Research Statement

Anshuman Suri

Machine learning models are susceptible to several privacy risks, including leaking sensitive information related to their training data. Much of the work on machine learning privacy has focused on *membership inference* [1], the task in which an adversary tries to infer whether some particular record is used in training a given model. However, other granularities of data-centric privacy are often ignored and not as well understood. My research focuses on one such granularity: the distribution level. As an example of distribution inference, consider an adversary probing a face-image-based model [2]. The adversary would more likely be interested in inferring the presence of (any) photos of an individual, not one particular image of them. The former is covered by distribution inference, while the latter corresponds to membership inference. Other examples include inferring the accent of speakers in voice recognition models [3] and learning ratios of characteristics like gender labels [4] which are relevant to fairness auditing [5, 6]. Recent works in the literature further show how these risks are not specific to "standard" models and apply to novel settings and models as well, such as prompt-learning [7] and large language models [8].

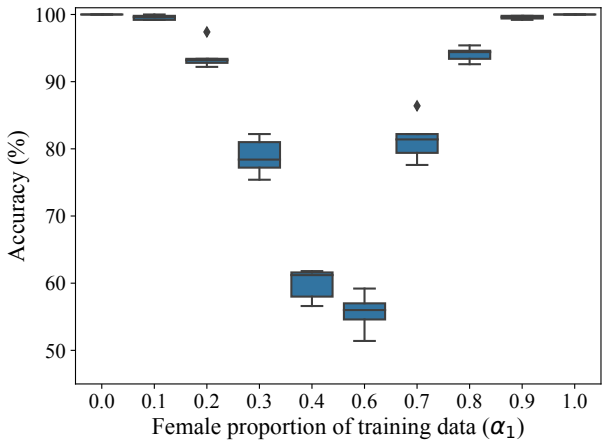
In addition to my privacy-focused research which I will describe in more detail next, I also maintain a general interest in adversarial machine learning (having had my first research experiences in adversarial examples [9] and poisoning attacks [10]). Our most recent work on studying black-box adversarial attacks [11], for instance, shows how current attack evaluations unnecessarily clip the number of iterations in gradient-based attacks, and how attack success can nearly doubled when these arbitrary iteration limits are removed. I also participated[1] in the Trojan Detection Challenge[2] recently, to get some hands-on experience with LLM jailbreaking and trojan detection. The task was to detect triggers corresponding to certain target strings in a given model, while maximizing the recall of finding these Trojans.

1 Dissertation Research: Distribution Inference

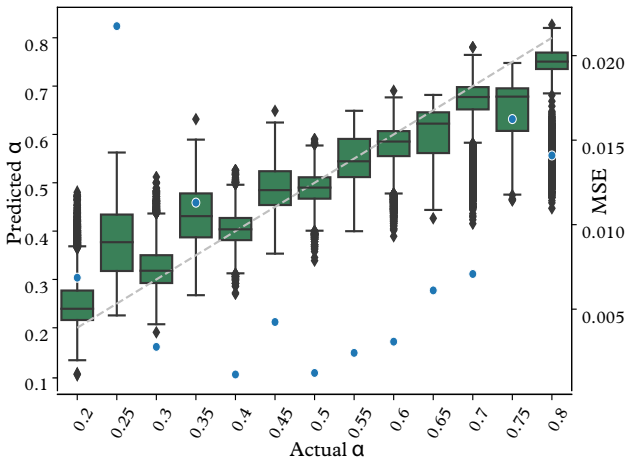
Since the main purpose of machine learning is to learn properties of a distribution, we need a way to define problematic distribution inference. In our initial work [12], I began by formalizing distribution inference. Our definitions standardize the adversary's setup by setting it up as a cryptographic game similar to membership inference [13], where the adversary's goal is to distinguish between potential distributions $\mathcal{G}_0(\mathcal{D})$ or $\mathcal{G}_1(\mathcal{D})$ derived from a common public distribution $\mathcal{D}$. The distribution transformation functions $\mathcal{G}_0$ and $\mathcal{G}_1$ can be used to model differences between the distributions to be distinguished, such as a difference in the ratios of females, or the presence of a group of users. Our recent work on studying inference threat models in machine-learning models, with collaborators at Microsoft, further underlines how distribution inference is very different from other threats like attribute inference and model inversion [14].

