SoK: Let the Privacy Games Begin! A Unified Treatment of Data Inference Privacy in Machine Learning

Ahmed Salem*‡, Giovanni Cherubin*, David Evans†, Boris Köpf*
Andrew Paverd*, Anshuman Suri†, Shruti Tople*, Santiago Zanella-Béguelin*‡
*Microsoft
{t-salem.ahmed, giovanni.cherubin, boris.koepf, andrew.paverd, shruti.tople, santiago}@microsoft.com
†University of Virginia
{evans, as9rw}@virginia.edu

***Abstract*—Deploying machine learning models in production may allow adversaries to infer sensitive information about training data. There is a vast literature analyzing different types of inference risks, ranging from membership inference to reconstruction attacks. Inspired by the success of games (i.e. probabilistic experiments) to study security properties in cryptography, some authors describe privacy inference risks in machine learning using a similar game-based style. However, adversary capabilities and goals are often stated in subtly different ways from one presentation to the other, which makes it hard to relate and compose results. In this paper, we present a game-based framework to systematize the body of knowledge on privacy inference risks in machine learning. We use this framework to (1) provide a unifying structure for definitions of inference risks, (2) formally establish known relations among definitions, and (3) to uncover hitherto unknown relations that would have been difficult to spot otherwise.**

***Index Terms*—privacy, machine learning, differential privacy, membership inference, attribute inference, property inference**

I. Introduction

Since the pioneering studies of attribute inference [22, 69] and membership inference [37, 56], research on the inference risks of deploying machine learning (ML) models has bloomed. There is a growing interest in understanding and mitigating the leakage of information about training data under various threat models that capture different adversarial capabilities (e.g., observing model outputs, model parameters, or transcripts of iterative optimization methods) and goals (e.g., membership inference [56], attribute inference [22, 69], property inference [23, 42, 58, 74], and data reconstruction [4, 11]).

An emerging trend in the literature is to capture threat models using *privacy games*. This originates from the seminal work of Wu et al. [69] on formalizing attribute inference. A privacy game is a probabilistic experiment where an *adversary* interacts with a *challenger*. The challenger drives the experiment, invoking the adversary to provide them with information and to allow them to make certain choices, possibly while interacting with oracles controlled by the challenger. The adversary eventually produces a guess for a confidential value. This experiment defines a probability space where the success of the adversary can be measured in terms of the probability of their guess being correct.

The use of games for privacy in ML is inspired by the well-established use of games to define and reason about security properties in cryptography. Cryptographic games are used to standardize and compare security definitions [25, 57], and to structure [6] and even mechanize proofs of security [5, 9]. In comparison, the use of privacy games in the ML literature is still in its infancy:

(1) there are no well-established standards for game-based definitions,

(2) relationships between different privacy games have only been partially explored, and

(3) games are rarely used as an integral part of proofs, despite being especially convenient for this task.

This has resulted in many game variants in the literature that attempt to formalize the same adversary goal but have subtle yet important differences. This fragmentation leads to confusion and hinders progress—for membership inference alone, we found variants that differ in details that can change their meaning and substantially alter results. To address this problem, we present the first systematization of knowledge about privacy inference risks in machine learning, going above and beyond the problem left open since 2016 by Wu et al. [69] of merely devising rigorous game-based definitions. Concretely,

- We break down the *anatomy* of game-based privacy definitions for ML systems into individual components: adversary's capabilities and goals, ways of choosing datasets and challenges, and measures of success (Section II).
- Based on this anatomy, we propose a *unified representation* of five fundamental privacy risks as games: membership inference, attribute inference, property inference, differential privacy distinguishability, and data reconstruction (Section III).
- Using the game-based framework, we *establish and rigorously prove relationships* between the above risks. Similarly to the study of *concrete security* in cryptography [7], we define a quantitative notion of *reduction* between privacy properties.

‡ Corresponding author

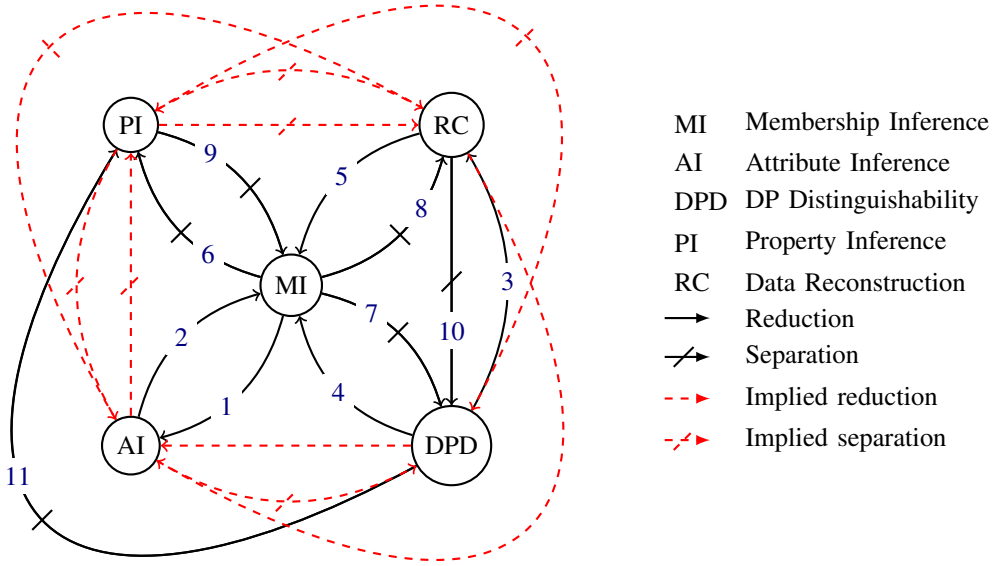

Fig. 1. Relations among adversary goals (under selected threat models). A solid arrow from node A to B means that security against A (i.e. a nontrivial advantage bound) implies security against B. A struck-through arrow from A to B means that security against A does not imply in general security against B; we show this separation with a construction that is secure against A but completely insecure against B. Dashed arrows are implied by solid arrows. Labels over solid arrows refer to the theorem showing the relationship. Some separations stem from differences in adversary capabilities, e.g. $\mathsf{MI} \not\to \mathsf{RC}$.

Using this notion, we prove a set of relations among the above five privacy risks. This allows us to establish, for every possible ordered pair of risks A, B, either a reduction showing that security against A implies security against B, or a separation result showing the impossibility of a generic reduction from A to B. Figure 1 summarizes the conclusions of this systematization effort for selected games.

- We present a *case study* (Section V), where we prove that a scenario described as a variant of membership inference in the literature can actually be decomposed into a combination of membership and property inference. Importantly, in this case we exploit *code-based* reductions, structured as a sequence of games; i.e., our arguments rely on transforming code with a formal semantics. This way of conducting proofs has seen great success in cryptography. However, before our work, it had not reached the same level of rigor when reasoning about privacy inference risks in ML.

Scope The focus of this SoK is to formalize and systematize game-based definitions that capture the risk of leaking information about the training data of ML models. We used the following methodology to identify existing game-based definitions from the literature: starting from the seminal works of Wu et al. [69] and Yeom et al. [70], we surveyed all peer-reviewed publications in Google Scholar as of August 2022 that cite either of these works. We examined these publications and collected all game-based definitions of attacks that aim to infer information about the training data of ML models. Our primary objective is to systematize games appearing in the literature. However, we also demonstrate the versatility of our framework by presenting new game-based definitions of attacks that have not been previously formulated as games.

Summary of contributions We propose a unifying game-based framework for formalizing privacy inference risks of training data in ML, which we use to systematize definitions from the literature and to establish relations between them. Our work aims to reduce ambiguity and increase rigor when reasoning and communicating about ML privacy, and gives a solid foundation to future research and decision-making.

II. Anatomy of a Privacy Game

Privacy games are parametrized by an adversary ($\mathcal{A}$) and a training pipeline that specifies the training algorithm ($\mathcal{T}$), data distribution ($\mathcal{D}$), and the size of the training dataset (n). A challenger simulates the ML system. The adversary uses their capabilities—defined by a threat model—to interact with the system and infer information about the training dataset.

Game 1: Membership Inference

Input: $\mathcal{A}, \mathcal{T}, n, \mathcal{D}$

1 $S \sim \mathcal{D}^n$ // sample n i.i.d. points from distribution $\mathcal{D}$
2 $b \sim \{0, 1\}$ // flip a fair coin
3 **if** $b = 0$ **then**
4 $\quad z \sim S$ // sample a challenge point uniformly from S
5 **else**
6 $\quad z \sim \mathcal{D}$ // sample a challenge point from $\mathcal{D}$
7 **end**
8 $\theta \leftarrow \mathcal{T}(S)$ // train a model θ
9 $\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \theta, z)$ // adversary guesses $b = \tilde{b}$

Game 1 formalizes the membership inference experiment of Yeom et al. [70], which we use as a running example. The challenger samples a training dataset S (line 1) and flips a fair coin b (line 2). Depending on the outcome, they either sample a challenge point z from the training dataset S, or from the data

distribution $\mathcal{D}$ (lines 3–7). We discuss alternatives for choosing training datasets and challenges in Section II-B. The challenger then trains a target model θ (line 8), and asks the adversary to make a guess $\tilde{b}$ for b (line 9). In this game, the adversary is given the training algorithm ($\mathcal{T}$), data distribution ($\mathcal{D}$), dataset size (n), target model (θ), and the challenge point (z). We discuss alternatives for adversary's capabilities in Section II-B and Section II-C. The success of the adversary in making a correct guess ($\tilde{b} = b$) is measured with respect to the baseline of a random guess. Any advantage over this baseline indicates leakage of membership information. We discuss other ways to quantify the adversary's success in Section II-D.

We now discuss in more detail the building blocks of games described above and highlight common choices.

A. Adversary Goals

We identify five adversary goals from the literature that enable an adversary to directly infer information about the training dataset of an ML model. We describe these goals informally below and formalize them as games in Section III.

Membership Inference (MI) The adversary aims to determine whether a specific *record* [56, 70] or *subject* [41, 59] (an entity who may contribute more than one record) was present in the training dataset of the target model. For example, a successful MI attack against a model trained on clinical records of patients with an infective disease can reveal that a target patient was infected.

Attribute Inference (AI) The adversary aims to use the model to infer unknown attributes of a record in the training dataset given partial information about the record [70]. A successful AI attack can result in the reconstruction of sensitive attributes of a target individual.

Property Inference (PI) The adversary aims to learn sensitive *statistical* properties of the target model's training distribution. For example, in a malware classifier, the training dataset may have been generated using a particular testing environment, and it may benefit the adversary to learn certain properties of this environment [23]. From an auditing perspective, property inference could be used to assess the training dataset for harms (e.g., under-representation) [74].

Differential Privacy Distinguishability (DPD) The adversary aims to determine which of a pair of adjacent datasets (e.g. differing in the data of one record) of their choosing was used to train the target model. This goal recasts differential privacy in a game-based setting by making the adversary explicit. This connection can be used to estimate the differential privacy budget of training pipelines [43, 46, 72].

Data Reconstruction (RC) The adversary aims to reconstruct samples from the training dataset of a target model [4, 10, 11]. A successful attack can partially reconstruct the training dataset, potentially violating confidentiality requirements.

Beyond training data inference Other adversary goals, such as model stealing [48, 62] and hyperparameter stealing [66] are beyond the scope of this SoK because they do not enable the adversary to directly infer information about the training data. However, the *effects* of these other goals are readily captured by our game-based analysis. For example, a successful model stealing attack that is used as a precursor to membership inference can be represented by changing the adversary access from black-box to white-box (Section II-C).

B. Selecting Challenges and Datasets

An important aspect of any privacy game is how the challenges and datasets are selected. In Game 1, the challenge point is a single record z; in other games, the challenge could comprise multiple points or even a data distribution. For the discussion below, we simplify the language by talking about a single challenge point. We discuss below three methods commonly used in the literature.

Randomly sampled The challenge is sampled from a distribution by the challenger as part of the game [31, 67, 70]. A randomly sampled challenge provides a measure of *average case* privacy. While average case privacy measures the risk for average users, the risk for outliers can be significantly higher.

Externally provided The challenge is provided as a parameter of the game [31, 41]. This may be used to measure privacy of specific points, i.e., it provides *individual case* privacy.

Adversarially chosen The challenge is selected by the adversary during the game [13, 43, 46]. Since the adversary can select the most advantageous challenge based on the information provided, this provides a measure of *worst case* privacy, i.e., measuring the risks for all users including outliers. For example, a strong membership inference adversary could choose a challenge that is an outlier w.r.t. the training data distribution, so that a target classification model is unlikely to classify it correctly unless it is included in the training dataset. This setting is usually considered when auditing a system to identify risks.

Additional considerations When the challenge is externally provided or adversarially chosen, the parameters of the game cannot completely determine a correct adversary guess. Otherwise, security statements that universally quantify over adversaries are void because the quantification includes adversaries with a hardcoded correct guess. This is similar to the difficulty of defining collision resistance of hash functions [51].

Selecting datasets The training dataset can also be selected using any of the three options above: it can be randomly sampled by the challenger, externally provided, or (partially) chosen by the adversary. The latter can be used to represent the case where the model has been trained on (poisoned) data contributed by potentially malicious users [42, 64].

C. Adversary Access

Depending on the scenario, the adversary may have different levels of access to the target model, training algorithm, training distribution, and training dataset. This allows the game to capture different threat models, which should ideally match the known or assumed capabilities of real-world adversaries. Most games assume one of two settings: *black-box* or *white-box* access.

