aug. 2017, pp. 394–401.
- [3] B. Bichsel, T. Gehr, D. Drachler-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.
- [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, pp.158–166.
- [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.
- [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](https://doi.org/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.
- [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. Venkatasubramanian, " 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://doi.org/10.14722/ndss.2018.23183).
- [40] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. 2nd ed. Springer, 2004.
- [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>

AQ:6

AQ:7

AQ:5