# Using Rényi-divergence and Arimoto-Rényi Information to Quantify Membership Information Leakage

Farhad Farokhi
Department of Electrical and Electronic Engineering
The University of Melbourne
Parkville, VIC 3010, Australia
Email: farhad.farokhi@unimelb.edu.au

***Abstract*—Membership inference attacks, i.e., adversarial attacks inferring whether a data record is used for training a machine learning model, has been recently shown to pose a legitimate privacy risk in machine learning literature. In this paper, we propose two measures of information leakage for investigating membership inference attacks backed by results on binary hypothesis testing in information theory literature. The first measure of information leakage is defined using Rényi $\alpha$-divergence of the distribution of output of a machine learning model for data records that are in and out of the training dataset. The second measure of information leakage is based on Arimoto-Rényi $\alpha$-information between the membership random variable (whether the data record is in or out of the training dataset) and the output of the machine learning model. These measures of leakage are shown to be related to each other. We compare the proposed measures of information leakage with $\alpha$-leakage from the information-theoretic privacy literature to establish some useful properties. We establish an upper bound for $\alpha$-divergence information leakage as a function of the privacy budget for differentially-private machine learning models.**

## I. Introduction

Membership inference attack (MIA), i.e., adversarial attacks employed to infer whether a given individual's data record is used for training a machine learning (ML) model, has been recently shown to pose a legitimate privacy concern [1]–[4]. The knowledge of belonging to the training dataset of an ML model, although potentially only leaking one bit of information to an adversary, can have devastating consequences for the privacy of that individual. For instance, belonging to the training dataset used for investigating efficacy of a new drug on a disease or for determining the relationship between certain genetic markers and severity of a disease can illustrate that a person has that disease. The threat poses a more significant threat to vulnerable subgroups of society [5].

MIAs often rely on that ML models behave differently on the training dataset in comparison to the test dataset. For instance, in the presence of over-fitting [6], ML models used for classification can demonstrate a higher level of confidence on the training dataset in comparison to data never encountered earlier. MIAs have been deployed and shown to be effective on various ML models [7]–[11].

After demonstrating the success of such attacks, several studies have proposed defense mechanisms for securing privacy of training datasets. For instance, the accuracy of MIAs was used as a regularization term when training ML models to generate models that are robust against MIAs [12]. MIA has shown to be over-fitting [6] and therefore traditional methods deployed in ML training to combat over-fitting, such as regularization [13], [14], have been touted for a successful strategies in mitigating MIAs. Differential privacy has shown significant promises in combating MIAs [1] however at the cost of significantly reducing the utility [15], [16].

In this paper, we propose two tailored measures of information leakage for MIA. We use results from binary hypothesis testing, particularly those in [17], [18], to develop these measures. The first measure of information leakage is based on Rényi $\alpha$-divergence of the distribution of the output of ML model for data records that are in and out of the training dataset. The second measure of information leakage is based on Arimoto-Rényi $\alpha$-information between the membership variable (signifying whether the data record is in or out of the training dataset) and the output of the ML model. We show that these measures of information leakage are related to each other upon selecting $\alpha$ correctly. We compare these measures of information leakage with $\alpha$-leakage from the privacy literature [19], [20]. This comparison allows us to establish quasi-convexity of $\alpha$-information MIA information leakage. Although $\alpha$-information MIA information leakage is proved to be equal $\alpha$-leakage, $\alpha$-divergence MIA information leakage is novel and useful for investigating the effects of differential privacy. Particularly, we establish an upper bound for $\alpha$-divergence MIA information leakage as a function of the privacy budget for differentially-private machine learning models. We prove that the MIA information leakage vanishes as the privacy budget tends to zero. Therefore, by using small enough privacy budgets, we can effectively combat MIA. This however comes at the cost of significant utility degradation as also empirically observed in [15], [16].

The work of F. Farokhi is funded by a startup grant from the Melbourne School of Engineering at the University of Melbourne.

The rest of the paper is organized as follows. Rényi $\alpha$-divergence and Arimoto-Rényi $\alpha$-information are briefly introduced in Section II. The measures of MIA information leakage are presented in Section III. Section IV compares measures of MIA information leakage with $\alpha$-leakage in privacy literature. The effects of differential privacy is investigated in Sections V. Finally, the paper is concluded in Section VI.

## II. Rényi $\alpha$-Divergence and Information

In this paper, we concentrate on random variables with finite support set or alphabet. For (discrete) random variable $Z$ distributed according to the probability distribution $Q$, $\mathbb{P}_{Z\sim Q}\{Z\in\mathcal{F}\}$ denotes the probability of event $\mathcal{F}\in\mathfrak{F}$, where $\mathfrak{F}$ denotes the event space (i.e., the set of all possible events). For the sake of convenience and brevity of notation, $Q(\mathcal{F})$ is often used instead of $\mathbb{P}_{Z\sim Q}\{Z\in\mathcal{F}\}$. When the event is a singleton, i.e., $\mathcal{F}=\{z\}$, we use $Q(z)$ instead of $Q(\mathcal{F})=Q(\{z\})$. The distribution $P$ is absolutely continuous with respect to $Q$, denoted by $P\ll Q$, if there does not exists any event $\mathcal{F}\in\mathfrak{F}$ such that $P(\mathcal{F})>0$ while $Q(\mathcal{F})=0$.