Black-box In this scenario, the adversary only has query access to the target model (e.g., a cloud-hosted model with an

inference API) [12]. To formalize this setting, we give the adversary access to the model through an oracle $\textbf{Oracle}\ \mathcal{O}^\theta(x)$: **return** $\theta(x)$. This allows the adversary to query the model θ on inputs of their choosing and observe the responses, but does not reveal internal workings of the model, such as its architecture or weights. Depending on the scenario, the oracle can return a confidence for each label, or only the highest-confidence label [17, 38]. The latter setting matches inference APIs that do not reveal confidence values, like some email spam classifiers or auto-completion systems. Additionally, the oracle can be instrumented to post-process responses, or to only emit responses for queries satisfying a (stateful) predicate, e.g., to enforce a bound N on the number of allowed queries the challenge can initialize $q_0 = 0$ and provide

Oracle $\mathcal{O}_N^\theta(x)$
$q_\theta \leftarrow q_\theta + 1$
if $q_\theta \leq N$ **then return** $\arg\max \theta(x)$ **else return** $\perp$

White-box The white-box setting represents the strongest adversary, who has full direct access to the target model i.e., $\mathcal{A}(\theta, \ldots)$. This obviously provides the adversary with all the capabilities of the black-box setting, but also allows the adversary to inspect the internals of the model including its trained weights [36, 52]. For instance, a model deployed on clients' devices gives white-box access to malicious clients. Alternatively, a successful black-box model stealing attack would enable an adversary to operate in a white-box setting.

Grey-box In between the black-box and white-box settings, there is a range of *grey-box* threat models in which the adversary has more than black-box but less than full white-box access to the target model. For example, the adversary could know the architecture of a target model, some of its training hyperparameters, or the public model from which the model has been fine-tuned [53, 56]. Such extra information can be the output of a hyperparameter stealing attack [66].

Auxiliary information In addition to having access to the target model, an adversary may have auxiliary information that could be useful for certain attacks. For example, most MI attacks assume the adversary has access to auxiliary data distributed similarly to the target model's training data, e.g., for building shadow models. This is captured in games by giving the adversary the distribution from which the training data was sampled.

Resource constraints Most game-based formulations do not explicitly limit the resources available to an adversary, i.e., they consider information-theoretic adversaries. It could be important to consider resource-limited adversaries that can only issue a specific number of queries to an oracle, or can use a certain amount of memory, or are otherwise computationally bounded. Intuitively, limiting these resources can reduce the effectiveness of an attack. These limitations can be specified outside the game as constraints on the adversary, enforced by instrumenting the code of the game (as in Oracle $\mathcal{O}_N^\theta$ above), or incorporated into the measure of success.

D. Measuring Adversary Success

There are various ways of quantifying the adversary's success in games. We discuss commonly used metrics next.

Attack Success Rate: The *attack success rate* (ASR) measures the expected number of times the adversary succeeds (i.e., wins the game) over multiple runs. ASR is arguably the most intuitive and widespread metric for quantifying adversary success; for example, it matches the attacker's *accuracy* in membership inference.

However, the main drawback of ASR is that it does not take into account the baseline success probability for a given task. For example, if we evaluate an ML model's resilience to attribute inference, the prior distribution of that attribute will play a role in the adversary's success. For instance, if the attribute can only take one value, it is trivial for an adversary to achieve 100% ASR, but this will not be a meaningful measure. Similarly, the prior probability that an example belongs to the training set affects membership inference accuracy.

Ideally, the metric should quantify the success of an adversary relative to a suitable *baseline*. The baseline should represent the *a priori* adversary success rate; that is, it should quantify the adversary's success rate if they used only their prior knowledge and had no access to the model.

Adversary Advantage: The notion of *advantage* is a commonly used metric in cryptography, which relates an adversary's success rate to a baseline. This gives a better intuition of how much an adversary gains by having access to the model (in any of the forms defined in Section II-C). In general terms, suppose the adversary is trying to infer some variable p; this could be the membership of a data record or the value of a coin toss. If $\Pr[\mathcal{A} = p]$ is the adversary's success rate (probability to guess p correctly), and G is the baseline success rate, the advantage can be expressed as $\mathsf{Adv}(\mathcal{A}) = {}^{\Pr[\mathcal{A}=p]-G}/_{1-G}$. Assuming $\Pr[\mathcal{A} = p] \geq G$, this metric quantifies the adversary's advantage on a scale of $[0, 1]$ relative to the baseline G; 0 represents no advantage over the baseline and 1 is a perfect attack. When the secret information p is binary with a uniform prior, $G = {}^1/_2$. This leads to the familiar expression $\mathsf{Adv}(\mathcal{A}) = 2\Pr[\mathcal{A} = p] - 1$. Advantage is commonly used as a metric for ML privacy attacks. For example, Yeom et al. [70] define the MI advantage for an adversary $\mathcal{A}$ as follows:

$$\mathsf{Adv}_{\mathsf{MI}}(\mathcal{A}, \mathcal{T}, n, \mathcal{D}) = 2\Pr\Big[\mathsf{MI}(\mathcal{A}, \mathcal{T}, n, \mathcal{D}) : \tilde{b} = b\Big] - 1,$$

where MI is the membership inference experiment in Game 1, and $\Pr[G : E]$ denotes the probability of event E in the probability space defined by game G.

Providing an adequate baseline may be difficult because it may not be possible to accurately model the adversary's knowledge. This issue can often be bypassed by careful design of the game. For example, instead of asking the adversary to reconstruct an arbitrary attribute's value, the game can be designed such that the adversary must distinguish between two equally-likely values of the attribute.

Beyond advantage: Average case metrics such as ASR fail to capture inference risks for individuals or subpopulations. For example, a MI attack against a model may achieve roughly 50% accuracy (with a 50% baseline) on average across the population, yet the same attack may perform better when targeting specific individuals or subpopulations [13, 35]. Having raised similar concerns, Carlini et al. [12] suggest that an adversary should be considered successful if it reliably succeeds even on small number of cases. For instance, a MI attack that achieves a high true positive rate (TPR) at some low false positive rate (FPR) could be consequential even if it has low accuracy.

In this paper, we focus on advantage as a metric, since it has the following benefits: (1) it has an easy interpretation—it represents the gain of an adversary from having access to the system under scrutiny versus an adversary with only prior knowledge; (2) it is directly related to other metrics, such as ASR (which can be derived directly from it), true and false positive rates (e.g., [70]), and Differential Privacy [14, 31]; (3) if the attacker's challenge is binary (e.g., distinguishing between members and nonmembers), the advantage computed when assuming the two choices have a uniform prior gives a bound for any other prior [14]. Nevertheless, given a game formulation, one can consider other metrics of interest: e.g., area under the ROC curve (AUC-ROC), F1-score, and TPR at fixed FPR thresholds [12].

E. Consequences of Attacks

The anatomy we presented can be used to specify threat models and quantify the chances that an adversary successfully achieves their goal. However, the *consequences* of a successful attack depend less on the threat model but rather on the adversary's goal (Section II-A) and on the design of the ML system, e.g., the sensitivity of the training data. For example, the consequences of successful membership inference will be the same irrespective of whether it was performed in a black-box or white-box setting.

III. Game-based Formalization of Inference Risks

In this section we present privacy games for the five adversary goals introduced in Section II-A. We summarize the notation in Table I and the threat models considered in all games in Table II.

TABLE I
SUMMARY OF NOTATION

Notation	Description
$\mathcal{T}$	A stochastic training algorithm
$\mathcal{D}$	A distribution over examples
$\mathcal{D}^n$	Distribution of n independent examples from $\mathcal{D}$
$\mathcal{A}$, $\mathcal{A}'$	Adversary procedures sharing mutable state
$z \sim \mathcal{D}$	Draw an example z from $\mathcal{D}$
$S \sim \mathcal{D}^n$	Draw n examples S independently from $\mathcal{D}$
$b \sim \{0,1\}$	Sample a bit b uniformly
$b \sim 0 \oplus_p 1$	Sample 0 with probability p, 1 with probability $1-p$
$y \leftarrow \mathcal{P}(\vec{x})$	Call $\mathcal{P}$ with arguments $\vec{x}$ and assign result to y

A. Membership Inference

Membership inference aims to predict the participation of an entity in the training dataset of the model. The first (record-level) membership inference attack on supervised learning was proposed by Shokri et al. [56] against ML-based classifiers. Subsequent work has explored membership inference attacks with differing degrees of access to the model (e.g., white-box [36, 52] or label-only attacks [17, 38]), against different types of models (e.g., generative models [16, 27, 29], image segmentation [28], contrastive learning [40], recommender systems [73], and Graph Neural Networks (GNN) [68]), and under entirely different threat models [30, 53, 55].

We present MI variants that have been formalized as games. We divide the games into two categories depending on whether they focus on a single record (*record-level*) or a *user* represented by a collection of records (*user-level*).

Record-level Membership Inference: The most common interpretation of record-level membership inference is given by the game introduced by Yeom et al. [70], which we presented as Game 1 in Section II. Game 2 below presents a semantically equivalent reformulation MI. The reader can verify that b, θ, z_0 are distributed identically to b, θ, z in Game 1 and thus the joint distribution of $b, \tilde{b}$ is the same in both games. This game considers an adversary with white-box access to the model—they have the model at their disposal and can query it freely, analyze its architecture and parameters, and observe its dynamic behavior. Since the training dataset and the challenge z_0 are sampled from $\mathcal{D}$, this game measures average case MI resilience.

Game 2: MI | $\mathsf{MI}^{\mathsf{skew}}$ | $\mathsf{MI}^{\mathsf{Adv}}$

Input: $\mathcal{T}, \mathcal{D}, n,$ p, $\mathcal{A}', \mathcal{A}$

1 $S \sim \mathcal{D}^{n-1}$
2 $b \sim \{0,1\}$ | $0 \oplus_p 1$ | $\{0,1\}$
3 $z_0 \sim \mathcal{D}$ | $\mathcal{D}$ | $\mathcal{A}'(\mathcal{T}, \mathcal{D}, n)$
4 $z_1 \sim \mathcal{D}$
5 $\theta \leftarrow \mathcal{T}(S \cup \{z_b\})$
6 $\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n,$ p $, \theta, z_0)$

Several variants of the basic MI game have been considered in the literature; some are semantically equivalent (e.g., [31, 35]) whilst others alter its semantics. We next systematize these latter variants using the anatomy presented in Section II.

Jayaraman et al. [34] consider game $\mathsf{MI}^{\mathsf{skew}}$ which generalizes MI by introducing a parameter p representing the prior membership probability (Game 2, line 2). The original MI game assumes a balanced prior and is recovered as a special case when $p = 1/2$.

Chang and Shokri [13] consider game $\mathsf{MI}^{\mathsf{Adv}}$ in Game 2 which strengthens the adversary by allowing them to select the challenge point (line 3). This game measures worst case MI resilience for an average dataset, i.e., resilience against this variant protects all records—even outliers—against MI.

See SMI in Game 10 for an even stronger attack where S is adversarially chosen.

Carlini et al. [12] consider game $\mathsf{MI}^{\mathsf{BB}}$ which differs in two aspects from MI. Firstly, it assumes a black-box adversary who is given only inference access to the model through an oracle, **Oracle** $\mathcal{O}^\theta(x)$: **return** $\theta(x)$ (modifying line 9 in Game 1). This is appropriate when the target model is hosted in the cloud or in a trusted execution environment that ensures its confidentiality. Secondly, rather than sampling the challenge point from $\mathcal{D}$ when $b = 1$, the challenger samples it from $\mathcal{D} \setminus S$ (modifying line 6 in Game 1), thus excluding the case where the challenge happens to be in S by chance. This is in contrast to game MI, where nonmembers are sampled from the complete distribution and *may* be contained in S. While doing this seems intuitive, Yeom et al. [70, p.41] note that it is problematic since an adversary could gain advantage not through access to the model but rather by analyzing $\mathcal{D}$ to infer which points are more likely to have been sampled into S. For instance, consider a distribution $\mathcal{D}$ with support $\{x_0, \ldots, x_m\}$ that assigns probability $1/2$ to x_0 and $1/2m$ to each of $x_1, \ldots, x_m$. An adversary that ignores θ and guesses $\tilde{b} = 0$ if and only if $z = x_0$ has advantage greater than $1/2 - 1/2^n$.

Tramèr et al. [64] introduce a generic privacy game where the goal of the adversary is to guess which point from a universe $\mathcal{U}$ has been included in the training dataset of the target model. They present variants with ($\mathsf{MI}^{\mathsf{Pois}}$) and without ($\mathsf{MI}^{\mathsf{Diff}}$) poisoning, shown in Game 3. $\mathsf{MI}^{\mathsf{Pois}}$ lets the adversary statically poison part of the training dataset (Section II-B). By considering $\mathcal{U} = \{\hat{z}, \perp\}$, where $\perp$ indicates the absence of an example, the generic game can represent a black-box membership inference attack for a fixed externally provided target example $\hat{z}$. Compared to variants of membership inference discussed previously, this results in training datasets of different sizes depending on the outcome of sampling the challenge z: e.g. in $\mathsf{MI}^{\mathsf{Diff}}$ the model may be trained on $S \cup \{\hat{z}\}$ or just on S. This usually does not make a significant difference as training datasets are large and models do not leak the size of their training dataset. As in $\mathsf{MI}^{\mathsf{BB}}$, values in S are excluded when sampling z, which leads to similar problems.

