Towards Differentially Private Machine Learning Models and their Robustness To Adversaries

Alberto Carlo Maria Mancino[1] and Tommaso Di Noia[2]

[1] Politecnico di Bari, Italy alberto.mancino@poliba.it
[2] Politecnico di Bari, Italy tommaso.dinoia@poliba.it

Abstract. The pervasiveness of modern machine learning algorithms exposes users to new vulnerabilities: violation of sensitive information stored in the training data and wrong model behaviors caused by adversaries. State-of-the-art approaches to prevent such behaviors are usually based on Differential Privacy (DP) and Adversarial Training (AT). DP is a rigorous formulation of privacy in probabilist terms to prevent information leakages that could reveal private information about the users, while AT algorithms empirically increase the system's robustness, injecting adversarial examples during the training process. Both techniques involve achieving their goal by modeling noise introduced into the system. We propose analyzing the relationship between these two techniques, studying how one affects the other. Our objective is to design a mechanism that guarantees DP and robustness against adversarial attacks, injecting modeled noise into the system. We propose Recommender Systems as an application scenario because of the severe risks to user privacy and system sensitivity to adversaries.

Keywords: Differential Privacy · Adversarial Training · Recommender Systems · Privacy Preservation · System Robustness.

1 Introduction

In the Big Data era, machine learning (ML) models are widely used in heterogeneous scenarios, enhancing most of the massive data available on the Web. Their exceptional performance on many learning tasks has enabled the emergence of several new services that users can benefit from in their everyday digital lives. However, to provide personalized services, they need access to a sizeable fine-grained collection of user data, critically raising the risks concerning their privacy.

Despite the benefits of legitimate ML applications, the potential risk of infringing user privacy must not be underestimated, and nowadays, people are increasingly aware of the risks they run by sharing their data. This common feeling of insecurity led the governments to legislate to regulate the usage and manipulation of users' sensible data. Some examples are the GDPR by the European Union [8], the CCPA in California [5], and the Cybersecurity Law in China [20].

Among the most widely used ML models on the Web, **Recommender Systems** (RSs) play a central role due to their ability to intercept and imitate user preferences. Online platforms benefit from the ability of these systems to exploit side information [4] and extrapolate behavioral patterns from historical users' interactions and assist them in facing the overwhelming quantity of products, with lists of personalized recommendations. However, everything has two sides, so do RSs.

RSs require users to provide their personal preferences for different items to compute tailored recommendations. The primary efforts of industry and academia have focused on improving recommendation accuracy while finding a suitable tradeoff between privacy and personalization is still an open question. Numerous privacy-preserving recommenders have been proposed in recent years, including **differential privacy** (DP) [7].

DP is a rigorous and provable formulation of privacy in probabilistic terms. The output of a model that respects the DP paradigm is insensitive to adding or removing a particular record. DP injects noise into sensitive data or the computation to achieve its goal. The strength of DP is the ability to quantify the privacy budget ϵ and the lower computational effort required concerning other classic privacy-preserving techniques.

Admitting protecting the users' privacy is a critical concern for Recommender Systems, their exposure to adversaries able to exploit their vulnerabilities must also be taken into account. RSs are affected by two main risks: integrity and availability. Compromising the integrity means inducing the system to produce an output different from the original one. While compromising the availability means reducing the recommendation performance.

Adversarial Training (AT) [11] is receiving considerable attention from the research community as a proven method against adversarial attacks [16]. The basic idea behind AT is to inject adversarial examples, slightly perturbated data aiming to fool the system, during the training loops. A system trained as just mentioned can face adversaries attacks without sensibly changing its behavior.

Recent works have shown the fragility of recommenders, like BPR, to adversarial perturbations [12], i.e., small perturbation added to the recommender model parameters. *He et al.* [12] proposed a defence strategy named *adversarial regularizer*, a solution based on the adversarial training procedure. Their work showed that it is possible to build a recommender system robust to adversarial noise and paved the way for different novel models that improved the robustness of recommender systems.

Given the vulnerabilities of recommender systems, concerning both privacy preservation, and adversarial robustness, and see the diversity of the different solutions proposed in the literature, my research aims to investigate the following open question: **is there a rigorous way to implement a recommender system able to protect users' privacy under the DP paradigm and be robust against adversarial perturbation?**

Recent papers like [14] demonstrate a connection between differential privacy and robustness against adversarial examples. Nevertheless, it is still an under-investigated field despite the fundamental role of the two aspects in our daily life. Furthermore, to the best of my knowledge, very few works have focused on the specific field of RSs.

However, an initial study [10] underlines the risk of applying DP techniques when facing adversarial attacks: the noise introduced in the system could help the attacker remain undetected, potentially facilitating him/her ability to degrade the system performances. Thus, a further research question arises: **could differential privacy impact the robustness of a ML system? Which is the interplay between DP and AT?**

In addition, studies on DP and AT indicate that the separate application of both techniques weakens recommender systems' accuracy. So, it is crucial to consider the following question: **how much does DP and AT impact separately the accuracy of the RSs? How do the performances vary when applied both?**

2 Research Overview

This section briefly summarizes the latest research contributions about differential privacy and adversarial training in machine learning, particularly in recommender systems. Then describes the earliest solutions proposed in the literature that combine both techniques and analyze the consequences.