The Rényi divergence (see, e.g., [21]) of positive order $\alpha\in(0,1)\cup(1,\infty)$, also referred to as $\alpha$-divergence, between probability distribution $P,Q$ such that $P\ll Q$ is

$$D_\alpha(P\|Q):=\frac{1}{\alpha-1}\log\left(\mathbb{E}_{Z\sim Q}\left\{\frac{P(Z)^\alpha}{Q(Z)^\alpha}\right\}\right)$$

with the conventions that $0/0=0$ and $a/0=\infty$ for $a>0$. The Rényi divergence is a non-decreasing function of $\alpha$ [21]. For $\alpha=1$, the Rényi divergence is given by

$$D_1(P\|Q):=\mathrm{KL}(P\|Q)=\sup_{0<\alpha<1}D_\alpha(P\|Q),$$

where $\mathrm{KL}(P\|Q)$ denotes the Kullback-Leibler divergence between $P$ and $Q$ defined as

$$\mathrm{KL}(P\|Q):=\mathbb{E}_{Z\sim P}\left\{\log\left(\frac{P(Z)}{Q(Z)}\right)\right\}.$$

If $\mathrm{KL}(P\|Q)<\infty$, $D_1(P\|Q):=\lim_{\alpha\uparrow 1}D_\alpha(P\|Q)$. For $\alpha=0$,

$$\begin{aligned}D_0(P\|Q)&:=\max_{\mathcal{F}:P(\mathcal{F})=1}\log\left(\frac{1}{Q(\mathcal{F})}\right)\\&=\inf_{0<\alpha<1}D_\alpha(P\|Q).\end{aligned}$$

Also,

$$\begin{aligned}D_\infty(P\|Q)&:=\log\left(\sup_{\mathcal{F}\in\mathfrak{F}}\frac{P(\mathcal{F})}{Q(\mathcal{F})}\right)\\&=\log\left(\sup_{z\in\mathrm{supp}(Q)}\frac{P(z)}{Q(z)}\right)\\&=\sup_{\alpha>1}D_\alpha(P\|Q).\end{aligned}$$

Note that, because $P\ll Q$, $\mathrm{supp}(P)\subseteq\mathrm{supp}(Q)$, where $\mathrm{supp}(\cdot)$ denotes the support set of the distribution, i.e., $\mathrm{supp}(P):=\{z:P(z)>0\}$. If $P\not\ll Q$, the definition can be extended to have $D_\infty(P\|Q)=+\infty$.

Let random variable $\mathbf{x}\in\mathcal{X}$ admit the probability distribution $P_{\mathbf{x}}$. Assume that $\mathcal{X}$ is a finite set. The Rényi entropy (see, e.g., [22]) of order $\alpha\in(0,1)\cup(1,\infty)$ for random variable $\mathbf{x}$ is given by

$$\begin{aligned}H_\alpha(\mathbf{x})&:=\frac{1}{1-\alpha}\log\left(\sum_{x\in\mathcal{X}}P_{\mathbf{x}}(x)^\alpha\right)\\&=\frac{1}{1-\alpha}\log\left(\mathbb{E}\{P_{\mathbf{x}}(\mathbf{x})^{\alpha-1}\}\right).\end{aligned}$$

By continuous extension,

$$\begin{aligned}H_0(\mathbf{x})&:=\log(|\mathrm{supp}(P_{\mathbf{x}})|),\\H_1(\mathbf{x})&:=H(\mathbf{x}),\\H_\infty(\mathbf{x})&:=-\log\left(\max_{x\in\mathrm{supp}(P_{\mathbf{x}})}P_{\mathbf{x}}(x)\right),\end{aligned}$$

where $H(\mathbf{x}):=\mathbb{E}\{\log(P_{\mathbf{x}}(\mathbf{x}))\}$ denotes the Shannon entropy. Let random variable $(\mathbf{x},\mathbf{y})\in\mathcal{X}\times\mathcal{Y}$ admit the probability distribution $P_{\mathbf{xy}}$ with marginals $P_{\mathbf{x}},P_{\mathbf{y}}$ and conditional distribution $P_{\mathbf{x}|\mathbf{y}}$. Again, assume that $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. The Arimoto-Rényi conditional entropy (see, e.g., [17], [23]) of order $\alpha\in(0,1)\cup(1,\infty)$ for random variable $\mathbf{x}$ given $\mathbf{y}$ is

$$\begin{aligned}&H_\alpha(\mathbf{x}|\mathbf{y})\\&:=\frac{\alpha}{1-\alpha}\log\left(\mathbb{E}\left\{\left(\sum_{x\in\mathcal{X}}P_{\mathbf{x}|\mathbf{y}}(x|\mathbf{y})^\alpha\right)^{\frac{1}{\alpha}}\right\}\right)\\&=\frac{\alpha}{1-\alpha}\log\left(\sum_{y\in\mathcal{Y}}P_{\mathbf{y}}(y)\exp\left(\frac{1-\alpha}{\alpha}H_\alpha(\mathbf{x}|\mathbf{y}=y)\right)\right).\end{aligned}$$

Again, by its continuous extension, we have

$$\begin{aligned}H_0(\mathbf{x}|\mathbf{y})&:=\log\left(\max_{y\in\mathcal{Y}}\left|\mathrm{supp}(P_{\mathbf{x}|\mathbf{y}}(\cdot|\mathbf{y}=y))\right|\right)\\&=\max_{y\in\mathcal{Y}}H_0(\mathbf{x}|\mathbf{y}=y),\\H_1(\mathbf{x}|\mathbf{y})&:=H(\mathbf{x}|\mathbf{y}),\\H_\infty(\mathbf{x}|\mathbf{y})&:=-\log\left(\mathbb{E}\left\{\max_{x\in\mathcal{X}}P_{\mathbf{x}|\mathbf{y}}(x|\mathbf{y})\right\}\right),\end{aligned}$$

where $H(\mathbf{x}|\mathbf{y})$ is the Shannon conditional entropy. The Arimoto-Rényi conditional entropy $H_\alpha(\mathbf{x}|\mathbf{y})$ is monotonically decreasing in $\alpha$ throughout the real line [17]. Furthermore, $((\alpha-1)/\alpha)H_\alpha(\mathbf{x}|\mathbf{y})$ is monotonically increasing in $\alpha$ on $(0,+\infty)$ [17].

