Privacy Enhancing Technologies
Lecture 12 – Membership Inference Attacks

Florian Tramèr

AGENDA

1. Recap
2. Reconstruction vs Distinguishing Attacks
3. Membership Inference Attacks
4. DP Lower-Bounds from Membership Inference

1 Recap on Data Privacy

We have started our discussion of data privacy by showing how things could go very, very wrong: if we answer too many queries too precisely, an adversary can *reconstruct* a huge chunk of the database... We definitely didn't want that, and so we introduced a formal notion of privacy, differential privacy, and showed how to achieve it by appropriately perturbing our queries. We now have a way to answer queries while defeating these reconstruction attacks.

2 From Reconstruction Attacks to Distinguishing Attacks

What type of attacks does DP actually protect against? We'll start by showing that if all that we care about is to make blatant reconstruction attacks infeasible, then we don't really need DP. At the same time, the mechanism we'll introduce is not "private" at all. This motivates us to consider *weaker attacks* than reconstruction attacks, that capture the protections offered by DP more accurately.

Defeating reconstruction through sampling. Consider a very simple mechanism: on input a database D of size n, create a new database X by sampling $n/2$ individuals uniformly at random from D. Then answer all queries using X.

This mechanism trivially defeats blatant reconstruction attacks: the attacker can't hope to reconstruct more than $n/2$ individuals from D, and has to resort to guessing for the rest. So there's very little chance that the attacker's guess would match D in all but $o(n)$ entries.

This mechanism is also quite useful: if we want to estimate the mean of our database D, then releasing the mean over X is a good estimate, with error on the order of $1/\sqrt{n}$.

And yet, this mechanism does not really help "privacy" (in particular, it certainly is not differentially private). If you're unlucky and your data happens to be in the sample X, the mechanism provides no protection whatsoever.

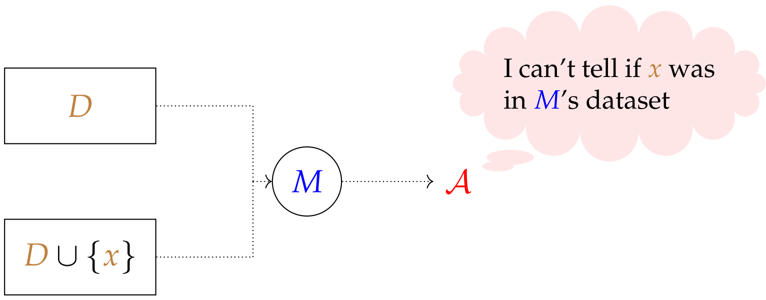
So the type of attacks that DP protects against is somehow broader than full reconstruction. What exactly are those attacks?

A detour through cryptography. Consider how we define privacy in cryptography. In many cases, the privacy definition is directly linked to an attack. For example, we can define security of encryption schemes by saying that no adversary can distinguish the ciphertexts of two messages (i.e., IND-CPA security).

Notice that this is *not* a reconstruction attack. We could have also defined security of an encryption scheme by asking for *plaintext recovery* attacks to be infeasible. But this is not strong enough.[1] Instead, we define privacy in terms of a *distinguishing attack*: an attack cannot even recover a single bit of the plaintext to distinguish it from another plaintext.

Differential privacy as hardness of inferring membership. What would an analogous attack look like for differential privacy? The definition of differential privacy does not directly refer to an adversary, or to an attack—primarily because it is an information-theoretic notion.

But differential privacy does have an indistinguishability flavor: it says that the distribution of outputs of the mechanism on database D is close to the distribution of outputs on a neighboring database D'. Informally speaking, this is equivalent to saying that no adversary (in this case even a computationally unbounded one) can easily distinguish whether some output was generated by the mechanism on D or on D'. And since D and D' only differ in one entry[2] (say the presence of some individual x), this is equivalent to saying that no adversary can distinguish whether x was in the database or not. This is what we call a *membership inference attack*.



3 Membership Inference Attacks

Membership inference attacks were originally introduced in an influential empirical paper by Homer et al. [HSR+08], which showed that it may be possible to infer an individual's presence in a medical study from aggregate genetic statistics. These attacks were later used to derive formal upper bounds on differential privacy [BUV14, DSS+15], in settings where reconstruction attacks do not yield tight bounds. In recent years, membership inference attacks have seen a lot of study in machine learning [SSSS17, CCN+22], where they serve as an empirical test for privacy leakage [JUO20, NST+21].