Our setup also allows for more generic properties, such as graph-related properties [15] like the mean node-degree. Motivated by this formalization, I also proposed a metric n_{leaked} to capture inference leakage for certain cases of distributions. The metric is designed to capture inherent distinguishability of distributions by relating observed inference accuracy with the number of samples which, when used to launch a Bayes-optimal attack, would yield the same observed inference accuracy. I derived theorems for computing n_{leaked} for certain settings: distributions with different priors over certain attributes, as well as the average node-degree of a graph, for both binary-classification and direct regression over these distributions. For instance, distinguishing between distributions that have all males or no makes is intuitively easier than distinguishing ones that have 50% or 60% males, as we observe empirically (Figure 1a). Our metric n_{leaked} formalizes a way to capture this nuance. Our evaluations reveal how direct inference of underlying distributional properties via regression can be extremely potent (Figure 1b). In a follow-up work [16], I proposed and evaluated a potent black-box attack that leaks information even when certain assumptions about the adversary's knowledge are relaxed, demonstrating the vulnerabilities related to distribution inference even in practical settings. The proposed KL attack uses shadow models and computes KL-divergence between model predictions from different models, using the divergence value to compare sets of shadow models and thus inferring which distribution the target model is closer to. The attack uses available shadow models by comparing pair-wise scores between shadow models, as opposed to a meta-classifier that adds an additional layer of obfuscation and does not explicitly capture relationships between pairs of models. The KL attack outperforms the best white-box attacks while using a fraction of shadow models.

[1] `https://www.anshumansuri.me/post/tdc/`
[2] `https://trojandetection.ai/`



(a) Census-19



(b) RSNA BoneAge

Figure 1: (a) Classification accuracy for distinguishing proportion of females in training data for Census-19 dataset, using the KL attack. Performance of the attacks increases as the distributions diverge (α_1 moves away from 0.5). (b) Predicted α values (left y-axis) for models with training distributions for varying α values (x-axis), for all victim models and regression meta-classifier experiments (green box-plots), along with mean squared error (right y-axis labels, with different scales on the two graphs, and blue dots) RSNA BoneAge dataset, using the Permutation-Invariant Network as a meta-classifier for inference. The diagonal gray dashed line represents the ideal case, where the regression classifier would perfectly predict α.

Our results also demonstrate how techniques like differentially-private training (DP-SGD [17]), meant to defend against membership inference, completely fail against distribution inference. These observations further illustrate how privacy notions related to record-level privacy, such as Differential Privacy and membership inference, are not sufficient for studying privacy leakage.

Our work also highlights how these inference risks are not limited to single-party settings and extend to other scenarios such as inferring the membership of subjects (individual persons) across clients in federated learning setups [18], a setup that is used by prominent services such as Google's Gboard keyboard auto-completion. Similarly, active attacks via attempts to introduce Trojans into models can also increase the risk of distribution inference in downstream fine-tuned models [19], accentuating the extent of this risk in different training setups. Even when considering adaptive detection methods, our experiments show how these Trojans can be rendered nearly undetectable by any known methods while retaining most of their utility in boosting inference risk in downstream models.

2 Future Directions

As a researcher, I am driven to address challenges at the intersection of theory and experimentation within real-world systems, prioritizing impactful solutions over empirical trial and error. I believe in addressing specific problems that have a direct impact on how ML-based systems function and how users experience them, with a special focus on privacy and security. My approach involves selecting and proposing projects based on their potential for both immediate and long-term relevance.

2.1 Auditing

Current privacy auditing tools rely on membership inference [20] for quantifying privacy leakage. My recent participation[3] in the Membership-Inference Competition (MICO[4]) helped me see more clearly how even for membership inference, attacks can be very brittle and sensitive to changes in training setups. Work on distribution inference, however, has made it clear how distribution inference is a relevant inference concern, and is separate from membership inference. Given this disconnect between distribution inference and other useful notions of privacy, I wish to understand how distribution inference can be used to create better privacy auditing tools that go beyond simply evaluating existing attacks on a given model. Distribution inference is especially relevant in modern-day settings, where users contribute multiple records to datasets, and testing for record-based membership requires access to records to begin with, which can be a strong assumption in many cases. We envision an auditing tool that utilizes all available information, including data and white-box access to the model, to predict leakage with a usable and quantitative metric.

[3] https://www.anshumansuri.me/post/mico/
[4] https://microsoft.github.io/MICO/

While black-box access to models (under certain assumptions) is sufficient for membership inference [21], it is unclear whether the same holds for user-level inference or distribution inference, or how these conclusions change when assumptions (such as using SGLD instead of SGD [22]) are modified to include realistic machine-learning designs, such as regularization and batch-normalization [23]. At the same time, attacks that use white-box access should be efficient enough for thorough privacy analysis for a given model and data in its pipeline. Thus, straightforward extensions of meta-classifier-based attacks may not be feasible for auditing, and there is a need for specialized auditing-oriented attacks. Auditing tools can help model trainers get a more holistic view of potential leakage, which is more relevant at the distribution level than record level in today's data paradigm, where data sources contain multiple records per user, and users care about leakage of *any* of their data, not just one particular record.