Game 3: $\mathsf{MI}^{\mathsf{Diff}}$ $\boxed{\mathsf{MI}^{\mathsf{Pois}}}$

Input: $\mathcal{T}, \mathcal{D}, \mathcal{U}, n, \mathcal{A}, \boxed{\mathcal{A}', n'}$
$S \sim \mathcal{D}^n$
$z \sim \mathcal{U} \setminus S$
$\boxed{S' \leftarrow \mathcal{A}'(\mathcal{T}, \mathcal{D}, \mathcal{U}, n')}$ // $|S'| = n'$
$\theta \leftarrow \mathcal{T}(S \cup \{z\} \ \boxed{\cup S'})$
$\tilde{z} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, \mathcal{U}, n, \mathcal{O}^\theta(\cdot), \boxed{S'})$
Oracle $\mathcal{O}^\theta(x)$: **return** $\theta(x)$

Other variants Humphries et al. [31] sample the training dataset and challenge point from different distributions (Game 11); we use this variant as the basis for our case study in Section V. Tang et al. [61] define single-query membership inference games where the adversary is only given the output of the trained model on the challenge point, but where the adversary selects the set of examples from where the training dataset is subsampled (see Section A in the Appendix). Gao et al. [24] consider *deletion inference*, a variant of membership inference in the setting of *machine unlearning*, where the adversary is given access to a model before and after one of two examples is deleted and is asked to guess which example was deleted.

User-level Membership Inference: Privacy laws such as GDPR require generalizing the goal of MI. Instead of focusing on a single record, the interest is now the complete data of an individual. For instance, an auditor would be interested in learning if a user's data—usually modeled as a collection of records—was used to train a target model. User-level membership inference was introduced to model such scenarios. Mahloujifar et al. [41] formalize user-level MI as in Game 4. They consider a meta-distribution $\mathcal{D}$ from where m user distributions are sampled. The adversary targets a particular user contributing a dataset S^*. This game presents the adversary with a task easier than Game 1 since they must infer whether an entire group of records is within the training dataset, i.e., it measures *group* privacy.

Game 4: $\mathsf{MI}^{\mathsf{User}}$

Input: $\mathcal{T}, \mathcal{D}, n, \mathcal{A}, S^*, m$
$b \sim \{0, 1\}$
$\mathcal{D}_1, \ldots, \mathcal{D}_m \sim \mathcal{D}$
for $i = 1, \ldots, m-1$ **do**
  $S_i \leftarrow \mathcal{D}_i^n$
end
if $b = 0$ **then**
  $S_m = S^*$
else
  $S_m \leftarrow \mathcal{D}_m^n$
end
$\theta \leftarrow \mathcal{T}(\bigcup_{i=1}^m S_i)$
$\tilde{b} \leftarrow \mathcal{A}\left(\mathcal{T}, \mathcal{D}, n, \mathcal{O}^\theta(\cdot), S^*, m\right)$
Oracle $\mathcal{O}^\theta(x)$: **return** $\theta(x)$

B. Attribute Inference

In attribute inference (AI) attacks, the adversary aims to infer a sensitive attribute of a target record. Wu et al. [69] were the first to formalize AI, confusingly under the name of *model inversion*. We follow here the more general formalization given by Yeom et al. [70] shown in Game 5. Recently the scope of AI expanded to other settings [33, 75].

In the AI game, $\varphi(z)$ denotes the adversary's knowledge about the challenge z, and π a function that extracts the information targeted by the attack, e.g., if t represents the target sensitive attributes, then $\pi(z) = t$. The experiment is similar to the basic membership inference experiment (Game 1) except for the information that the adversary is given and the winning condition. The adversary is given $\varphi(z)$ and aims to infer $\pi(z)$.

Game 5: AI Inv

Input: $\mathcal{T}, \mathcal{D}, n, \mathcal{A}, \varphi, \pi$
$S \sim \mathcal{D}^n$
$b \sim \{0,1\}$
if $b = 0$ **then**
  $z \sim S$ $\mathcal{D}$
else
  $z \sim \mathcal{D}$
end
$\theta \leftarrow \mathcal{T}(S)$
$\tilde{a} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \theta, \varphi(z))$

The adversary wins if it correctly predicts these attributes, i.e., $\tilde{a} = \pi(z)$. Training data poisoning can be considered by including adversarially chosen data when training the target model as done for MI in Game 3 ($\mathsf{MI}^{\mathsf{Pois}}$), an instance of the generic game of Tramèr et al. [64].

Model inversion Another adversary goal with a similar aim to AI is model inversion [67]. Model inversion attacks were introduced by Fredrikson et al. [22] and subsequently formalized by Wang et al. [67] (Inv in Game 5). The difference between attribute inference and model inversion according to Wang et al. [67] is in how the challenge is sampled: in AI it is sampled from the training dataset, while in Inv it is sampled from the distribution $\mathcal{D}$. While AI measures privacy risk for members of a model's training dataset, model inversion measures the privacy loss of publishing the model for members of the underlying population. Whether this is considered a privacy risk is up to debate: a successful attack may lead to the adversary learning information from records that are not part of the training dataset or that do not even exist. Model owners concerned only with the privacy of the training dataset would use the AI game, whilst those concerned about population privacy would prefer Inv.

C. Reconstruction

Reconstruction attacks aim to recover entire examples in the training dataset of a model. Reconstruction has been studied in various settings, including Graph Neural Networks [75], image classification [54], and text generation [10, 11, 71]. A distilled scenario, where the adversary learns the training data of the target model except for a target example was first formalized by Balle et al. [4] as experiment RC in Game 6.

Game 6: RC $\mathsf{RC}^{\mathsf{Ran}}$

Input: S, $\mathcal{D}, n$, $\pi, \mathcal{T}, \mathcal{A}$
$S \sim \mathcal{D}^{n-1}$
$z \sim \pi$
$\theta \leftarrow \mathcal{T}(S \cup \{z\})$
$\tilde{z} \leftarrow \mathcal{A}(\mathcal{T}, \theta, \mathcal{D}, n, S)$

Reconstruction robustness is parametrized by bounds on the error and success probability and defined as follows.

Definition 1 (Balle et al. [4], Definition 2)**.** *A training pipeline is (η, γ)-reconstruction robust with respect to a prior π and reconstruction loss ℓ if for any dataset S and any reconstruction adversary $\mathcal{A}$,*

$$\Pr[\mathsf{RC} : \ell(z, \tilde{z}) \leq \eta] \leq \gamma$$

The adversary is given the model θ, training algorithm $\mathcal{T}$, and the training dataset S except for one point z which they need to reconstruct. Game $\mathsf{RC}^{\mathsf{Ran}}$ models how other points in the training dataset are sampled, instead of considering a fixed dataset S. The advantage of an adversary $\mathcal{A}$ against RC w.r.t. a baseline that ignores θ and just samples $\tilde{z}$ from $\mathcal{D}$ is

$$\mathsf{Adv}_{\mathsf{RC}}(\mathcal{A}) = \Pr[\mathsf{RC} : \tilde{z} = z] - \Pr[z, \tilde{z} \sim \mathcal{D} : \tilde{z} = z]$$

Alternatively, one can consider the baseline success of an adversary that picks $\tilde{z}$ according to π,

$$\sup_{\tilde{z} \in \mathrm{supp}(\pi)} \Pr[z \sim \pi : \ell(z, \tilde{z}) \leq \eta] \tag{1}$$

Both games can be adapted to consider a poisoning-capable adversary as demonstrated in Game 3.

Reconstruction in language models Recent work focused on large language models and evaluated reconstruction attacks against them. Attacks can be categorized as untargeted [11] or targeted [10]. Untargeted attacks aim to reconstruct *any* training data from the generative model, whilst targeted attacks aim to reconstruct *specific* training data records, which may have been inserted as canaries during training. To demonstrate the flexibility of privacy games, we formalize an example from each category, as shown in Game 7.

We formalize a black-box untargeted data reconstruction attack by Carlini et al. [11] tailored to large generative language models as $\mathsf{RC}^{\mathsf{Untarg}}$. The authors measure the success of an attack by its true positive rate or recall, that is, the fraction of examples in $\tilde{S}$ that are in the training dataset S.

We formalize a black-box targeted reconstruction attack by Carlini et al. [10] as $\mathsf{RC}^{\mathsf{Targ}}$. The authors insert a *canary* multiple times into the training data as a way to measure unintended memorization in generative models. Canaries are specified by a format sequence $s[\cdot]$ that fixes some tokens and leaves *holes* to be filled with secrets sampled from a randomness space $\mathcal{R}$. For example, $s =$ "the PIN is ⌀⌀⌀⌀" with $\mathcal{R}$ being the space of 4-digit decimal numbers. Carlini et al. [10] measure the success of targeted canary reconstruction as the reduction in the guessing entropy of secrets in canaries given the model.

Selecting a game Game RC is appropriate when evaluating the worst case risk of reconstructing an example in the training dataset. It conservatively considers an informed adversary that knows all examples but the target, and incorporates the adversary's background knowledge in a prior. Game $\mathsf{RC}^{\mathsf{Ran}}$ considers an equally informed adversary, but averages the reconstruction risk over the choice of other training examples. Game $\mathsf{RC}^{\mathsf{Untarg}}$ represents a more realistic threat model and

Game 7: $\mathsf{RC}^{\mathsf{Untarg}}$ $\boxed{\mathsf{RC}^{\mathsf{Targ}}}$

Input: $\mathcal{T}, \mathcal{D}, n, \mathcal{A}, \boxed{\mathcal{R}, s, m}$
$S \sim \mathcal{D}^n$
$\boxed{r \sim \mathcal{R}}$
$\theta \leftarrow \mathcal{T}(S \ \boxed{\cup\{s[r]\}^m}\ )$
$\widetilde{S} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \mathcal{O}^\theta(\cdot), \boxed{\mathcal{R}, s})$
Oracle $\mathcal{O}^\theta(\mathrm{x})$: **return** $\theta(x)$

should be chosen when evaluating the risk of indiscriminately reconstructing training data, while $\mathsf{RC}^{\mathsf{Targ}}$ is appropriate for auditing the risk of extracting data following certain patterns.

Other variants Similar to MI, reconstruction attacks have been adapted to the machine unlearning setting. Gao et al. [24] consider *deletion reconstruction*, where an adversary is given access to a model before and after a random training example is deleted and is asked to reconstruct it.

D. Distribution Inference

Distribution inference attacks do not focus on specific data records, but instead aim at inferring properties about the training data distribution. We next describe two variants of distribution inference. The first is property inference, e.g., where the adversary is interested in learning about the prevalence of specific sensitive attributes in the training data, such as sex or ethnicity. The second is subject-level distribution inference, where the training data is sampled from a mixture of distributions, each corresponding to a subject that may participate in training. The adversary's goal is to infer whether a subject has participated knowing the subject's data distribution rather than concrete samples like in game $\mathsf{MI}^{\mathsf{User}}$ in Game 4.

Property Inference: Property inference attacks were first proposed by Ganju et al. [23] in the white-box setting and by Zhang et al. [74] in the black-box setting. Zhou et al. [76] showed them to be effective against generative models and GANs specifically. Suri and Evans [58] formalized property inference attacks as PI in Game 8, parametrized by two functions $\mathcal{G}_0, \mathcal{G}_1$ that transform an underlying distribution.

Game 8: $\boxed{\mathsf{PI}}$ $\boxed{\mathsf{PI}^{\mathsf{Gen}}}$

Input: $\boxed{\mathcal{D}, \mathcal{G}_0, \mathcal{G}_1}$ $\boxed{\mathcal{D}_0, \mathcal{D}_1}$, $n, \mathcal{T}, \mathcal{A}$
$b \sim \{0,1\}$
$S \sim \boxed{\mathcal{G}_b(\mathcal{D})^n}$ $\boxed{\mathcal{D}_b^n}$
$\theta \leftarrow \mathcal{T}(S)$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \boxed{\mathcal{D}, \mathcal{G}_0, \mathcal{G}_1}$ $\boxed{\mathcal{D}_0, \mathcal{D}_1}, n, \theta)$

$\mathsf{PI}^{\mathsf{Gen}}$ is an equivalent formulation parametrized by two distributions corresponding to the application of $\mathcal{G}_0$, $\mathcal{G}_1$ to the base distribution $\mathcal{D}$ in PI. Hartmann et al. [26] generalize this to more than two distributions.

Similarly to MI and AI, poisoning can be modelled as in Game 3 by letting the adversary choose part of the training dataset of the target model. Mahloujifar et al. [42] and Chaudhari et al. [15] show that poisoning increases inference risk by injecting data to maximize leakage of properties of the training dataset. For instance, in multi-party learning, a malicious participant may contribute poisoned data crafted to amplify property leakage of data from other participants.

Subject-level Distribution Inference: Subject-level distribution inference broadens the scope of user-level membership inference by not assuming access to the user's exact data that may have been used to train a model. Instead, it only requires the adversary know the distribution from which the target user's data is sampled. Suri et al. [59] present subject membership inference as a special case of distribution inference. We similarly formalize subject-level inference in Game 9.

Game 9: $\mathsf{MI}^{\mathsf{Subj}}$

Input: $\mathcal{T}, \mathcal{D}, \mathcal{D}_*, n, m, \mathcal{A}$
$b \sim \{0,1\}$
$\mathcal{D}_1, \ldots, \mathcal{D}_m \sim \mathcal{D}$
for $i = 1, \ldots, m-1$ **do**
  $S_i \sim \mathcal{D}_i^n$
end
if $b = 0$ **then**
  $S_m \sim \mathcal{D}_*^n$
else
  $S_m \sim \mathcal{D}_m^n$
end
$\theta \leftarrow \mathcal{T}(\bigcup_{i=1}^m S_i)$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, \mathcal{D}_*, n, m, \theta)$

The training data distribution is structured as a mixture of distributions corresponding to a set of subjects. This is a property inference attack because the adversary seeks to infer which of two distributions the training data is sampled from. However, conceptually, the adversary's goal is to infer membership of a subject's data since the only difference between the two distributions is the presence of the target subject in the mixture.