2.1 Privacy Preservation with Differential Privacy (DP)

Differential Privacy (DP) [7] is a mathematical definition of privacy related to the quantity of information of an individual an attacker could disclose. Given a generic computation over the data we want to protect, DP proves that the structured injection of noise can hide the characteristics of the individuals of the system.

Consider a randomized mechanism $\mathcal{M}$ that takes as input a dataset d and returns a value in a space $\mathcal{O}$. $\mathcal{M}$ is said to satisfy DP if given any two adjacent datasets d_1 and d_2, that differ by only one record, and for any subset of possible outputs $\mathcal{S} \subseteq \mathcal{O}$, we have

$$P(\mathcal{M}(d_1) \in \mathcal{S}) < e^{\epsilon} P(\mathcal{M}(d_2) \in \mathcal{S}) + \delta \tag{1}$$

Where $e > 0$ and $\delta \in [0, 1]$ are parameters that define the privacy strength of the randomized mechanism. Concretely DP is a rigorous and quantifiable guarantee for removing or adding a record in the dataset without sensibly altering the algorithm's outcome. Consequentially, a malicious attacker can not distinguish the presence or absence of an individual in the dataset.

In recommender systems (RS), a formal application of DP was introduced by McSherry et al. [17] with a collaborative filtering model. They propose randomizing the users' ratings before sharing them with the system and factor the

learning algorithm into two phases, aggregation/learning and individual recommendation. Then they analyze the impact of privacy adaptations on the accuracy performance. Other works [13][22] propose to inject noise, usually laplacian, directly into the objective function, satisfying the differential privacy. At the same time, Friedman et al. [9] studied an approach with noisy ratings and a strategy that exploits a perturbed version of the stochastic gradient descent. De Montjove [19] also proposed to perturb the input of the recommender, but in the specific area of point-of-interest recommendation.

2.2 Robustness to Adversaries with Adversarial Training (AT)

Researchers have shown that ML models are susceptible to small perturbations [21][3]. Adversaries could exploit this behavior to compromise the system's proper functioning with imperceptible perturbations. Figure 1 of [11] is a classic example of how it is possible to lead a system to a misclassification. Goodfellow et al. [11] proposed *adversarial training*, a defense based on the minimax learning strategy to address this drawback, defined as in eq. 2.

$$\min_{\Theta}[\underbrace{\mathcal{L}(f(\mathbf{x};\Theta),y) + \lambda \underbrace{\max_{\delta:\|\delta\|\leq\epsilon} \mathcal{L}(f(\mathbf{x}+\delta;\Theta),y)}_{\text{Adversarial Regularizer}}}_{\text{Adversarial Regularized Loss}}] \tag{2}$$

The *adversarial regularizer* mitigates the attack surface considering an adversary's behavior who aims to maximize the system's loss.

The same behavior also applies to Recommender Systems, where a malicious can destroy the accuracy of the systems with injected noise [12][18][6][15].

A recommender's three main components could be principally perturbed by an adversary: the interactions (e.g., injecting fake records), the side-information data, and the model parameters. The adversarial training techniques have also been applied to recommenders to improve their *robustness to adversaries*, i.e., their capacity to not significantly change the output when adversary noise is instilled into the system.

2.3 The Interplay between DP and AT

Sections 2.1 and 2.2 underlined the growing need to consider the vulnerabilities of ML models, particularly with RSs. Differential privacy is necessary to guarantee the protection of user data while adversarial training robustifies the system against malicious attacks.

Although it is fair that a system should guarantee both the cited aspects, the literature still lacks studies that clarify how to design such a model. Lecuyer et al. [14] proved the existence of a connection between DP and AT. Their study provides a formal definition of adversarial robustness in mathematical terms. Then applying DP to the input, they derived stability bounds for the expected

output and combined them with their definition of robustness. In this way, they derived a certified defense and proved its efficacy against adversarial attacks.

Another research direction studied how an attacker could leverage the DP noise to compromise the system's integrity, acting undetected. Giraldo et al. showed that an attacker could fool the adversarial classifier when DP noise is injected into the system, with the possibility of biasing the model more than he/she could do without the application of DP. They evaluate their hypotesis over a traffic congestion problem

These works demonstrate that a connection between DP and AT exists, but it is not obvious that it could benefit the privacy and robustness performance. Furthermore, both impacts on the system's accuracy must be taken into account. These aspects, and the open possibilities and needs in Recommender Systems, motivate my research in studying the interplay between these two crucial techniques. It is still under-discussed the relationship of both techniques with a privacy-by-design technique as Federated Learning [2].