The Arimoto information (see, e.g., [17], [24]) of order $\alpha$, which we refer to in this paper as $\alpha$-mutual information[1], is

$$I_\alpha(\mathbf{x};\mathbf{y}):=H_\alpha(\mathbf{x})-H_\alpha(\mathbf{x}|\mathbf{y}).$$

Note that $I_\alpha(\mathbf{x};\mathbf{y})$ is equal to the conventional mutual information for $\alpha=1$. Finally, if $\mathbf{x}$ is equiprobable on $\mathrm{supp}(P_{\mathbf{x}})$, we get $I_\alpha(\mathbf{x};\mathbf{y})=\log(|\mathrm{supp}(P_{\mathbf{x}})|)-H_\alpha(\mathbf{x}|\mathbf{y})$ [17].

[1] Note that $\alpha$-mutual information often refers to the Sibson mutual information [25]. This is not to be mistaken with Arimoto's definition adopted in this paper.

## III. Membership Inference

We follow the approach of [5] for investigating membership inference. Hence, we consider the binary classification with input $\mathbf{x}$ and label $\mathbf{y} \in \{-1,+1\}$. A binary classifier $\mathfrak{M}$, which is the subject of the membership inference attack, is trained on the training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$. The classifier is trained to output the confidence for belonging to the positive class. We assume that $\mathfrak{M}(\mathbf{x}) \in \mathcal{P}$, where $\mathcal{P}$ is a finite set (i.e., quantized confidence levels) to avoid issues pertaining to mixtures of discrete and continuous random variables.

An adversary wants to determine whether some data point $(\mathbf{x}, \mathbf{y})$ belongs to the training set of the classifier or not. This is referred to as membership inference attack (MIA). Similar to [5], we assume that the adversary has black-box access to the classifier, i.e., the adversary can only request classification queries. Random variable $\mathbf{m} \in \{0,1\}$ denotes whether $(\mathbf{x}, \mathbf{y})$ belongs to the training dataset or not.

Membership inference [26], [27], tracing attack [28], and privacy [29] studies sometimes use distinguishability cryptographic games to evaluate the ability of the adversary. The game is played between the adversary and a challenger (a fictitious character). The challenger flips a fair coin to realize random variable $\mathbf{m}$. If $\mathbf{m} = 1$, the challenger selects $(\mathbf{x}, \mathbf{y})$ arbitrarily from the training dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$. On the other hand, if $\mathbf{m} = 0$, the adversary selects $(\mathbf{x}, \mathbf{y})$ independently from the training dataset but from the same distribution. The adversary needs to distinguish the data points, provided by the challenger, that are from the training dataset. The adversary's guess is denoted by $\widehat{\mathbf{m}}$.

**Assumption 1.** *The following assumptions hold:*

1) $\mathbb{P}\{\mathbf{m} = 1\} = \mathbb{P}\{\mathbf{m} = 0\}$*;*
2) $\mathbb{P}\{\mathbf{y} = y | \mathbf{m} = 1\} = \mathbb{P}\{\mathbf{y} = y | \mathbf{m} = 0\}$*,* $\forall y \in \{-1,+1\}$*.*

Assumption 1.1 indicates that the adversary does not have any prior knowledge on whether $(\mathbf{x}, \mathbf{y})$ is included in the training dataset, c.f. [5]. This means that, in the distinguishability game introduced above, the challenger uses a fair coin for realizing $\mathbf{m}$ and the adversary knows this. Assumption 1.2 implies that the challenger does not favor any class above the other when selecting $(\mathbf{x}, \mathbf{y})$. Therefore, the class does not possess any information about the membership. In the next definition, we use the notation $[0,\infty]$ to denote $[0,\infty) \cup \{+\infty\}$.

**Definition 1** ($\alpha$-Divergence MIA Privacy Loss)**.** *For* $\alpha \in [0,\infty]$*, the* $\alpha$*-divergence MIA information leakage is given by*

$$\Gamma_\alpha(\mathfrak{M}) := \sup_{\substack{y \in \{-1,+1\}, \\ m, m' \in \{0,1\}, \\ m \neq m'}} D_\alpha(P_{y,m} \| P_{y,m'}), \tag{1}$$

*where* $P_{y,m}$ *and* $P_{y,m'}$ *are, respectively, the probability distribution of* $\mathfrak{M}(\mathbf{x})$ *given* $\mathbf{y} = y, \mathbf{m} = m$ *and* $\mathbf{y} = y, \mathbf{m} = m'$*, i.e.,* $P_{\mathfrak{M}(\mathbf{x})|\mathbf{y},\mathbf{m}}(\cdot|y,m)$ *and* $P_{\mathfrak{M}(\mathbf{x})|\mathbf{y},\mathbf{m}}(\cdot|y,m')$*.*

In [5], the following worst-case notion of MIA information leakage was introduced:

$$L^{\mathrm{MIA}} := \sup_{\substack{p \in \mathcal{P}, \\ y \in \{-1,+1\}}} \left| \log\left( \frac{P_{y,1}(p)}{P_{y,0}(p)} \right) \right|.$$

We can easily see that $L^{\mathrm{MIA}} = \Gamma_\infty(\mathfrak{M})$. Hence, $\Gamma_\infty(\mathfrak{M}) = 0$ implies perfect MIA-indistinguishability in the sense of [5]. Since $\Gamma_\alpha(\mathfrak{M})$ is non-decreasing in $\alpha$, $\Gamma_\infty(\mathfrak{M}) = 0$ implies that $\Gamma_\alpha(\mathfrak{M}) = 0$ for all $\alpha \in [0,\infty)$. In the next theorem, we show that the $\alpha$-divergence MIA information leakage is also related to the adversary's ability to correctly distinguish the data points belonging to the training dataset in the distinguishability game described earlier.

