I. Additional Results for Disguised Data Poisoning Attacks

A. Frequency Estimation

Impact of d. Fig. 1 reveals the impact of the parameter d on the overall gains in various datasets. Notably, we can only vary the parameter d on the two synthetic datasets, i.e., ZIPF and Uniform. We observe that the overall gains of the three attacks do not change with the varying d essentially. This is because that the overall gains of the target items are independent of d. In addition, the results indicate that our attack is comparable with MGA, outperforms RIA and RPA due to their randomness.

Moreover, we further study the disguise of the proposed attack by investigating the frequencies of neighboring items. Table I exhibits the impact of the varying d on the neighboring items. We find that our attack achieves the larger frequency gains of the neighboring items, which is similar to the impact of the privacy budget ϵ on the frequency gains of the neighboring items. It is because our attack focuses on increasing the frequencies of neighboring items, rather than randomly choosing items and maximal strategy.

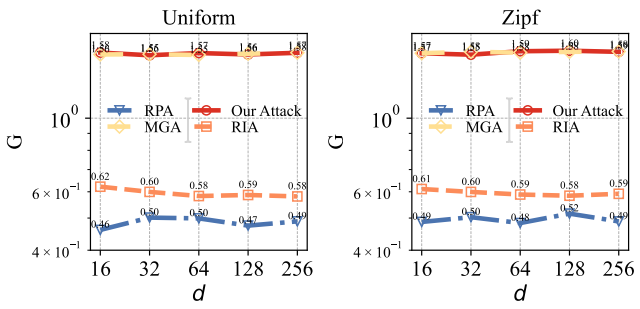

Fig. 1. The impact of the number of items in the domain d on the overall gains of the target items for the protocol OUE.

TABLE I
The impact of the number of items on the frequency gains of the neighboring items for protocol OUE.

Dataset	Attack	Index of neighboring item is 100				
		$d = 16$	$d = 32$	$d = 64$	$d = 128$	$d = 256$
ZIPF	Our Attack	0.15132 ↑	0.14608 ↑	0.15588 ↑	0.15661 ↑	0.15928 ↑
	MGA	−0.06106 ↓	−0.05948 ↓	−0.05669 ↓	−0.03155 ↓	−0.03291 ↓
	RIA	−0.06278 ↓	−0.05650 ↓	−0.06549 ↓	−0.05217 ↓	−0.06095 ↓
	RPA	0.04672 ↑	0.04497 ↑	0.04778 ↑	0.05008 ↑	0.05261 ↑
uniform	Our Attack	0.15831 ↑	0.15905 ↑	0.15442 ↑	0.15854 ↑	0.15653 ↑
	MGA	−0.06908 ↓	−0.06218 ↓	−0.05921 ↓	−0.03471 ↓	−0.03288 ↓
	RIA	−0.06021 ↓	−0.05902 ↓	−0.06118 ↓	−0.05986 ↓	−0.05983 ↓
	RPA	0.05112 ↑	0.05142 ↑	0.04942 ↑	0.04971 ↑	0.05008 ↑

B. Heavy Hitter Identification

Impact of top-k. Fig. 2 shows the impact of the parameter top-k on the success rates in multiple datasets. We observe that the success rates of our attack are low when top-k is small, which is identical with that of MGA. It is because during the iterations, the number of bits in per perturbation may not be enough to significantly improve the success rates of all target items when the top-k value is small. Specifically, our attack promotes all the target items to be in the top-20 heavy hitters with 5% of fake users.

Additionally, we assess the disguise of the proposed attack by investigating the promotion of the neighboring items' rankings. Table II shows the impact of the varying top-k on the neighboring items' rankings. We derive a similar conclusion regarding the superiority of our attack as analyzed above.

TABLE II
The impact of the number of heavy hitters top-k on the promotion of rankings of neighboring items for PEM.

Dataset	Attack	Index of neighboring item is 100			
		top-$k = 10$	top-$k = 15$	top-$k = 20$	top-$k = 25$
Fire	Our Attack	50 ↑	53 ↑	50 ↑	43 ↑
	MGA	−1 ↓	−3 ↓	0 -	−2 ↓
	RIA	0 -	−9 ↓	0 -	0 -
	RPA	0 -	12 ↑	0 -	0 -
IPUMS	Our Attack	32 ↑	41 ↑	40 ↑	40 ↑
	MGA	0 -	−3 ↓	2 ↑	2 ↑
	RIA	−4 ↓	0 -	0 -	0 -
	RPA	1 ↑	4 ↑	0 -	0 -
Compensation	Our Attack	27 ↑	30 ↑	37 ↑	37 ↑
	MGA	−1 ↓	4 ↑	4 ↑	3 ↑
	RIA	0 -	−9 ↓	0 -	−15 ↓
	RPA	1 ↑	3 ↑	−7 ↓	0 -
ZIPF	Our Attack	25 ↑	52 ↑	60 ↑	55 ↑
	MGA	6 ↑	6 ↑	7 ↑	6 ↑
	RIA	0 -	5 ↑	0 -	10 ↑
	RPA	0 -	−10 ↓	0 -	0 -
uniform	Our Attack	19 ↑	19 ↑	20 ↑	23 ↑
	MGA	0 -	−2 ↓	0 -	0 -
	RIA	0 -	1 ↑	0 -	−7 ↓
	RPA	−8 ↓	0 -	0 -	7 ↑

