Privacy-Preserving Synthetic Educational Data Generation

Jill-Jênn Vie Tomas Rigaux Sein Minn

Inria

September 15, 2022

Goal

- It is hard to get access to educational data for research, considered too sensitive
- Open data is great! But a dataset posted online may be archived forever (privacy issues)
- How about having instead access to a fake dataset? (ex. for reproducibility of experiments)

Outline

- Privacy issues
- Format of educational tabular data
- Framework for assessing privacy leaks in data generation
  - Membership inference
  - Metrics: utility and re-identification
- We present: generative models, attack model, results

Removing names / pseudonymizing does not ensure privacy

Using the pseudonymized Netflix dataset of ratings given by users (identified by ID) on movies, Narayanan and Shmatikov (2008) matched:

- some public ratings on IMDb of some user's public profile
- with all their private ratings in the Netflix dataset

revealing their political & sexual preferences or religious views.

Arvind Narayanan and Vitaly Shmatikov (2008). "Robust de-anonymization of large sparse datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, pp. 111–125

Few points are enough to uniquely identify users

4 timestamp-location points are needed to uniquely identify 95% of individual trajectories in a dataset of 1.5M rows

Yves-Alexandre De Montjoye et al. (2013). "Unique in the crowd: The privacy bounds of human mobility". In: *Scientific reports* 3.1, pp. 1–5

15 demographic points are enough to re-identify 99.96% of Americans

Luc Rocher, Julien M Hendrickx, and Yves-Alexandre De Montjoye (2019). "Estimating the success of re-identifications in incomplete datasets using generative models". In: *Nature communications* 10.1, pp. 1–9

Intuition

Knowledge parameters should be safe to be shared

User parameters should not be the true ones, but drawn from the same distribution (or blurred)

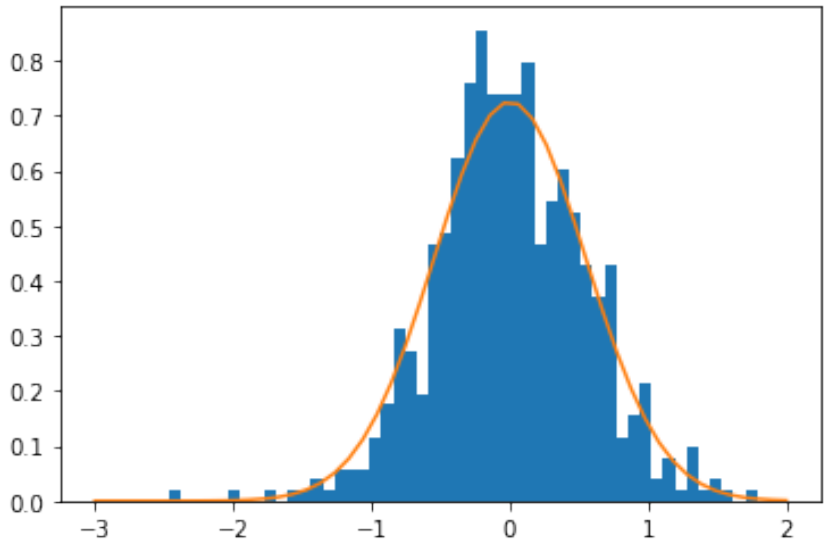


Figure 1: User ability parameters for an educational dataset

Example: the Duolingo SLAM dataset (Settles et al. 2018)

	PRON	VERB	PRON	NOUN	CONJ	PRON	VERB	PRON	NOUN
correct:	She	is	my	mother	and	he	is	my	father
student:	she	is		mader	and	he	is		fhader
label:	O	O	✖	✖	O	O	O	✖	✖

user ID	action ID	outcome	description
2487	384	1	user 2487 got token “I” correct
2487	242	0	user 2487 got token “ate” incorrect
2487	39	1	user 2487 got token “an” correct
2487	65	1	user 2487 got token “apple” correct

We want to generate data under this format, using existing data.