2.2 Mitigating Inference Risks

The standard privacy defense of adding noise to enforce Differential Privacy does not work, and the only known effective defenses that work are vulnerable to adaptive attacks [16]. Chen and Ohrimenko [24] recently proposed a theoretically-grounded defense mechanism, but it only applies to statistical queries and not machine-learning models. Current works [25] suggest causality-aligned learning as a potential defense against distribution inference. Our preliminary experiments with CyCNN [26], a learning technique designed to provide better domain generalization, suggest that this indeed might be the case. I wish to further explore this connection with techniques explicitly designed with causality in mind like MatchDG [27] and explore theoretical connections between distribution inference and causality.

Some aspects of a distribution (such as the types of digits in an MNIST classifier) may be tied closely to the given task of a machine learning model and hence, may be unavoidable for a good classification model. At the same time, some properties may not be very useful for an inference adversary (such as the average color of images). Having a formal and principled distinction between such properties can not only help focus on relevant inference cases, but also lead to principled defenses against distribution inference. A good classifier should ignore style-specific features, but it remains unclear how much of content-related properties the classifier may actually memorize, and what portion of such memorization may be unavoidable. Recent analyses for the case of distributional membership inference [25] shows how under perfect invariance-learning assumptions, user-level membership inference under black-box access is impossible. I would like to further extend these analyses to better understand leakage for other distributional properties, and work on relaxing assumptions around perfect invariance learning.

2.3 Large Language Models (LLMs)

There is an widespread interest in LLMs, with users interacting with such machine learning models at an unprecedented scale. These models are trained on titanic volumes of data scraped from the Internet and conversations via agents like ChatGPT, blowing up privacy and security-related concerns. Our recent exploration of LLMs [28], in collaboration with EPFL and others, studies how memorization of information is inherently different from traditional machine-learning models. The interactive nature of these models also creates new risks and scope for attacks such as clean-text prompt instruction modification [29].

I am interested in extending auditing techniques for measuring privacy leakage in LLMs. Techniques that require any form of shadow-model training are immediately inapplicable for such large models, given their scale and the time and resource required to train them. The key thus lies in having efficient techniques that can audit these models without having to train a lot of large shadow models. In a recent collaboration with the University of Washington [30], we studied leakage via membership-inference attacks and how under proper evaluation settings, leakage is near-zero (attack performance close to random guess) for nearly all data and models. This further underlines how the current approach of privacy auditing using membership-inference would not work well, as the absence of detectable membership inference in this case is not conclusive, and may as well arise from a lack of potent attacks specifically designed for LLMs. Recent exploration on interpretability [31] shows promise, and it may be possible to modify/extend such techniques to measure leakage via "interpretability".

In my research approach, I prioritize problems that captivate me and hold intrinsic value for the research community, steering clear of excessive influence from current trends. For instance, our work on distribution inference was inspired by a concept introduced several years ago, showcasing my inclination to explore beyond immediate "hot" topics. While my expertise lies in ML privacy and security, I actively try to diversify my knowledge, as I did in my recent explorations into Large Language Models (LLMs) and causal learning. Regarding research areas, I favor a focused approach, committing time to make substantial contributions to progress within a specific domain, while encouraging other researchers to bring their ideas and perspectives in their research on that topic. Emphasizing the importance of proper coding practices, I view it as an opportunity to enhance research utility rather than a mere task for "releasing code for some paper." Leading a research group, I envision clusters of students tackling interconnected yet distinct problems, fostering meaningful contributions to each other's projects.

References

[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy*. IEEE, 2017.

[2] M. Chen, Z. Zhang, T. Wang, M. Backes, and Y. Zhang, "Face-auditor: Data auditing in facial recognition systems," in *USENIX Security Symposium*, 2023.

[3] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," *International Journal of Security and Networks*, 2015.

[4] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov, "Property inference attacks on fully connected neural networks using permutation invariant representations," in *ACM Conference on Computer and Communications Security*, 2018.

[5] M. Juárez, S. Yeom, and M. Fredrikson, "Black-box audits for group distribution shifts," *arXiv preprint arXiv:2209.03620*, 2022.

[6] V. Duddu, A. Das, N. Khayata, H. Yalame, T. Schneider, and N. Asokan, "Attesting distributional properties of training data for machine learning," *arXiv preprint arXiv:2308.09552*, 2023.

