Evaluations of Machine Learning Privacy Defenses are Misleading

Anonymous Authors[1]

Abstract

Empirical defenses for machine learning privacy aim to achieve high utility while resisting realistic adversaries. Those defenses require a rigorous audit due to the absence of provable guarantees such as differential privacy. However, we identify severe pitfalls in existing empirical privacy evaluations that result in misleading conclusions. In particular, we show that prior evaluations fail to characterize the privacy leakage of the *most vulnerable samples*, use *weak attacks*, and avoid comparisons with *practical differential privacy baselines*. In 5 case studies of empirical privacy defenses, we find that prior evaluations underestimate privacy leakage by an order of magnitude. Under our stronger evaluation, none of the empirical defenses we study are competitive with a properly tuned, high-utility DP-SGD baseline (with vacuous provable guarantees).

1. Introduction

Machine learning models can memorize sensitive information in their training data. This behavior enables privacy attacks such as membership inference (Shokri et al., 2017) and data extraction (Carlini et al., 2021). A provable defense against such attacks is differential privacy (Dwork et al., 2006)—particularly DP-SGD (Abadi et al., 2016). Yet, achieving strong provable guarantees with good utility remains a challenge (Feldman, 2020). This challenge has led to growing interest in heuristic privacy defenses, which might offer a better privacy-utility tradeoff against practical attacks but no formal guarantees (Nasr et al., 2018; Jia et al., 2019; Yang et al., 2020; Tang et al., 2022; Salem et al., 2019; Chen et al., 2022; Chen & Pattabiraman, 2024).

Without guarantees, evaluations of such empirical defenses use *membership inference attacks* (Shokri et al., 2017) as the canonical approach to obtain a bound on privacy leakage. Under the notion of membership privacy, many heuristic defenses claim to achieve a better privacy-utility tradeoff than DP-SGD against state-of-the-art attacks (Jia et al., 2019; Tang et al., 2022; Chen et al., 2022; Chen & Pattabiraman, 2024). However, we find that such empirical evaluations can be severely misleading (see Figure 1):

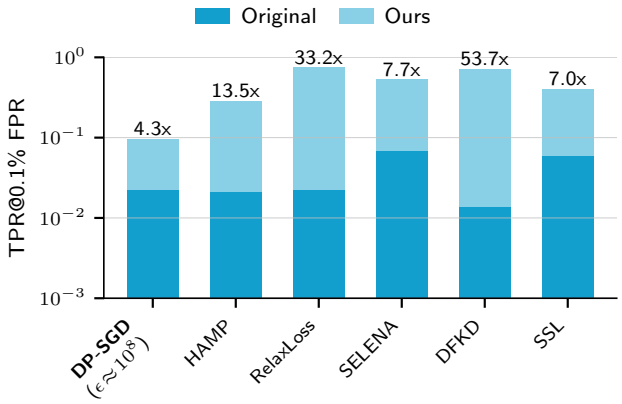

Figure 1. Empirical privacy evaluations provide a false sense of security. We study five heuristic defenses and a properly tuned DP-SGD baseline that all achieve similar high utility (88% test accuracy on CIFAR-10). We first perform a standard privacy evaluation ("Original") and report the attack's TPR at a low FPR across the dataset (following Carlini et al. (2022a)). Our new evaluation methodology ("Ours"), which adapts the attack to each defense and targets the least-private samples, reveals an order-of-magnitude higher privacy leakage. Our DP-SGD baseline provides better privacy (at similar utility) than all the empirical defenses.

1. Current membership inference evaluations fail to reflect a model's privacy *on the most vulnerable data*, and instead aggregate the attack success over a population. **But privacy is not an average-case metric!** (Steinke & Ullman, 2020) We find that population-level privacy fails to reflect individual privacy, even with recent proposals to report an attack's true positive rate at low false positive rates (Carlini et al., 2022a).

2. Many evaluations apply either a *weak* attack that does not reflect the state-of-the-art (Carlini et al., 2022a; Ye et al., 2022), or fail to properly *adapt* the attack to account for unusual defense components or learning paradigms. This issue is reminiscent of well-known pitfalls for non-adaptive evaluations of machine learning robustness (Athalye et al., 2018; Tramer et al., 2020).

3. Empirical defenses are typically compared to *weak DP-SGD baselines* (Jia et al., 2019; Tang et al., 2022; Chen et al., 2022; Chen & Pattabiraman, 2024) with utility below the state-of-the-art.

