Attribute Privacy: Framework and Mechanisms

Wanrong Zhang
Harvard University
Cambridge, MA, USA
wanrongzhang@fas.harvard.edu

Olga Ohrimenko
The University of Melbourne
Melbourne, Australia

Rachel Cummings
Columbia University
New York, NY, Arunachal Pradesh
USA

ABSTRACT

Ensuring the privacy of training data is a growing concern since many machine learning models are trained on confidential and potentially sensitive data. Much attention has been devoted to methods for protecting individual privacy during analyses of large datasets. However in many settings, global properties of the dataset may also be sensitive (e.g., mortality rate in a hospital rather than presence of a particular patient in the dataset). In this work, we depart from individual privacy to initiate the study of attribute privacy, where a data owner is concerned about revealing sensitive properties of a whole dataset during analysis. We propose definitions to capture *attribute privacy* in two relevant cases where global attributes may need to be protected: (1) properties of a specific dataset and (2) parameters of the underlying distribution from which dataset is sampled. We also provide two efficient mechanisms for specific data distributions and one general but inefficient mechanism that satisfy attribute privacy for these settings. We base our results on a novel and non-trivial use of the Pufferfish framework to account for correlations across attributes in the data, thus addressing "the challenging problem of developing Pufferfish instantiations and algorithms for general aggregate secrets" that was left open by Kifer and Machanavajjhala in 2014 [15].

KEYWORDS

attribute privacy, Pufferfish privacy, formal privacy frameworks, privacy-preserving mechanisms

ACM Reference Format:
Wanrong Zhang, Olga Ohrimenko, and Rachel Cummings. 2021. Attribute Privacy: Framework and Mechanisms. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3411764.3445100

1 INTRODUCTION

Privacy in the computer science literature has generally been defined at the *individual level*, such as differential privacy [7], which protects the value of an individual's data within analysis of a larger dataset. However, there are many settings where confidential information contained in the data goes beyond the presence or absence of an individual in the data and instead relates to attributes at the *dataset level*. Global properties about attributes revealed from data analysis may leak trade secrets, intellectual property and other valuable information pertaining to the data owner, even if differential privacy is applied [5].

In this paper, we are interested in privacy of *attributes in a dataset*, where an analyst must prevent *global properties* of sensitive attributes in her dataset from leaking during analysis. For example, insurance quotes generated by a machine-learned model might leak information about how many female and male drivers are insured by the company that trained the model; voice and facial recognition models may leak the distribution of race and gender among users in the training dataset [1, 4]. Under certain circumstances, even releasing the distribution from which the data were sampled may be sensitive. For example, experimental findings by a pharmaceutical company measuring the efficacy of a new drug would be considered proprietary information. It is important to note that the problem we consider here departs from individual-level attribute privacy where one wishes to protect attribute value of a record (e.g., person's race) as opposed to a function over all values of this attribute in the dataset (e.g., race distribution in a dataset).

Several recent attacks show that global properties about a dataset can indeed be leaked from machine learning model APIs [10, 20, 23, 26]. In fact, these works show that models learn sensitive attributes even when censorship is applied or when the attributes are deemed irrelevant for the actual learning task. Hence, the naive solution of removing sensitive attributes from the dataset is insufficient, as attributes are often correlated, and protected information can still be leaked by releasing non-sensitive information about the data. Though differential privacy (DP) can be used to protect sensitive attributes at the individual level (e.g., in the algorithmic fairness literature [6]), the study of attribute privacy at the dataset or distribution level is limited, both in terms of a framework for reasoning about it and mechanisms for protecting it.

1.1 Our Contributions

Problem formulation. We initiate the study of *attribute privacy* at the dataset and distribution level and establish the first formal framework for reasoning about these privacy notions. Our work aims to protect *aggregate properties of a dataset* and not individual properties that make up this dataset as the as the latter can be already captured by DP. For example, contrast protecting trade secrets of an organization based on their customer base versus the privacy of their individual customers.

We identify two cases where information about global properties of a dataset may need to be protected: (1) properties of a specific dataset and (2) parameters of the underlying distribution from which dataset is sampled. We refer to the first setting as *dataset attribute privacy*, where the data owner wishes to protect properties of her sample from a distribution, but is not concerned about

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445100

revealing the distribution. For example, even though the overall prevalence of a disease may be known, a hospital may wish to protect the fraction of its patients with that disease. We refer to the second setting as *distributional attribute privacy*, which considers the distribution parameter itself a secret. For example, demographic information of the population targeted by a company may reveal information about its proprietary marketing strategy. These two definitions distinguish between protecting a sample and protecting the distribution from which the dataset is sampled.

Definitions of Attribute Privacy. We propose definitions for capturing dataset and distributional attribute privacy by instantiating a general privacy framework called the Pufferfish framework [15]. This framework was originally introduced to handle correlations across individual entries in a database. Instantiating this framework for attribute privacy is non-trivial as it requires reasoning about secrets and parameters at a dataset level.

For dataset attribute privacy, our definition considers the setting where individual records are independent of each other while correlations may exist between attribute values of each record. Then, to be able to capture general global properties of a dataset that need to be protected, we choose to express secrets as functions over attribute values across all records in a dataset. For example, this allows one to express that the average income of individuals in a dataset being below or above \$50K is secret information.

Our second definition also instantiates the Pufferfish framework while explicitly capturing the random variables used to generate attribute values of a record. Here, the parameters of the distribution of protected attributes are treated as confidential information. For example, in a dataset where records capture trials in a stochastic chemical environment, one can express that determining whether the probability with which a certain compound is added in each trial is 0.2 or 0.8 is a secret.

Interestingly, there are some settings where dataset and distributional privacy are equivalent. In Section 3 we describe technical conditions on the secrets when this occurs.

Mechanisms to Protect Attribute Privacy. Our definitions allow an analyst to specify secrets about global properties of a dataset that they wish to protect. In order to satisfy these definitions the analyst can use a general tool for providing Pufferfish privacy called the Wasserstein mechanism proposed by Song *et al.* [24]. However, this mechanism is computationally expensive and may require computing an exponential number of pairwise Wasserstein distances. Wasserstein distance is generally hard to compute, which makes the algorithm not feasible in most practical settings. To this end, we propose more practical mechanisms in the following two settings.

For dataset attribute privacy, we consider a special class of functions and attribute properties and propose a mechanism based on Gaussian noise. Though the nature of the noise is added from the same family of distributions as the differentially private Gaussian Mechanism, in Section 4 we articulate that the similarity between the two is based solely on the nature of the noise. In Section 4.1, we show that the mechanism can be applied to datasets where (1) attributes follow a multivariate Gaussian distribution and (2) the function to be computed on the data and the attribute property to be protected are linear in the number of records in the dataset (e.g., mean). We empirically evaluate the noise added by our mechanism and show that there are settings where no noise needs to be added due to randomness present in the data.

We note that with the help of variational auto-encoders (VAEs) [16], one can obtain a Gaussian representation of the data even if a dataset does not come from a Gaussian distribution. Moreover, such disentangled representations can be based on interpretable attributes [12] that are easier for specifying which attributes require protection, particularly when the original data are complex (e.g., pixels on an image vs. the gender of the person in it).[1] Nevertheless, we also consider the case where data may not follow Gaussian distribution. Specifically, in Appendix B, we relax the Gaussian assumption and show that our mechanism can still provide dataset attribute privacy by leveraging Gaussian approximations.

For distributional attribute privacy, we consider a model where dependencies between the attributes form a Bayesian network. This model helps us capture the extent to which a sensitive attribute parameter affects parameters of attributes in the query, and we add noise proportional to this influence. Although our mechanism is inspired by the Markov Quilt mechanism [24], the difference in settings prompts several changes, including a different metric for measuring influence between the variables.