A successful subject-level distribution inference attack can identify if a user's data was used to train the target model without knowing which exact examples were used; i.e., with access to only the user's data distribution and not the sampled dataset as in Game 4.

E. Differential Privacy Distinguishability

Differential Privacy Distinguishability (DPD) formalizes the threat model underlying the definition of DP, where the adversary aims to distinguish between models trained on adjacent datasets. We formalize as game DPD in Game 10 the variant corresponding to the *substitute one* adjacency relation, where two datasets are adjacent if one can be obtained from the other by substituting a single record. The DPD game represents a worst-case variant of the membership inference game

MI where the training data and challenges are adversarially chosen.

Prior work used DP distinguishing attacks to statistically estimate or audit the privacy of training pipelines [32, 46, 63, 72]. Marathe and Kanani [44] define subject-level differential privacy by considering datasets as adjacent when they differ in the data of a user, which can be seen as a counterpart to user-level membership inference. Humphries et al. [31] and Balle et al. [4] discuss strong membership inference, a threat model in between DPD and MI. In this game, formalized as SMI in Game 10, the adversary knows but does not choose the two adjacent datasets. As mentioned in Section II-B this narrows the scope of the measured privacy, e.g., from worst to individual case privacy.

Game 10: DPD SMI

Input: $\mathcal{T}, \mathcal{A}, \boxed{\mathcal{A}', n}, S, z_0, z_1$

$S, z_0, z_1 \leftarrow \mathcal{A}'(\mathcal{T}, n)$ // $|S| = n - 1$

$b \sim \{0, 1\}$

$\theta \leftarrow \mathcal{T}(S \cup \{z_b\})$

$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \theta, S, z_0, z_1)$

IV. Relations and Proofs

In this section we establish relationships between privacy games. To this end, we define a notion of *reduction* and use it to translate attacks and guarantees between the five fundamental games from the previous section, or show that no generic connection can exist.

A. Reductions for Privacy Games

Inspired by notions of reduction from complexity theory and cryptography [1], we introduce reductions between privacy games as a means of comparing the various inference risks. Whilst reductions in cryptography are traditionally based on asymptotic behavior governed by a security parameter, the reductions we define here are closer to those used in *concrete security* proofs, in that the constants underlying the loss incurred in the reduction are made explicit.

Definition 2. *We say that game G_1 is reducible to game G_2 if there is a constant $c > 0$ such that, for any adversary $\mathcal{A}$ against G_2, there exists an adversary $\mathcal{B}$ against G_1 such that*

$$\mathsf{Adv}_{G_1}(\mathcal{B}) \geq c \cdot \mathsf{Adv}_{G_2}(\mathcal{A})$$

We denote this using the shorthand $G_1 \preceq_c G_2$ and sometimes drop the constant c.

The intuition behind the shorthand is that game G_1 is at most as hard to win as G_2—modulo the constant c. This intuition holds for c around or larger than 1. For $c \ll 1$, however, the lower bound on $\mathsf{Adv}_{G_1}(\mathcal{B})$ can get close to 0, in which case the intuition may be misleading.

Resilience to attacks Reductions between privacy games imply that attacks against one game translate into attacks against the other. An equivalent reading is the contrapositive, that resilience against attacks in one game implies resilience against attacks in the other.

Definition 3. *A game G is* p-resilient *if for all adversaries $\mathcal{A}$ against G,*

$$\mathsf{Adv}_G(\mathcal{A}) < p$$

Proposition 1. *If $G_1 \preceq_c G_2$ and G_1 is p-resilient then G_2 is p/c-resilient.*

Proof. By contradiction: If there is an attack on G_2 with advantage more than p/c, then there is one on G_1 with advantage more than p. □

Proofs of resilience are rare in the literature. Prime examples are results that establish upper bounds on the advantage of a DP distinguisher when the model is trained with differential privacy [31, 70]. The tightest known bound is given in the following proposition.

Proposition 2 (Humphries et al. [31, Theorem 3.1])**.** *Let $\mathcal{T}$ be an (ε, δ)-differentially private training algorithm. Then*

$$\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}$$

Therefore, any game the DP distinguisher inference game can be reduced to (see Figure 1 for an overview) inherits the security benefits of training with differential privacy via Propositions 1 and 2.

Separation Results No reductions exist between several games. For them, we show separation results of the form $G_1 \not\preceq G_2$. We establish such results by showing that there is an instance of G_1 that is resilient to attacks whereas its G_2 counterpart is not, and use Proposition 1 to conclude that no reduction exists.

B. Overview of Relations between Games

Figure 1 shows the relations between the five fundamental privacy games. Each node in the figure and in the following theorems refers to the basic game-based definition of the corresponding inference risk, i.e., MI, AI, RC, DPD, and PI.

As expected, PI is fully disconnected: there exists a separation result between it and every other game. This can be attributed to the PI adversary's goal of learning properties of the training data distribution rather than about individual records as in the other games. RC and DPD have the strongest threat models, where the adversary controls the entire training dataset except for one example, and hence are unsurprisingly the hardest to reduce from other games. Finally, MI and AI are reducible to each other and their relatively weak threat models make both RC and DPD reducible to them. For this reason, we use the MI game as the anchor for our proofs. We next present results for a set of edges (solid lines) in Figure 1 that imply all other relations. We defer the proofs to the Appendix.

TABLE II
AN OVERVIEW OF DIFFERENT GAMES AND FEATURES OF THEIR CORRESPONDING THREAT MODELS.
✓ indicates the game has this feature, – indicates the game does not have this feature, × indicates that the feature is not applicable.

		Adversary Access		Challenge			Training Dataset			Adversary Interest		
Game	Definition	Black-box	White-box	Rand	Adv	Param	Rand	Adv	Param	Record	Subject	Distribution
Membership Inference												
MI	Game 2 [31, 35, 70]	–	✓	✓	–	–	✓	–	–	✓	–	–
$\mathsf{MI}^{\mathsf{skew}}$	Game 2 [34]	–	✓	✓	–	–	✓	–	–	✓	–	–
$\mathsf{MI}^{\mathsf{BB}}$	Game 2 [12]	✓	–	✓	–	–	✓	–	–	✓	–	–
$\mathsf{MI}^{\mathsf{Adv}}$	Game 2 [13]	–	✓	–	✓	–	✓	–	–	✓	–	–
$\mathsf{MI}^{\mathsf{Diff}}$	Game 3 [64]	✓	–	✓	–	–	✓	–	–	✓	–	–
$\mathsf{MI}^{\mathsf{Pois}}$	Game 3 [64]	✓	–	✓	–	–	✓	✓	–	✓	–	–
$\mathsf{MI}^{\mathsf{User}}$	Game 4 [41]	✓	–	–	–	✓	✓	–	–	–	✓	–
MM	Game 11 [31]	–	✓	✓	–	–	✓	–	–	✓	–	✓
$\mathsf{MI}^{\mathsf{SQ}}$	Game 17 [61]	✓	–	✓	–	–	–	✓	–	✓	–	–
Attribute Inference and Model Inversion												
AI	Game 5 [70]	–	✓	✓	–	–	✓	–	–	✓	–	–
Inv	Game 5 [67]	–	✓	✓	–	–	✓	–	–	✓	–	–
Data Reconstruction												
RC	Game 6 [4]	–	✓	✓	–	–	–	–	✓	✓	–	–
$\mathsf{RC}^{\mathsf{Untarg}}$	Game 7 [11]	✓	–	×	×	×	✓	–	–	✓	–	–
$\mathsf{RC}^{\mathsf{Targ}}$	Game 7 [10]	✓	–	✓	–	–	✓	–	–	✓	–	–
Distribution Inference												
PI	Game 8 [58]	–	✓	×	×	×	✓	–	–	–	–	✓
$\mathsf{MI}^{\mathsf{Subj}}$	Game 9 [59]	–	✓	✓	–	–	✓	–	–	–	✓	✓
Differential Privacy Distinguishability												
DPD	Game 10 [43, 46]	–	✓	–	✓	–	–	✓	–	✓	–	–
SMI	Game 10 [4, 31]	–	✓	–	–	✓	–	–	✓	✓	–	–

C. Reductions

Despite reductions in either direction, MI and AI are separable by constants in the reductions, with resilience against AI easier to achieve than resilience against MI. The following theorems proved by Yeom et al. [70] relate MI and AI.

Theorem 1 ($\mathsf{MI} \preceq_1 \mathsf{AI}$ [70, Theorem 6])**.** *For any adversary $\mathcal{A}_{\mathsf{AI}}$ against attribute inference, there exists an adversary $\mathcal{A}_{\mathsf{MI}}$ against membership inference such that*

$$\mathsf{Adv}_{\mathsf{MI}}(\mathcal{A}_{\mathsf{MI}}) = \mathsf{Adv}_{\mathsf{AI}}(\mathcal{A}_{\mathsf{AI}})$$

Theorem 2 ($\mathsf{AI} \preceq_{1/m} \mathsf{MI}$ [70, Theorem 7])**.** *Assume that for all $z \in \mathrm{supp}(\mathcal{D})$, $\varphi(z)$ and $\pi(z)$ uniquely determine z. For any adversary $\mathcal{A}_{\mathsf{MI}}$ against membership inference, there exists an adversary $\mathcal{A}_{\mathsf{AI}}$ against attribute inference such that*

$$\mathsf{Adv}_{\mathsf{AI}}(\mathcal{A}_{\mathsf{AI}}) = \frac{1}{m} \cdot \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A}_{\mathsf{MI}})$$

where m is the number of possible values for the target attribute $\pi(z)$.

Resilience against DPD implies resilience against all other attacks except PI. We present the necessary theorems below. The remaining reductions ($\mathsf{RC} \preceq \mathsf{AI}, \mathsf{DPD} \preceq \mathsf{AI}$) are implied by the ones we show.

Balle et al. [4, Theorem 3] show that training pipelines satisfying Rényi DP (and thus (ε, δ)-DP) enjoy resilience against reconstruction attacks. In contrast, a bound on $\mathsf{Adv}_{\mathsf{DPD}}$ does not imply a nontrivial bound on ε in (ε, δ). In fact, $\mathsf{Adv}_{\mathsf{DPD}} \leq \delta$ is equivalent to $(0, \delta)$-DP. Thus, we require an anti-concentration bound on the prior π and that reconstruction succeeds with probability at least ½ to reduce DPD to RC.

Theorem 3 ($\mathsf{DPD} \preceq \mathsf{RC}$)**.** *Let π be a prior over samples, S a dataset of $n-1$ samples, and ℓ a symmetric reconstruction loss satisfying the triangle inequality. Let $\mathcal{A}$ be an adversary against data reconstruction (RC) w.r.t. S and π that reconstructs its challenge within error η with probability $\gamma \geq 1/2$. Let*

$$\alpha = \inf_{z_0 \in \mathrm{supp}(\pi)} \Pr[z_1 \sim \pi : \ell(z_0, z_1) > 2\eta]$$

There exists a DP distinguisher $\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}$ such that

$$\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}) \geq 2\alpha\left(\gamma - \frac{1}{2}\right)$$

DP distinguishability can be reduced to membership inference. This is an example of a generic class of reductions: In both games the adversary has the same goal and their advantage is identically defined, but in game MI the adversary has strictly fewer capabilities than in DPD. Thus, any adversary against MI can be turned into a valid adversary against DPD with the same advantage. In general, a more informed/capable adversary, such as a DP distinguisher, can be used to build a reduction to games with a less informed/capable adversary.

Theorem 4 ($\mathsf{DPD} \preceq \mathsf{MI}$)**.** *For any adversary $\mathcal{A}_{\mathsf{MI}}$ against membership inference, there exists a DP distinguisher $\mathcal{A}_{\mathsf{DPD}}$ such that*

$$\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}_{\mathsf{DPD}}) = \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A}_{\mathsf{MI}})$$

Finally, we show that a membership inference attack can be turned into a reconstruction attack, with a constant depending on the size of the support of the training data distribution.

Theorem 5 (RC $\preceq_{1/|\operatorname{supp}(\mathcal{D})|}$ MI)**.** *For any membership inference adversary $\mathcal{A}$ against $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n)$ there exists a reconstruction adversary $\mathcal{B}$ against $\mathsf{RC}^{\mathsf{Ran}}(\mathcal{D}, n, \mathcal{D}, \mathcal{T})$ (i.e., with prior $\pi = \mathcal{D}$) such that*

$$\mathsf{Adv}_{\mathsf{RC}^{\mathsf{Ran}}}(\mathcal{B}) = \frac{1}{|\operatorname{supp}(\mathcal{D})|} \cdot \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A})$$

D. Separation Results

Theorem 6 (MI $\not\preceq$ PI)**.** *Resilience against membership inference does not imply resilience against property inference.*

Theorem 7 (MI $\not\preceq$ DPD)**.** *Resilience against membership inference does not imply resilience against DP distinguishability.*

Theorem 8 (MI $\not\preceq$ RC)**.** *Resilience against membership inference does not imply resilience against reconstruction.*

This last counterintuitive separation result stems from a discrepancy between adversary capabilities: The MI game is based on an average case scenario, while the reconstruction game assumes a more informed worst-case adversary. By considering a membership adversary matching the capabilities of the adversary in the RC game, we can build a reduction to data reconstruction. We show this in Theorem 13 in the Appendix, which reduces the strong membership inference game SMI (Game 10) to game RC.

Theorem 9 (PI $\not\preceq$ MI)**.** *Resilience against property inference does not imply resilience against membership inference.*

Theorem 10 (RC $\not\preceq$ DPD)**.** *Resilience against reconstruction does not imply resilience against DP distinguishability.*

Theorem 11 (DPD $\not\preceq$ PI)**.** *Resilience against DP distinguishability does not imply resilience against property inference.*

V. Case Study: Mixture Model Membership Inference

We present a case study where we showcase the expressive power and rigor of privacy games. In particular, we show that a novel variant of membership inference can be decomposed into a combination of membership and property inference. This complex relationship goes beyond the direct reductions presented in Section IV. In our proofs, we exploit *code-based* reductions structured as a sequence of games; i.e., our arguments rely on transforming code with a formal semantics.

