14 Membership Inference Attacks

Membership inference attacks [Hom+08], also known as *tracing* or *fingerprinting* attacks, play a major role in the theory and practice of privacy in statistics and machine learning. In the simplest setting, a curator holds a dataset S consisting of records for individuals from some underlying population. The curator publishes some summary $\hat{\theta} = \mathcal{A}(S)$, such as a predictive or generative model trained on S, marginal tables for S, or synthetic data that mimics statistical properties of S. A membership inference attacker, given a data record Z and the published output $\hat{\theta}$, aims to determine if Z is a member of the dataset S or, alternatively, a random member of the population.

From the attacker's point of view, membership inference can be modeled as a hypothesis testing problem (Figure 8). We model the population as a probability distribution $\mathcal{D}$, and model the sample $S = \{X_1, \ldots, X_n\}$ as consisting of n records drawn independently and identically from $\mathcal{D}$. The target record Z is chosen in one of two ways: In the IN scenario, Z is a uniformly random record from S. In the OUT scenario, $Z = X_0$ is an independent sample from $\mathcal{D}$. The attacker (or tester) must distinguish these two possibilities. The attack's *false positive rate* (FPR) is the probability that it outputs "IN" in the OUT scenario, and the *true positive rate* (TPR) is the probability of saying "IN" in the IN scenario.

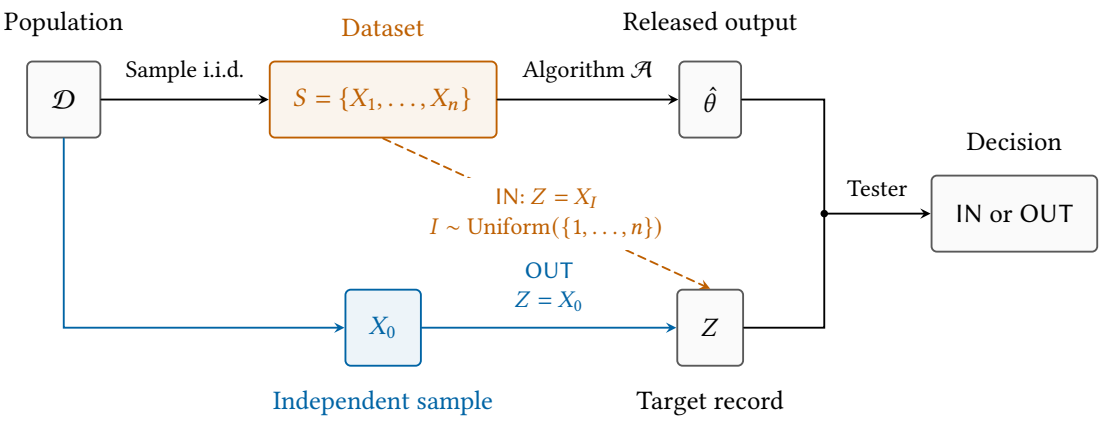


Figure 8: The basic membership inference setting. The tester observes the released output $\hat{\theta}$ and a target record Z. In the IN scenario, $Z = X_I$ for a uniformly random $I \in \{1, \ldots, n\}$; in the OUT scenario, $Z = X_0$ is drawn independently from $\mathcal{D}$. The two paths into Z represent alternative scenarios.

One might ask why membership inference is a privacy problem at all, especially given that the attacker already has the data record Z. First, the data may be split into groups based on some sensitive attribute, and learning membership in one of the groups is a clear privacy violation. For example, in the original work on membership inference [Hom+08, SOJH09], the data was split into a case group and a control group for a medical study, and membership in the case group indicates having a rare medical condition. Second, membership inference is the simplest meaningful privacy attack, and thus the easiest and most broadly applicable way to identify potential privacy issues that indicate more serious problems such as data reconstruction or training data extraction. As a result, membership inference is often used as a diagnostic tool for training data leakage in machine learning, either as an end goal [SSSS17] or to provide empirical lower bounds on the privacy parameters for differential privacy.