CERIAS

The Center for Education and Research in Information Assurance and Security

Diffstats: Mitigating Data Poisoning Attacks to Local Differential Privacy

Xiaolin Li and Wenhai Sun

Department of Computer and Information Technology, Purdue University

Overview

Question

Local differential privacy (LDP) is a promising privacy-enhancing tool and widely used in many application. For instance, Google deployed LDP in the Chrome browser to collect users' default homepages for Chrome; and Apple adopted LDP on iOS to identify popular emojis, which are subsequently recommended to users. Recent studies show that LDP is vulnerable to data poisoning attacks in frequency estimation scenario, i.e., In Maximal Gain Attack (MGA) small portion of malicious local users can increase the estimated frequencies of a set of r target items $\mathrm{T} = \{t_1, t_2, \dots, t_r\}$, by carefully crafts their reports $\hat{Y}$ to maximize the overall frequency of target items. MGA remains an **unresolved issue in LDP estimation**.

Our Results

- We use the bits set to "1" in the users report as the additional information and by looking at the statistical differences between the benign and malicious users to detect the attackers' report precisely.
- We quantify the statistics differences between the benign and malicious users and proof the local error $E_{sq}(k)$ and global error E_{freq} to Indicates the relationship between MGA attacks and local as well as global statistical error.
- We propose a general framework by utilizing $E_{sq}(k)$ and E_{freq}, it allows the Server to constantly achieve high F-1 scores without any knowledge about the attacker. In the popular emoji identification task, our method achieves a fake user detection F1-score of **79.96%,** a **53.13%** improvement over the state-of-the-art Frequent itemset anomaly detection (FIAD) approach. Moreover, removing the reports from detected fake users effectively restores benign users' statistical results.

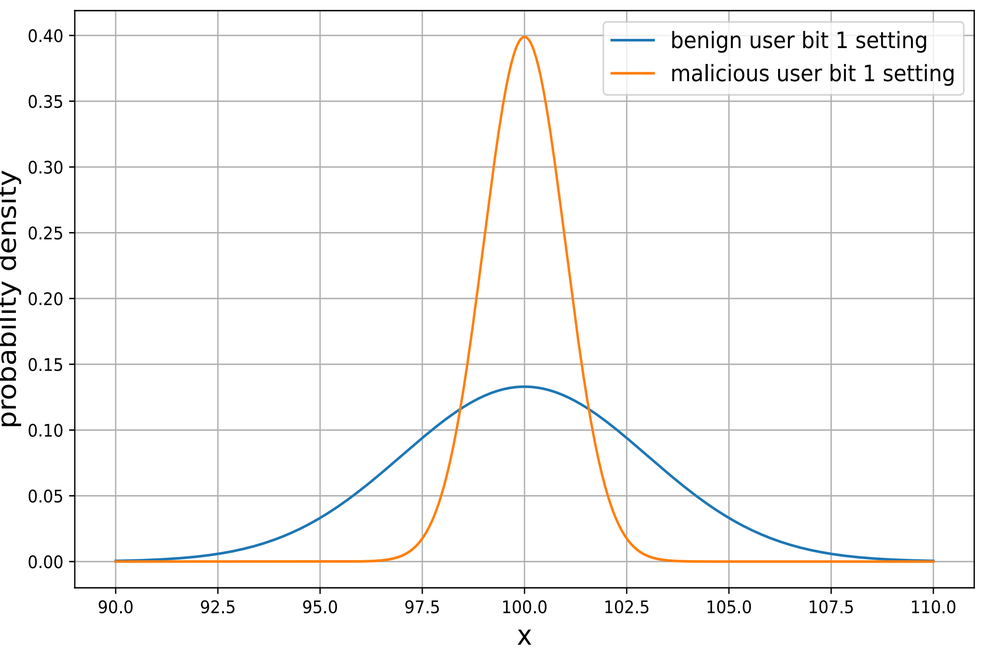
Our Result in Details

What is LDP

LDP enables n users to share their data $v \in [0, d]$ with an untrusted server through a local perturbation function $\Psi(\cdot)$ such that only obfuscated items $\Psi(v)$ is obtained by the server. Formally,

DEFINITION 1. *(ϵ-Local Differential Privacy). An algorithm $\Psi(\cdot) : \mathcal{D} \to \hat{\mathcal{D}}$ satisfies ϵ-LDP if for any $v_1, v_2 \in \mathcal{D}$ and for $y \in \hat{\mathcal{D}}$,* $\Pr[\Psi(v_1) = y] \leq e^{\epsilon} \Pr[\Psi(v_2) = y]$.

