PORE: Provably Robust Recommender Systems against Data Poisoning Attacks

Jinyuan Jia[1], Yupei Liu[2], Yuepeng Hu[2], Neil Zhenqiang Gong[2]
[1]Penn State University
[2]Duke University

08/09/2023

The first two authors made equal contribution.

Recommender Systems

- Widely deployed to engage users
  - Amazon, YouTube, TikTok, eBay
- Recommender system
  - Input: Rating-score matrix
  - Output: Recommended top-*N* items for each user
- Recommender system algorithm
  - Bayesian Personalized Ranking (BPR)
  - Item-based Recommendation (IR)
  - Neural Collaborative Filtering (NCF)

Recommender Systems

$$\begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \\ u_4 \end{array} \begin{array}{c} \begin{array}{ccccc} i_1 & i_2 & i_3 & i_4 & i_5 \end{array} \\ \begin{pmatrix} 5 & 4 & 5 & 0 & 0 \\ 4 & 0 & 5 & 0 & 0 \\ 0 & 0 & 5 & 1 & 4 \\ 4 & 2 & 0 & 0 & 4 \end{pmatrix} \end{array} \xrightarrow{\text{Recommender system algorithm}} \begin{pmatrix} u_1\colon \{i_5\} \\ u_2\colon \{i_5\} \\ u_3\colon \{i_1\} \\ u_4\colon \{i_3\} \end{pmatrix}$$

Rating-score matrix

Top-*1* recommendation

Recommender Systems are Vulnerable to Data Poisoning Attacks

- An attacker could inject fake users
  - By registering and maintaining fake accounts
- At most e fake users
  - Give an arbitrary rating score to an item
  - Rate as many items as the fake user wishes
- A poisoned recommender system makes attacker-desired, arbitrary recommendations.

Recommender Systems are Vulnerable to Data Poisoning Attacks

$$
\begin{array}{c|c}
 & \begin{matrix} i_1 & i_2 & i_3 & i_4 & i_5 \end{matrix} \\
\begin{matrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ \color{red}{u_5} \end{matrix} & \begin{pmatrix} 5 & 4 & 5 & 0 & 0 \\ 4 & 0 & 5 & 0 & 0 \\ 0 & 0 & 5 & 1 & 4 \\ 4 & 2 & 0 & 0 & 4 \\ \color{red}{0} & \color{red}{4} & \color{red}{5} & \color{red}{4} & \color{red}{0} \end{pmatrix}
\end{array}
\xrightarrow{\text{Recommender system algorithm}}
\begin{pmatrix} \color{red}{u_1: \{i_4\}} \\ u_2: \{i_5\} \\ \color{red}{u_3: \{i_2\}} \\ u_4: \{i_3\} \\ u_5: \{i_1\} \end{pmatrix}
$$

Poisoned rating-score matrix

Top-*1* recommendation

Limitations of Existing Defenses

- Empirical defenses
    - Cannot provide formal robustness guarantee

- Provable defenses
    - Designed for classifiers: Bagging
    - Suboptimal provable robustness guarantees

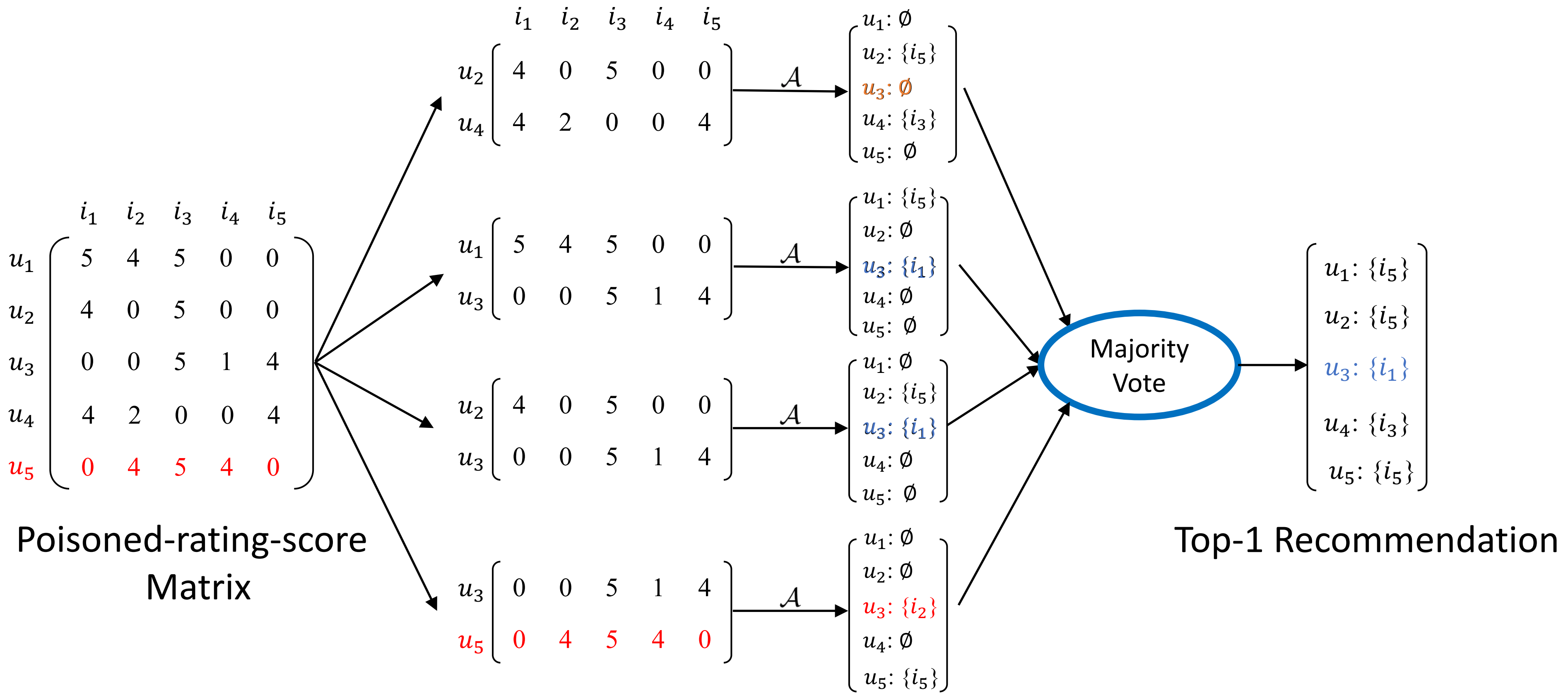
PORE: First Framework to Build Provably Robust Recommender Systems

