When Does Data Augmentation Help With Membership Inference Attacks? Supplementary Material

A Impact of N on the Membership Inference Attack Success

The MIAs we used assume that the adversary is in the possession of N data samples from $\mathcal{S}$ (the training set) and N data samples from $\mathcal{D}$ (the testing set). We set $N = 100$ as a reasonable assumption, in our main paper. In this section, we evalaute the impact of lower ($N = 50$) or higher ($N = 250$) values of N on the MIA success. In Table 5, we present the results on CIFAR-100 for the maximum accuracy setting, i.e., augmentation is applied to boost the model's accuracy. We see that reducing N to 50, from 100, reduces the attack success at most by 7% in the case of Gaussian augmentation (*GA*). Similarly, increasing N to 250, from 100, increases the attack success at most by 13% in the case of Mixup (*MU*). These results show that after a certain amount of samples, e.g., $N = 50$, having more samples lead to diminishing returns on the MIA success. As decreasing N makes the assumption becomes weaker and more realistic, this also highlights the practicality of black-box MIAs.

Table 5: **The impact of the # of attacker's samples (N) on the MIA success in the maximum accuracy setting in Table 1.** Each line presents Adv_{std} (*left*) and Adv_{pow} (*right*). The models are trained on CIFAR-100.

MECH.	$N = 50$	$N = 100$	$N = 250$
SL	20.6 / 39.1	20.8 / 39.7	21.4 / 40.0
LS	57.0 / 74.1	61.4 / 75.6	59.6 / 75.8
DL	46.0 / 63.1	49.7 / 64.0	50.2 / 64.1
RC	33.0 / 30.6	32.7 / 32.0	32.7 / 33.1
CO	33.3 / 33.3	34.9 / 33.9	34.9 / 34.3
GA	54.1 / 62.4	58.7 / 62.5	58.8 / 62.2
MU	43.6 / 49.1	45.1 / 57.0	51.4 / 55.5

B Augmentation-Aware MIAs

The black-box MIAs we applied use the victim model's output on a data sample to infer whether this samples was in $\mathcal{S}$. Because our MIAs are unaware of how the model is trained, the adversary uses the original, unaugmented, data sample ((x_t, y_t)) to query the model and to infer membership. On the other hand, Yu et al. (2020) shows that an *augmentation-aware* MIA can boost the attack success when the victim model is trained with augmentation. Compared to an unaware MIA, their attacks have ∼15% higher success rate, which leads them to argue that unaware MIAs underestimate the risk. As we apply unaware attacks against the augmentation methods, we also might have underestimated of the risk. The main intuition behind these attacks is that a model trained with augmentation can overfit on the augmented samples. For example, a model trained with Gaussian augmentation becomes much more resilient to noise (Cohen et al., 2019). This overfitting effect gives more leverage to augmentation-aware MIAs.

In this section, we apply *augmentation-aware* attacks to evalaute whether they change the trends we highlighted in our paper. These attacks, instead of querying the model with the original sample, query the model with a set of samples generated based on the victim's augmentation strategy. For example, against the random cropping (*RC*), the attacker creates M different versions of (x_t, y_t), each randomly cropped based on the victim's *RC* parameter ($\mathcal{P}$). Further, against soft labels (*SL*), the attacker can query the teacher model, which generated victim's training labels, and computes the victim's loss using these soft labels. After collecting the set loss values on the augmented samples, the attacker then computes simple statistics on this set, e.g., taking the average or the median. Finally, the attacker compares this statistics with a tuned threshold τ to infer the membership of (x_t, y_t), similar to the unaware attacks.

In Table 6, we compare the success of the powerful unaware MIA, Adv_{pow}, to the success of augmentation aware MIA, Adv_{awa}. We see that almost always Adv_{pow} outperforming Adv_{awa} by 5%-50%. On the other hand, against *GA* and *RC* mechanisms, we see that Adv_{awa} is higher by up 20%. This implies that when a model is augmented using these mechanisms, it has a risk of overfitting on the augmented training set samples, which amplifies the Adv_{awa}. Overall, these differences between augmentation-aware and unaware MIAs are not significant enough to change the general trends.

Table 6: **Comparing augmentation aware (Adv_{awa}) and unaware Adv_{pow} MIAs in maximum accuracy, RAD<10% and RAD<25% settings.** Each line presents Adv_{awa} (*left*) and Adv_{pow} (*right*). The models are trained on CIFAR-100.

MECH.	Max Acc	RAD<10%	RAD<25%
SL	26.6 / 39.7	21.1 / 20.1	12.9 / 14.3
LS	67.6 / 75.6	39.6 / 45.0	39.6 / 45.0
DL	64.0 / 64.0	27.3 / 32.0	4.9 / 7.0
RC	37.3 / 32.0	7.1 / 7.1	4.1 / 2.9
CO	31.7 / 33.9	6.4 / 8.2	4.8 / 5.0
GA	63.5 / 62.5	61.0 / 49.9	62.9 / 42.3
MU	57.0 / 25.4	10.5 / 13.7	9.7 / 11.9

References

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pp. 1310–1320. PMLR, 2019.

Yu, D., Zhang, H., Chen, W., Yin, J., and Liu, T.-Y. How does data augmentation affect privacy in machine learning?, 2020.