3 Research Direction

Here I summarize the direction of my research proposal being in the first year of my Ph.D., given the potentialities and the limits briefly described in section 2 about modeling a recommender that grants differential privacy (DP) and robustness to adversaries, with adversarial training (AT).

Different studies highlighted how much both DP and AT could degrade the accuracy performances [17][1]. Consequently, I propose first analyzing the trade-off between **utility**, **privacy** and **security** in RSs. The study aims to evaluate how applying differential privacy simultaneously with adversarial training on the same recommender affects its accuracy and beyond-accuracy metrics. The performance could be evaluated varying the nature of the recommender system (such as latent factor, neural, graph-based), the DP and AT algorithms, and the privacy/perturbation budget.

Then it is necessary to deepen when and how differential privacy and adversarial training influence the performances of each other. The nature of the relationship between these two methods is still undisclosed.

Finally, the findings could be formalized to propose a rigorous formulation on modeling a secure and private recommender system quantifying the amount of sacrificed accuracy.

References

1. Anelli, V.W., Bellogín, A., Deldjoo, Y., Noia, T.D., Merra, F.A.: MSAP: multi-step adversarial perturbations on recommender systems embeddings. In: FLAIRS Conference (2021)
2. Anelli, V.W., Deldjoo, Y., Noia, T.D., Ferrara, A., Narducci, F.: How to put users in control of their data in federated top-n recommendation with learning to rank. In: SAC. pp. 1359–1362. ACM (2021)

3. Anelli, V.W., Noia, T.D., Malitesta, D., Merra, F.A.: Assessing perceptual and recommendation mutation of adversarially-poisoned visual recommenders (short paper). In: DP@AI*IA. CEUR Workshop Proceedings, vol. 2776, pp. 49–56. CEUR-WS.org (2020)
4. Anelli, V.W., Noia, T.D., Sciascio, E.D., Ferrara, A., Mancino, A.C.M.: Sparse feature factorization for recommender systems with knowledge graphs. In: RecSys. pp. 154–165. ACM (2021)
5. California State Legislature: The california consumer privacy act of 2018 (2018), https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375
6. Deldjoo, Y., Noia, T.D., Sciascio, E.D., Merra, F.A.: How dataset characteristics affect the robustness of collaborative recommendation models. In: SIGIR. pp. 951–960. ACM (2020)
7. Dwork, C.: Differential privacy. In: ICALP (2). Lecture Notes in Computer Science, vol. 4052, pp. 1–12. Springer (2006)
8. European Commission: 2018 reform of eu data protection rules (2018), https://ec.europa.eu/info/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules/eu-data-protection-rules_en
9. Friedman, A., Berkovsky, S., Kâafar, M.A.: A differential privacy framework for matrix factorization recommender systems. User Model. User Adapt. Interact. **26**(5), 425–458 (2016)
10. Giraldo, J., Cárdenas, A.A., Kantarcioglu, M., Katz, J.: Adversarial classification under differential privacy. In: NDSS. The Internet Society (2020)
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (Poster) (2015)
12. He, X., He, Z., Du, X., Chua, T.: Adversarial personalized ranking for recommendation. In: SIGIR. pp. 355–364. ACM (2018)
13. Hua, J., Xia, C., Zhong, S.: Differentially private matrix factorization. In: IJCAI. pp. 1763–1770. AAAI Press (2015)
14. Lécuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: IEEE Symposium on Security and Privacy. pp. 656–672. IEEE (2019)
15. Li, B., Wang, Y., Singh, A., Vorobeychik, Y.: Data poisoning attacks on factorization-based collaborative filtering. In: NIPS. pp. 1885–1893 (2016)
16. Maini, P., Wong, E., Kolter, J.Z.: Adversarial robustness against the union of multiple perturbation models. In: ICML. Proceedings of Machine Learning Research, vol. 119, pp. 6640–6650. PMLR (2020)
17. McSherry, F., Mironov, I.: Differentially private recommender systems: Building privacy into the netflix prize contenders. In: KDD. pp. 627–636. ACM (2009)
18. O'Mahony, M.P., Hurley, N.J., Silvestre, G.C.M.: Recommender systems: Attack types and strategies. In: AAAI. pp. 334–339. AAAI Press / The MIT Press (2005)
19. Song, Y., Dahlmeier, D., Bressan, S.: Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In: PIR@SIGIR. CEUR Workshop Proceedings, vol. 1225, pp. 19–24. CEUR-WS.org (2014)
20. Standing Committee of the National People's Congress of Popular Republic of China: China internet security law (2017), http://www.npc.gov.cn/npc/c1481/201507/82ce4cb5549c4f56be8a6744cf2b3273.shtml
21. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: ICLR (Poster) (2014)
22. Zhang, F., Lee, V.E., Choo, K.R.: Jo-dpmf: Differentially private matrix factorization learning through joint optimization. Inf. Sci. **467**, 271–281 (2018)