Finally, we note that although [15] identified that "there is little focus in the literature on rigorous and formal privacy guarantees for business data", they leave "the challenging problem of developing Pufferfish instantiations and algorithms for general aggregate secrets" as future work.

1.2 Examples

Dataset privacy. Consider a hospital releasing statistics about their adult patients who have been diagnosed with a certain virus. The hospital is careful not to release information about the gender of the patients as there is not enough evidence yet to suggest whether the virus is more prevalent among certain gender groups. Among other biometric statistics about the patients receiving treatment, the information released by the hospital includes patients' average height of 67.7 inches and average weight 185 pounds. According to https://www.worlddata.info/average-bodyheight.php, average height and weight of men and women is 68.9 inches and 200 pounds and 64.2 inches and 169 pounds, respectively. Any observer comparing these public numbers to the statistics released by the hospital, may notice that it is likely that the dataset contained more male patients than other genders, leading to unfounded conclusions about correlation of the virus and gender.

Distributional privacy. An area council requests all schools to submit statistics about their students including their height, weight, SAT score, and gender. Since the schools are concerned with privacy of their students, they agree to submit statistics from only a random sample of their students. Furthermore, in order to preserve anonymity of the schools (e.g., single vs. mixed gender schools), the schools opt out from releasing the gender distribution along with their statistics. Due to correlations among gender, height, and weight, naive release of the sampled data by schools would give an

[1] Though naive use of VAEs may not provide end-to-end privacy guarantees, it serves as an example that it is possible to obtain a representation of non-Gaussian data with interpretable Gaussian features. We leave it as an interesting open question on how to provide end-to-end privacy-preserving feature disentanglement.

indication about the gender distribution of the students where the samples were taken from. As a result, schools' identities could be revealed.

We see that these two examples contrasting the information that needs to be concealed under both privacy notions. In the former it is the distribution of gender among patients, while in the latter it is the distribution of the students from which the sample was taken.

We also note that our examples are reminiscent of algorithmic fairness, and indeed, our definitions are related to group fairness. Our attribute privacy definitions and mechanisms provide statistics that are de-correlated from protected attributes in the dataset, and can be used to satisfy *group-blind* fairness notions such as demographic parity, where the output of a mechanism (such as labels generated by a classifier) should be independent of protected attributes. We note, however, that our privacy notions would not be suitable for satisfying fairness notions that explicitly incorporate protected attributes to make up for historical injustices, such as affirmative action or fairness through awareness.

1.3 Related work

Machine learning models have been shown to memorize and leak data used to train them, raising questions about the release and use of these models in practice. For example, membership attacks [22] show that models can leak whether certain records (e.g., patient data) were part of the training dataset or not. Attribute (or feature) privacy attacks, on the other hand, consider leakage of attribute values at an individual level [9, 23], and property inference attacks show that global properties about datasets can be leaked [1, 10, 20].

Differential privacy (DP) [7, 8] guarantees individual-level privacy when publishing an output computed on a database, by bounding the influence of any record on the output and adding noise. Importantly, DP does not aim to protect population-level information, and was designed to learn global properties of a dataset without sacrificing individual privacy. DP does provide *group privacy* guarantees for groups of k correlated records, but these quantitative guarantees are only meaningful when k is small relative to the size of the dataset. Syntactically, DP guarantees that if any individual record were to be changed—including all attributes of that record—the result of the analysis would be approximately the same. For attribute privacy, we seek similar guarantees if an entire attribute of the dataset were to be changed—including all individuals' values for that attribute. Several works [2, 3] consider settings where data come from a fixed distribution, as opposed to the set of all possible neighboring datasets as in DP. The corresponding definitions and mechanisms also explicitly specify these data distributions. However, similar to differential privacy, these works focus on hiding the presence or absence of a single record, rather than dataset properties as we consider here.

The Pufferfish framework [15] that we instantiate and describe in detail in the following sections, can be seen as a generalization of differential privacy that explicitly states the information that needs to be kept secret and the adversary's background knowledge about the data. Blowfish privacy [11] also allows one to express secrets and publicly known information about the data, but expressed as constraints on the data rather than distributions over data. We adapt the Markov Quilt Mechanism from [24], who also employ the Pufferfish framework [15] for private analysis of correlated data, although they also focus on individual-level privacy. Our focus instead on privacy of dataset properties and distributions leads to a substantially different instantiation of the Pufferfish framework where the secrets are defined over attribute values rather than individual records in the dataset.

Research on algorithmic fairness has proposed several definitions formalizing the idea that machine learning models should not exhibit discrimination based on protected attributes (e.g., gender or race). Demographic parity formalizes fairness by requiring that a classifier's predicted label is independent of an individual's protected attributes. Our notion of dataset attribute privacy is a general framework where one can specify what information about attributes need to be protected, with attribute independence being one such scenario. However, our attribute privacy definitions would not be useful for satisfying other fairness notions that explicitly incorporate protected attributes, such as affirmative action or fairness through awareness [6]. Moreover, techniques proposed to obtain fair representations of the training data [18, 25] have been shown to still leak sensitive attributes [23] when applied in the privacy context.

2 PRELIMINARIES

Pufferfish Privacy. The Pufferfish privacy framework [15] consists of three components: a set S of secrets, a set $\mathcal{Q} \subseteq S \times S$ of discriminative pairs, and a class Θ of data distributions. S is a set of possible facts about the database that we might wish to hide. $\mathcal{Q}$ is the set of secret pairs (s_i, s_j), $s_i, s_j \in S$, that we wish to be indistinguishable. The class of data distributions Θ can be viewed as a set of conservative assumptions about the underlying distribution that generates the database.

Definition 1 ((ϵ, δ)-Pufferfish Privacy [15, 24][2]). *A mechanism $\mathcal{M}$ is (ϵ, δ)-Pufferfish private in a framework $(S, \mathcal{Q}, \Theta)$ if for all $\theta \in \Theta$ with $X \sim \theta$, for all secret pairs $(s_i, s_j) \in \mathcal{Q}$ such that $P(s_i|\theta) \neq 0$ and $P(s_j|\theta) \neq 0$, and for all $T \subseteq Range(\mathcal{M})$, we have*

$$P_{\mathcal{M},\theta}(\mathcal{M}(X) \in T | s_i, \theta) \leq \exp(\epsilon) P_{\mathcal{M},\theta}(\mathcal{M}(X) \in T | s_j, \theta) + \delta.$$

The Wasserstein Mechanism proposed in [24] and defined formally in Appendix C is the first general mechanism for satisfying instantiations of Pufferfish privacy framework. It defines sensitivity of a function F as the maximum Wasserstein distance between the distribution of $F(X)$ given two different realizations of secrets s_i and s_j for $(s_i, s_j) \in \mathcal{Q}$. The mechanism then instantiates the Laplace mechanism by outputting $F(X)$ plus Laplace noise that scales with this sensitivity. Although this mechanism works in general for any instantiation of the Pufferfish framework, computing Wasserstein distance for all pairs of secrets is computationally expensive, and will typically not be feasible in practice.

Song *et al.* [24] also gave the Markov Quilt Mechanism (Algorithm 3 in Appendix A) for some special structures of data dependence. It is computationally less expensive than the Wasserstein Mechanism and also guarantees $(\epsilon, 0)$-Pufferfish privacy. The Markov Quilt Mechanism of [24] assumes that the entries in the input database Y form a Bayesian network, as defined below. These

[2]The original definition [15] and the one considered in [24] is $(\epsilon, 0)$-Pufferfish. We extend the definition to (ϵ, δ)-Pufferfish in the natural way.

entries could either be: (1) the multiple attributes of a single record when the database contained only one record, or (2) the attribute values across multiple records for a single-fixed attribute when the database contained multiple attributes. Hence, the original Markov Quilt Mechanism aims to protect the privacy of a single entry, and it could not accommodate correlations across multiple attributes in multiple records, as we study in this work. Full details of this algorithm are given in Appendix A.