To address those issues, we propose a rigorous protocol for evaluating an empirical privacy defense:

1. Evaluate membership inference success (specifically TPR at low FPR) for *the most vulnerable sample in a dataset*, instead of an aggregate over all samples. To make this process computationally efficient, audit a set of *canaries* whose privacy leakage approximates that of the most vulnerable sample.

2. Use a *strong* membership inference attack that is properly *adapted* to the specifics of the defense.

3. Compare to DP baselines (e.g., DP-SGD) that *use state-of-the-art techniques* and reach similar *high utility* to the defense.

2. Case Study

We demonstrate the importance and effectiveness of our protocol in a case study of five diverse empirical privacy defenses.[1] In Figure 1, we compare a standard LiRA attack and evaluation (Carlini et al., 2022a) to properly adapted attacks targeted at the most vulnerable samples. Our evaluation reveals that none of the five defenses effectively protect the most vulnerable sample in the dataset and that a realistic DP-SGD baseline (without meaningful provable guarantees) obtains the *best empirical privacy-utility tradeoff*.

Defenses. We study four peer-reviewed defenses that explicitly aim to protect privacy (HAMP (Chen & Pattabiraman, 2024), RelaxLoss (Chen et al., 2022), SELENA (Tang et al., 2022), DFKD (Fang et al., 2022)) and one heuristic based on self-supervised learning (SSL). For SSL, we first train a SimCLR (Chen et al., 2020) encoder on private images and then fit a linear classifier using the private labels on top. We tune all defenses to maximize privacy, subject to $\geq 88\%$ CIFAR-10 test accuracy whenever possible.

Evaluation setup. For each defense, we train 64 shadow models on CIFAR-10 (Krizhevsky et al., 2009). However, rather than subsampling the entire training set as in (Carlini et al., 2022a) or evaluating each sample in isolation, we follow Steinke et al. (2023): we designate 500 random data points as "audit samples" on which we evaluate membership inference and always include the remaining 49,500 samples in every model's training data. Inspired by work on worst-case privacy auditing (Carlini et al., 2019; Jagielski et al., 2020), our proposed evaluation ("Ours" in Figure 1) replaces the 500 audit samples with appropriately chosen canaries. Importantly, we design the canaries to mimic the most vulnerable samples in the dataset. We then perform a leave-one-out attack over shadow models (as in (Tramèr et al., 2022)), and report the TPR and FPR over the $64 \cdot 500$ attacker guesses.

Canary choices. We identify canaries that mimic the most vulnerable samples for each defense, so that the average privacy over canaries approximates the worst-case individual privacy in the dataset. For HAMP, RelaxLoss, and DFKD, we find that mislabeled samples are appropriate canaries. SELENA is more vulnerable if a mislabeled sample has a (near)-duplicate in the training data; we hence duplicate half of the original audit set, and mislabel the duplicates. In contrast, the SSL defense is robust against label noise but memorizes atypical features. We thus use ImageNet images as SSL canaries.

Adaptive attacks. We find that the standard LiRA attack achieves strong privacy leakage for RelaxLoss, SELENA, and DFKD. Since HAMP performs confidence masking, we develop a straightforward label-only attack (Choquette-Choo et al., 2021). Finally, we adapt the attack on the SSL defense to the contrastive loss: we extract the features of different augmented version of the same image, and derive a LiRA score from their cosine similarity.

Strong DP-SGD baselines. We use state-of-the-art DP-SGD training tricks (De et al., 2022; Sander et al., 2023), including a modified WRN16-4 architecture, an exponential moving average of model weights, and augmentation multiplicity (De et al., 2022) using the modified Opacus (Yousefpour et al., 2021) library of Sander et al. (2023). We tune the baseline to maximize privacy while achieving higher utility than most of the case study defenses (91% CIFAR-10 test accuracy). This results in essentially vacuous worst-case privacy guarantees ($\epsilon \approx 1.8 \cdot 10^8$ for $\delta = 10^{-5}$). We attack the DP-SGD baseline using the standard LiRA attack on atypical images (the strongest adversary we could find).

3. Conclusion

Our work adds to the growing literature on pitfalls in evaluations of ML privacy defenses (Choquette-Choo et al., 2021; Tramer et al., 2022; Carlini et al., 2022b). As we show, the exact way we measure the privacy of a defense matters a lot. Before evaluating a defense—or an attack—we thus need to clearly define the privacy *semantics* that the evaluation targets (e.g., do we care about the proportion of vulnerable individuals, or if *any* individual is vulnerable).

Our insights transfer directly to generative models. For such models, privacy attacks and defenses are still a highly active area of research (e.g., (Duan et al., 2024; Dubinski et al., 2024)). We thus hope that our rigorous evaluation protocol helps future work to avoid repeating the mistakes of the past—thereby accelerating the development of trustworthy privacy defenses for generative models.