Item response theory (IRT) for response pattern generation

Well known model (Rasch, 1961) denoted by IRT

Ex. r_{ij} is 1 if user i gets a positive outcome on action (item) j

$$p_{ij} = \Pr(R_{ij} = 1) = \sigma(\theta_i - d_j)$$

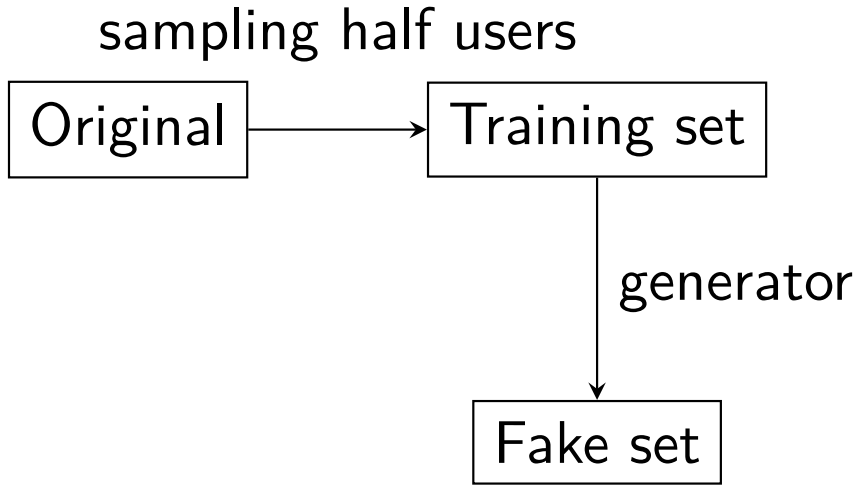
where θ_i is ability of user i and d_j is difficulty of action j

Trained using Newton's method: minimize log-loss
$\mathcal{L} = \sum_{i,j}(1 - r_{ij}) \log(1 - p_{ij}) + r_{ij} \log p_{ij}$

Generation

We select users from the original dataset to form a training dataset, either by random sampling or following a criterion (such as students who went to a given school, or students with special needs)

Then we use a generative model to make a fake dataset



To generate educational data, we can have two generative models:

- Sequence generation: Predicting the next action ID
- Response pattern generation: Predicting the outcome given user parameter and action ID

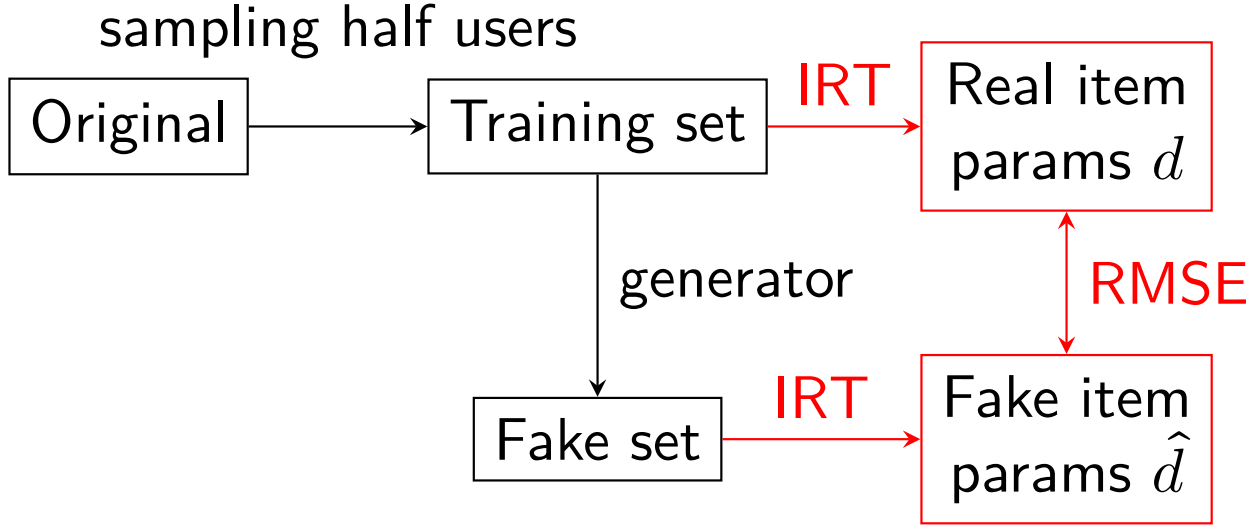
Utility: fake dataset should be useful

Practitioners who conduct study on the real and fake dataset should have similar findings

↓

Trained IRT model on original dataset should have parameters that are not too far in $RMSE = \sqrt{\sum_{j=1}^{N}(d_j - \widehat{d_j})^2}$