**Theorem 1.** *For any* $\alpha \in [0,1]$*,*

$$\inf_{\widehat{\mathbf{m}}} \mathbb{P}\{\widehat{\mathbf{m}} \neq \mathbf{m}\} \leq \frac{1}{2}\exp((\alpha-1)\Gamma_\alpha(\mathfrak{M})) \tag{2}$$

*Proof.* The proof follows from the application of Theorem 1 in [18] or Theorem 13 in [17]. □

The upper bound for the probability of error $\inf_{\widehat{\mathbf{m}}} \mathbb{P}\{\widehat{\mathbf{m}} \neq \mathbf{m}\}$ in Theorem 1 is a decreasing function of $\Gamma_\alpha(\mathfrak{M})$ for $\alpha \in [0,1]$. This implies that $\mathbb{P}\{\widehat{\mathbf{m}} \neq \mathbf{m}\}$ can be made small for large information leakage $\Gamma_\alpha(\mathfrak{M})$. Hence, in order to make the adversary's task harder, we need to make sure that $\Gamma_\alpha(\mathfrak{M})$ is small for $\alpha \in [0,1]$. Noting that $\Gamma_\alpha(\mathfrak{M})$ is non-decreasing in $\alpha$, therefore, we may use $\Gamma_\alpha(\mathfrak{M})$ for $\alpha \in (1,\infty]$ as upper bound of $\Gamma_\alpha(\mathfrak{M})$ for $\alpha \in [0,1]$. Another way to measure membership information leakage is to use Arimoto $\alpha$-information. This is pursued in the next definition.

**Definition 2** ($\alpha$-Information MIA Privacy Loss)**.** *For* $\alpha \in [0,\infty]$*, the* $\alpha$*-information MIA information leakage is given by*

$$\Xi_\alpha(\mathfrak{M}) := I_\alpha(\mathbf{m}; \mathbf{y}, \mathfrak{M}(\mathbf{x})). \tag{3}$$

In the next theorem, we show that the $\alpha$-information MIA information leakage is in fact related to the adversary's ability to correctly distinguish the data points belonging to the training dataset in the distinguishability game.

**Theorem 2.** *For all* $\alpha \in (1,\infty)$*,*

$$\inf_{\widehat{\mathbf{m}}} \mathbb{P}\{\widehat{\mathbf{m}} \neq \mathbf{m}\} \geq 1 - \exp\left( \frac{1-\alpha}{\alpha} H_\alpha(\mathbf{m}|\mathbf{y}, \mathfrak{M}(\mathbf{x})) \right). \tag{4}$$

*Proof.* The proof follows from the application of Theorem 7 in [17]. □

Noting that $\inf_{\widehat{\mathbf{m}}} \mathbb{P}\{\widehat{\mathbf{m}} \neq \mathbf{m}\} = \inf_{\widehat{\mathbf{m}}}(1 - \mathbb{P}\{\widehat{\mathbf{m}} = \mathbf{m}\}) = 1 - \sup_{\widehat{\mathbf{m}}} \mathbb{P}\{\widehat{\mathbf{m}} = \mathbf{m}\}$, Theorem 2 shows that

$$\sup_{\widehat{\mathbf{m}}} \mathbb{P}\{\widehat{\mathbf{m}} = \mathbf{m}\} \leq \exp\left( \frac{1-\alpha}{\alpha} H_\alpha(\mathbf{m}|\mathbf{y}, \mathfrak{M}(\mathbf{x})) \right).$$

This shows that the upper bound for the probability of success by the adversary, i.e., $\mathbb{P}\{\widehat{\mathbf{m}} = \mathbf{m}\}$, is a decreasing function of $H_\alpha(\mathbf{m}|\mathbf{y}, \mathfrak{M}(\mathbf{x}))$. This shows that the larger $H_\alpha(\mathbf{m}|\mathbf{y}, \mathfrak{M}(\mathbf{x}))$, the more difficult the adversary's task is to correctly distinguish the data points belonging to the training dataset in the distinguishability game. Also, by definition of Arimoto's information, the lower bound for the probability of error is a decreasing function of the information $I_\alpha(\mathbf{m};\mathbf{y},\mathfrak{M}(\mathbf{x})) = H_\alpha(\mathbf{m}) - H_\alpha(\mathbf{m}|\mathbf{y},\mathfrak{M}(\mathbf{x}))$. Therefore, the $\alpha$ information is a better indicator of MIA information leakage or information leakage.

These two notions of MIA information leakage, i.e., $\alpha$-divergence MIA information leakage and $\alpha$-information MIA information leakage are in fact related to each other. This is shown in the next theorem.