C. Mean-Variance Estimation

Impact of n_e. Fig. 3 shows the impact of the number of users n_e on the attack error MSE in two synthetic datasets, i.e., ZIPF [1] and Uniform [2]. We observe that the performance of our attack is better than IPA because of the introduction of more LDP noise in IPA. Additionally, the performance of our attack is worse than OPA, ultimately because the accuracy is sacrificed for the disguise.

II. Additional Results for Frequency Analysis-Based Defense

A. Details of Defense

As depicted in Algorithm 1, it applies the Discrete Cosine Transform to convert data into frequency components, focusing on low-frequency elements that carry significant information (cf. Lines 1-5). It then employs HDBSCAN clustering to identify and remove noise, ensuring the integrity of data for reliable analytics (cf. Lines 6-15).

B. Frequency Estimation

Impact of r. Fig. 4 shows the impact of the parameter r on the overall gains on diverse datasets for our defense and two baselines. The results indicate that our defense reaches the smallest increment of the frequencies, because of our defense's capability in handling noise and crafted data. Specifically, our defense can reduce the overall gain to a maximum of $e-7$ on the uniform dataset. The reasons are that the features of crafted

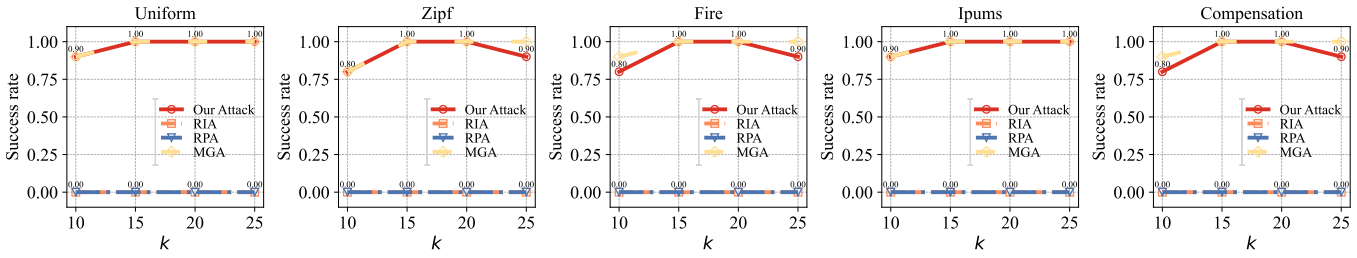

Fig. 2. The impact of number of heavy hitters top-k on success rates of the target items for PEM in our attack, MGA, RIA and RPA.

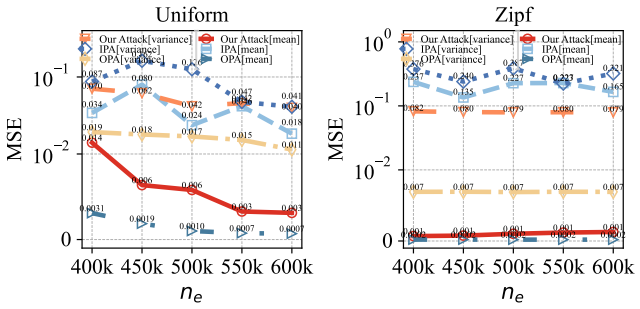

Fig. 3. The impact of the estimated number of genuine users n_e on the MSE of estimating mean and variance for PM.