Definition 2 (Bayesian Networks). *A Bayesian network is described by a set of variables $Y = \{Y_1, \ldots, Y_n\}$ and a directed acyclic graph $G = (Y, E)$ whose vertices are variables in Y. The probabilistic dependence on Y induced by the network can be written as: $\Pr(Y_1, \ldots, Y_n) = \Pi_{i=1}^{n} \Pr(Y_i|\text{parent}(Y_i))$, where the parent of Y_i is the vertex connected to Y_i on the path to the root.*

Accuracy. We will measure accuracy of our mechanisms with the following definition. For real-valued outputs, this definition says that the mechanism must output an answer that is at most an additive α away from the true answer with probability $1 - \beta$. For vector-valued outputs, this notion can be naturally extended using the appropriate norm.

Definition 3 ((α, β)-accuracy). *A mechanism $\mathcal{M}$ with real-valued outputs is (α, β)-accurate for a function F if for all databases X,*

$$P(|\mathcal{M}(X) - F(X)| > \alpha) \leq \beta.$$

3 ATTRIBUTE PRIVACY DEFINITIONS

Data model and representation. The dataset X contains n records, where each record consists of m attributes. We view the dataset X as an $n \times m$ matrix. In this work, we are interested in privacy of the *columns*, which represent attributes that a data owner wishes to protect. Thus we refer to the matrix X as $X = [X_1, \ldots, X_m]$, where X_i is the column vector related to the ith attribute (column). In contrast, traditional differential privacy [7, 8] is concerned with privacy of the *rows* of the dataset matrix. We let X_i^j denote ith attribute value of the jth record. One might think of applying DP to the rows of the transpose of the data matrix, however, using group privacy to handle the correlation across the attributes would give poor utility, because the set of correlated attributes to be covered under group privacy might be large relative to the total number of attributes. Moreover, directly applying Pufferfish privacy or Blowfish privacy can only capture the correlation across attributes for a specific individual, rather than the global properties over n individuals.

Each record is assumed to be sampled i.i.d. from an unknown distribution, where attributes within a single record can be correlated (e.g., consider height and weight). We use $C \subseteq [m]$ to denote a set of indices of the sensitive attributes that require privacy protection (e.g., race and gender may be sensitive attributes; hair color may be non-sensitive). The data owner wishes to compute a function F over her dataset and release the value (or estimate of the value) $F(X)$ while protecting some information about the sensitive attributes.

Privacy notions. We distinguish between three kinds of attribute privacy, corresponding to three different types of information the data owner may wish to protect.

Individual attribute privacy protects X_i^j for sensitive attribute i when $F(X)$ is released. Note that differential privacy provides individual attribute privacy simultaneously for all individuals and all attributes [7], but does not protect against individual-level inferences from population-level statistics [5]. For example, if a DP result shows a correlation between lung disease and smoking, one may infer that a known-smoker in the dataset has an elevated likelihood of lung disease.

Dataset attribute privacy is applicable when the owner wishes to reveal $F(X)$ while protecting the value of some function $g(X_i)$ for sensitive attribute $i \in C$ (e.g., whether there were more Caucasians or Asians present in the dataset).

Distribution attribute privacy protects privacy of a parameter ϕ_i that governs the distribution of ith sensitive attribute in the underlying population from which the data are sampled.

The last two notions are the ones put forward in this paper and studied in detail. The difference between them may be subtle depending on g and ϕ. For example, consider one setting where the sensitive attribute is binary and g is the fraction of records where this attribute is 1, and another setting where the sensitive attribute is a Bernoulli random variable with parameter ϕ. In this case, g can be seen as an estimate of ϕ based on a sample. The difference becomes particularly relevant in settings where privacy is required for realizations of the dataset that are unlikely under the data distribution, or settings with small datasets where g is a poor estimate of ϕ.

Formal framework for attribute privacy. The standard notion of differential privacy is not directly applicable to our setting since we are interested in protecting population-level information. Instead, we formalize our attribute privacy definitions using the Pufferfish privacy framework of Definition 1 by specifying the three components $(S, \mathcal{Q}, \Theta)$. The distributional assumptions of this framework are additionally useful for formalizing correlation across attributes.

Definition 4 (Dataset Attribute Privacy). *Let $(X_1^j, X_2^j, \ldots, X_m^j)$ be a record with m attributes that is sampled from an unknown distribution $\mathcal{D}$, and let $X = [X_1, \ldots, X_m]$ be a dataset of n records sampled i.i.d. from $\mathcal{D}$ where X_i denotes the (column) vector containing values of the ith attribute of every record. Let $C \subseteq [m]$ be the set of indices of sensitive attributes, and for each $i \in C$, let $g_i(X_i)$ be a function with codomain $\mathcal{U}^i$.*

A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-dataset attribute privacy *if it is (ϵ, δ)-Pufferfish private for the following framework $(S, \mathcal{Q}, \Theta)$:*

> *Set of secrets: $S = \{s_a^i := \mathbb{1}[g_i(X_i) \in \mathcal{U}_a^i] : \mathcal{U}_a^i \subseteq \mathcal{U}^i, i \in C\}$, where $\mathcal{U}_a^i$ is a certain specified subset of $\mathcal{U}^i$.*
> *Set of secret pairs: $\mathcal{Q} = \{(s_a^i, s_b^i) \in S \times S, i \in C\}$, which consists of certain pairs of interest.*
> *Distribution: Θ is a set of possible distributions θ over the dataset X. For each possible distribution $\mathcal{D}$ over records, there exists a $\theta_{\mathcal{D}} \in \Theta$ that corresponds to the distribution over n i.i.d. samples from $\mathcal{D}$.*

This definition defines each secret s_a^i as the event that $g_i(X_i)$ takes a value in a particular set $\mathcal{U}_a^i$, and the set of secrets S is the collection of all such secrets for all sensitive attributes. This collection may include all possible subsets of $\mathcal{U}^i$, or it may include

only application-relevant events. For example, if all $\mathcal{U}_a^i$ are singletons, this corresponds to protecting any realization of $g_i(X_i)$. Alternatively, the data owner may only wish to protect whether $g_i(X_i)$ is positive or negative, which requires only $\mathcal{U}_a^i = (-\infty, 0)$ and $\mathcal{U}_b^i = [0, \infty)$. The set of secret pairs $\mathcal{Q}$ that must be protected includes all pairs of the events on the same sensitive attribute. The Pufferfish framework considers distributions θ over the entire dataset X, whereas we require distributions $\mathcal{D}$ over records. We resolve this by defining Θ to be the collection of distributions over datasets induced by the allowable i.i.d. distributions over records.

Determining which functions g_i to consider is an interesting question. For example, in [19] the authors show that it is tractable to check whether the output of certain classes of functions evaluated on a dataset reveals information about the output of another query evaluated on the same dataset. Hence, given a function F whose output a data owner wishes to release, the owner may consider either those g_i's about which F reveals information, or those for which verifying perfect privacy w.r.t. F is infeasible.