(where $d_j, \widehat{d_j}$ are item j's inferred difficulty from the real and fake datasets)

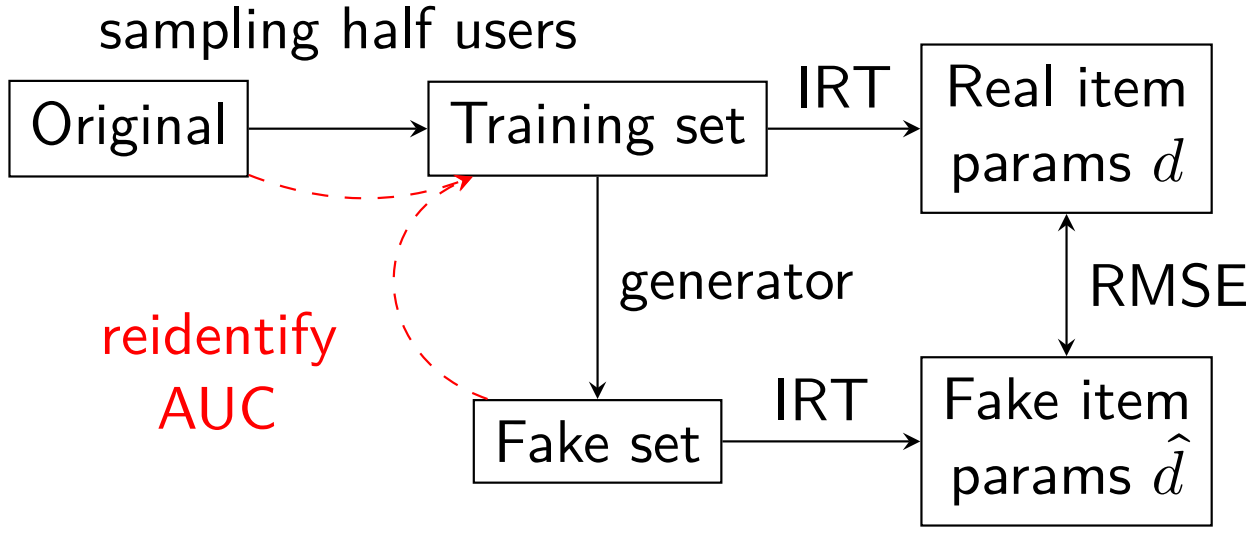


Membership inference: reidentification task

It should not be easy to re-identify people / the fake dataset should not leak too much information about participants

↓

An attacker has to guess, from the original and fake sets, who was in the training set (predict 1 if in training, 0 otherwise)



(framework inspired by NeurIPS "Hide and Seek" challenge in healthcare by Jordon et al., 2020)

Example scenarios of membership inference

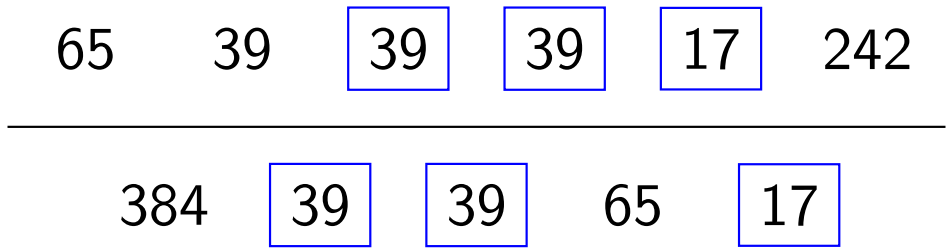
Membership inference seems innocuous, but could lead to privacy issues.

For instance, if we want to publish a dataset of test results from students with special needs using an anonymizing method, it shouldn't be possible to guess who was selected (= has special needs) in the training dataset.

More generally, any leak of information is potentially bad.

Reidentification: attack model

We use a heuristic based on Longest Common Subsequence (LCS) to reidentify



LCS: $39 - 39 - 17$ with length 3

For each user pair in the original $\times$ fake datasets, we compute the LCS between them; it gives a matching score which is the normalized maximum LCS on all fake users.

Original users with highest matching score are expected to be in the training set. To evaluate, we compute the Area under the ROC curve (AUC) associated with those scores.

