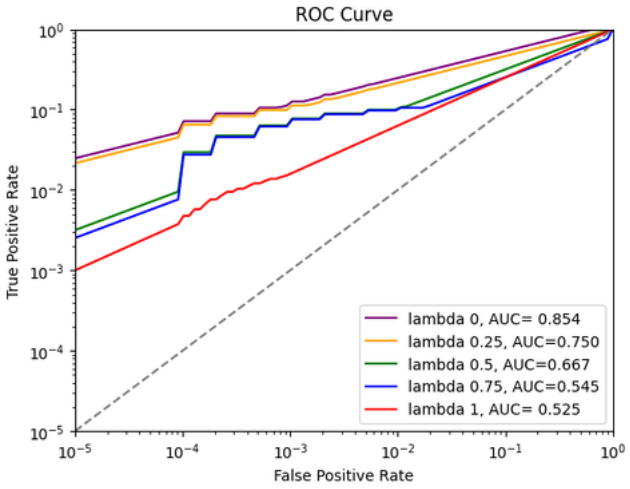

Figure 13: Inf^{2}Guard against LiRA on CIFAR10.

We use the source code of LiRA and test its effectiveness against our Inf^{2}Guard on CIFAR10. Specifically, we treat the data representations and membership network learnt by Inf^{2}Guard as the input data and target model for LiRA. We then follow LiRA's setting and train 16 *white-box* shadow models (i.e., assume LiRA knows the membership network used in Inf^{2}Guard) on the data representations of the 25K members, and report the TPR vs. FPR (on the 5K member and 5K non-member that are not seen by the encoder in Inf^{2}Guard) in Figure 13.

We can observe that: 1) With lambda=0, which means the representations are learnt without privacy protection, the TPR at a low FPR are relatively large (and is similar to Figure 1 of LiRA's paper). 2) By increasing lambda's value, the TPR can be largely reduced. This implies Inf^{2}Guard indeed learns the representations that can (to some extent) defend against the state-of-the-art LiRA.

We will test results on the other two datasets used in MIAs.