- Create multiple sub-rating-score matrices
  - Each sub-rating-score matrix: rating scores of s randomly sampled users

- Build a base recommender system upon each sub-rating-score matrix
  - Use an arbitrary recommender system algorithm

- Build an ensemble recommender system
  - Majority vote

An Example for PORE



The Provable Robustness Guarantee of PORE

$\mathcal{E}_u$ A set of ground-truth items for a user *u*

$\mathcal{L}(\boldsymbol{M}, e)$ A set of all possible poisoned rating-score matrices

With a probability at least $1 - \alpha$, we have:

The set of recommended items for user u by our ensemble recommender system

$$\min_{\boldsymbol{M}' \in \mathcal{L}(\boldsymbol{M}, e)} |\mathcal{E}_u \cap \mathcal{A}(\boldsymbol{M}', u)| \geq r_u$$

certified intersection size

poisoned rating-score matrix

Computing the Robustness Guarantee

- Formulating the computation of r_u as the following optimization problem:

$$r_u = \underset{r' \in \{1,2,\cdots,\min(k,N)\}}{\operatorname{argmax}} \quad r'$$

$$s.t.\ \underline{p^*_{\mu_{r'}}} > \min\left(\min_{c=1}^{N-r'+1} \frac{N' \cdot (\overline{p}^*_{\mathcal{H}_c} + \sigma)}{c}, \overline{p}^*_{v_1} + \sigma\right)$$

Recommender System Setup

- MovieLens-1M
  - 1,000,209 rating scores
  - 6,040 users and 3,952 items
- Base recommender system algorithm
  - BPR
- Parameter setting
  - ***N'*=1** (number of items recommended by a base recommender system)
  - ***N*=10** (number of items recommended by our ensemble recommender system)
  - ***T*=100,000** (total number of base recommender systems)
  - **$\boldsymbol{\alpha}$=0.001** (1- α is the confidence score)
  - ***s*=500** (number of users in each sub-rating-score matrix)

Evaluation Metrics

- Precision@*N*
  - The fraction of recommended items that are in the ground truth set of a user
- Recall@*N*
  - The fraction of items in the ground truth set that are recommended
- F1-Score@*N*
  - Tradeoff between Precision@*N* and Recall@*N*