Algorithm 1 FREQUENCY ANALYSIS-BASED DEFENSE

Input Poisoned_Data, a min cluster size, α low frequency ratio
Output clean_data

Apply Discrete Cosine Transform (DCT) to extract low-frequency components.
    $data\_dct \leftarrow \text{DCT}(Poisoned\_Data)$
        ▷ Use DCT-II for the transformation.
    $data\_low\_freq \leftarrow \text{Filtering}(data\_dct, \alpha)$
        ▷ Take the α of DCT as low-frequency components.
Convert low-frequency components into a two-dimensional array for clustering purposes.
    $data\_low\_freq\_2d \leftarrow \text{Reshape}(data\_low\_freq)$
▷ Reshape the array to $(-1, 1)$ so each sample has one feature.
Perform HDBSCAN clustering on the low-frequency components.
    $cluster \leftarrow \text{HDBSCAN}(\text{min\_cluster\_size} = a)$
    $cluster.\text{fit}(data\_low\_freq\_2d)$
    $cluster\_labels \leftarrow cluster.\text{labels\_}$
        ▷ Input Poisoned_Data to clusterer.
if No noise points are present **then**
    Print the number of clusters (excluding noise).
else
    Noise points are present in the data.
$noise\_indices \leftarrow \text{Find}(cluster\_labels == -1)$
        ▷ Tick out the noise points.
Remove noise points from the data.
    $clean\_data \leftarrow \text{Delete}(data, noise\_indices)$

data can be highlighted in the frequency domain and thus the crafted data can be effectively identified by clustering.

Impact of d. Fig. 5 shows the impact of the parameter d on the overall gains on two datasets for our defense and two baselines. We observe that the advantages of our attack remain apparent, because that the features of the crafted data are prominent in the frequency domain, enabling the effective identification through clustering.

Impact of α. Since the selection of low frequency components is also an important part of the proposed defense, we adjust the low-frequency ratio α within $[0.01, 0.3]$. Fig. 6 shows the impact of the parameter α on the overall gains in various datasets. We observe that our defense has a relatively good performance when $\alpha = 0.01$. Furthermore, our defense reaches the best performance when $\alpha = 0.1$ and 0.2. This is because that the low frequencies already contain enough information when the α is in $(0.1, 0.2)$.

C. Mean-Variance Estimation

Impact of n_e. Fig. 7 shows the impact of the number of users n_e on the accuracy gain (AG) in two different synthetic datasets for our defense and the baseline. We find that the AG of our attack is superior to that of the baseline because of the effectiveness of our defense. Additionally, performance of our defense does not affected by the the variations in the number of users. This is because even if n_e changes, the low-frequency component containing the main information can still provide stable and representative information about the data.

III. OTHER DEFENSES

Cao et al. [3] proposed statistical analysis based detection, and conditional probability based detection for a single target item. The work [4] designed an adaptive defense strategy to recover the exact aggregation frequency from a poisoning attack and the study [5] proposed a defense framework to estimate fake users specifically for MGA, both of which are compared with our defense. Wu et al. [6] investigated the detection of poisoning attacks against interactive LDP protocols for key-value data. However, it is not suitable for non-interactive LDP protocols and numerical data. The follow-up work [7] employed the multi-party computation to limit the attacker's ability to influence the heavy hitter outcomes. Likewise, Song et al. [8] proposed a verifiable LDP protocol based on an interactive framework to enhance the security of data aggregation by using Pedersen commitments. Nonetheless, the

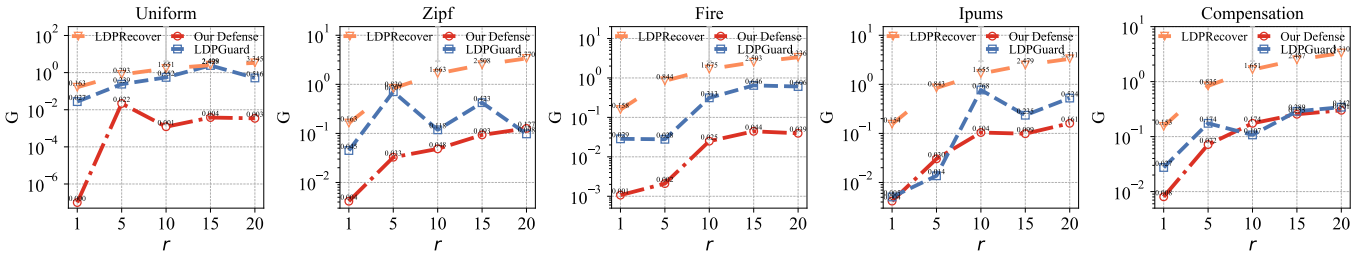

Fig. 4. The impact of number of target items r on overall gains of the target items in our defense, LDPGuard and LDPRecover.

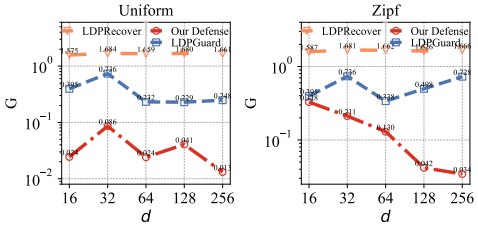

Fig. 5. The impact of the number of items in the domain d on the overall gains of the target items in our defense, LDPGuard and LDPRecover on two synthetic datasets.