Let us instantiate the first example in the introduction using the definition above. Suppose a patient's record consists of 5 attributes: gender, height, weight, average blood pressure, and temperature. Exemplar records could be:

$$X^1 = (\texttt{male}, 69, 200, 119{:}79, 100) \quad X^2 = (\texttt{female}, 59, 200, 115{:}83, 98)$$
$$X^3 = (\texttt{male}, 79, 225, 119{:}81, 99) \quad X^4 = (\texttt{other}, 71, 180, 114{:}79, 102)$$

The hospital wishes to release an average of every column of the dataset except for gender, X_1, which it deems as the sensitive attribute. That is, the hospital wishes to release a tuple of four values denoting average height, weight, blood pressure, and temperature. In order to provide privacy for the gender attribute, the hospital instantiates Definition 4 as follows. Given n, the size of the dataset, it defines g_1 to be a function that counts the fraction of patients in the dataset with gender `female`. It then declares a set of secrets $S = \{\mathbb{1}[g_1(X_1) = n/2], \mathbb{1}[g_1(X_1) = n/4], \mathbb{1}[g_1(X_1) = n/8], \mathbb{1}[g_1(X_1) = n/8]\}$, capturing the fact that any mechanism $\mathcal{M}$ that operates on this data should not allow an attacker to distinguish whether number of females in the dataset is half, a quarter, an eighth or a tenth of the dataset.

Definition 5 (Distributional Attribute Privacy). *Let $(X_1^j, X_2^j, \ldots, X_m^j)$ be a record with m attributes that is sampled from an unknown distribution described by a vector of random variables $(\phi_1, \ldots, \phi_m)$, where ϕ_i parameterizes the marginal distribution of X_i^j conditioned on the values of all ϕ_k for $k \neq i$. The $(\phi_1, \ldots, \phi_m)$ are drawn from a known joint distribution P, and each ϕ_i has support Φ^i. Let $X = [X_1, \ldots, X_m]$ be a dataset of n records sampled i.i.d. from the distribution described by $(\phi_1, \ldots, \phi_m)$ where X_i denotes the (column) vector containing values of ith attribute of every record. Let $C \subseteq [m]$ be the set of indices of sensitive attributes.*

*A mechanism $\mathcal{M}$ satisfies (ϵ, δ)-*distributional attribute privacy *if it is (ϵ, δ)-Pufferfish private for the following framework $(S, \mathcal{Q}, \Theta)$:*

> *Set of secrets: $S = \{s_a^i := \mathbb{1}[\phi_i \in \Phi_a^i] : \Phi_a^i \subset \Phi^i, i \in C\}$, where Φ_a^i is certain specified subset of Φ^i.*
> *Set of secret pairs: $\mathcal{Q} = \{(s_a^i, s_b^i) \in S \times S, i \in C\}$, which consists of certain pairs of interest.*
> *Distribution: Θ is a set of possible distributions θ over the dataset X. For each possible $\phi = (\phi_1, \ldots, \phi_m)$ describing the conditional marginal distributions for all attributes, there exists a $\theta_\phi \in \Theta$ that corresponds to the distribution over n i.i.d. samples from the distribution over records described by ϕ.*

This definition naturally parallels Definition 4, with the attribute-specific random variable ϕ_i taking the place of the attribute-specific function $g_i(X_i)$. Although it might seem natural for ϕ_i to define the *marginal* distribution of the ith attribute, this would not capture the correlation across attributes that we wish to study. Instead, ϕ_i defines the *conditional marginal* distribution of the ith attribute given all other $\phi_{\neq i}$, which does capture such correlation. This also allows the distribution θ over datasets to be fully specified given these parameters and the size of the dataset.

More specifically, we model attribute distributions using standard notion of Bayesian hierarchical modeling. The $(\phi_1, \ldots, \phi_m)$ can be viewed as a set of hyperparameters of the distributions of the attributes, and P as hyper-priors of the hyperparameters. The distribution P is captured in Θ, and the distribution of attribute X_i is governed by a realization of the random variable ϕ_i. The ϕ_i describes the conditional marginal distribution for attribute i: it is the hyperparameter of the probability of X_i given hyperparameters of all other attributes $P(X_i|\phi_1, \ldots, \phi_{i-1}, \phi_{i+1}, \ldots, \phi_m)$. We make the "naive" conditional independence assumption that all attributes X_i are mutually independent conditional on the set of parameters $(\phi_1, \ldots, \phi_m)$, hence, $(\phi_1, \ldots, \phi_m)$ fully capture the distribution of a record. The "naive" conditional independence is a common assumption in probabilistic models, and naive Bayes is a simple example that employs this assumption.

Let us instantiate the second example in the introduction using the definition above. Suppose an all-girls school takes a sample of their students, where each student record contains 4 attributes: gender, height, weight, and SAT score. Exemplar records could be:

$$X^1 = (\texttt{female}, 69, 200, 1250) \quad X^2 = (\texttt{female}, 59, 200, 1300)$$
$$X^3 = (\texttt{female}, 79, 225, 1400) \quad X^4 = (\texttt{female}, 71, 180, 1410).$$

The school is asked to release an average of every column of the dataset except for gender, X_1, which it regards as a sensitive attribute. That is, the school wishes to release a tuple of three values denoting average height, weight, and SAT score. In order to provide privacy for the gender attribute, the school instantiates Definition 5 as follows. It models X_1 as a Bernoulli random variable where ϕ_1 is the probability of a record having a `female` attribute. It then declares a set of secrets corresponding to protected values of ϕ_1 to be $S = \{\mathbb{1}[\phi_1 = 0], \mathbb{1}[\phi_1 = 0.5], \mathbb{1}[\phi_1 = 1]\}$. This captures the fact that any mechanism $\mathcal{M}$ that operates on this data should not allow an attacker to distinguish whether the data was sampled from the population with no, 50%, or all female students. The school may consider the set of possible distributions Θ to contain all gender distributions represented at schools in the district, or it may consider all possible distributions with support $[0, 1]$.

Dataset vs. Distributional privacy. Suppose there exists an unbiased estimator for the distribution parameter ϕ_i corresponding to the secrets. Then we can achieve distributional attribute privacy by using dataset attribute private mechanisms (such as those given in Section 4) with the secrets defined as values of this unbiased estimator. In the last example, a function g_1 that computes an average

number of female patients in the dataset is an unbiased estimator for ϕ_1. Hence, one could also instantiate this example using the dataset attribute privacy with secrets based on the value of g_1. However, such unbiased estimators do not always exist. In these cases, it is necessary to use mechanisms designed for distributional attribute privacy (such as those given in Section 5). In general, unbiased estimators do not exist for quantities that cannot be written as a polynomial of degree less than the population size n [17].

Mechanisms for attribute privacy. Since both of our attribute privacy definitions are instantiations of the Pufferfish privacy framework, one could easily apply the Wasserstein Mechanism [24] to satisfy $(\epsilon, 0)$-attribute privacy for either of our definitions. The Wasserstein distance metric has also been used to calibrate noise in prior work on distributional variants of differential (individual-level) privacy [13, 14]. However, as described in Section 2, implementing this mechanism requires computing Wasserstein distance between the conditional distribution on $F(X)$ for all pairs of secrets in $\boldsymbol{Q}$. Computing exact Wasserstein distance is known to be computationally expensive, and our settings may require exponentially many computations in the worst case. In the remainder of the paper, we provide algorithms that satisfy each of these privacy definitions, focusing on dataset attribute privacy in Section 4 and distributional attribute privacy in Section 5. We formally discuss the (computationally more expensive) Wasserstein Mechanism along with concrete examples of its instantiation in Appendix C.

4 THE GAUSSIAN MECHANISM FOR DATASET ATTRIBUTE PRIVACY

In this section we consider dataset attribute privacy as introduced in Definition 4. In this setting, an analyst wants to publish a function F evaluated on her dataset X, but is concerned about an adversary observing $F(X)$ and performing a Bayesian update to make inferences about a protected quantity $g_i(X_i)$. We propose a variant of the Gaussian Mechanism [8] that satisfies dataset attribute privacy when $F(X)$ conditioned on $g_i(X_i)$ follows a Gaussian distribution, with constant variance conditioned on $g_i(X_i) = a$ for all a.