[1] Our case study considers image classification, because most researched focused on that area until recently. Nevertheless, our evaluation protocol directly applies to generative models as well.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, pp. 267–284, 2019.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramer, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914. IEEE, 2022a.

Carlini, N., Feldman, V., and Nasr, M. No free lunch in" privacy for free: How does dataset condensation help privacy". *arXiv preprint arXiv:2209.14987*, 2022b.

Chen, D., Yu, N., and Fritz, M. RelaxLoss: Defending Membership Inference Attacks without Losing Utility. In *International Conference on Learning Representations*, 2022.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607, 2020.

Chen, Z. and Pattabiraman, K. Overconfidence is a Dangerous Thing: Mitigating Membership Inference Attacks by Enforcing Less Confident Prediction. In *NDSS Symposium*, 2024.

Choquette-Choo, C. A., Tramer, F., Carlini, N., and Papernot, N. Label-only membership inference attacks. In *International Conference on Machine Learning*, pp. 1964–1974. PMLR, 2021.

De, S., Berrada, L., Hayes, J., Smith, S. L., and Balle, B. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Duan, M., Suri, A., Mireshghallah, N., Min, S., Shi, W., Zettlemoyer, L., Tsvetkov, Y., Choi, Y., Evans, D., and Hajishirzi, H. Do membership inference attacks work on large language models?, 2024.

Dubinski, J., Kowalczuk, A., Pawlak, S., Rokita, P., Trzcinski, T., and Morawiecki, P. Towards more realistic membership inference attacks on large diffusion models. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4848–4857, 2024.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Fang, G., Mo, K., Wang, X., Song, J., Bei, S., Zhang, H., and Song, M. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6597–6604, 2022.

Feldman, V. Does learning require memorization? A short tale about a long tail. In *ACM SIGACT Symposium on Theory of Computing*, pp. 954–959, 2020.

Jagielski, M., Ullman, J., and Oprea, A. Auditing differentially private machine learning: How private is private sgd? *Advances in Neural Information Processing Systems*, 33:22205–22216, 2020.

Jia, J., Salem, A., Backes, M., Zhang, Y., and Gong, N. Z. MemGuard: Defending against black-box membership inference attacks via adversarial examples. In *ACM SIGSAC Conference on Computer and Communications Security*, pp. 259–274, 2019.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Nasr, M., Shokri, R., and Houmansadr, A. Machine learning with membership privacy using adversarial regularization. In *ACM SIGSAC Conference on Computer and Communications Security*, pp. 634–646, 2018.

Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., and Backes, M. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. 2 2019. doi: 10.60882/cispa.24612846.v1. URL https://publications.cispa.de/articles/conference_contribution/ML-Leaks_Model_and_Data_Independent_Membership_Inference_Attacks_and_Defenses_on_Machine_Learning_Models/24612846.

Sander, T., Stock, P., and Sablayrolles, A. TAN without a burn: Scaling laws of DP-SGD. In *Proceedings of the International Conference on Machine Learning*, volume 202, pp. 29937–29949, 2023.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pp. 3–18. IEEE, 2017.

Steinke, T. and Ullman, J. The pitfalls of average-case differential privacy. DifferentialPrivacy.org, 07 2020. https://differentialprivacy.org/average-case-dp/.

Steinke, T., Nasr, M., and Jagielski, M. Privacy Auditing with One (1) Training Run. In *Advances in Neural Information Processing Systems*, pp. 49268–49280, 2023.

Tang, X., Mahloujifar, S., Song, L., Shejwalkar, V., Nasr, M., Houmansadr, A., and Mittal, P. Mitigating Membership Inference Attacks by Self-Distillation Through a Novel Ensemble Architecture. In *USENIX Security Symposium*, pp. 1433–1450, 2022.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Tramèr, F., Shokri, R., San Joaquin, A., Le, H., Jagielski, M., Hong, S., and Carlini, N. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2779–2792, 2022.

Tramer, F., Terzis, A., Steinke, T., Song, S., Jagielski, M., and Carlini, N. Debugging differential privacy: A case study for privacy auditing. *arXiv preprint arXiv:2202.12219*, 2022.

Yang, Z., Shao, B., Xuan, B., Chang, E.-C., and Zhang, F. Defending model inversion and membership inference attacks via prediction purification, 2020.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., and Shokri, R. Enhanced membership inference attacks against machine learning models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3093–3106, 2022.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*, 2021.