Users with too few actions (in the information entropy sense) are excluded.

Experiments

Baseline: "Drop $p\%$" is dropping $p\%$ of rows and renumbering the user IDs

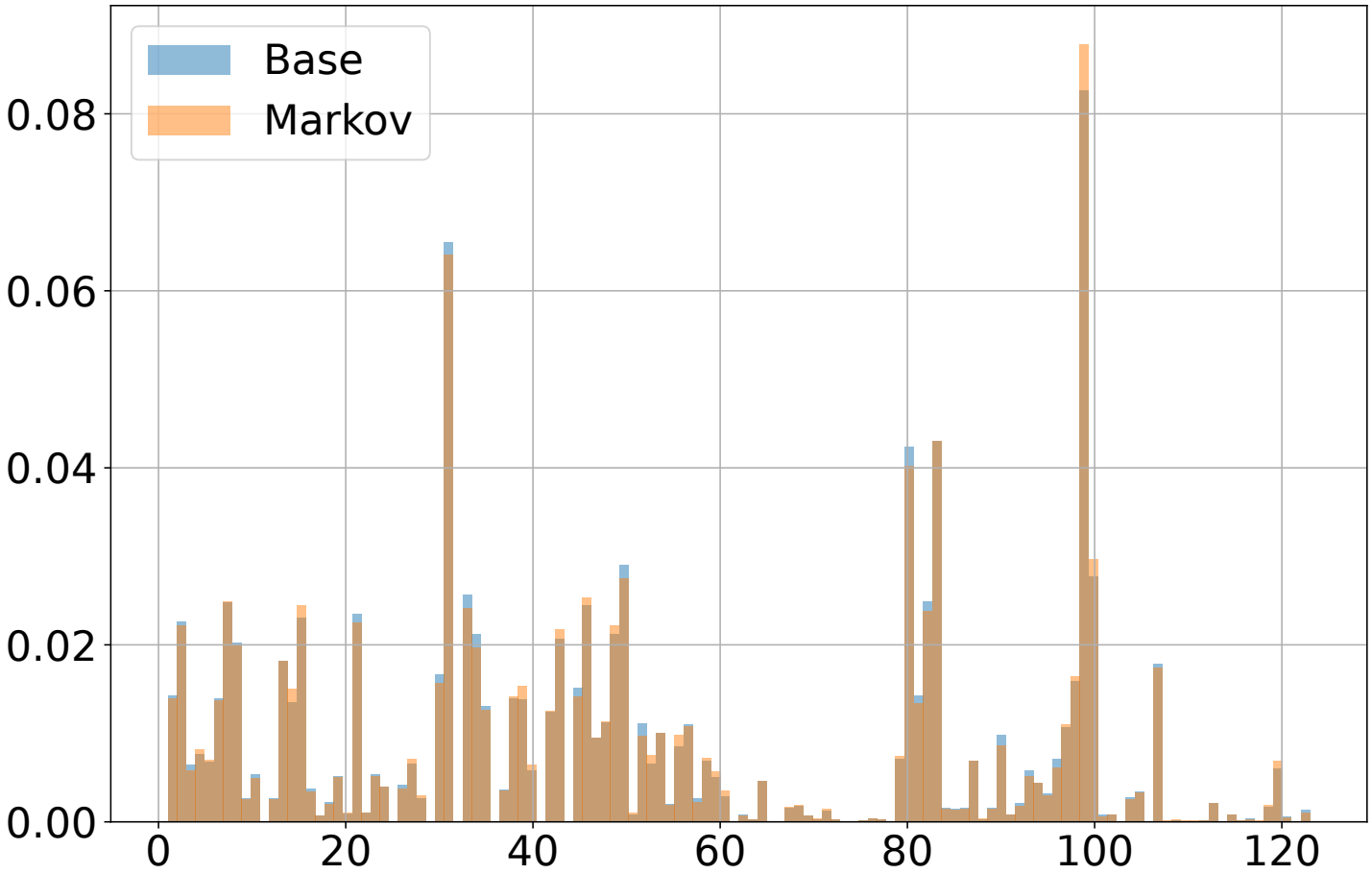
Sequence generation model: recurrent neural network (RNN) or Markov chain (probability to jump from an action to another)

Predicting the outcome: Rasch model (IRT)

Datasets (publicly available)