Although this setting is more restrictive, it is still of practical interest. For example, it can be applied when X follows a multivariate Gaussian distribution and g_i and F are linear with respect to the entries of X, as we show in the instantiation of our mechanism in Section 4.2. We also note that using variational auto-encoders (VAEs) [12, 16], it is possible to encode data from other distributions using a Gaussian representation with interpretable features. This would then allow an analyst to specify which latent features are deemed sensitive for the data, even if the original features are less descriptive (e.g., pixels on an image vs. the gender of the person in it).

In Appendix B, we propose the Attribute-Private Gaussian Mechanism for non-Gaussian data that does not make the above assumptions. In particular, the mechanism allows the analyst to use Gaussian approximations to characterize the conditional distribution of $F(X)$ given $g_i(X_i)$, while still providing formal dataset attribute privacy guarantees. Further details are deferred to the appendix.

4.1 Attribute-Private Gaussian Mechanism

Algorithm 1 presents the Attribute-Private Gaussian Mechanism for answering a real-valued query $F(X)$ while protecting the values of $g_i(X_i)$ for $i \in C$. Much like the Gaussian Mechanism for differential privacy [8], the Attribute-Private Gaussian Mechanism first computes the true value $F(X)$, and then adds a Gaussian noise term with mean zero and standard deviation that scales with the sensitivity of the function. However, *sensitivity* of F in the attribute privacy setting is defined with respect to each secret attribute X_i as,

$$\Delta_i F = \max_{\theta \in \Theta} \max_{(s_a^i, s_b^i) \in \boldsymbol{Q}} \left| \mathbb{E}\left[F(X) | s_a^i, \theta\right] - \mathbb{E}\left[F(X) | s_b^i, \theta\right] \right|. \quad (1)$$

This differs from the sensitivity notion used in differential privacy in two key ways. First, we are concerned with measuring changes to the value of $F(X)$ caused by changing secrets s_a^i corresponding to realizations of $g_i(X_i)$, rather than by changing an individual's data. Second, we assume our data are drawn from an unknown underlying distribution θ, so $F(X)$ is a random variable. Our attribute privacy sensitivity bounds the maximum change in posterior expected value of $F(X)$ in the worst case over all distributions and pairs of secrets for each attribute. We note that if $F(X)$ is independent of the protected attribute X_i, then $\Delta_i F = 0$ and no additional noise is needed for privacy. The Attribute-Private Gaussian Mechanism of Algorithm 1 further benefits from the inherent randomness of the output $F(X)$. In particular, it reduces the variance σ^2 of the noise added by the conditional variance of $F(X)$ given $g_i(X_i)$ and θ, as the sampling noise can mask some of the correlation. Hence, privacy also comes for free if the function of interest has low correlation with the protected attributes. A similar observation is made by "noiseless" mechanisms in [2, 3] that show that the uncertainly that comes from some data distributions can be leveraged to protect individual record privacy without additional noise.

Algorithm 1 can be easily extended to handle vector-valued queries with $F(X) \in \mathbb{R}^k$ and sensitive functions g_i over multiple attributes by changing $\Delta_i F$ in Equation (1) to be the maximum ℓ_2 distance rather than absolute value. Additionally, the noise adjustment for each attribute should be based on the conditional covariance matrix of $F(X)$ rather than the conditional variance.

Algorithm 1 Attribute-Private Gaussian Mechanism, APGM$(X, F, \{g_i\}, C, \{S, \boldsymbol{Q}, \Theta\}, \epsilon, \delta)$ for dataset attribute privacy.

Input: dataset X, query F, functions g_i for protected attributes $i \in C$, framework $\{S, \boldsymbol{Q}, \Theta\}$, privacy parameters ϵ, δ
Set $\sigma^2 = 0$, $c = \sqrt{2\log(1.25/\delta)}$.
for each $i \in C$ **do**
  Set $\Delta_i F = \max_{\theta \in \Theta} \max_{(s_a^i, s_b^i) \in \boldsymbol{Q}} \left| \mathbb{E}\left[F(X) | s_i^a\right] - \mathbb{E}\left[F(X) | s_i^b\right] \right|$.
  if $(c\Delta_i F/\epsilon)^2 - \min_{\theta \in \Theta} \mathrm{Var}(F(X)|g_i(X_i), \theta) \geq \sigma^2$ **then**
    Set $\sigma^2 = (c\Delta_i F/\epsilon)^2 - \min_{\theta \in \Theta} \mathrm{Var}(F(X)|g_i(X_i), \theta)$.
if $\sigma^2 > 0$ **then**
  Sample $Z \sim \mathcal{N}(0, \sigma^2)$.
  Return $F(X) + Z$.
else Return $F(X)$.

Theorem 1. *The Attribute-Private Gaussian Mechanism APGM($X, F, \{g_i\}, C, \{S, \mathcal{Q}, \Theta\}, \epsilon, \delta$) in Algorithm 1 is (ϵ, δ)-dataset attribute private when $F(X)|g_i(X_i)$ is Gaussian distributed for any $\theta \in \Theta$ and $i \in C$.*

Privacy follows from the observation that the summation of $F(X)$ and the Gaussian noise Z is Gaussian distributed conditioned on any secrets, and the probabilities of the output conditioned on any pairs of secrets have the same variance with mean difference $\Delta_i F$. Since we bound the ratio of the two probabilities caused by shifting this variable, the analysis reduces to the proof of Gaussian mechanism in differential privacy. The full proof of the theorem appears in Appendix D.

High probability additive accuracy bounds on the output of Algorithm 1 can be derived using tail bounds on the noise term Z based on its variance σ^2. The formal accuracy guarantee is stated in Theorem 2, which follows immediately from tail bounds of a Gaussian.

Theorem 2. *The Attribute-Private Gaussian Mechanism APGM($X, F, \{g_i\}, C, \{S, \mathcal{Q}, \Theta\}, \epsilon, \delta$) in Algorithm 1 is (α, β)-accurate for any $\beta > 0$ and*

$$\alpha = \sqrt{\max\{0, \max_{i \in C}\{(c\Delta_i F/\epsilon)^2 - \min_{\theta \in \Theta} \mathrm{Var}(F(X)|g_i(X_i), \theta)\}\}}\Phi^{-1}(1 - \frac{\beta}{2}),$$

where $c = \sqrt{2\log(1.25/\delta)}$ and Φ is the CDF of the standard normal distribution.

In general, if $F(X)$ is independent of, or only weakly correlated with the protected functions $g_i(X_i)$, then no noise is needed is preserve dataset attribute privacy, and the mechanism can output the exact answer $F(X)$. On the other hand, if $F(X)$ is highly correlated with $g_i(X_i)$, we then consider a tradeoff between the sensitivity and the variance of $F(X)$. If the variance of $F(X)$ is relatively large, then $F(X)$ is inherently private, and less noise is required. If the variance of $F(X)$ is small and the sensitivity of $F(X)$ is large, the mechanism must add a noise term with large σ^2, resulting in low accuracy with respect to the true answer. In contrast to individual privacy, the accuracy of the mechanism will decrease as the number of records n goes to ∞, because more records will reveal the global information more accurately (the variance of $F(X)$ decreases). To make these statements more concrete and understandable, Section 4.2 provides a concrete instantiation of Algorithm 1.

4.2 Instantiation with Gaussian distributed data

In this section, we show an instantiation of our Attribute-Private Gaussian Mechanism when the joint distribution of the m attributes is multivariate Gaussian. The privacy guarantee of this mechanism requires that $F(X)|g_i(X_i)$ is Gaussian distributed, which is satisfied when g_i and F are linear with respect to the entries of X. For simplicity of illustration, we will choose both $F(X)$ and all $g_i(X_i)$ to compute averages.