**Theorem 3.** *For $\alpha \in [1,\infty]$, $\Xi_{1/\alpha}(\mathfrak{M}) \leq \Xi_1(\mathfrak{M}) \leq \Gamma_1(\mathfrak{M}) \leq \Gamma_\alpha(\mathfrak{M})$.*

*Proof.* For $\alpha \in [1,\infty]$, $H_{1/\alpha}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}) \geq H(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})$. Further, $H_{1/\alpha}(\mathbf{m}) = 1 = H(\mathbf{m})$. Hence,

$$\begin{aligned}
\Xi_{1/\alpha}(\mathfrak{M}) =& I_{1/\alpha}(\mathbf{m};\mathbf{y},\mathfrak{M}(\mathbf{x}))\\
\leq& I(\mathbf{m};\mathbf{y},\mathfrak{M}(\mathbf{x}))\\
=& \mathbb{E}\left\{\log\left(\frac{P_{\mathbf{m},\mathbf{y},\mathfrak{M}(\mathbf{x})}(\mathbf{m},\mathbf{y},\mathfrak{M}(\mathbf{x}))}{P_{\mathbf{m}}(\mathbf{m})P_{\mathbf{y},\mathfrak{M}(\mathbf{x})}(\mathbf{y},\mathfrak{M}(\mathbf{x}))}\right)\right\}\\
=& \mathbb{E}\left\{\log\left(\frac{P_{\mathbf{y},\mathfrak{M}(\mathbf{x})|\mathbf{m}}(\mathbf{y},\mathfrak{M}(\mathbf{x})|\mathbf{m})}{P_{\mathbf{y},\mathfrak{M}(\mathbf{x})}(\mathbf{y},\mathfrak{M}(\mathbf{x}))}\right)\right\}\\
=& \mathbb{E}\left\{\log\left(\frac{P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})P_{\mathbf{y}|\mathbf{m}}(\mathbf{y}|\mathbf{m})}{P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{y}}(\mathbf{y})}\right)\right\}\\
=& \mathbb{E}\left\{\log\left(\frac{P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})}{P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{m}}(\mathbf{m})}\right)\right\}.
\end{aligned}$$

Now, note that

$$\begin{aligned}
&\mathbb{E}\left\{\log\left(\frac{P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})}{P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{m}}(\mathbf{m})}\right)\middle|\mathbf{m},\mathbf{y}\right\}\\
&= \mathrm{KL}(P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})\|P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{m}}(\mathbf{m})).
\end{aligned}$$

Noting that KL divergence is jointly convex and that

$$\begin{aligned}
&P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{m}}(\mathbf{m})\\
&\qquad=P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})P_{\mathbf{m},\mathbf{y}}(0,\mathbf{y})\\
&\qquad\quad+P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})P_{\mathbf{m},\mathbf{y}}(1,\mathbf{y}),
\end{aligned}$$

we get

$$\begin{aligned}
&\mathbb{E}\left\{\log\left(\frac{P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})}{P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{m}}(\mathbf{m})}\right)\middle|\mathbf{m},\mathbf{y}\right\}\\
\leq&\mathrm{KL}(P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})\|P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|0,\mathbf{y}))\\
&\quad\times P_{\mathbf{m},\mathbf{y}}(0,\mathbf{y})\\
&+\mathrm{KL}(P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})\|P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|1,\mathbf{y}))\\
&\quad\times P_{\mathbf{m},\mathbf{y}}(1,\mathbf{y})\\
\leq&\Gamma_1(\mathfrak{M})(P_{\mathbf{m},\mathbf{y}}(0,\mathbf{y})+P_{\mathbf{m},\mathbf{y}}(1,\mathbf{y}))\\
=&\Gamma_1(\mathfrak{M})P_{\mathbf{y}}(\mathbf{y}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\left\{\log\left(\frac{P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})}{P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{m}}(\mathbf{m})}\right)\right\}\\
&\quad=\mathbb{E}\left\{\mathbb{E}\left\{\log\left(\frac{P_{\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{m},\mathbf{y})}{P_{\mathfrak{M}(\mathbf{x})|\mathbf{y}}(\mathfrak{M}(\mathbf{x})|\mathbf{y})P_{\mathbf{m}}(\mathbf{m})}\right)\middle|\mathbf{m},\mathbf{y}\right\}\right\}\\
&\quad\leq\Gamma_1(\mathfrak{M})\mathbb{E}\{P_{\mathbf{y}}(\mathbf{y})\}\\
&\quad\leq\Gamma_1(\mathfrak{M}).
\end{aligned}$$

Finally, we conclude the proof by noting the relationship $\Gamma_1(\mathfrak{M}) \leq \Gamma_\alpha(\mathfrak{M})$. □

In the remainder of this section, we investigate post-processing the outcome of the machine learning model $\mathcal{M}(\mathbf{x})$ to make MIA more difficult. Hence, the output of the machine learning model is obfuscated by the mapping $\mathfrak{P}:\mathcal{P}\rightarrow\mathcal{P}$. This implies that, the ML model reports $\mathfrak{P}\circ\mathcal{M}(\mathbf{x})$.

**Theorem 4.** *For all $\alpha[0,\infty]$, the following holds:*

1) *$\Gamma_\alpha(\mathfrak{P}\circ\mathfrak{M}) \leq \Gamma_\alpha(\mathfrak{M})$;*
2) *$\Xi_\alpha(\mathfrak{P}\circ\mathfrak{M}) \leq \Xi_\alpha(\mathfrak{M})$.*

*Proof.* The proof for the first part follows from the data processing inequality Rényi $\alpha$-divergence [21] and the proof for the second part follows from the data processing inequality for Arimoto-Rényi conditional entropy [30]. □

Theorem 4 shows that post processing or obfuscation cannot increase MIA information leakage (in either sense). We can cast the problem finding the optimal mapping $\mathfrak{P}$ as an optimization problem with a bound on the performance of the obfuscated ML model. This is however an integer problem due to the structure of $\mathfrak{P}$ and can be computationally cumbersome. If we allow randomization at the ML output, this problem becomes more computationally friendly specially noting that $\alpha$-information MIA information leakage is quasi-convex, see Corollary 6 in the next section.

