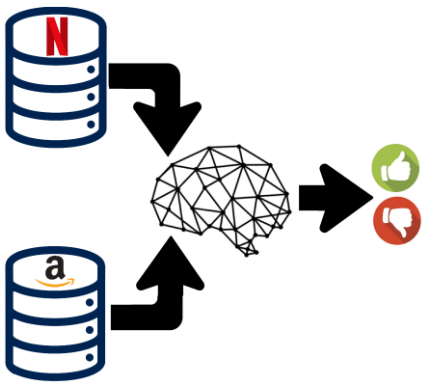
Together or Alone:
The Price of Privacy in Collaborative Learning

Balázs Pejó (SnT) & Gergely Biczók (CrySys) & Qiang Tang (LIST)

Collaborative Learning



Few data holders would like to train a machine learning model together in order to achieve higher accuracy.

Example

Two recommendation system (*RecSys*) providers (such as Netflix and Amazon) would like to collaborate to provide their users with better predictions.

Privacy Issue

In the training process the privacy of the data holders may be compromised. To protect their data, they can apply a privacy preserving mechanism before the training which inevitably will affect the trained model's accuracy.

Research Questions

- What are the possible privacy parameters that make the collaboration more accurate than training alone?
- What is the optimal privacy parameter?
- How much accuracy is lost due to the privacy-preserving mechanism?

Game Theoretic Model

Variable	Description
$\mathbf{p_n} \in [0,1]$	Privacy Parameter
$\mathbf{C_n} \in R^+$	Privacy Weight
$\mathbf{B_n} \in R^+$	Accuracy Weight
$\boldsymbol{\theta_n}$	Error by Training Alone
$\boldsymbol{\phi_n}(p_1,p_2)$	Error by Training Together
$\mathbf{b}(\theta_n,\phi_n)$	Benefit Function
$\mathbf{c}(p_n)$	Privacy Loss Function

- $p_n = 0$: no protection
- $p_n = 1$: maximal protection
- $C_n = 0$: Privacy unconcerned
- $C_n > 0$: Privacy concerned
- Collaborate if $\theta_n > \phi_n$

The normal form representation of the **Collaborative Learning Game** is a tuple (N,Σ,U) where the set of the players is $N=\{1,2\}$, their actions are $\Sigma=\{p_1,p_2\}$ while their utility functions are $U=\{u_1,u_2\}$ such that for each $n \in N$:

$$u_n(p_1,p_2) = B_n \cdot b(\theta_n,\phi_n) - C_n \cdot c(p_n)$$

To measure the hypothetical loss in accuracy due to privacy constraints, we define the **Price of Privacy**:

$$1 - \sum_n b(\theta_n,\phi_n(p_1,p_2)) \,/\, \sum_n b(\theta_n,\phi_n(0,0))$$

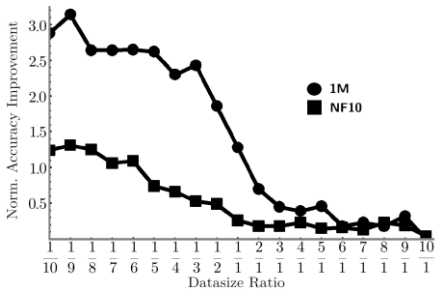
Equilibria

- In the **privacy unconcerned case** (i.e., $C_2=0$), the Nash Equilibrium (NE) is one of the following:
  - $(p_1^*,p_2^*) = (0,0)$
  - $(p_1^*,p_2^*) = ([\partial_{p1} b \cdot \partial_{p1} \phi_1 / \partial_{p1} c]^{-1}(C_1/B_1),0)$
  - $(p_1^*,p_2^*) = (1,1)$
- No collaboration is **trivial NE:** $(p_1^*,p_2^*)=(1,1)$
- A **non-trivial NE** exists, if $\partial^i_{p1}\phi_1=\partial^i_{p2}\phi_2$ holds for $i=\{1,2\}$.
  - This condition is quite natural, and indeed true in our RecSys case (*Figure: Training with an Unconcerned*)