We continue with our example from Section 3 with a dataset that consists of patients' information including gender X_{gndr}, heights X_{h}, weights X_{w}, blood pressure X_{b}, and temperature X_{t}. The hospital wishes to release the average weight of its patients, so $F(X) = \frac{1}{n}\sum_{j=1}^{n} X^j_{\text{w}}$. The hospital also wants to prevent an adversary from inferring the proportion of females among their patients, so $C = \{\text{gndr}\}$ and $g(X_{\text{gndr}}) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}[X^j_{\text{gndr}} = \texttt{female}]$.

To instantiate our framework, let s^i_a denote the event that $g(X_i) = a$. If $g(X_i)$ has support $\mathcal{U}^i$, then the set of secrets is $S = \{s^i_a : a \in \mathcal{U}^i, i \in C\}$, and the set of secret pairs is $\mathcal{Q} = \{(s^i_a, s^i_b) : a, b \in \mathcal{U}^i, a \neq b, i \in C\}$. In the above example, setting a pair of secrets $\left(s^{\text{gndr}}_{n/2}, s^{\text{gndr}}_{n/4}\right)$ to $\left(g(X_{\text{gndr}}) = n/2, g(X_{\text{gndr}}) = n/4\right)$ captures that the hospital wishes to hide whether the proportion of female patients is 50% or 25%.

Each $\theta \in \Theta$ is a distribution over n i.i.d. samples from an underlying multivariate Gaussian distribution with mean $(\mu_1, \ldots, \mu_m)^T$ and covariance matrix (V_{ij}), $i, j \in [m]$, where $V_{ij} = V_{ji}$ is the covariance between X_i and X_j if $i \neq j$, and V_{ii} is the variance of X_i. We note that the height, weight, and gender attributes may not be Gaussian distributed in practice. Hence, the choice of whether to use the Attribute-Private Gaussian Mechanism for Gaussian or non-Gaussian data should be determined by the practitioner. See Appendix B for a version of Algorithm 1 that does not rely on assumptions of Gaussian-distributed data.

Suppose we want to guarantee (ϵ, δ)-dataset attribute privacy through the Attribute-Private Gaussian Mechanism. Then we need to first compute $\mathbb{E}\left[F(X)|s^i_a\right]$ and $\mathrm{Var}(F(X)|s^i_a)$ for each $i \in C$. Let j denote the index of the attribute averaged in $F(X)$. By the properties of a multivariate Gaussian distribution, the distribution of $F(X)$ conditional on $g(X_i) = a$ is Gaussian $\mathcal{N}(\bar{\mu}_a, \bar{V})$, where $\bar{\mu}_a = \mu_j + \frac{V_{ij}}{V_{ii}}(a - \mu_i)$ and $\bar{V} = \frac{1}{n}(V_{jj} - \frac{V^2_{ij}}{V_{ii}})$. We define the diameter of $\mathcal{U}$ as $d(\mathcal{U}) = \max_{a,b \in \mathcal{U}} |a - b|$. The sensitivity is: $\Delta_i F = \max_{(s^i_a, s^i_b) \in \mathcal{Q}} |\bar{\mu}_a - \bar{\mu}_b| = \frac{V_{ij}}{V_{ii}} \max_{a,b \in \mathcal{U}} |a - b| = \frac{V_{ij}}{V_{ii}} d(\mathcal{U})$. To ensure (ϵ, δ)-dataset attribute privacy for protected attribute X_i, the variance of the Gaussian noise must be at least $(c\frac{V_{ij} d(\mathcal{U})}{V_{ii}\epsilon})^2 - \frac{1}{n}(V_{jj} - \frac{V^2_{ij}}{V_{ii}})$ for $c = \sqrt{2\log(1.25/\delta)}$ as in Algorithm 1. Adding Gaussian noise with variance $\sigma^2 = \max_{i \in C}\{(c\frac{V_{ij} d(\mathcal{U})}{V_{ii}\epsilon})^2 - \frac{1}{n}(V_{jj} - \frac{V^2_{ij}}{V_{ii}})\}$ will provide (ϵ, δ)-dataset attribute privacy for all protected attributes.

We note that σ^2 is monotonically increasing with respect to V_{ij}. That is, our Attribute-Private Gaussian Mechanism will add less noise to the output if the query F is about an attribute which has a low correlation with the protected attributes.

So far we have discussed about the case when Θ only consists of one distribution, in order to show the impact of V_{ij}. For the general case, the sensitivity $\Delta_i F$ is $\max_{\theta \in \Theta} \frac{V_{ij}}{V_{ii}} d(\mathcal{U})$, and the noise is scaled with variance $\sigma^2 = \max_{i \in C}\{(c \max_{\theta \in \Theta} \frac{V_{ij} d(\mathcal{U})}{V_{ii}\epsilon})^2 - \min_{\theta \in \Theta} \frac{1}{n}(V_{jj} - \frac{V^2_{ij}}{V_{ii}})\}$.

Experiments. In this section we evaluate our Algorithm 1 in the above Gaussian setting. The goal of the experiments is to empirically evaluate how sensitivity affects accuracy of the results, and to demonstrate cases where no noise needs to be injected. We generate a synthetic dataset of 50 patient records as follows. Gender X_{gndr} is generated by *Bernoulli*(.6) with 1 representing female and

0 representing all other genders. If female, then we generate heights X_{h} from $N(65, 3)$ and heights X_{h} from $N(70, 3)$, otherwise. For any given heights X_{h}, we generate the corresponding weights X_{w} by $N(3.6x_{\mathsf{w}} - 90, 15)$. Since temperature is irrelevant to gender, we generate the temperature column by $N(98.6, 0.5)$, and we generate the blood pressure by $N(100, 3)$ for female [21] and $N(106, 3)$, otherwise.

We now describe the set Θ of possible distributions. We assume the possible conditional distributions of weights is $\{N(k_1 x_{\mathsf{w}} - k_2, 15) : k_1 \in (3, 3.5), k_2 \in (80, 90)\}$; the possible conditional distribution of heights is $\{N(\mu_F, 3), \mu_F \in (64, 66)\}$ if female and $\{N(\mu_O, 8), \mu_O \in (69, 71)\}$ otherwise; the possible conditional distribution of blood pressure is $N(b_F, 10)$ if female and $N(b_O, 10)$ otherwise with $6 \leq b_F - b_O \leq 10$; the possible distribution of temperature $N(t_F, 2)$ if female and $N(t_O, 2)$ otherwise with $0 \leq t_F - t_O \leq 0.3$. The possible distributions of gender is $\{Bernoulli(p), p \in (.25, .75)\}$.

An analyst wishes to release aggregate statistics of the record attributes while hiding whether proportion of female patients is above 50% or below 50%. Therefore, the set of secrets is defined as $Q = \{(s_a^{\mathsf{gndr}}, s_b^{\mathsf{gndr}}) : a \in (.25, .5), b \in (.5, .75)\}$. We now evaluate Algorithm 1 on three scenarios where the analyst wants to release one of the following statistics: the average weight, the average blood pressure and the average temperature. We set $\epsilon = 1$ and $\delta = 10^{-3}$. We report accuracy results over an average of 100 runs.

Releasing average weight F_1: Given the above parameters, the sensitivity of F_1 is $\Delta_{\mathsf{gndr}} F_1 = 3.5 \max |\mu_F - \mu_O| \max |a - b| = 6.125$ and the variance is $\mathrm{Var}(F_1 | s_a^{\mathsf{gndr}}) = 4.5$. The true average of weights in the simulated dataset is 161.33, and the average absolute error of Algorithm 1 is 18.90.