## IV. Relationship with the $\alpha$-Leakage

Recently, maximal leakage was introduced as an operational measure of information leakage for privacy analysis [31]. This measure was defined following an axiomatic framework for measuring information leakage requiring minimal assumptions, being interpretability, and satisfying data-processing, independence, and additivity properties. Maximal leakage was recently generalized to a family of leakage that can be fine-tuned to specific applications [19], [20]. In this pursuit, $\alpha$-leakage (when the adversary's target is known) and maximal $\alpha$-leakage (when the adversary's target is not known) was introduced. Since, in this paper, we know that the adversary aims at guessing membership (i.e., correctly guessing $\mathbf{m}$), $\alpha$-leakage is an appropriate measure of information leakage. In this section, we investigate the relationship between this tunable measure of information leakage and MIA information leakage.

**Definition 3** ($\alpha$-Leakage)**.** *For $\alpha \in (1,\infty)$, the $\alpha$-leakage from $\mathbf{m}$ to $(\mathbf{y},\mathfrak{M}(\mathbf{x}))$ is defined as*

$$\begin{aligned}&\mathcal{L}_\alpha(\mathbf{m}\to(\mathbf{y},\mathfrak{M}(\mathbf{x})))\\ &:= \frac{\alpha}{\alpha-1}\log\left(\frac{\max\limits_{P_{\widehat{\mathbf{m}}|\mathbf{y},\mathfrak{M}(\mathbf{x})}} \mathbb{E}\{P_{\widehat{\mathbf{m}}|\mathbf{y},\mathfrak{M}(\mathbf{x})}(\mathbf{m}|\mathbf{y},\mathfrak{M}(\mathbf{x}))\}^{\frac{\alpha-1}{\alpha}}}{\max\limits_{P_{\widehat{\mathbf{m}}}}\mathbb{E}\{P_{\widehat{\mathbf{m}}}(\mathbf{m})\}^{\frac{\alpha-1}{\alpha}}}\right),\end{aligned}$$

*where $\widehat{\mathbf{m}}$ is an estimate of $\mathbf{m}$ with the same support set. The definition for $\alpha = 1$ and $\alpha = \infty$ is done by continuous extension.*

Here, $\alpha$-leakage measures the improvement in the adversary's ability to guess realization of $\mathbf{m}$ with access to the realizations of $\mathbf{y}$ and $\mathfrak{M}(\mathbf{x})$.

**Theorem 5.** *For $\alpha \in [1,\infty]$, $\Xi_\alpha(\mathfrak{M}) = \mathcal{L}_\alpha(\mathbf{m}\to(\mathbf{y},\mathfrak{M}(\mathbf{x})))$.*

*Proof.* Following [19, Theorem 1], we know that $\mathcal{L}_\alpha(\mathbf{m}\to(\mathbf{y},\mathfrak{M}(\mathbf{x}))) = I_\alpha(\mathbf{m};(\mathbf{y},\mathfrak{M}(\mathbf{x})))$. The rest follows from the definition of $\Gamma_\alpha(\mathfrak{M})$ and Theorem 3. □

**Corollary 6.** *For $\alpha \in [1,\infty]$, $\Xi_\alpha(\mathfrak{M})$ is quasi-convex in $P_{\mathfrak{M}(\mathbf{x})|\mathbf{y},\mathbf{m}}$.*

*Proof.* According to [19, § IV], $\Xi_\alpha(\mathfrak{M})$ is quasi-convex in $P_{\mathfrak{M}(\mathbf{x}),\mathbf{y}|\mathbf{m}}$. This follows from the applications of the results of [32, § 3.5]. Note that $P_{\mathfrak{M}(\mathbf{x}),\mathbf{y}|\mathbf{m}} = P_{\mathfrak{M}(\mathbf{x})|\mathbf{y},\mathbf{m}}P_{\mathbf{y}|\mathbf{m}}$. □

## V. Differential Privacy in MIA

The ML training algorithm is denoted by $\mathcal{A}$, i.e., $\mathcal{A}:\mathcal{D}\mapsto\mathfrak{M}$. For instance, $\mathcal{A}$ can denote an stochastic gradient algorithm with privacy-preserving noise added to the gradients used for training ML models.

**Definition 4** (Differential Privacy)**.** *A training algorithm $\mathcal{A}$ is $(\epsilon,\delta)$-differentially private if for all training datasets $\mathcal{D}$ and $\mathcal{D}'$ that differ in only one entry and all measurable sets $\overline{\mathcal{M}} \subseteq \mathcal{M}$,*

$$\mathbb{P}\{\mathcal{A}(\mathcal{D}')\in\overline{\mathcal{M}}\} \leq \exp(\epsilon)\mathbb{P}\{\mathcal{A}(\mathcal{D})\in\overline{\mathcal{M}}\} + \delta,$$

*where $\mathcal{M}$ is the set of all ML models. We refer to $(\epsilon,0)$-differentially privacy as $\epsilon$-differentially privacy.*

**Remark 1** (Parameterized ML Models)**.** *Most often, we deal with parameterized ML models, such as deep neural networks, regression models, and support vector machines. For instance, in logistic regression, the model $\mathfrak{M}(x) = 1/(1+\exp(-\theta^\top x))$ is parameterized by $\theta\in\mathbb{R}^{p_x}$, where $p_x$ is the dimension of the input vector. Therefore, the set of all ML models $\mathcal{M}$ is isomorphic to $\mathbb{R}^{p_x}$.*