[7] Y. Wu, R. Wen, M. Backes, P. Berrang, M. Humbert, Y. Shen, and Y. Zhang, "Quantifying privacy risks of prompts in visual prompt learning," in *USENIX Security Symposium*, 2024.

[8] N. Kandpal, K. Pillutla, A. Oprea, P. Kairouz, C. A. Choquette-Choo, and Z. Xu, "User inference attacks on large language models," *arXiv preprint arXiv:2310.09266*, 2023.

[9] D. Vijaykeerthy*, **A. Suri***, S. Mehta, and P. Kumaraguru, "Hardening deep neural networks via adversarial model cascades," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.

[10] F. Suya, S. Mahloujifar, **A. Suri**, D. Evans, and Y. Tian, "Model-targeted poisoning attacks with provable convergence," in *International Conference on Machine Learning*. PMLR, 2021.

[11] F. Suya*, **A. Suri***, T. Zhang, J. Hong, Y. Tian, and D. Evans, "SoK: Pitfalls in evaluating black-box attacks," in *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

[12] **A. Suri** and D. Evans, "Formalizing and estimating distribution inference risks," *Proceedings on Privacy Enhancing Technologies*, 2022.

[13] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE Computer Security Foundations Symposium*, 2018.

[14] A. Salem, G. Cherubin, D. Evans, B. Köpf, A. Paverd, **A. Suri**, S. Tople, and S. Zanella-Béguelin, "SoK: Let the privacy games begin! a unified treatment of data inference privacy in machine learning," in *IEEE Symposium on Security and Privacy (SP)*, 2023.

[15] Z. Zhang, M. Chen, M. Backes, Y. Shen, and Y. Zhang, "Inference attacks against graph neural networks," in *USENIX Security Symposium*, 2022.

[16] **A. Suri**, Y. Lu, Y. Chen, and D. Evans, "Dissecting distribution inference," in *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2022.

[17] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM Conference on Computer and Communications Security (CCS)*, 2016.

[18] **A. Suri**, P. Kanani, V. J. Marathe, and D. W. Peterson, "Subject membership inference attacks in federated learning," *arXiv:2206.03317*, 2022.

[19] Y. Tian, F. Suya, **A. Suri**, F. Xu, and D. Evans, "Manipulating transfer learning for property inference," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[20] F. Lu, J. Munoz, M. Fuchs, T. LeBlond, E. Zaresky-Williams, E. Raff, F. Ferraro, and B. Testa, "A general framework for auditing differentially private machine learning," *Advances in Neural Information Processing Systems*, 2022.

[21] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *International Conference on Machine Learning*. PMLR, 2019.

[22] I. Sato and H. Nakagawa, "Approximation analysis of stochastic gradient langevin dynamics by using fokker-planck equation and ito process," in *International Conference on Machine Learning*. PMLR, 2014, pp. 982–990.

[23] S. Lange, W. Deng, Q. Ye, and G. Lin, "Batch normalization preconditioning for stochastic gradient langevin dynamics," *Journal of Machine Learning*, vol. 2, no. 1, pp. 65–82, 2023.

[24] M. Chen and O. Ohrimenko, "Protecting global properties of datasets with distribution privacy mechanisms," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023.

[25] V. Hartmann, L. Meynent, M. Peyrard, D. Dimitriadis, S. Tople, and R. West, "Distribution inference risks: Identifying and mitigating sources of leakage," in *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2023.

[26] J. Kim, W. Jung, H. Kim, and J. Lee, "Cycnn: A rotation invariant cnn using polar mapping and cylindrical convolution layers," *arXiv:2007.10588*, 2020.

[27] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *International Conference on Machine Learning*. PMLR, 2021.

[28] V. Hartmann, **A. Suri**, V. Bindschaedler, D. Evans, S. Tople, and R. West, "SoK: Memorization in general-purpose large language models," *arXiv preprint arXiv:2310.18362*, 2023.

[29] R. Shah, Q. Feuillade-Montixi, S. Pour, A. Tagade, S. Casper, and J. Rando, "Scalable and transferable black-box jailbreaks for language models via persona modulation," *arXiv preprint 2311.03348*, 2023.

[30] M. Duan*, **A. Suri***, N. Mireshghallah, S. Min, W. Shi, L. Zettlemoyer, Y. Tsvetkov, Y. Choi, D. Evans, and H. Hajishirzi, "Do membership inference attacks work on large language models?" *arXiv:2402.07841*, 2024.

[31] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah, "Towards monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits Thread*, 2023. [Online]. Available: https://transformer-circuits.pub/2023/monosemantic-features/index.html