The membership inference game. We can setup the problem of membership inference as a game between an *adversary* and a *challenger*. In its strictest form, the game might look

[1] E.g., consider a scheme that encrypts half of the bits of the plaintext with a one-time-pad, and leaves the rest in the clear. Full plaintext recovery is impossible, but the scheme is clearly insecure. This is somewhat analogous to the sampling mechanism we considered above, which defeats reconstruction attacks, but doesn't help privacy.

[2] We consider here the definition of neighboring databases where the two databases differ in the presence of a single individual x. So D and D' have different size (i.e., $|D'| = |D| + 1$).

like this: the adversary chooses a database D and a data point x, and the challenger runs a mechanism M (e.g., a machine learning training algorithm) on one of two neighboring databases D or $D \cup \{x\}$. The adversary then tries to guess which database the mechanism was run on.

Membership Inference Game		
Adversary		**Challenger**
	$\xrightarrow{D,\, x}$	$D_0 = D, D_1 = D \cup \{x\}$
	$\xleftarrow{M(D_b)}$	$b \leftarrow^{\$} \{0,1\}$
output $b' \in \{0,1\}$		
Adversary wins if $b' = b$		

This is a "worst-case" variant of the game, where the adversary freely gets to choose both the database and the target data point. There are many weaker variants of this game in the research literature, where the database might instead by sampled from some distribution, and the target data point might also be randomly sampled from this distribution.

As we will see, the worst-case variant is very closely tied to the definition of differential privacy. For the distributional variants, it is a bit harder to characterize what privacy notion is captured, but they are useful nonetheless.

Success metrics. How should we measure the success of an adversary in the membership inference game? This is actually a bit subtle, and one of the reasons why the research literature on membership inference attacks has considered so many different variants.

As a first attempt, we might take inspiration from cryptography and define the advantage of the adversary as

$$\mathrm{Adv}_{\mathrm{MI}}(\mathcal{A}) = \left|\Pr[b' = b] - 1/2\right| .$$

This is how we would measure the success of the adversary in a typical cryptographic game, e.g., an IND-CPA game. In cryptography, such a success metric is enough, because we typically want this to be negligibly small.

But in our setting there, the adversary's advantage must always be non-negligible if we want the mechanism M to be useful (can you see why?). And so a more granular analysis of the attacker's success rate can be useful.

One way is to take a closer look at the attacker's "confusion matrix", i.e., the rates at which the adversary guesses correctly or incorrectly whether the mechanism used the target data point x or not. Define

$$\begin{aligned} \mathrm{TPR} &= \Pr[b' = 1 \mid b = 1] & \qquad \mathrm{TNR} &= \Pr[b' = 0 \mid b = 0] \\ \mathrm{FPR} &= \Pr[b' = 1 \mid b = 0] & \qquad \mathrm{FNR} &= \Pr[b' = 0 \mid b = 1] \end{aligned}$$

as the true positive rate, true negative rate, false positive rate, and false negative rate, respectively (i.e., the probability that the adversary guesses correctly/incorrectly that x is/isn't a member of the database).

It is then also natural to measure the success of the adversary in the membership inference game via some comparison of the attack's positive and negative rates at predicting membership (or non-membership). It turns out that this view is directly connected to the definition of differential privacy:

Theorem 1 ([KOV15]). Let $\varepsilon > 0$ and $\delta \in [0,1]$. A mechanism M is (ε, δ)-DP *if and only if* for all adversaries in the membership inference game,

$$\begin{aligned} e^{\varepsilon} \cdot \text{FNR} \geq \text{TNR} - \delta \quad \text{and} \\ e^{\varepsilon} \cdot \text{FPR} \geq \text{TPR} - \delta \end{aligned} \tag{1}$$

So a mechanism being DP provides much stronger guarantees than just against data reconstruction attacks. It says that any adversary that tries to distinguish the presence or absence of one individual in the database will necessarily make true and false inferences with roughly the same probability.