Statistical difference



The bit "1" settings of fake users tend to be more centralized than those of normal users

Statistical Error

Theorem 1 *For m fake users and $k \in [0, d]$, the expected error $\mathbb{E}[E_{sq}(k)]$ of the MGA and MGA-A is*

$$\mathbb{E}\left[E_{sq}^{MGA}(k)\right] = \begin{cases} m^2 \cdot (P(X=k)-1)^2 + Var(O_{MGA}^k), & \text{if } k = l_g, \\ m^2 \cdot (P(X=k))^2 + Var(O_{MGA}^k), & \text{otherwise,} \end{cases}$$

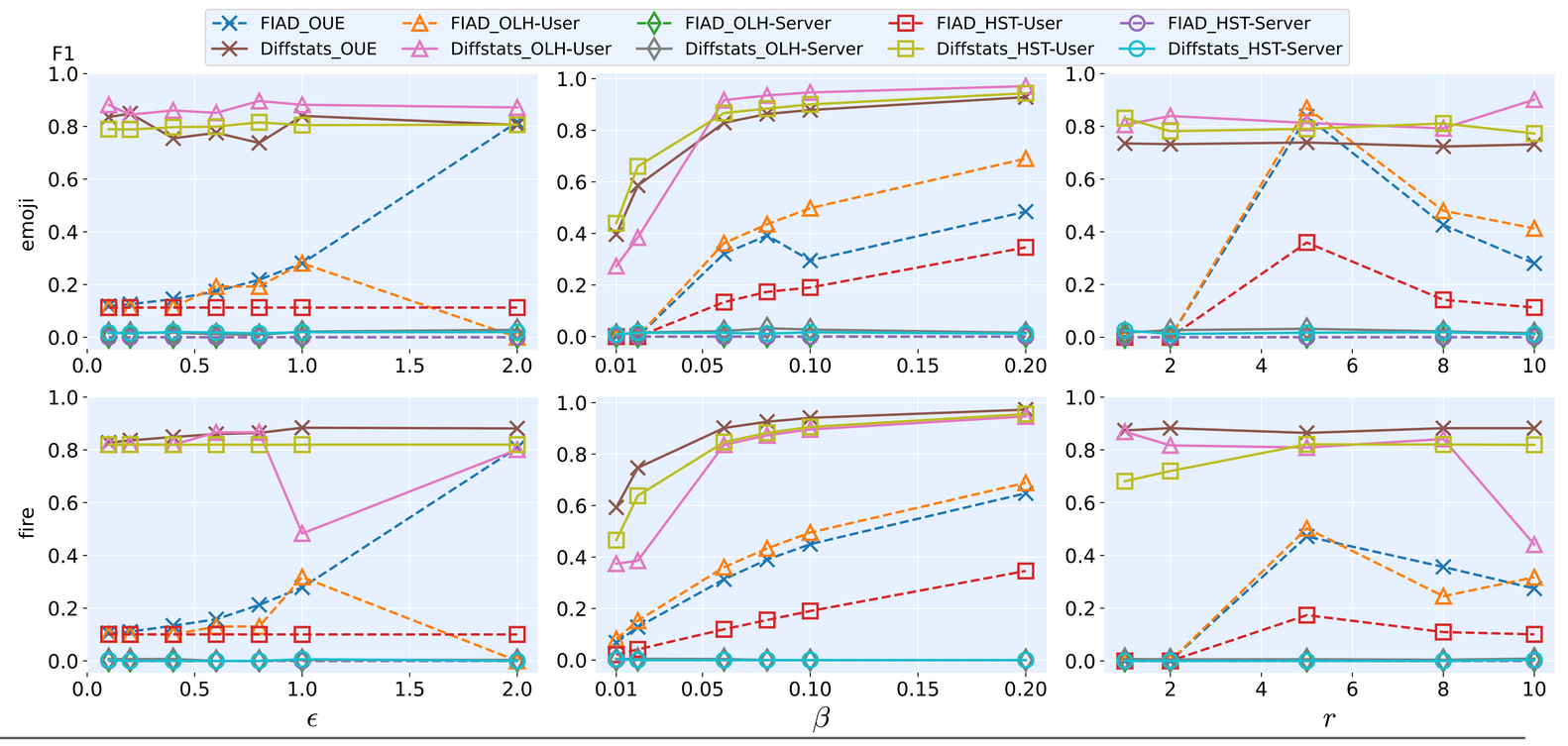
where $l_g = \lfloor p + (d-1)q \rfloor$ is the expected number of "1" in a genuine user's report and $Var(O_{MGA}^k) = (n-m)P(X=k)(1-P(X=k))$ denotes the variance of the observed frequency O_{MGA}^k under the MGA attack.

Theorem 2 *The error $E_{freq}(O_{MGA}^k, Y^k)$ between the observed frequency O_{MGA}^k under MGA or MGA-A attack and the expected frequency Y^k is*

$$E_{freq}(O_{MGA}^k, Y^k) = \frac{m^2}{n}\left(\frac{1}{P(X=l_g)} - 1\right) + \frac{(n-m)\cdot d}{n}$$

Evaluation on real dataset

Comparison with state-of-the-art fake user detection method FIAD on F1-score



PURDUE UNIVERSITY

CERIAS