The game we target is due to Humphries et al. [31], who use it to model membership inference attacks in the presence of dependencies in the training data. In their game (MM in Game 11), the training data follows a two-stage *mixture model*. Examples in the training dataset and the target example are chosen independently from two data distributions, $\mathcal{D}_k$ and $\mathcal{D}_{k'}$, which are chosen uniformly at random without replacement from K possible distributions $\boldsymbol{\mathcal{D}} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$.

Game 11: MM G_0

Input: $\mathcal{T}, \boldsymbol{\mathcal{D}}, n, \mathcal{A}$
$k \sim [K]$
$k' \sim [K] \setminus \{k\}$
$S \sim \mathcal{D}_k^n$
$\theta \leftarrow \mathcal{T}(S)$
$b \sim \{0, 1\}$
if $b = 0$ **then**
  $z \sim S$ $\quad z \sim \mathcal{D}_k$
else
  $z \sim \mathcal{D}_{k'}$
end
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \boldsymbol{\mathcal{D}}, n, \theta, z)$

Game 12: G_1

Input: $\mathcal{T}, \boldsymbol{\mathcal{D}}, n, \mathcal{A}$
$k \sim [K]$
$k' \sim [K] \setminus \{k\}$
$z \sim \mathcal{D}_k$
$b \sim \{0, 1\}$
if $b = 0$ **then**
  $S \sim \mathcal{D}_k^n$
else
  $S \sim \mathcal{D}_{k'}^n$
end
$\theta \leftarrow \mathcal{T}(S)$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \boldsymbol{\mathcal{D}}, n, \theta, z)$

We show that MM can be decomposed into a property inference goal (inferring the training data distribution) and a membership inference goal (inferring whether a target example has been sampled from the training data distribution $\mathcal{D}_k$ or from the training dataset S).

Theorem 12. *For any adversary $\mathcal{A}$ against* MM*, there exist adversaries $\mathcal{A}^i_{\mathsf{MI}}$ and $\mathcal{A}^{i,j}_{\mathsf{PI}}$ such that*

$$\mathsf{Adv}_{\mathsf{MM}}(\mathcal{A}) \leq \max_{i \in [K]} \mathsf{Adv}_{\mathsf{MI}_i}(\mathcal{A}^i_{\mathsf{MI}}) + \max_{i \neq j \in [K]} \mathsf{Adv}_{\mathsf{PI}_{i,j}}(\mathcal{A}^{i,j}_{\mathsf{PI}})$$

where MI_i is the membership inference game with training data distribution $\mathcal{D}_i$, and in $\mathsf{PI}_{i,j}$ the property to infer is whether the training data distribution is $\mathcal{D}_i$ or $\mathcal{D}_j$.

Proof. Let $\mathcal{A}$ be an adversary against MM. Consider G_0 shown alongside MM in Game 11. Its only difference w.r.t. MM is that when $b = 0$, the example z is freshly sampled from the training data distribution $\mathcal{D}_k$ rather than from the training dataset S. Conditioned on $b = 0, k = i$, distinguishing between games G_0 and MM is as difficult as winning a membership inference game. We show this using a black-box reduction: fixing $k = i$, we construct an adversary $\mathcal{A}^i_{\mathsf{MI}}$ that uses $\mathcal{A}$ as an oracle to guess the challenge bit b in game MI_i (see Game 13). $\mathcal{A}^i_{\mathsf{MI}}$ simply forwards its inputs $\mathcal{T}, n, \theta, z$ to $\mathcal{A}$, passing to it in addition the distribution set $\boldsymbol{\mathcal{D}}$.

Game 13: MI_i

Input: $\mathcal{T}, \mathcal{D}_i, n, \mathcal{A}$
$S \sim \mathcal{D}_i^n$
$\theta \leftarrow \mathcal{T}(S)$
$b \sim \{0,1\}$
if $b = 0$ **then** $z \sim S$ **else** $z \sim \mathcal{D}_i$
$\tilde{b} \leftarrow \mathcal{A}^i_{\mathsf{MI}}(\mathcal{T}, \mathcal{D}_i, n, \theta, z)$

Adversary 14: $\mathcal{A}^i_{\mathsf{MI}}$

Input: $\mathcal{T}, \mathcal{D}_i, n, \theta, z$
return $\mathcal{A}(\mathcal{T}, \boldsymbol{\mathcal{D}}, n, \theta, z)$

Game MM conditioned on $b = 0, k = i$ is equivalent to MI_i conditioned on $b = 0$. Likewise, game G_0 conditioned on $b = 0, k = i$ is equivalent to MI_i conditioned on $b = 1$. Hence,

$$\begin{aligned}\mathsf{Adv}_{\mathsf{MI}_i}(\mathcal{A}^i_{\mathsf{MI}}) &= \Pr\left[\mathsf{MI}_i : \neg\tilde{b} \mid \neg b\right] - \Pr\left[\mathsf{MI}_i : \neg\tilde{b} \mid b\right] \\ &= \Pr\left[\mathsf{MM} : \neg\tilde{b} \mid \neg b, k = i\right] - \Pr\left[G_0 : \neg\tilde{b} \mid \neg b, k = i\right] \quad (2)\end{aligned}$$

Game MM conditioned on $b = 1$ is equivalent to G_0 conditioned on $b = 1$, and so we have

$$\begin{aligned}\mathsf{Adv}_{\mathsf{MM}}(\mathcal{A}) &= \Pr\left[\mathsf{MM} : \neg\tilde{b} \mid \neg b\right] - \Pr\left[\mathsf{MM} : \neg\tilde{b} \mid b\right] \\ &= \frac{1}{K}\sum_{i=1}^{K} \Pr\left[\mathsf{MM} : \neg\tilde{b} \mid \neg b, k = i\right] - \Pr\left[\mathsf{MM} : \neg\tilde{b} \mid b, k = i\right] \\ &= \frac{1}{K}\sum_{i=1}^{K} \mathsf{Adv}_{\mathsf{MI}_i}(\mathcal{A}^i_{\mathsf{MI}}) + \Pr\left[G_0 : \neg\tilde{b} \mid \neg b\right] - \Pr\left[G_0 : \neg\tilde{b} \mid b\right] \quad (3)\end{aligned}$$

where the last equation follows from (2) and the fact that b and k are independent.

We reformulate G_0 as G_1 (see Game 12). To see why both formulations are equivalent, note that conditioned on $b = 0$, in both games S and z are sampled from the same distribution chosen uniformly from $\boldsymbol{\mathcal{D}}$, while conditioned on $b = 1$, S and z are sampled each from one of two distributions sampled without replacement from $\boldsymbol{\mathcal{D}}$. Since b is independently sampled in the same way, both games result in the same joint distribution of θ, z, b, and therefore $\tilde{b}, b$:

$$\Pr\left[G_0 : \neg\tilde{b} \mid \neg b\right] = \Pr\left[G_1 : \neg\tilde{b} \mid \neg b\right] \quad (4)$$

$$\Pr\left[G_0 : \neg\tilde{b} \mid b\right] = \Pr\left[G_1 : \neg\tilde{b} \mid b\right] \quad (5)$$

Next, we show using a black-box reduction that distinguishing between the case when $b = 0$ and $b = 1$ in G_1 conditioned on $k = i, k' = j$ is as hard as guessing the challenge bit in the property inference experiment $\mathsf{PI}_{i,j}$ shown in Game 15. To do this, we construct an adversary $\mathcal{A}^{i,j}_{\mathsf{PI}}$ that uses $\mathcal{A}$ as a black-box. $\mathcal{A}^{i,j}_{\mathsf{PI}}$ perfectly simulates the inputs to $\mathcal{A}$ in G_1 by forwarding its own inputs and freshly sampling z from $\mathcal{D}_i$.

$$\begin{aligned}\mathsf{Adv}_{\mathsf{PI}_{i,j}}(\mathcal{A}^{i,j}_{\mathsf{PI}}) &= \Pr\left[\mathsf{PI}_{i,j} : \neg\tilde{b} \mid \neg b\right] - \Pr\left[\mathsf{PI}_{i,j} : \neg\tilde{b} \mid b\right] \\ &= \Pr\left[G_1 : \neg\tilde{b} \mid \neg b, k = i, k' = j\right] - \Pr\left[G_1 : \neg\tilde{b} \mid b, k = i, k' = j\right]\end{aligned}$$

Putting this and (3)–(5) together we obtain

$$\begin{aligned}\mathsf{Adv}_{\mathsf{MM}}(\mathcal{A}) &= \frac{1}{K}\sum_{i=1}^{K} \mathsf{Adv}_{\mathsf{MI}_i}(\mathcal{A}^i_{\mathsf{MI}}) \\ &\quad + \frac{1}{K(K-1)}\sum_{i=1}^{K}\sum_{j=1, i\neq j}^{K} \mathsf{Adv}_{\mathsf{PI}_{i,j}}(\mathcal{A}^{i,j}_{\mathsf{PI}}) \\ &\leq \max_{i\in[K]} \mathsf{Adv}_{\mathsf{MI}_i}(\mathcal{A}^i_{\mathsf{MI}}) + \max_{i\neq j\in[K]} \mathsf{Adv}_{\mathsf{PI}_{i,j}}(\mathcal{A}^{i,j}_{\mathsf{PI}}) \;\square\end{aligned}$$

Game 15: $\mathsf{PI}_{i,j}$

Input: $\mathcal{T}, \mathcal{D}_i, \mathcal{D}_j, n, \mathcal{A}$
$b \sim \{0,1\}$
if $b = 0$ **then**
  $S \sim \mathcal{D}_i^n$
else
  $S \sim \mathcal{D}_j^n$
end
$\theta \leftarrow \mathcal{T}(S)$
$\tilde{b} \leftarrow \mathcal{A}^{i,j}_{\mathsf{PI}}(\mathcal{T}, \mathcal{D}_i, \mathcal{D}_j, n, \theta)$

Adversary 16: $\mathcal{A}^{i,j}_{\mathsf{PI}}$

Input: $\mathcal{T}, \mathcal{D}_i, \mathcal{D}_j, n, \theta$
$z \sim \mathcal{D}_i$
return $\mathcal{A}(\mathcal{T}, \boldsymbol{\mathcal{D}}, n, \theta, z)$

VI. Discussion

We discuss strategies for choosing privacy games, their current and future uses, and their limitations.

A. Selecting Games

With the variety of privacy games in the literature, it is natural to ask whether there is a *canonical* game that should be used instead of others. We believe this is not the case, i.e., no single game is the best choice in all circumstances because subtle differences in threat scenarios can lead to vastly different privacy evaluations (see, e.g., [46]). Instead, we recommend that users of games leverage the building blocks we provide in this paper to design games that accurately capture their application-specific threat models. For a given threat model, however, some differences between modelling choices are less important (e.g., in MI, whether one samples nonmembers from the full distribution or excluding the training set), and we highlight this distinction throughout the paper.

B. Current Uses of Privacy Games

The use of privacy games has become prevalent in the literature on machine learning privacy. As of today, there have been two main applications: 1) supporting the *empirical evaluation* of machine learning systems against a variety of threats, and 2) *comparing* the strength of privacy properties and attacks. Reductions enable a third application: translating *provable guarantees* from one property to another.

Game-based definitions of inference risks are presented in often inconsistent form fragmented across the literature and only a few of the reductions and separation results in Figure 1 were made explicit. We present a common vocabulary for game-based definitions, formalize games for five fundamental inference risks, and establish connections between them.

C. Prospective Uses of Privacy Games

We highlight two other promising uses for games.

Communicating privacy properties Reasoning about ML privacy risks is not the exclusive purview of researchers. Other personas, e.g., privacy managers and auditors, need to make decisions about the compliance of training pipelines with regulatory or contractual constraints. Based on our experience, privacy managers currently base their reasoning on (1) empirical privacy evaluations, (2) formal guarantees of mechanisms such as DP-SGD, and (3) informal texts such as the *Opinion 05/2014* [2] of the European Commission's Article 29 Working Party. They are then faced with the daunting task of combining these pieces into a coherent picture to assess the privacy risks of specific applications. Privacy games can help with this task: by making the threat model and assumptions about dataset creation and training explicit, they can disambiguate interpretations and can abstract an application scenario with respect to its (provable and empirical) privacy properties. Indeed, based on our initial experience, games facilitate discussing privacy goals and guarantees with stakeholders making guidelines and decisions around ML privacy.

Mechanization of proofs An advantage of the game-based formalism is that games can be given an unambiguous semantics as probabilistic programs. This enables reasoning about games using program logics and manipulating them using program transformations. Reusable program transformations (e.g., procedure inlining) and proof techniques (e.g., conditioning on events) arise naturally and make proofs more amenable. As we show in Section V, our proofs exhibit some of these patterns.

We envisage techniques and frameworks to reason about game-based cryptographic proofs (e.g., EasyCrypt, FCF) being repurposed to reason about privacy games. The apparent complexity of privacy games compared to cryptographic games is not an obstacle since most proofs manipulate training algorithms, models, and data as abstract objects with minimal structure. For instance, we think it is possible to formalize the proof in Section V in a tool like EasyCrypt. The main challenge for mechanizing proofs about privacy games is that, unlike cryptographic games, privacy games sometimes require reasoning about continuous distributions (e.g., Gaussian noise in DP-SGD), but logics implemented in existing frameworks often assume a discrete probability space.