Measuring ϕ for a RecSys

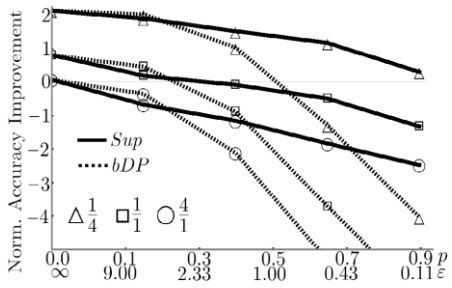
Dataset: Movilens (1 million ratings) / Netflix (10 million ratings)
Training Algorithm:
Matrix Factorization via Stochastic Gradient Descent
Privacy Method:
Suppression (removing data) / **D**ifferential **P**rivacy (adding noise)

Alone vs Together



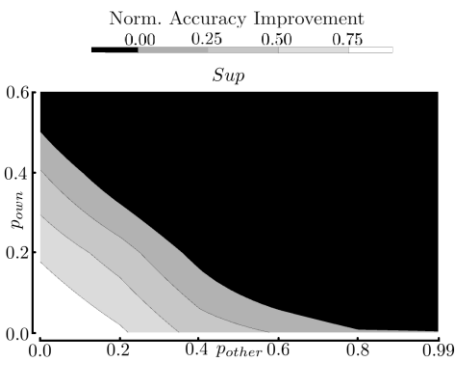
Training together is superior to training alone for both datasets and all size ratios.

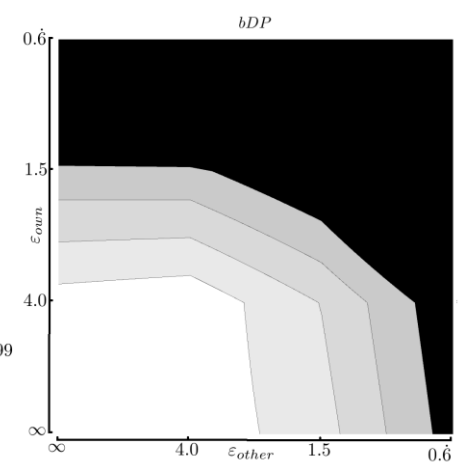
Training with an Unconcerned



The dataset size only effects the accuracy through a constant factor, i.e., there existence a non-trivial pure strategic NE.

Training with a Concerned

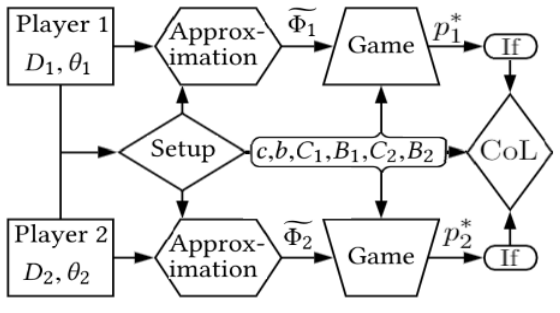




By degrading the quality of a given player's data, this player's accuracy will be more effected.

Big Picture

The exact NE (p_1^*,p_2^*) depends on the function ϕ which need to be calculated in advance.



Approximating ϕ

- **Self-Division:** Imitating CoL Game by splitting the local dataset to mimic collaboration
- $(\Theta_n - \phi_n) \leq (\vartheta_n - \varphi_n)$
  If D_n is too small
- $(\Theta_n - \phi_n) \geq (\vartheta_n - \varphi_n)$
  If D_n is too large

Variable	Description
$\mathbf{D_n}$	Dataset
$\mathbf{d}$	Density
$\boldsymbol{\vartheta_n}$	Approximated θ_n
$\boldsymbol{\varphi_n}$	Approximated ϕ_n

$$(\Theta_n - \phi_n) \approx (\vartheta_n - \varphi_n) \;\leftrightarrow\; 100000 \approx d \cdot |Dn|$$

	P1 - Sup	P2 - Sup	P1 - bDP	P2 - bDP
Avg. Approx. Error	0.001867	0.001055	0.001731	0.000917

Conclusion

Collaborative Learning is only feasible when either one of the data holders is privacy unconcerned or they have approximately the same dataset sizes with very low privacy weights.

Reference

Balázs Pejó, Gergely Biczók and Qiang Tang: *Together or Alone: The Price of Privacy in Collaborative learning* https://arxiv.org/abs/1712.00270