Proof. We first prove that (ε, δ)-differential privacy implies eq. (1). We'll assume here for simplicity that the attacker is deterministic. Let S be the set of outputs of M for which the attacker outputs $b' = 0$. From the definition of differential privacy, we know that:

$$\underbrace{\Pr[M(D_0) \in S]}_{\text{TNR}} \leq e^{\varepsilon} \cdot \underbrace{\Pr[M(D_1) \in S]}_{\text{FNR}} + \delta.$$

This is the first inequality in (1). The second one follows similarly by considering the complement set $\bar{S}$ for which the attacker outputs $b' = 1$.

For the converse statement, assume that for any choice of set S, if the attacker outputs $b' = 0$ whenever the mechanism output is in S we get that eq. (1) holds. This implies that for all S, we have

$$\Pr[M(D_0) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D_1) \in S] + \delta\ .$$

If we instead consider sets S that make the attacker output $b' = 1$, we get

$$\Pr[M(D_1) \in S] \leq e^{\varepsilon} \cdot \Pr[M(D_0) \in S] + \delta\ .$$

Together, these imply (ε, δ)-differential privacy. □

Corollary 1. A mechanism M is (ε, δ)-DP *if and only if* for all adversaries in the membership inference game,

$$e^{\varepsilon} \geq \max\left(\frac{\text{TNR} - \delta}{\text{FNR}}, \frac{\text{TPR} - \delta}{\text{FPR}}\right)$$

So, for example, if an attacker has a TPR of 90% and an FPR of 1% (i.e., the attacker can correctly infer membership with 90% probability, and falsely accuses someone of membership with only 1% probability), then the mechanism is (at best) DP for $\varepsilon = \ln 90 \approx 4.5$. If the best attack is much weaker, and only achieves 5% TPR and 1% FPR, then the mechanism is ε-DP for $\varepsilon \geq \ln 5 \approx 1.6$.

Interpreting membership inference. How should we "think" about membership inference? On one hand, we can view it as a concrete attack that an adversary might try to mount against a system. This is how it was originally motivated in the context of genomic data [HSR+08], where an attacker might be able to determine that your genetic data was used in a study of some sensitive condition, say, a rare disease.

But this is a rather weak attack with a very strong pre-condition: we assume that the attacker already knows *all* your genetic data, and just needs to learn one extra bit of information, namely whether you have a specific disease or not. From a privacy standpoint, the fact that the attacker already knows all your genetic data is maybe the bigger issue.

So I tend to not think of membership inference as a concrete attack, but rather as a way to *lower-bound* the privacy of a system. That is, if we can show that a membership inference attack is successful, then we know there is *something* wrong with the system's privacy, even if the distinguishing attack itself is not that worrisome. This is a bit like a chosen ciphertext attack in cryptography: the ability to distinguish two plaintexts from their encryptions might not be a very strong attack in itself (although it can sometimes be used to mount much stronger key- or plaintext-recovery attacks); however, the fact that it is possible in the first place tells us that the encryption scheme is not good.

As we will see in this lecture and the next, this interpretation of membership inference can be very useful both in theory (to derive lower-bounds for differentially private mechanisms), and in practice (to empirically audit or debug privacy-preserving systems).

4 DP Lower-bounds from Membership Inference Attacks

We'll first show how to use membership inference attacks to prove lower-bounds on how many counting queries we can answer with the Gaussian mechanism.

Recall that our analysis of the Gaussian mechanism from lecture 10 showed that if we add noise of standard deviation $\sigma = \tilde{O}(\sqrt{k}/n)$ to each query, then we get DP with $\varepsilon = O(1)$. We now show that this is also necessary: if we add slightly less noise, the mechanism does not satisfy DP with a constant ε (and $\delta = o(1)$).

Theorem 2. Let M be a Gaussian mechanism for computing the mean of a database $D = \{\vec{x}_1, \dots, \vec{x}_n\}$ of n vectors from the domain $\{0,1\}^k$. I.e.,

$$M(D) = \frac{1}{n} \sum_{i=1}^{n} \vec{x}_i + \mathcal{N}(\vec{0}, \sigma^2 \cdot I_{k \times k})$$

Assume $\sigma = o(\sqrt{k}/n)$ and $\delta = o(1)$. Then this mechanism is <u>*not*</u> $(O(1), \delta)$-DP.