**Theorem 7.** *Let $\mathcal{A}$ be $\epsilon$-differentially private. For all $\alpha \in [0,1)$, $\Gamma_\alpha(\mathcal{A}(\mathcal{D})) \leq \log(\max(2-\exp(\epsilon),0))/(\alpha-1)$.*

*Proof.* First, note that $\inf_{\widehat{\mathbf{m}}}\mathbb{P}\{\widehat{\mathbf{m}}\neq\mathbf{m}\} = \inf_{\widehat{\mathbf{m}}}(1-\mathbb{P}\{\widehat{\mathbf{m}}=\mathbf{m}\}) = 1-\sup_{\widehat{\mathbf{m}}}\mathbb{P}\{\widehat{\mathbf{m}}=\mathbf{m}\} \leq \exp((\alpha-1)\Gamma_\alpha(\mathcal{A}(\mathcal{D})))/2$. Theorem 1 in [33] implies that $\exp(\epsilon)-1 \geq 2\sup_{\widehat{\mathbf{m}}}\mathbb{P}\{\widehat{\mathbf{m}}=\mathbf{m}\}-1$ and therefore, $\exp(\epsilon)-1 \geq 1-\exp((\alpha-1)\Gamma_\alpha(\mathcal{A}(\mathcal{D})))$. Noting that $\exp((\alpha-1)\Gamma_\alpha(\mathcal{A}(\mathcal{D})))$ cannot be smaller than zero, we get $\exp((\alpha-1)\Gamma_\alpha(\mathcal{A}(\mathcal{D}))) \geq \max(2-\exp(\epsilon),0)$. Taking logarithm from both sides proves the results. □

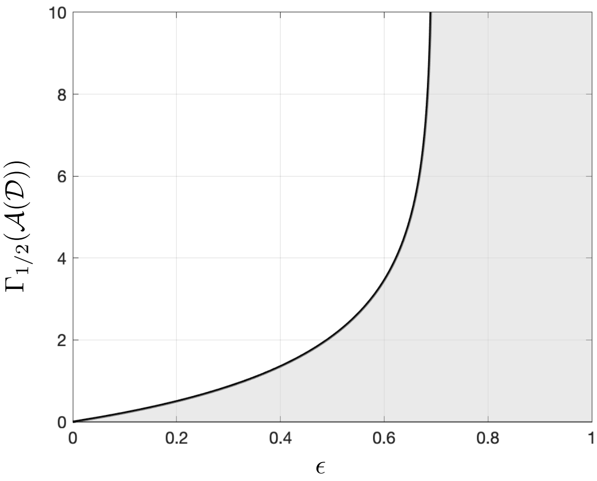

Fig. 1. The upper bound of $\Gamma_{1/2}(\mathcal{A}(\mathcal{D}))$ for $\epsilon$-differentially private $\mathcal{A}$.

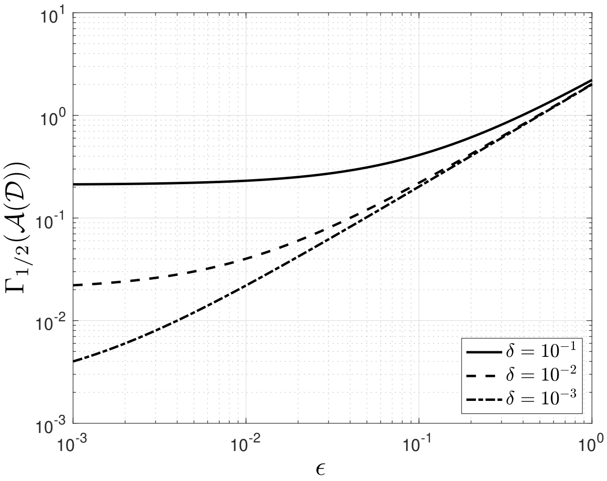

Fig. 2. The upper bound of $\Gamma_{1/2}(\mathcal{A}(\mathcal{D}))$ for $\epsilon$-differentially private $\mathcal{A}$.

**Theorem 8.** *Let $\mathcal{A}$ be $(\epsilon,\delta)$-differentially private. For all $\alpha \in [0,1)$, $\Gamma_\alpha(\mathcal{A}(\mathcal{D})) \leq (-\epsilon+\log(1-\delta))/(\alpha-1)$.*

*Proof.* Proposition 2 in [34] implies that $2\sup_{\widehat{\mathbf{m}}}\mathbb{P}\{\widehat{\mathbf{m}}=\mathbf{m}\}-1 \leq 1-\exp(-\epsilon)(1-\delta)$. Therefore, $1-\exp(-\epsilon)(1-\delta) \geq 1-\exp((\alpha-1)\Gamma_\alpha(\mathcal{A}(\mathcal{D})))$, and, as a result, $\exp((\alpha-1)\Gamma_\alpha(\mathcal{A}(\mathcal{D}))) \geq \exp(-\epsilon)(1-\delta)$. Taking logarithm from both sides proves the results. □

Figure 1 shows the upper bound of $\Gamma_{1/2}(\mathcal{A}(\mathcal{D}))$ for $\epsilon$-differentially private $\mathcal{A}$. The gray area under the upper bound shown by the solid black curve is where $\Gamma_{1/2}(\mathcal{A}(\mathcal{D}))$ resides. As $\epsilon$ tends to zero, MIA information leakage vanishes. Therefore, by using small privacy budgets $\epsilon$, we can effectively combat MIA. This however comes at the cost of significant utility degradation. The upper bound is not tight (or useful) for large $\epsilon$ as it rapidly tends to infinity at $\log(2)$. Figure 2 shows the upper bound of $\Gamma_{1/2}(\mathcal{A}(\mathcal{D}))$ for $(\epsilon,\delta)$-differentially private $\mathcal{A}$. This bound is tighter specially for small $\delta$.