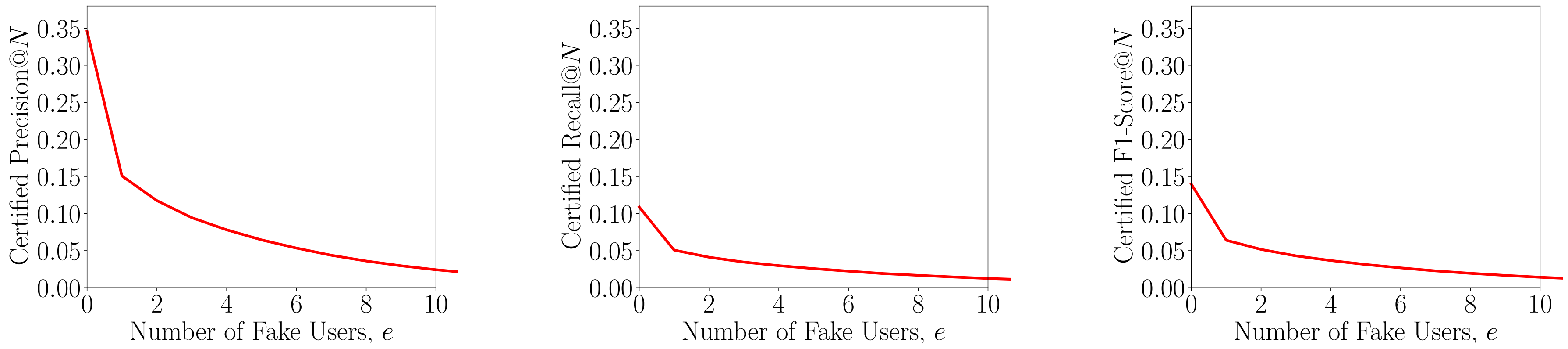
Evaluation Metrics

$$\text{Certified Precision@}N = \frac{r_u}{N}$$

$$\text{Certified Recall@}N = \frac{r_u}{|\mathcal{E}_u|}$$

$$\text{Certified F1-Score@}N = \frac{2 \cdot r_u}{|\mathcal{E}_u| + N}$$

PORE is Provably Robust against Data Poisoning Attacks



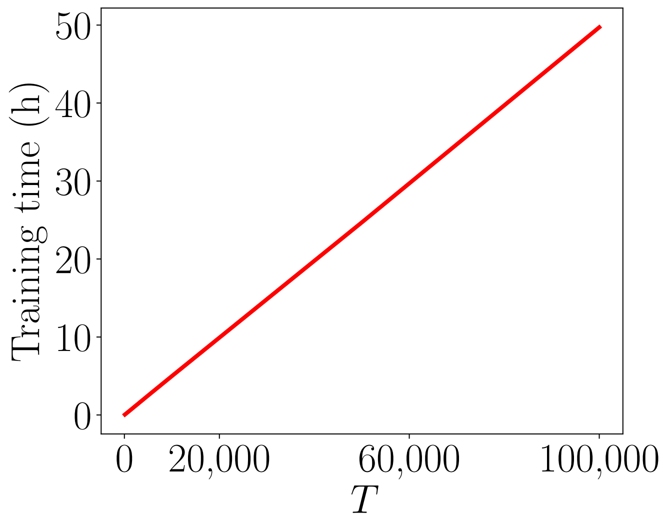
PORE Maintains Utility

PORE

Algorithm	Precision@10	Recall@10	F1-Score@10
BPR	0.324449	0.118385	0.144765
Ensemble BPR	0.362945	0.119441	0.151509

Our PORE maintains utility without attacks.

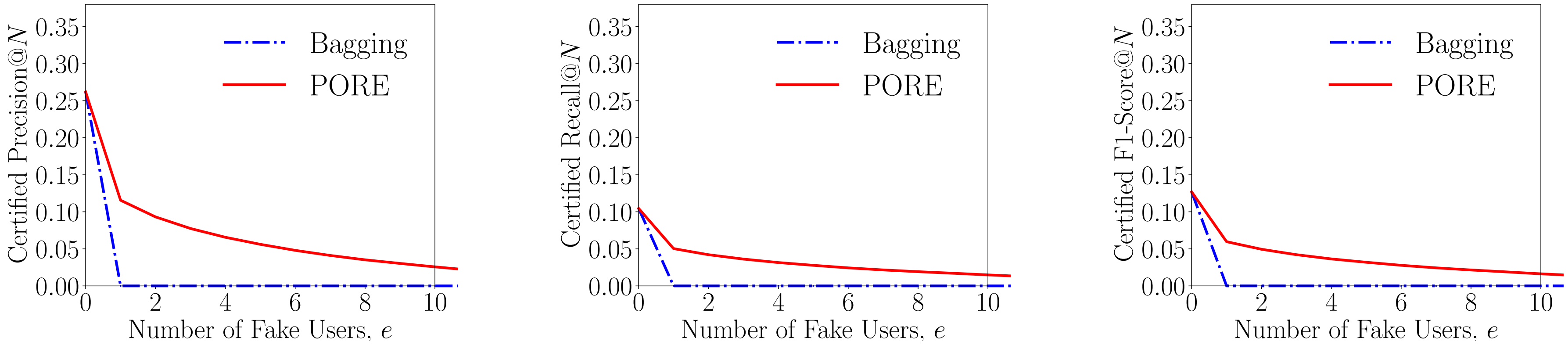
Time Complexity of PORE



Compared Method

- Bagging
  - State-of-the-art method to build provably robust classifier

PORE Outperforms the Existing Method



Summary

- We propose the first framework to build provably secure recommender systems
- Our PORE could be applied to an arbitrary recommender system algorithm
- Our PORE outperforms existing method extended from classifiers