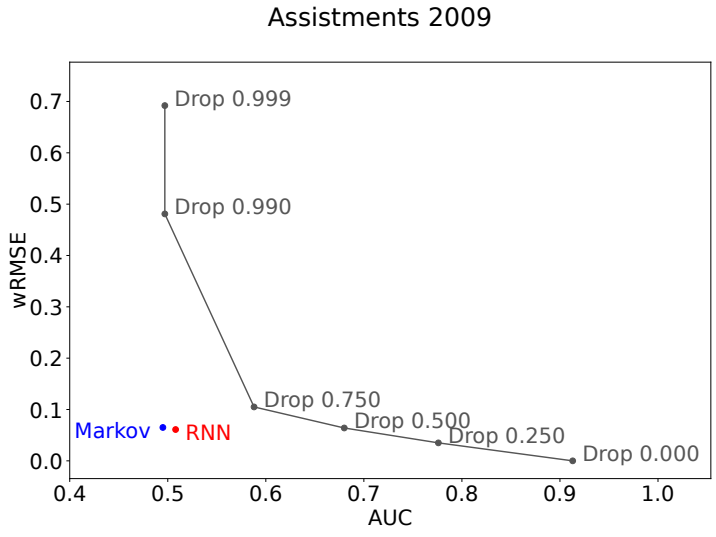
- the Duolingo dataset described above, 1M rows of English people learning French
- ASSISTments 2009 dataset (action types are mathematical skills that are accessed)

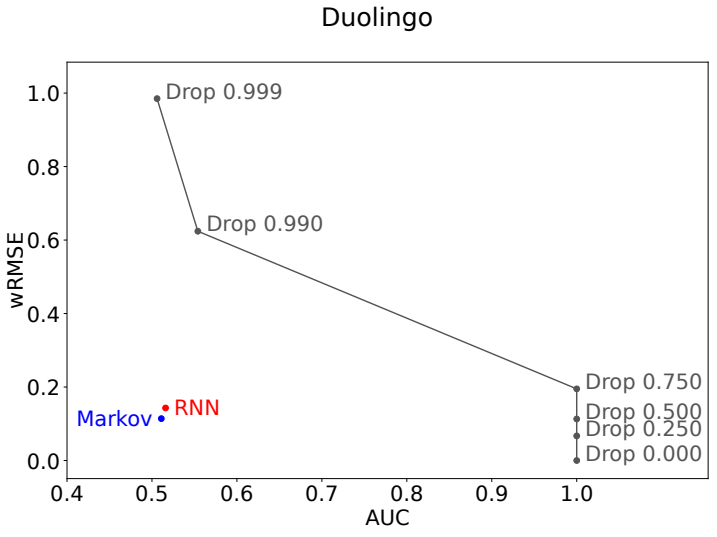
Histogram of actions (y-axis: frequency)



Actions in the ASSISTments dataset

Quantitative results





$\downarrow$ low distance between real and fake parameters, lower is better (high utility)

$\leftarrow$ low reidentification score, lower is better (hard to identify)

- Both Markov and RNN generate datasets with high utility and low reidentification score
- RNN generates better sentences (Duolingo dataset), see the paper for examples

Take home message

We managed to generate fake datasets that are:

- useful for practitioners (because item difficulties can be estimated similarly)
- hard to re-identify (because membership inference is not possible)

Extensions:

- Our approach can be easily generalized to more complicated datasets
- With more columns it is even easier to re-identify

Let's share the data of people who do not exist!
Generating synthetic datasets for reproducing experiments

Thanks! Questions?

- Slides on jjv.ie/slides/ectel2022.pdf
- Code on github.com/Akulen/PrivGen

- De Montjoye, Yves-Alexandre et al. (2013). "Unique in the crowd: The privacy bounds of human mobility". In: *Scientific reports* 3.1, pp. 1–5.
- Jordon, James et al. (2020). "Hide-and-seek privacy challenge". In: *arXiv preprint arXiv:2007.12087*.
- Narayanan, Arvind and Vitaly Shmatikov (2008). "Robust de-anonymization of large sparse datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, pp. 111–125.
- Rocher, Luc, Julien M Hendrickx, and Yves-Alexandre De Montjoye (2019). "Estimating the success of re-identifications in incomplete datasets using generative models". In: *Nature communications* 10.1, pp. 1–9.