D. Limitations of Privacy Games

Privacy games are sequential probabilistic programs; they are not an immediate fit for expressing concurrent computations. This prevents the direct application of games to important scenarios such as federated learning (FL). Intuitively, this is due to the hardness of modeling the various possible parallel interactions between the different parties. The situation is similar for cryptographic games, where process calculi are used instead of games for modeling more complex multi-party interactions [8, 45]. It is an open question whether these calculi could also be used in the context of concurrent ML scenarios such as FL.

VII. Related Work

Alternatives We discuss below informal and formal alternatives to games to express privacy properties.

A key example of a *formal* property is Differential Privacy [21]. The definition of Differential privacy is relational, in that it compares the probability of events in two alternative worlds. DP abstracts from many details that are relevant for threat modelling, such as adversary capabilities, goals, and background knowledge, as well as the way datasets are created. This has led to disagreements in the literature about the consequences of differential privacy (see [65]).

A key example of an *informal* account of privacy properties is the *Opinion 05/2014 on Anonymization Techniques* [2] that complements the EU General Data Protection Regulation (GDPR) with practical recommendations for the use of anonymization techniques to meet the requirements set out by the regulator. In this influential document, the authors identify three privacy risks: *singling out*, *linkability*, and *inference*. They analyze the suitability of different anonymization techniques—including k-anonymity and DP—for mitigating these risks, but the discussion remains inconclusive due to the lack of precise definitions. Subsequent research [18] rigorously revisited the notion of singling out and suggested reconsidering the Opinion recommendations.

Game-based definitions address shortcomings of both alternatives: They make the threat model and assumptions explicit and precise, which helps disambiguate interpretations.

Game-based privacy proofs Nissim et al. [47] construct a privacy game that reflects the requirements of the U.S. Family Educational Rights and Privacy Act (FERPA) for protecting privacy in releases of education records, and show in a proof structured as a sequence of games that DP is enough to satisfy these requirements. While constructing the game, they identify dimensions similar to our anatomy in Section II.

Surveys and taxonomies on privacy Several papers propose taxonomies of privacy attacks against machine learning systems [19, 39, 50]. Papernot et al. [49] focus on systematizing the possible attack surfaces of standard machine learning pipelines. Desfontaines and Pejó [20] systematically study variants and extensions of differential privacy. Before attacks against ML systems were demonstrated, Li et al. [37] proposed a unifying framework for membership and differential privacy definitions mainly applicable to database systems.

Acknowledgment

This work was partially funded by the U.S. National Science Foundation through the Center for Trustworthy Machine Learning (#1804603).

References

[1] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. Cited on p. 9.

[2] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymization techniques, Nov 2014. Cited on p. 13.

[3] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. J. Secur. Netw.*, 10(3): 137–150, 2015. Cited on p. 20.

[4] B. Balle, G. Cherubin, and J. Hayes. Reconstructing training data with informed adversaries. In *43rd IEEE Symposium on Security and Privacy (S&P)*, pp. 1138–1156. IEEE, 2022. Cited on pp. 1, 3, 7, 9, 10, and 19

[5] G. Barthe, B. Grégoire, and S. Zanella-Béguelin. Formal certification of code-based cryptographic proofs. In *36th annual ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages (POPL)*, pp. 90–101. ACM, 2009. Cited on p. 1.

[6] M. Bellare and P. Rogaway. The security of triple encryption and a framework for code-based game-playing proofs. In *Advances in Cryptology – EUROCRYPT*, pp. 409–426. Springer, 2006. Cited on p. 1.

[7] M. Bellare, A. Desai, E. Jokipii, and P. Rogaway. A concrete security treatment of symmetric encryption. In *38th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 394–403. IEEE, 1997. Cited on p. 1.

[8] B. Blanchet. Computationally sound mechanized proofs of correspondence assertions. In *20th IEEE Computer Security Foundations Symposium (CSF)*, pp. 97–111. IEEE, 2007. Cited on p. 13.

[9] B. Blanchet. Mechanizing game-based proofs of security protocols. *Software Safety and Security*, 33:1–25, 2012. Cited on p. 1.

[10] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium*, pp. 267–284. USENIX Association, 2019. Cited on pp. 3, 7, and 10

[11] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, Ú. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models. In *30th USENIX Security Symposium*, pp. 2633–2650. USENIX Association, 2021. Cited on pp. 1, 3, 7, and 10

[12] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership inference attacks from first principles. In *43rd IEEE Symposium on Security and Privacy (S&P)*, pp. 1546–1564. IEEE, 2022. Cited on pp. 4, 5, 6, and 10

[13] H. Chang and R. Shokri. On the privacy risks of algorithmic fairness. In *6th IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 292–303. IEEE, 2021. Cited on pp. 3, 5, and 10

[14] K. Chatzikokolakis, G. Cherubin, C. Palamidessi, and C. Troncoso. The Bayes security measure. *arXiv preprint arXiv:2011.03396 [cs.CR]*, 2020. Cited on p. 5.

[15] H. Chaudhari, J. Abascal, A. Oprea, M. Jagielski, F. Tramèr, and J. Ullman. SNAP: Efficient extraction of private properties with poisoning. In *44th IEEE Symposium on Security and Privacy (S&P)*, pp. 1935–1952. IEEE, 2023. Cited on p. 8.

[16] D. Chen, N. Yu, Y. Zhang, and M. Fritz. GAN-leaks: A taxonomy of membership inference attacks against generative models. In *27th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 343–362. ACM, 2020. Cited on p. 5.

[17] C. A. Choquette Choo, F. Tramèr, N. Carlini, and N. Papernot. Label-only membership inference attacks. In *International Conference on Machine Learning (ICML)*, vol. 139, pp. 1964–1974. PMLR, 2021. Cited on pp. 4 and 5

[18] A. Cohen and K. Nissim. Towards formalizing the GDPR's notion of singling out. *Proc. Natl. Acad. Sci. USA*, 117(15):8344–8352, 2020. Cited on p. 13.

[19] E. De Cristofaro. A critical overview of privacy in machine learning. *IEEE Security & Privacy*, 19(4):19–27, 2021. Cited on p. 13.

[20] D. Desfontaines and B. Pejó. SoK: Differential privacies. *Priv. Enhancing Technol.*, 2020(2):288–313, 2020. Cited on p. 13.

[21] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *3rd Theory of Cryptography Conference (TCC)*, pp. 265–284. Springer, 2006. Cited on p. 13.

[22] M. Fredrikson, S. Jha, and T. Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1322–1333. ACM, 2015. Cited on pp. 1 and 7

[23] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov. Property inference attacks on fully connected neural networks using permutation invariant representations. In *26th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 619–633. ACM, 2018. Cited on pp. 1, 3, and 8

[24] J. Gao, S. Garg, M. Mahmoody, and P. N. Vasudevan. Deletion inference, reconstruction, and compliance in machine (un)learning. *Priv. Enhancing Technol.*, pp. 415–436, 2022. Cited on pp. 6 and 8

[25] S. Goldwasser and S. Micali. Probabilistic encryption & how to play mental poker keeping secret all partial information. In *14th Annual ACM Symposium on Theory of Computing (STOC)*, pp. 365–377. ACM, 1982. Cited on p. 1.

[26] V. Hartmann, L. Meynent, M. Peyrard, D. Dimitriadis, S. Tople, and R. West. Distribution inference risks: Identifying and mitigating sources of leakage. In *IEEE*

Conference on Secure and Trustworthy Machine Learning (SaTML), 2023. To appear. Cited on p. 8.

[27] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro. LOGAN: Evaluating privacy leakage of generative models using generative adversarial networks. *Priv. Enhancing Technol.*, 2019:133–152, 2019. Cited on p. 5.

[28] Y. He, S. Rahimian, B. Schiele1, and M. Fritz. Segmentations-Leak: Membership inference attacks and defenses in semantic image segmentation. In *European Conference on Computer Vision (ECCV)*, pp. 519–535. Springer, 2020. Cited on p. 5.

[29] B. Hilprecht, M. Härterich, and D. Bernau. Monte Carlo and reconstruction membership inference attacks against generative models. *Priv. Enhancing Technol.*, 2019:232–249, 2019. Cited on p. 5.

[30] B. Hui, Y. Yang, H. Yuan, P. Burlina, N. Z. Gong, and Y. Cao. Practical blind membership inference attack via differential comparisons. In *28th Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2021. Cited on p. 5.

[31] T. Humphries, S. Oya, L. Tulloch, M. Rafuse, I. Goldberg, U. Hengartner, and F. Kerschbaum. Investigating membership inference attacks under data dependencies. *arXiv preprint arXiv:2010.12112 [cs.CR]*, 2020. Cited on pp. 3, 5, 6, 9, 10, and 11

[32] M. Jagielski, J. Ullman, and A. Oprea. Auditing differentially private machine learning: How private is private SGD? In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 22205–22216. Curran Associates, Inc., 2020. Cited on p. 9.

[33] B. Jayaraman and D. Evans. Are attribute inference attacks just imputation? In *29th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1569–1582. ACM, 2022. Cited on p. 6.

[34] B. Jayaraman, L. Wang, K. Knipmeyer, Q. Gu, and D. Evans. Revisiting membership inference under realistic assumptions. *Priv. Enhancing Technol.*, pp. 348–368, 2021. Cited on pp. 5 and 10

[35] B. Kulynych, M. Yaghini, G. Cherubin, M. Veale, and C. Troncoso. Disparate vulnerability to membership inference attacks. *Priv. Enhancing Technol.*, pp. 460–480, 2022. Cited on pp. 5 and 10

[36] K. Leino and M. Fredrikson. Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In *29th USENIX Security Symposium*, pp. 1605–1622. USENIX Association, 2020. Cited on pp. 4 and 5

[37] N. Li, W. Qardaji, D. Su, Y. Wu, and W. Yang. Membership privacy: A unifying framework for privacy definitions. In *20th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 889–900. ACM, 2013. Cited on pp. 1 and 13

[38] Z. Li and Y. Zhang. Membership leakage in label-only exposures. In *28th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 880–895. ACM, 2021. Cited on pp. 4 and 5

[39] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin. When machine learning meets privacy: A survey and outlook. *ACM Comput. Surv.*, 54(2):1–36, 2021. Cited on p. 13.

[40] H. Liu, J. Jia, W. Qu, and N. Z. Gong. EncoderMI: Membership inference against pre-trained encoders in contrastive learning. In *28th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 2081–2095. ACM, 2021. Cited on p. 5.

[41] S. Mahloujifar, H. A. Inan, M. Chase, E. Ghosh, and M. Hasegawa. Membership inference on word embedding and beyond. *arXiv preprint arXiv:2106.11384 [cs.CL]*, 2021. Cited on pp. 3, 6, and 10

[42] S. Mahloujifar, E. Ghosh, and M. Chase. Property inference from poisoning. In *43rd IEEE Symposium on Security and Privacy (S&P)*, pp. 1120–1137. IEEE, 2022. Cited on pp. 1, 3, and 8

[43] M. Malek Esmaeili, I. Mironov, K. Prasad, I. Shilov, and F. Tramèr. Antipodes of label differential privacy: PATE and ALIBI. In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 6934–6945. Curran Associates, Inc., 2021. Cited on pp. 3 and 10

[44] V. J. Marathe and P. Kanani. Subject granular differential privacy in federated learning. *arXiv preprint arXiv:2206.03617 [cs.LG]*, 2022. Cited on p. 9.

[45] S. Meier, B. Schmidt, C. Cremers, and D. Basin. The TAMARIN prover for the symbolic analysis of security protocols. In *Computer Aided Verification (CAV)*, pp. 696–701. Springer, 2013. Cited on p. 13.

[46] M. Nasr, S. Songi, A. Thakurta, N. Papernot, and N. Carlini. Adversary instantiation: Lower bounds for differentially private machine learning. In *42nd IEEE Symposium on Security and Privacy (S&P)*, pp. 866–882. IEEE, 2021. Cited on pp. 3, 9, 10, and 12

[47] K. Nissim, A. Bembenek, A. Wood, M. Bun, M. Gaboardi, U. Gasser, D. O'Brien, S. P. Vadhan, and T. Steinke. Bridging the gap between computer science and legal approaches to privacy. *Harvard Journal of Law & Technology*, 31(2):687–780, 2018. Cited on p. 13.

[48] T. Orekondy, B. Schiele, and M. Fritz. Knockoff nets: Stealing functionality of black-box models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4949–4958. IEEE, 2019. Cited on p. 3.

[49] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman. SoK: Security and privacy in machine learning. In *3rd IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 399–414. IEEE, 2018. Cited on p. 13.

[50] M. Rigaki and S. Garcia. A survey of privacy attacks in machine learning. *arXiv preprint arXiv:2007.07646 [cs.CR]*, 2020. Cited on p. 13.

[51] P. Rogaway. Formalizing human ignorance. In *Progress in Cryptology – VIETCRYPT*, pp. 211–228. Springer, 2006. Cited on p. 3.

[52] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International*

Conference on Machine Learning (ICML), vol. 97, pp. 5558–5567. PMLR, 2019. Cited on pp. 4 and 5

[53] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *26th Annual Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019. Cited on pp. 4 and 5

[54] A. Salem, A. Bhattacharya, M. Backes, M. Fritz, and Y. Zhang. Updates-Leak: Data set inference and reconstruction attacks in online learning. In *29th USENIX Security Symposium*, pp. 1291–1308. USENIX Association, 2020. Cited on p. 7.

[55] V. Shejwalkar and A. Houmansadr. Membership privacy for machine learning models through knowledge transfer. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 9549–9557. AAAI Press, 2021. Cited on p. 5.