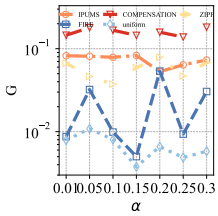

Fig. 6. The impact of low-frequency ratio α on overall gains of target items in our our defense and LDPRecover on two synthetic datasets.

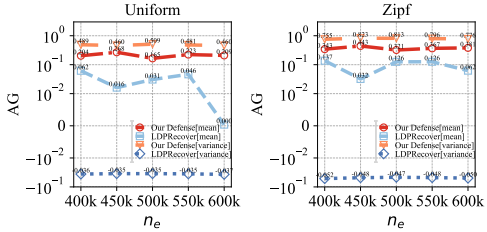

Fig. 7. The impact of the estimated number of genuine users n_e on the accuracy gain of estimating mean and variance in our defense and LDPRecover on two synthetic datasets.

communication overhead in the two works [7], [8] can be significant, which is not suitable for the case of high requirements for large-scale data processing and real-time performance. The approach in [9] integrated clustering on sampled data to mitigate data poisoning attack against LDP. Nevertheless, it is challenging to choose an appropriate sampling rate, as more bias can be introduced if the sampling rate is too low. Tong et al. [10] proposed three defense strategies against attacks on LDP frequent itemset mining protocols, filtering and limiting the number of items, introducing randomness into protocol parameters, and using cryptographic techniques. However, these methods have limited adaptability to attacks and may also affect the practicality of protocols. The latest works [11], [12] utilized basic mathematical principles to effectively reduce the attack gain. Work [11] proposed a novel zero-shot attack detection that identifies poisoning attacks against LDP exploiting distributional differences within the data. Moreover, Zheng et al. [12] formulated the problem as a zero-sum Stackelberg game, using a log-likelihood ratio test to identify and mitigate the impact of malicious workers. However, these methods may lack sensitivity to data distribution when the number of fake users is limited. Overall, to mitigate the proposed attack, new defenses in the further are needed.

REFERENCES

[1] T. Wang, J. Blocki, N. Li, and S. Jha, "Locally differentially private protocols for frequency estimation," in *Proc. 26th USENIX Secur. Symp.*, 2017, pp. 729–745.

[2] K. Huang, G. Ouyang, Q. Ye, H. Hu, B. Zheng, X. Zhao, R. Zhang, and X. Zhou. Ldpguardcode. 2023. [Online]. Available: https://github.com/TechReport2023/LDPGuard/blob/main/LDPGuardCode

[3] X. Cao, J. Jia, and N. Z. Gong, "Data poisoning attacks to local differential privacy protocols," in *Proc. 30th USENIX Secur. Symp.*, 2021, pp. 947–964.

[4] X. Sun, Q. Ye, H. Hu, J. Duan, T. Wo, J. Xu, and R. Yang, "Ldprecover: Recovering frequencies from poisoning attacks against local differential privacy," in *Proc. IEEE 40th Int. Conf. Data Eng.*, 2024, pp. 1619–1631.

[5] K. Huang, G. Ouyang, Q. Ye, H. Hu, B. Zheng, X. Zhao, and Zhang, "Ldpguard: Defenses against data poisoning attacks to local differential privacy protocols," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 7, pp. 3195–3209, 2024.

[6] Y. Wu, X. Cao, J. Jia, and N. Z. Gong, "Poisoning attacks to local differential privacy protocols for key-value data," in *Proc. 31st USENIX Secur. Symp*, 2022, p. 519–536.

[7] M. Naor, B. Pinkas, and E. Ronen, "How to (not) share a password: Privacy preserving protocols for finding heavy hitters with adversarial behavior," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2019, pp. 1369–1386.

[8] S. Song, L. Xu, and L. Zhu, "Efficient defenses against output poisoning attacks on local differential privacy," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 5506–5521, 2023.

[9] X. Li, N. Z. Gong, N. Li, W. Sun, and H. Li, "Fine-grained poisoning attacks to local differential privacy protocols for mean and variance estimation," in *Proc. 32nd USENIX Secur. Symp.*, 2023, pp. 1739–1756.

[10] W. Tong, H. Chen, J. Niu, and S. Zhong, "Data poisoning attacks to locally differentially private frequent itemset mining protocols," in *Proc. 31st ACM Conf. Comput. Commun. Secur.*, 2024.

[11] X. Li, Z. Li, N. Li, and W. Sun, "On the robustness of ldp protocols for numerical attributes under data poisoning attacks," in *Proc. 34th Netw. Distrib. Syst. Secur. Symp*, 2025.

[12] Z. Zheng, Z. Li, C. Huang, S. Long, and X. Shen, "Defending data poisoning attacks in dp-based crowdsensing: A game-theoretic approach," *IEEE Trans. Mobile Comput.*, vol. 24, no. 3, pp. 1859–1876, 2025.