Releasing average blood pressure F_2: The sensitivity is now $\Delta_{\mathsf{gndr}} F_2 = \max |b_F - b_O| \max |a - b| = 10 * 0.25 = 2.5$ and the variance is $\mathrm{Var}(F_2 | s_a^{\mathsf{gndr}}) = 2$. The true average blood pressure in the simulated dataset is 102.18, and the average absolute error of Algorithm 1 is 7.40.

Releasing average temperature F_3: For this query, the sensitivity is $\Delta_{\mathsf{gndr}} F_3 = \max |t_F - t_O| \max |a - b| = 0.3 * 0.25 = 0.075$ and the variance $\mathrm{Var}(F_3 | s_a^{\mathsf{gndr}}) = 0.08$. Note that Algorithm 1 does not inject noise in this case, and thus outputs the true average of weights in our simulated dataset, which is 98.58.

Algorithm 1 is computationally efficient and, similar to differential privacy, relies on knowing the sensitivity of the query $F(X)$ that may require additional analysis. The sensitivity can be computed through the possible conditional distribution information from Θ, which is dataset-independent, as we have seen in the above experiments.

5 THE MARKOV QUILT MECHANISM FOR DISTRIBUTIONAL ATTRIBUTE PRIVACY

In this section we consider distributional attribute privacy, as introduced in Definition 5, and develop a mechanism that satisfies this privacy definition. Recall that in this setting, an analyst aims to release $F(X)$ while protecting the realization of a random parameter ϕ_i, which describes the conditional marginal distribution of the ith attribute, given the realization of all ϕ_k for $k \neq i$ for all other attributes. This formalization implies that all (column) attribute vectors X_i are mutually independent, conditional on the set of parameters $(\phi_1, \ldots, \phi_m)$.

5.1 Attribute-Private Markov Quilt Mechanism

We base our mechanism on the idea of a *Markov Quilt*, which partitions a network of correlated random variables into those which are "near" (X_N) a particular variable X_i, and those which are "remote" (X_R). Intuitively, we will use this to partition attributes into those which are highly correlated (X_N) with our sensitive attributes, and those which are only weakly correlated (X_R).

Definition 6 (Markov Quilt). *A set of nodes X_Q in a Bayesian network $G = (X, E)$ is a Markov Quilt for a node X_i if deleting X_Q partitions G into parts X_N and X_R such that $X_i \in X_N$ and X_R is independent of X_i conditioned on X_Q.*

We quantify the effect that changing the distribution parameter ϕ_i of a sensitive attribute X_i has on a set of distribution parameters ϕ_A (corresponding to a set of attributes X_A) using the *max-influence*. Since attributes are mutually independent conditioned on the vector $(\phi_1, \ldots, \phi_m)$, the max-influence is sufficient to quantify how much a change of all values in attribute X_i will affect the values of X_A. If ϕ_i and ϕ_A are independent, then X_A and X_i are also independent, and the max-influence is 0.

Definition 7. *The* max-influence *of an attribute X_i on a set of attributes X_A under Θ is:*

$$e_\Theta(X_A | X_i) = \sup_{\theta \in \Theta} \max_{\phi_i^a, \phi_i^b \in \Phi_i} \max_{\phi_A \in \Phi_A} \log \frac{P(\phi_A | \phi_i^a, \theta)}{P(\phi_A | \phi_i^b, \theta)}.$$

The sensitivity of F with respect to a set of attributes $A \subseteq [m]$, denoted $\Delta_A F$, is defined as the maximum change that the value of $F(X)$ caused by changing all columns X_A. Formally, we say that two datasets X, X' are *A-column-neighbors* if they are identical except for the columns corresponding to attributes in A, which may be arbitrarily different. Then

$$\Delta_A F = \max_{X, X' \ A\text{-column-neighbors}} \left| F(X) - F(X') \right|.$$

Although changing X_A may lead to changes in other columns, these changes are governed by the max influence, and will not affect attributes that are nearly independent of X_i.

Observe that the event that X_R and X_i are independent conditional on X_N is equivalent to the event when ϕ_R and ϕ_i are independent conditional on ϕ_N, which is why we can define the Markov Quilt based on X_i. However, since the distribution of X_is are governed by ϕ_is, the max-influence score must be computed using ϕ_is rather than X_is.

The mechanism. We extend the idea of the Markov Quilt Mechanism in [24] to the attribute privacy setting as follows. Let $A \subseteq [m]$ be a set of attributes over which F is computed. For example, F may compute the average of a particular attribute or a regression on several attributes. At a high level, we add noise to the output of F scaled based on the sensitivity of F with respect to X_Ns. However, when computing the sensitivity of F we only need to consider sensitivity of F with respect to $A \cap N$, i.e., the queried set of attributes A that are in the "nearby" set of the protected attribute.

Algorithm 2 Attribute-Private Markov Quilt Mechanism, APMQM$(X, F, A, C, \{S, \boldsymbol{Q}, \Theta\}, \epsilon)$ for distributional attribute privacy.

Input: dataset X, query F, index set of queried attributes A, index set of sensitive attributes C, framework $\{S, \boldsymbol{Q}, \Theta\}$, privacy parameter ϵ.
for each $i \in C$ **do**
  Set $b_i = \Delta_A F/\epsilon$.
  Set $G_i := \{(X_Q, X_N, X_R) : e_\Theta(X_Q|X_i) \leq \epsilon\}$ to be all possible Markov quilts of X_i with max-influence less than ϵ.
  if $G_i \neq \emptyset$ **do**
    for each $(X_Q, X_N, X_R) \in G_i$ **do**
      if $\Delta_{A\cap N} F/(\epsilon - e_\Theta(X_Q|X_i)) \leq b_i$ **then**
        Set $b_i = \Delta_{A\cap N} F/(\epsilon - e_\Theta(X_Q|X_i))$.
Sample $Z \sim \text{Lap}(\max_{i\in C} b_i)$.
Return $F(X) + Z$

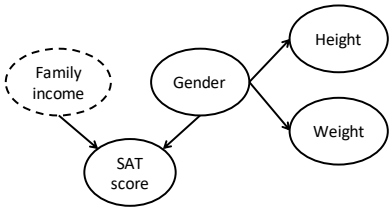

Figure 1: Bayesian Network of five attributes where income is a sensitive attribute.

If the query F is about attributes that are all in the "remote" set X_{R_i} and the max-influence on the corresponding Markov quilt is less than the privacy parameter ϵ, then $\Delta_{A\cap N} F$ is simply 0 and the mechanism will not add noise to the query answer. We note that the original Markov Quilt Mechanism was designed for protecting individual privacy, or one attribute record where the dataset only consists of one person's data. To extend the idea to protect dataset-level attribute information, non-trivial new data model and the corresponding max-influence definition is required.

Theorem 3. *The Attribute-Private Markov Quilt Mechanism APMQM$(X, F, A, C, \{S, \boldsymbol{Q}, \Theta\}, \epsilon)$ in Algorithm 2 is $(\epsilon, 0)$-distributional attribute private.*

The proof of this theorem appears in Appendix D. As before, the accuracy follows immediately from the tail bound on the noise term based on the Laplace distribution parameter. The accuracy depends on $A \cap N$ and $e_\Theta(X_Q|X_i)$, which measures the correlations between the sensitive attributes and the queried attributes.

Theorem 4. *The Attribute-Private Markov Quilt Mechanism APMQM$(X, F, A, C, \{S, \boldsymbol{Q}, \Theta\}, \epsilon)$ in Algorithm 2 is (α, β)-accurate for any $\beta > 0$ and*

$$\alpha = \max\{\max_{i\in C} \min_{(X_Q, X_N, X_R)\in G_i} \Delta_{A\cap N} F/(\epsilon - e_\Theta(X_Q|X_i)), \Delta_A F/\epsilon\} \log(\frac{1}{2\beta}),$$

where $G_i := \{(X_Q, X_N, X_R) : e_\Theta(X_Q|X_i) \leq \epsilon\}$ is the set of all possible Markov quilts of X_i with max-influence less than ϵ.