[56] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *37th IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18. IEEE, 2017. Cited on pp. 1, 3, 4, and 5

[57] J. Stern. Why provable security matters? In *Advances in Cryptology – EUROCRYPT*, pp. 449–461. Springer, 2003. Cited on p. 1.

[58] A. Suri and D. Evans. Formalizing and estimating distribution inference risks. *Priv. Enhancing Technol.*, pp. 528–551, 2022. Cited on pp. 1, 8, and 10

[59] A. Suri, P. Kanani, V. J. Marathe, and D. W. Peterson. Subject membership inference attacks in federated learning. *arXiv preprint arXiv:2206.03317 [cs.LG]*, 2022. Cited on pp. 3, 8, and 10

[60] A. Suri, Y. Lu, Y. Chen, and D. Evans. Dissecting distribution inference. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023. To appear. Cited on p. 20.

[61] X. Tang, S. Mahloujifar, L. Song, V. Shejwalkar, M. Nasr, A. Houmansadr, and P. Mittal. Mitigating membership inference attacks by self-distillation through a novel ensemble architecture. In *31st USENIX Security Symposium*, pp. 1433–1450. USENIX Association, 2022. Cited on pp. 6, 10, and 17

[62] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium*, pp. 601–618. USENIX Association, 2016. Cited on p. 3.

[63] F. Tramèr, A. Terzis, T. Steinke, S. Song, M. Jagielski, and N. Carlini. Debugging differential privacy: A case study for privacy auditing. *arXiv preprint arXiv:2202.12219 [cs.LG]*, 2022. Cited on p. 9.

[64] F. Tramèr, R. Shokri, A. S. Joaquin, H. Le, M. Jagielski, S. Hong, and N. Carlini. Truth serum: Poisoning machine learning models to reveal their secrets. In *29th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 2779–2792. ACM, 2022. Cited on pp. 3, 6, 7, and 10

[65] M. C. Tschantz, S. Sen, and A. Datta. SoK: Differential privacy as a causal property. In *41st IEEE Symposium on Security and Privacy (S&P)*, pp. 354–371. IEEE, 2020. Cited on p. 13.

[66] B. Wang and N. Z. Gong. Stealing hyperparameters in machine learning. In *39th IEEE Symposium on Security and Privacy (S&P)*, pp. 36–52. IEEE, 2018. Cited on pp. 3 and 4

[67] T. Wang, Y. Zhang, and R. Jia. Improving robustness to model inversion attacks via mutual information regularization. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 11666–11673. AAAI Press, 2021. Cited on pp. 3, 7, and 10

[68] B. Wu, X. Yang, S. Pan, and X. Yuan. Adapting membership inference attacks to GNN for graph classification: Approaches and implications. In *IEEE International Conference on Data Mining (ICDM)*, pp. 1421–1426. IEEE, 2021. Cited on p. 5.

[69] X. Wu, M. Fredrikson, S. Jha, and J. F. Naughton. A methodology for formalizing model-inversion attacks. In *29th IEEE Computer Security Foundations Symposium (CSF)*, pp. 355–370. IEEE, 2016. Cited on pp. 1, 2, and 6

[70] S. Yeom, I. Giacomelli, A. Menaged, M. Fredrikson, and S. Jha. Overfitting, robustness, and malicious algorithms: A study of potential causes of privacy risk in machine learning. *J. of Comput. Secur.*, 28(1):35–70, 2020. Cited on pp. 2, 3, 4, 5, 6, 9, and 10

[71] S. Zanella-Béguelin, L. Wutschitz, S. Tople, V. Rühle, A. Paverd, O. Ohrimenko, B. Köpf, and M. Brockschmidt. Analyzing information leakage of updates to natural language models. In *27th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 363–375. ACM, 2020. Cited on p. 7.

[72] S. Zanella-Béguelin, L. Wutschitz, S. Tople, A. Salem, V. Rühle, A. Paverd, M. Naseri, and B. Köpf. Bayesian estimation of differential privacy. *arXiv preprint arXiv:2206.05199 [cs.LG]*, 2022. Cited on pp. 3 and 9

[73] M. Zhang, Z. Ren, Z. Wang, P. Ren, Z. Chen, P. Hu, and Y. Zhang. Membership inference attacks against recommender systems. In *28th ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 864–879. ACM, 2021. Cited on p. 5.

[74] W. Zhang, S. Tople, and O. Ohrimenko. Leakage of dataset properties in Multi-Party machine learning. In *30th USENIX Security Symposium*, pp. 2687–2704. USENIX Association, 2021. Cited on pp. 1, 3, and 8

[75] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang. Inference attacks against graph neural networks. In *31st USENIX Security Symposium*, pp. 4543–4560. USENIX Association, 2022. Cited on pp. 6 and 7

[76] J. Zhou, Y. Chen, C. Shen, and Y. Zhang. Property inference attacks against GANs. In *29th Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2022. Cited on p. 8.

APPENDIX

A. Direct, Single-Query Membership Inference

Tang et al. [61] present single-query variants of membership inference where the adversary is given only the model output on the challenge point. In their base game (Game 17), the adversary selects a universe of $2n$ points from where n points are sub-sampled to construct the training dataset of the target model. The adversary goal is to infer whether a challenge z_j uniformly sampled from the initial $2n$ points was used to train the model, i.e., guess $B[j]$, given just the model output on z_j. They also consider variants where the set of $2n$ points is fixed externally, and a worst-case variant where the challenge z_j is selected by the adversary.

Game 17: $\mathsf{MI}^{\mathsf{SQ}}$

Input: $\mathcal{T}, n, \mathcal{A}, \mathcal{A}'$
$\{z_i\}_{i\in[2n]} \leftarrow \mathcal{A}'(\mathcal{T}, n)$
$B \sim \{0,1\}^{2n}$ s.t. $\sum_{i\in[2n]} B[i] = n$
$S \leftarrow \{z_i \,|\, B[i] = 0\}_{i\in[2n]}$
$\theta \leftarrow \mathcal{T}(S)$
$j \sim [2n]$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, n, \{z_i\}_{i\in[2n]}, j, \theta(z_j))$

B. Deferred Proofs

Theorem 13 ($\mathsf{SMI} \preceq \mathsf{RC}$). *Let z_0, z_1 be two samples, S a dataset of $n-1$ samples, and ℓ a symmetric reconstruction loss satisfying the triangle inequality. Let $\mathcal{A}$ be an adversary against data reconstruction (RC) w.r.t. S and the uniform prior on $\{z_0, z_1\}$ that reconstructs its challenge with error $\eta < \ell(z_0, z_1)/2$ with probability γ. Then, there exists a strong membership inference adversary $\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}$ such that*

$$\mathsf{Adv}_{\mathsf{SMI}}(\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}) \geq 2\gamma - 1$$

Adversary 18: $\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}$

Input: $\mathcal{T}, \theta, S, z_0, z_1$
$\tilde{z} \leftarrow \mathcal{A}(\mathcal{T}, \theta, S)$
if $\ell(z_0, \tilde{z}) < \ell(z_1, \tilde{z})$ **then**
  return 0
else
  return 1
end

Proof. Define $\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}$ as in Adversary 18. For any $\tilde{z}$, we have from the triangle inequality,

$$\ell(z_0, \tilde{z}) < \ell(z_0, z_1)/2 < (\ell(z_0, \tilde{z}) + \ell(\tilde{z}, z_1))/2 \implies \ell(z_0, \tilde{z}) < \ell(z_1, \tilde{z})$$

Therefore, when $b = 0$ in SMI and $\mathcal{A}$ succeeds in reconstructing z_0 within error η, $\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}$ guesses correctly. Similarly, when $b = 1$ and $\mathcal{A}$ succeeds in reconstructing z_1 within error η, $\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}$ guesses correctly. Thus, $\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}$ guesses b correctly at least with probability γ and

$$\mathsf{Adv}_{\mathsf{SMI}}(\mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}) = 2\Pr\left[\mathsf{SMI}(\cdots) : \tilde{b} = b\right] - 1 \geq 2\gamma - 1 \quad \square$$

Theorem 3 ($\mathsf{DPD} \preceq \mathsf{RC}$). *Let π be a prior over samples, S a dataset of $n-1$ samples, and ℓ a symmetric reconstruction loss satisfying the triangle inequality. Let $\mathcal{A}$ be an adversary against data reconstruction (RC) w.r.t. S and π that reconstructs its challenge within error η with probability $\gamma \geq 1/2$. Let*

$$\alpha = \inf_{z_0 \in \mathrm{supp}(\pi)} \Pr[z_1 \sim \pi : \ell(z_0, z_1) > 2\eta]$$

There exists a DP distinguisher $\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}$ such that

$$\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}) \geq 2\alpha\left(\gamma - \frac{1}{2}\right)$$

Proof. Observe that $1 - \alpha$ is the baseline success of a reconstruction adversary with error 2η (see Equation 1).

Define $\mathcal{A}'_{\mathsf{DPD}\to\mathsf{RC}}$ as in Adversary 19 and $\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}$ as in Adversary 20.

Adversary 19: $\mathcal{A}'_{\mathsf{DPD}\to\mathsf{RC}}$

Input: $\mathcal{T}, n$
$z_0, z_1 \sim \pi$
return S, z_0, z_1

Adversary 20: $\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}$

Input: $\mathcal{T}, \theta, S, z_0, z_1$
if $\ell(z_0, z_1) \leq 2\eta$ **then**
  $\tilde{b} \sim \{0, 1\}$
else
  $\tilde{b} \leftarrow \mathcal{A}_{\mathsf{SMI}\to\mathsf{RC}}(\mathcal{T}, \theta, S, z_0, z_1)$
end
return $\tilde{b}$

In the DPD game, when $\ell(z_0, z_1) > 2\eta$, which occurs with probability at least α, a similar analysis as in Theorem 13 shows that $\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}$ guesses b correctly whenever $\mathcal{A}$ succeeds in reconstructing its challenge within error η. Otherwise, the adversary guesses with probability $1/2$. Thus,

$$\begin{aligned}\Pr\left[\mathsf{DPD} : \tilde{b} = b\right] &\geq \Pr\left[\mathsf{DPD} : \tilde{b} = b | \ell(z_0, z_1) > 2\eta\right] \alpha + \\ &\quad \Pr\left[\mathsf{DPD} : \tilde{b} = b | \ell(z_0, z_1) \leq 2\eta\right] (1 - \alpha) \\ &= \gamma\alpha + \tfrac{1}{2}(1 - \alpha)\end{aligned}$$

The DPD advantage of $\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}$ is

$$\begin{aligned}\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}_{\mathsf{DPD}\to\mathsf{RC}}) &= 2\Pr\left[\mathsf{DPD} : \tilde{b} = b\right] - 1 \\ &\geq 2\alpha\left(\gamma - \frac{1}{2}\right) \quad \square\end{aligned}$$

Theorem 4 ($\mathsf{DPD} \preceq \mathsf{MI}$). *For any adversary $\mathcal{A}_{\mathsf{MI}}$ against membership inference, there exists a DP distinguisher $\mathcal{A}_{\mathsf{DPD}}$ such that*

$$\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}_{\mathsf{DPD}}) = \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A}_{\mathsf{MI}})$$

Proof. Let $\mathcal{A}$ be an adversary against $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n)$. We construct an adversary against $\mathsf{DPD}(\mathcal{T}, n)$ as in Adversary 21 and 22. These adversary procedures, when inlined in $\mathsf{DPD}(\mathcal{T}, n)$ (Game 10), result in an experiment semantically equivalent to $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n, \mathcal{A})$ (Game 2). Thus,

$$\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}_{\mathsf{DPD}\to\mathsf{MI}}) = \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A}) \qquad \square$$

Adversary 21: $\mathcal{A}'_{\mathsf{DPD}\to\mathsf{MI}}$

Input: $\mathcal{T}, n$
$S \sim \mathcal{D}^{n-1}$
$z_0, z_1 \sim \mathcal{D}$
return S, z_0, z_1

Adversary 22: $\mathcal{A}_{\mathsf{DPD}\to\mathsf{MI}}$

Input: $\mathcal{T}, \theta, S, z_0, z_1$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \theta, z_0)$
return $\tilde{b}$

Theorem 5 ($\mathsf{RC} \preceq_{1/|\operatorname{supp}(\mathcal{D})|} \mathsf{MI}$)**.** *For any membership inference adversary $\mathcal{A}$ against $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n)$ there exists a reconstruction adversary $\mathcal{B}$ against $\mathsf{RC}^{\mathsf{Ran}}(\mathcal{D}, n, \mathcal{D}, \mathcal{T})$ (i.e., with prior $\pi = \mathcal{D}$) such that*

$$\mathsf{Adv}_{\mathsf{RC}^{\mathsf{Ran}}}(\mathcal{B}) = \frac{1}{|\operatorname{supp}(\mathcal{D})|} \cdot \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A})$$

Proof. Consider Game 23, which is equivalent to AI except the adversary is also given S.

Game 23: AI'

Input: $\mathcal{T}, \mathcal{D}, n, \mathcal{B}, \varphi, \pi$
$S \sim \mathcal{D}^{n-1}$
$z_0, z_1 \sim \mathcal{D}$
$b \sim \{0, 1\}$
$\theta \leftarrow \mathcal{T}(S \cup \{z_b\})$
$\tilde{z} \leftarrow \mathcal{B}(\mathcal{T}, \mathcal{D}, n, \theta, S, \varphi(z_0))$

The reconstruction advantage of $\mathcal{B}$ coincides with its advantage in AI' in the special case where $\varphi(z) = \perp$ and $\pi(z) = z$, i.e., the adversary has to reconstruct all attributes. This is because $\varphi(z_0) = \perp$ and thus the guess $\tilde{z}$ is independent of z_0 conditioned on $b = 1$.

$$\mathsf{Adv}_{\mathsf{RC}^{\mathsf{Ran}}}(\mathcal{B}) = \Pr\left[\mathsf{AI}' : \tilde{z} = z_0 | b = 0\right] - \Pr\left[\mathsf{AI}' : \tilde{z} = z_0 | b = 1\right]$$

The rest of the proof is similar to the proof of Theorem 2, but we present it for the sake of completeness.