## VI. Conclusions and Future Work

We proposed two measures of information leakage for investigating MIA using Rényi $\alpha$-divergence and Arimoto-Rényi $\alpha$-information. We established an upper bound for $\alpha$-divergence information leakage as a function of the privacy budget for differentially-private machine learning models. Future work can focus on investigating the effects of fairness, over-fitting, and memorization in MIA attacks.

## References

[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 3–18, IEEE, 2017.

[2] S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei, "Towards demystifying membership inference attacks," *arXiv preprint arXiv:1807.09173*, 2018.

[3] A. Salem, Y. Zhang, M. Humbert, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019)*, 2019.

[4] A. Sablayrolles, M. Douze, Y. Ollivier, C. Schmid, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[5] M. Yaghini, B. Kulynych, and C. Troncoso, "Disparate vulnerability: On the unfairness of privacy attacks against machine learning," *arXiv preprint arXiv:1906.00389*, 2019.

[6] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, IEEE, 2018.

[7] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 1, pp. 133–152, 2019.

[8] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against GANs," *arXiv preprint arXiv:1909.03935*, 2019.

[9] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 4, pp. 232–249, 2019.

[10] G. Liu, C. Wang, K. Peng, H. Huang, Y. Li, and W. Cheng, "SocInf: Membership inference attacks on social media health data with machine learning," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 5, pp. 907–921, 2019.

[11] M. Backes, P. Berrang, M. Humbert, and P. Manoharan, "Membership privacy in microrna-based studies," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 319–330, ACM, 2016.

[12] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646, ACM, 2018.

[13] F. Farokhi and M. A. Kaafar, "Modelling and quantifying membership information leakage in machine learning," *arXiv preprint arXiv:2001.10648*, 2020.

[14] Y. Kaya, S. Hong, and T. Dumitras, "On the effectiveness of regularization against membership inference attacks," *arXiv preprint arXiv:2006.05336*, 2020.

[15] M. A. Rahman, T. Rahman, R. Laganiere, N. Mohammed, and Y. Wang, "Membership inference attack against differentially private deep learning model," *Transactions on Data Privacy*, vol. 11, no. 1, pp. 61–79, 2018.

[16] K. Leino and M. Fredrikson, "Stolen memories: Leveraging model memorization for calibrated white-box membership inference," *arXiv preprint arXiv:1906.11798*, 2019.

[17] I. Sason and S. Verdú, "Arimoto–Rényi conditional entropy and bayesian $m$-ary hypothesis testing," *IEEE Transactions on Information Theory*, vol. 64, no. 1, pp. 4–25, 2017.

[18] M. Hellman and J. Raviv, "Probability of error, equivocation, and the chernoff bound," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 368–372, 1970.

[19] J. Liao, O. Kosut, L. Sankar, and F. du Pin Calmon, "Tunable measures for information leakage and applications to privacy-utility tradeoffs," *IEEE Transactions on Information Theory*, vol. 65, no. 12, pp. 8043–8066, 2019.

[20] J. Liao, O. Kosut, L. Sankar, and F. P. Calmon, "A tunable measure for information leakage," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 701–705, IEEE, 2018.

[21] T. Van Erven and P. Harremos, "Rényi divergence and Kullback-Leibler divergence," *IEEE Transactions on Information Theory*, vol. 60, no. 7, pp. 3797–3820, 2014.

[22] A. Rényi, "On measures of entropy and information," in *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 1961.

[23] S. Arimoto, "Information measures and capacity of order $\alpha$ for discrete memoryless channels," in *Proceedings of the 2nd Colloquium on Topics on Information Theory, Keszthely, Hungary*, vol. 16, p. 1975.

[24] S. Verdú, "$\alpha$-mutual information," in *2015 Information Theory and Applications Workshop (ITA)*, pp. 1–6, 2015.

[25] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, 1969.

[26] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "Knock knock, who's there? membership inference on aggregate location data," *arXiv preprint arXiv:1708.06145*, 2017.

[27] A. Pyrgelis, C. Troncoso, and E. De Cristofaro, "Under the hood of membership inference attacks on aggregate location time-series," *arXiv preprint arXiv:1902.07456*, 2019.

[28] N. Buescher, S. Boukoros, S. Bauregger, and S. Katzenbeisser, "Two is not enough: Privacy assessment of aggregation schemes in smart metering," *Proceedings on Privacy Enhancing Technologies*, vol. 2017, no. 4, pp. 198–214, 2017.

[29] J.-M. Bohli, C. Sorge, and O. Ugus, "A privacy model for smart metering," in *2010 IEEE International Conference on Communications Workshops*, pp. 1–5, 2010.

[30] S. Fehr and S. Berens, "On the conditional Rényi entropy," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6801–6810, 2014.

[31] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, 2020.

[32] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[33] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018.

[34] Ú. Erlingsson, I. Mironov, A. Raghunathan, and S. Song, "That which we call private," *arXiv preprint arXiv:1908.03566*, 2019.

Author/s:
Farokhi, F

Title:
Using Renyi-divergence and Arimoto-Renyi Information to Quantify Membership Information Leakage

Date:
2021-01-01

Citation:
Farokhi, F. (2021). Using Renyi-divergence and Arimoto-Renyi Information to Quantify Membership Information Leakage. 2021 55TH ANNUAL CONFERENCE ON INFORMATION SCIENCES AND SYSTEMS (CISS), 00, IEEE. https://doi.org/10.1109/CISS50987.2021.9400316.