Example 1. *Consider the setting from Section 3 where a school is interested in releasing statistics about their students, while now wishing to protect the distribution of family income (instead of gender) among their students. The set of secrets is $S = \{s_a^i := \mathbb{1}[\phi_i \in \Phi_a^i] : \Phi_a^i \subset \Phi^i\}$, where ϕ_i denotes the distribution parameter of family income, and Φ_a^i can be a range or a particular value in Φ^i. Here, a dataset consists of students' SAT scores X_s, heights X_h, weights X_w, gender X_gndr, and their family income X_i, where these variables form a Bayesian network as in Figure 1. The school wishes to release the number of students that are taller than 66 inches, while protecting the distribution of family income of their students with privacy parameter ϵ. In this case, $C = \{\text{i}\}$, $A = \{\text{h}\}$ and $F(X) = \sum_{j=1}^n \mathbb{1}[X_\text{h}^j > 66]$. Consider a Markov quilt for X_i: $Q = \{\text{gndr}\}, N = \{\text{i}, \text{s}\}, R = \{\text{h}, \text{w}\}$. Then $A \cap N = \emptyset$, so we can safely release $F(X) = \sum_{i=1}^n \mathbb{1}[X_\text{h}^j > 66]$ without additional noise.*

Next consider the case when the school wishes to release the number of students that are taller than 66 inches and have SAT score > 1300. Then, $F(X) = \sum_{j=1}^n \mathbb{1}[(X_\text{h}^j > 66) \wedge (X_\text{s}^j > 1300)]$ and $A = \{\text{h}, \text{s}\}$. In this case we can still use the same Markov quilt as before, but now $A \cap N = \{\text{s}\}$. The mechanism will add Laplace noise scaled with $\Delta_{\{\text{s}\}} F/(\epsilon - e_\Theta(X_\text{gndr}|X_\text{i}))$.

It is instructive to contrast the above mechanism to the Markov Quilt Mechanism of [24], presented fully in Appendix A. The most important difference is that the mechanism in [24] was not designed to guarantee attribute privacy. It provides privacy of the values X_i^j but does not protect the distribution from which X_i^j is generated. This difference in high-level goals leads to three key technical differences. Firstly, the definition of max-influence in [24] measures influence of a *variable value on values of other variables*. This is insufficient when one wants to protect distributional information, as X_i may take a range of values while still following a particular distribution (e.g., hiding the gender of an individual in a dataset vs. hiding the proportion of females to males in this dataset.) Secondly, while it is natural to consider L-Lipschitz functions to bound sensitivity when one value changes (as is done in [24]), this is not applicable to settings where the distribution of data changes, since this may change all values in a column. As a result, we do not restrict F in this way. Finally, the mechanisms themselves are different as [24] consider answering query F over all attributes of an individual. As a result, they need to consider sensitivity of a function to all the "nearby" attributes. In contrast, we only consider sensitivity of those "nearby" attributes that happen to be in the query (i.e., those in A).

ACKNOWLEDGMENTS

W.Z. was supported by a Computing Innovation Fellowship from the Computing Research Association (CRA) and the Computing Community Consortium (CCC). R.C. was supported in part by NSF grants CNS-1850187 and CNS-1942772 (CAREER), a JPMorgan Chase Faculty Research Award, and an Apple Privacy-Preserving Machine Learning Award. Most of this work was completed while W.Z. and R.C. were at Georgia Institute of Technology. This work was initiated while W.Z. and O.O. were at Microsoft.

REFERENCES

[1] Giuseppe Ateniese, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. 2015. Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers. *International Journal of Security and Networks* 10, 3 (Sept. 2015), 137–150.

[2] Raef Bassily, Adam Groce, Jonathan Katz, and Adam Smith. 2013. Coupled-Worlds Privacy: Exploiting Adversarial Uncertainty in Statistical Data Privacy. In *Symposium on Foundations of Computer Science (FOCS)*. 439–448. https://doi.org/10.1109/FOCS.2013.54

[3] Raghav Bhaskar, Abhishek Bhowmick, Vipul Goyal, Srivatsan Laxman, and Abhradeep Thakurta. 2011. Noiseless Database Privacy. In *Proceedings of the 17th International Conference on The Theory and Application of Cryptology and Information Security*. Springer-Verlag, Berlin, Heidelberg, 215–232. https://doi.org/10.1007/978-3-642-25385-0_12

[4] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT* '18)*. 77–91.

[5] Graham Cormode. 2011. Personal Privacy vs Population Privacy: Learning to Attack Anonymization. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. 1253–1261.

[6] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. 214–226.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography (TCC '06)*. 265–284.

[8] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407. https://doi.org/10.1561/0400000042

[9] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM Conference on Computer and Communications Security (CCS '15)*. 1322–1333.

[10] Karan Ganju, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov. 2018. Property Inference Attacks on Fully Connected Neural Networks Using Permutation Invariant Representations. In *Proceedings of the 25th ACM Conference on Computer and Communications Security (CCS '18)*. 619–633.

[11] Xi He, Ashwin Machanavajjhala, and Bolin Ding. 2014. Blowfish privacy: tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD '14)*. 1447–1458.

[12] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *Proceedings of the 5th International Conference on Learning Representations (ICLR '17)*.

[13] Yusuke Kawamoto and Takao Murakami. 2019. Local Distribution Obfuscation via Probability Coupling. In *Proceedings of the 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton '19)*. 718–725.

[14] Yusuke Kawamoto and Takao Murakami. 2019. Local Obfuscation Mechanisms for Hiding Probability Distributions. In *Proceedings of the 24th European Symposium on Research in Computer Security (ESORICS '19)*. 128–148.

[15] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)* 39, 1 (2014), 1–36.

[16] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR '14)*. http://arxiv.org/abs/1312.6114

[17] Erich L Lehmann and George Casella. 2006. *Theory of point estimation*. Springer Science & Business Media.

[18] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The Variational Fair Autoencoder. In *Proceedings of the 4th International Conference on Learning Representations (ICLR '16)*. http://arxiv.org/abs/1511.00830

[19] Ashwin Machanavajjhala and Johannes Gehrke. 2006. On the Efficiency of Checking Perfect Privacy. In *Symposium on Principles of Database Systems (PODS)* (Chicago, IL, USA) *(PODS '06)*. Association for Computing Machinery, New York, NY, USA, 163–172. https://doi.org/10.1145/1142351.1142375

[20] Luca Melis, Congzheng Song, Emiliano De Cristofaro, and Vitaly Shmatikov. 2019. Exploiting Unintended Feature Leakage in Collaborative Learning. In *Proceedings of the 40th IEEE Symposium on Security and Privacy (S&P '19)*.

[21] Jane F Reckelhoff. 2001. Gender differences in the regulation of blood pressure. *Hypertension* 37, 5 (2001), 1199–1208.

[22] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks Against Machine Learning Models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P '17)*.

[23] Congzheng Song and Vitaly Shmatikov. 2020. Overlearning Reveals Sensitive Attributes. In *Proceedings of the 8th International Conference on Learning Representations (ICLR '20)*. https://openreview.net/forum?id=SJeNz04tDS

[24] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. 2017. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM SIGMOD International Conference on Management of Data (SIGMOD '17)*. 1291–1306.

[25] Richard Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML '13)*. 325–333.

[26] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. 2020. Dataset-Level Attribute Leakage in Collaborative Learning. arXiv:2006.07267 [cs.LG]