Proof. We're going to build a membership inference attack against this mechanism. Since the adversary gets to pick the database D and the target $\vec{x}$, let's consider the absolute worst case, where the adversary picks D to be the all-zero database, and $\vec{x}$ to be the all-one vector.

Our attack now proceeds as follows:

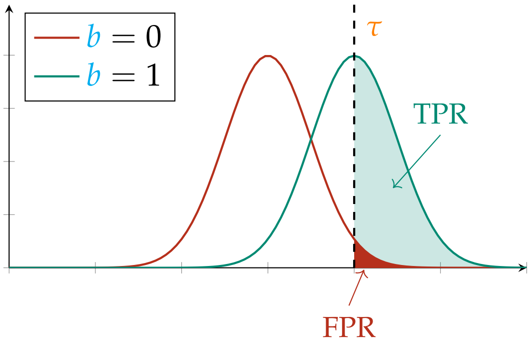
1. The adversary sends D and $\vec{x}$ to the challenger, and receives $\vec{y} = M(D_b)$.
2. The attacker computes the test statistic $z = \langle \vec{y}, \vec{x} \rangle$.
3. The attacker outputs 1 if $z \geq \tau$ and 0 otherwise, for some threshold τ to be chosen later.

Intuitively, the attacker tries to build a statistic z that takes on very different values depending on whether the mechanism was run on D_0 or D_1. Note that if $\vec{x}$ is not in the database, then the output of M is only noise, and so z is just the sum of k independent Gaussians

$\mathcal{N}(0, \sigma^2)$. And if $\vec{x}$ is in the database, then this sum of Gaussians is shifted by k/n (i.e., the value of $\langle \vec{x}, \vec{x}/n \rangle$). That is, we have:

$$z = \begin{cases} \sum_{i=1}^{k} \mathcal{N}(0, \sigma^2) = \mathcal{N}(0, k\sigma^2) & \text{if } b = 0 \\ \sum_{i=1}^{k} \mathcal{N}(1/n, \sigma^2) = \mathcal{N}(k/n, k\sigma^2) & \text{if } b = 1 \end{cases}$$

Here's how this would roughly look like:



Let's now fix a threshold $\tau = k/n$. We want to show that the attack can achieve high TPR at low FPR. So we'll lower-bound the TPR and upper-bound the FPR.

Computing the TPR is easy. By symmetry of the Gaussian, we have:

$$\text{TPR} = \Pr[z \geq \tau \mid b = 1] = \Pr[\mathcal{N}(k/n, k\sigma^2) \geq k/n] = 1/2 \,.$$

To upper-bound the FPR, we'll use a standard tail-bound on the Gaussian.

Let $X \sim \mathcal{N}(0, \sigma^2)$ and $t > 0$. Then:

$$\Pr[X \geq t] \leq \exp(-t^2/2\sigma^2)$$

So we get:

$$\begin{aligned} \text{FPR} &= \Pr[z \geq \tau \mid b = 0] = \Pr[\mathcal{N}(0, k\sigma^2) \geq k/n] \\ &\leq \exp\left(-\frac{(k/n)^2}{2k\sigma^2}\right) = o(1) \,, \end{aligned}$$

where we used that $\sigma = o(\sqrt{k}/n)$.

Using Corollary 1, we can now conclude that the mechanism's privacy budget ε is at least:

$$\varepsilon \geq \ln\left(\frac{\text{TPR} - \delta}{\text{FPR}}\right) = \ln\left(\frac{1/2 - o(1)}{o(1)}\right) = \omega(1) \,,$$

i.e., the mechanism is not DP for any constant ε. □

So we have shown that to answer k counting queries with the Gaussian mechanism, if we want a constant privacy budget ε, we need $\sigma = \Omega(\sqrt{k}/n)$, and thus error $\Omega(\sqrt{k}/n)$ per query. Using more involved arguments based on membership inference, one can show that this lower-bound applies to *any* mechanism that answers k counting queries, not just the Gaussian mechanism [BUV14, DSS+15].

References

[BUV14] Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 1–10, 2014.

[CCN$^+$22] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[DSS$^+$15] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

[HSR$^+$08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[JUO20] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. Auditing differentially private machine learning: How private is private SGD? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

[KOV15] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[NST$^+$21] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pages 866–882. IEEE, 2021.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.