Let $\mathcal{A}$ be an adversary against $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n)$. We construct an adversary $\mathcal{B}$ against $\mathsf{RC}^{\mathsf{Ran}}(\mathcal{D}, n, \mathcal{D}, \mathcal{T})$, shown in Game 24, which uses $\mathcal{A}$ to reconstruct its challenge.

Denote $\mathcal{D}(z_i)$ the quantity $\Pr[z \sim \mathcal{D} : z = z_i]$, i.e., the probability mass of $\mathcal{D}$ at z_i and let $m = |\operatorname{supp}(\mathcal{D})|$. In the following, we use RC to denote the game $\mathsf{RC}^{\mathsf{Ran}}(\mathcal{D}, n, \mathcal{D}, \mathcal{T}, \mathcal{B})$ and MI to denote $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n, \mathcal{A})$.

Adversary 24: $\mathcal{B}$

Input: $\mathcal{T}, \theta, S$
$z' \sim \operatorname{supp}(\mathcal{D})$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \theta)$
if $\tilde{b} = 0$ **then**
  return z'
else
  return $\perp$
end

Since $\mathcal{B}$ guesses $\tilde{z} = z$ if and only if $z' = z$ and $\tilde{b} = 0$, for any $z_i \in \operatorname{supp}(\mathcal{D})$ we have for $\hat{b} \in \{0, 1\}$

$$\Pr\left[\mathsf{AI}' : \tilde{z} = z | b = \hat{b}, z = z_i\right] = \frac{1}{m} \Pr\left[\mathsf{AI}' : \tilde{b} = 0 | b = \hat{b}, z = z_i\right] \tag{6}$$

Hence, the advantage of $\mathcal{B}$ is

$$\begin{aligned}
&\mathsf{Adv}_{\mathsf{RC}^{\mathsf{Ran}}}(\mathcal{B}) \\
&= \sum_{z_i \in \operatorname{supp}(\mathcal{D})} \mathcal{D}(z_i) \left(\Pr\left[\mathsf{AI}' : \tilde{z} = z_0 | b = 0, z = z_i\right]\right. \\
&\qquad \left. - \Pr\left[\mathsf{AI}' : \tilde{z} = z_0 | b = 1, z = z_i\right]\right) \\
&= \frac{1}{m} \sum_{z_i \in \operatorname{supp}(\mathcal{D})} \mathcal{D}(z_i) \left(\Pr\left[\mathsf{AI}' : \tilde{b} = 0 | b = 0, z = z_i\right]\right. \\
&\qquad \left. - \Pr\left[\mathsf{AI}' : \tilde{b} = 0 | b = 1, z = z_i\right]\right) \\
&= \frac{1}{m} \left(\Pr\left[\mathsf{AI}' : \tilde{b} = 0 | b = 0\right] - \Pr\left[\mathsf{AI}' : \tilde{b} = 0 | b = 1\right]\right) \\
&= \frac{1}{m} \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A})
\end{aligned}$$

The penultimate equality holds because b and z are independent. The last equality holds because game $\mathsf{AI}'(\mathcal{T}, \mathcal{D}, n, \mathcal{B})$ matches game $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n, \mathcal{A})$ and so the joint distribution of $\tilde{b}, b$ is the identical in both games. $\square$

Theorem 7 ($\mathsf{MI} \not\preceq \mathsf{DPD}$)**.** *Resilience against membership inference does not imply resilience against DP distinguishability.*

Proof. We show that there are training pipelines that are arbitrarily resilient against membership inference attacks but completely insecure against DP distinguishing attacks.

We construct a training pipeline $(\mathcal{T}, \mathcal{D}, n)$ such that the MI advantage of an adversary against it is at most $1/\sqrt{n}$, and so vanishes as n grows. Yet, we exhibit a DP distinguisher against the pipeline that achieves perfect advantage.

Game 25: MI'

Input: $\mathcal{T}, \mathcal{D}, n, \mathcal{A}$
$b \sim \{0, 1\}$
$S \sim \mathcal{D}^{n-1}$
$z_0, z_1 \sim \mathcal{D}$
$\theta \leftarrow \mathcal{T}(S \cup \{z_b\})$
$\tilde{b} \leftarrow \mathcal{A}(\mathcal{T}, \mathcal{D}, n, \theta, z_0, z_1)$

Let $\mathcal{D} = \text{Bernoulli}(p)$ and $\mathcal{T}(S) = \sum_{x \in S} x$. Consider Game 25. If the adversary were only given z_0, this game would be equivalent to the basic MI game (Game 1). Since the adversary is given strictly more information, any bound on its advantage in this game would also bound the MI advantage of adversaries against the training pipeline. The adversary must distinguish between two simple hypotheses:

- $H_0 : \theta \sim \text{Binomial}(n-1, p) + z_0$
- $H_1 : \theta \sim \text{Binomial}(n-1, p) + z_1$

When $z_0 = z_1$, these coincide and the advantage of the adversary is 0. Otherwise, without loss of generality, assume $z_b = b$. By the Neyman-Pearson lemma, a likelihood ratio test yields the most powerful test for a significance α (i.e., Type-I error, false positive rate). Let f and F be the probability mass and cumulative distribution function of $\text{Binomial}(n-1, p)$, respectively. The likelihood ratio is

$$\Lambda(\theta = k) = \begin{cases} \infty & \text{if } k = 0 \\ 0 & \text{if } k = n \\ \frac{f(k)}{f(k-1)} = \frac{(n-k)p}{k(1-p)} & \text{otherwise} \end{cases}$$

The test rejects H_0 when $\Lambda(\theta) < c$, for some c. The false positive rate α (the probability of rejecting H_0 when H_0 is true) is

$$\begin{aligned} \Pr_{H_0}(\Lambda(\theta) < c) &= \Pr_{H_0}\left(\frac{(n-k)p}{k(1-p)} < c\right) \\ &= \Pr_{H_0}\left(k > \frac{np}{p + c(1-p)}\right) \\ &= 1 - F\left(\frac{np}{p + c(1-p)}\right) \end{aligned}$$

The false negative rate β is

$$\begin{aligned} \Pr_{H_1}(\Lambda(\theta) \geq c) &= \Pr_{H_1}\left(\frac{(n-k)p}{k(1-p)} \geq c\right) \\ &= \Pr_{H_1}\left(k \leq \frac{np}{p + c(1-p)} - 1\right) \\ &= F\left(\frac{np}{p + c(1-p)} - 1\right) \end{aligned}$$

Now, take $p = 0.5$ and assume that $n \geq 4$ and that n is even so that the mode of $\text{Binomial}(n-1, p)$ is $n/2$. The MI advantage of the adversary is

$$\begin{aligned} \mathsf{Adv}_{\mathsf{MI}}(\mathcal{A}) &= \frac{1}{2}(f(0) + f(n-1) + (1 - \alpha - \beta)) \\ &= f(0) + \frac{1}{2} f\left(\frac{np}{p + c(1-p)}\right) \\ &\leq \frac{1}{2^{n-1}} + \frac{f(n/2)}{2} \\ &\leq \frac{1}{2\sqrt{n}} + \frac{1}{2\sqrt{n}} = \frac{1}{\sqrt{n}} \end{aligned}$$

On the other hand, a DP distinguisher $\mathcal{A}$ that chooses $z_0 = 0, z_1 = 1$, an arbitrary S, and that guesses $\tilde{b} = \theta - \sum_{x \in S} S$, has perfect advantage $\mathsf{Adv}_{\mathsf{DPD}}(\mathcal{A}) = 1$. □

Theorem 6 (MI $\not\preceq$ PI)**.** *Resilience against membership inference does not imply resilience against property inference.*

Proof. We construct a training pipeline $(\mathcal{T}, \mathcal{D}_b, n)$ that is arbitrarily resilient to membership inference for $b \in \{0, 1\}$. Yet, we exhibit a property inference attack against it that achieves perfect advantage.

Let $\mathcal{D}_b = \text{Bernoulli}(p_b)$ with $p_0 \neq p_1$ and $\mathcal{T}(S) = \sum_{x \in S} x$. As shown in Theorem 7, the advantage of a membership inference adversary against $(\mathcal{T}, \mathcal{D}_b, n)$ is at most $1/\sqrt{n}$. However, as n grows, $\mathcal{T}(S)/n$ is an unbiased estimator for the mean p_b, which allows a property inference adversary to easily distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$, particularly when p_0 and p_1 are far apart. □

Theorem 8 (MI $\not\preceq$ RC)**.** *Resilience against membership inference does not imply resilience against reconstruction.*

Proof. It suffices to show that the training pipeline from Theorem 7, which is resilient to membership inference attacks, admits a reconstruction attack. For this, recall that in RC the adversary knows the dataset S (but not the target sample z). For the pipeline $(\mathcal{T}, \mathcal{D}, n)$ given in Theorem 7, z can be perfectly reconstructed since $z = \theta - \sum_{x \in S} x$. □

Theorem 9 (PI $\not\preceq$ MI)**.** *Resilience against property inference does not imply resilience against membership inference.*

Proof. We exhibit a pipeline resilient to property inference that is completely vulnerable to a membership inference attack.

Let $\mathcal{D}$ be an arbitrary distribution and define $\mathcal{D}_b$ so that $z \sim \mathcal{D}_b \equiv x \sim \mathcal{D}; z \leftarrow (x, b)$. Let $n > 0$ and define

$$\mathcal{T}(S) = \{x \in S | (x, y) \in S\}$$

A membership inference adversary against $\mathsf{MI}(\mathcal{T}, \mathcal{D}, n)$ that given $\theta, z_0 = (x, y)$ returns 0 if and only if $x \in S$ achieves the maximum advantage, i.e.,

$$1 - \Pr[S \sim \mathcal{D}^n; x \sim \mathcal{D} : x \in S]$$

However, a property inference adversary gets no information about b as θ and b are independent, so its advantage is 0. □

Theorem 10 (RC $\not\preceq$ DPD)**.** *Resilience against reconstruction does not imply resilience against DP distinguishability.*

Proof. Balle et al. [4, Theorem 5] show that resilience against reconstruction w.r.t. all priors in a family of distributions concentrated on all ordered pairs of distinct examples implies (ε, δ)-DP, and hence via Proposition 2 resilience against DPD.

However, resilience against a single prior π, even if its support includes all possible examples, is clearly insufficient to guarantee resilience against DPD. To see why, consider a deterministic DPD adversary that picks S, z_0, z_1. Given error bound η and success probability γ, all reconstruction adversaries can have error larger than η when $z \notin \{z_0, z_1\}$ but reconstruct $z \in \{z_0, z_1\}$ perfectly, as long as $\Pr[z \sim \pi : z \in \{z_0, z_1\}] < \gamma$, i.e., the probability mass of the prior on $\{z_0, z_1\}$. The situation is worse when $z_0, z_1 \notin \text{supp}(\pi)$, where resilience against reconstruction for arbitrary η, γ is compatible with perfect DPD advantage. □

Theorem 11 (DPD $\not\preceq$ PI)**.** *Resilience against DP distinguishability does not imply resilience against property inference.*

Proof. Let $\varepsilon, \delta \in (0, 1)$. We build two training pipelines $(\mathcal{T}, \mathcal{D}_b, n)$, $b \in \{0, 1\}$, that satisfy (ε, δ)-DP and thus are resilient against DPD (see Proposition 2). We then show an adversary against $\mathsf{PI}(\mathcal{D}_0, \mathcal{D}_1, n, \mathcal{T})$ whose advantage grows with n.

Let $\mathcal{D}_b = \text{Bernoulli}(p_b)$ with $p_0 \neq p_1$ and define $\mathcal{T}(S) = \sum_{x \in S} x + \mathcal{N}(0, \sigma^2)$ where $\sigma^2 = 2\ln(1.25/\delta)\varepsilon^{-1}$. Since the sum above has sensitivity 1 and $\mathcal{T}$ is the standard Gaussian mechanism, the training pipeline is (ε, δ)-DP.

Note that for S sampled from $\mathcal{D}_b$, the random variable $\mathcal{T}(S)$ is distributed as $\text{Binomial}(n, p_b) + \mathcal{N}(0, \sigma^2)$. We can use Berry-Esséen theorem to approximate the binomial distribution with a normal distribution, so that approximately

$$\begin{aligned}\mathcal{T}(S) &\sim \mathcal{N}(np_b, np_b(1-p_b)) + \mathcal{N}(0, \sigma^2) \\ &\sim \mathcal{N}(np_b, np_b(1-p_b) + \sigma^2)\end{aligned}$$

The approximation error is $O(\sqrt{n})$. $\mathcal{T}(S)/n$ is an unbiased estimator for p_b with variance $p_b(1-p_b) + \sigma^2/n$. Since σ does not depend on n, as n grows the approximation error and the variance of the estimate decrease. This allows a property inference adversary to distinguish between $\mathcal{D}_0$ and $\mathcal{D}_1$, particularly when p_0 and p_1 are far apart. □

Ateniese et al. [3, Section 4.2] were the first to observe that differential privacy does not protect against property inference and provided a practical counterexample: a differentially private k-means network traffic classifier that nonetheless leaks the presence of traces from Google.com web traffic in their training dataset. However, their argument remains informal, appealing to visual differences in the centroids of just two trained models. Suri et al. [60, Section V] give a more compelling example where property inference risk remains high on neural networks trained with DP-SGD.