Challenging Differential Privacy
The Case of Non-interactive Mechanisms

Raghavendran Balu[1], Teddy Furon[1] and Sébastien Gambs[1,2]

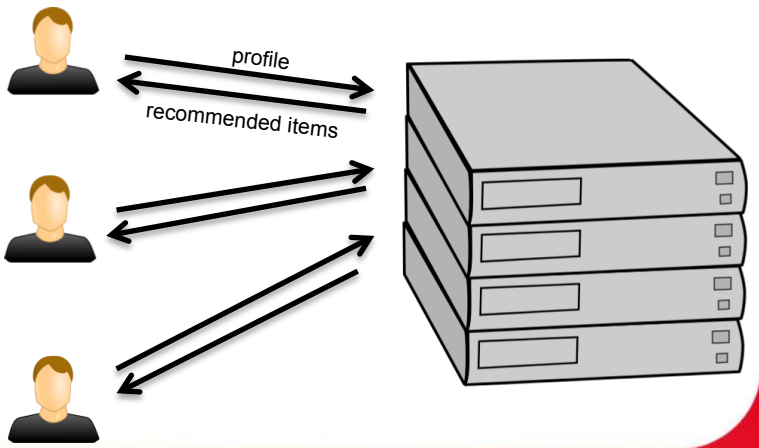
1. INRIA, Rennes, France
2. University of Rennes 1 / IRISA, France

Outline

1. Personalization and privacy
2. Theoretical analysis
3. Practical decoders
4. Experiments

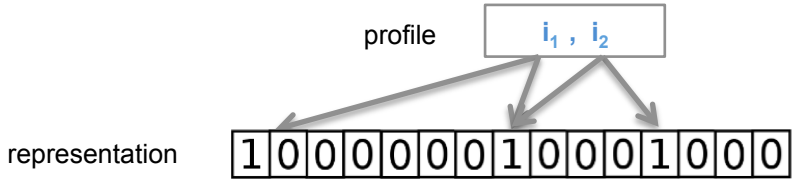
Personalized recommendation system

- Use user-item similarity for prediction and ranking
- Maintain a user profile
  - Composed of past items and preferences
- Aggregate user profiles for similarity computation
  - Needs profile information exchange

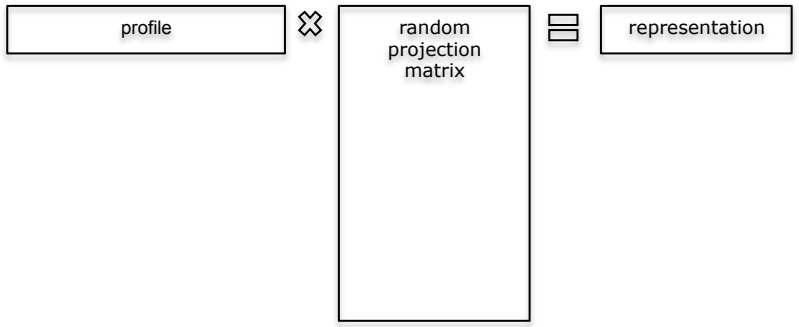


Profile representation

- Compact representation of user profile
- Examples
  - Bloom filter hash-based probabilistic data structure
    - Items stored as bits addressed by k-hash functions



  - Random projection in low dimensional space (Johnson–Lindenstrauss transformation)
    - Items are vector points



Privacy

- User profile is personal data
- Sanitization mechanism
  - Modify the representation before its disclosure
- Measures of privacy
  - *k*-anonymity, *l*-diversity, *t*-closeness, …
  - We choose differential privacy

Differential privacy [1]

- $\mathcal{F} : \mathcal{D}^n \rightarrow \mathcal{D}^n$: **ε**-differentially private randomized function if for all neighboring profiles **x'** of **x**, where $\mathbf{x}', \mathbf{x}, \mathbf{t} \in \mathcal{D}^n$

$$\mathbb{P}[\mathcal{F}(\mathbf{x}) = \mathbf{t}] \leq e^{\epsilon}.\mathbb{P}[\mathcal{F}(\mathbf{x}') = \mathbf{t}]$$

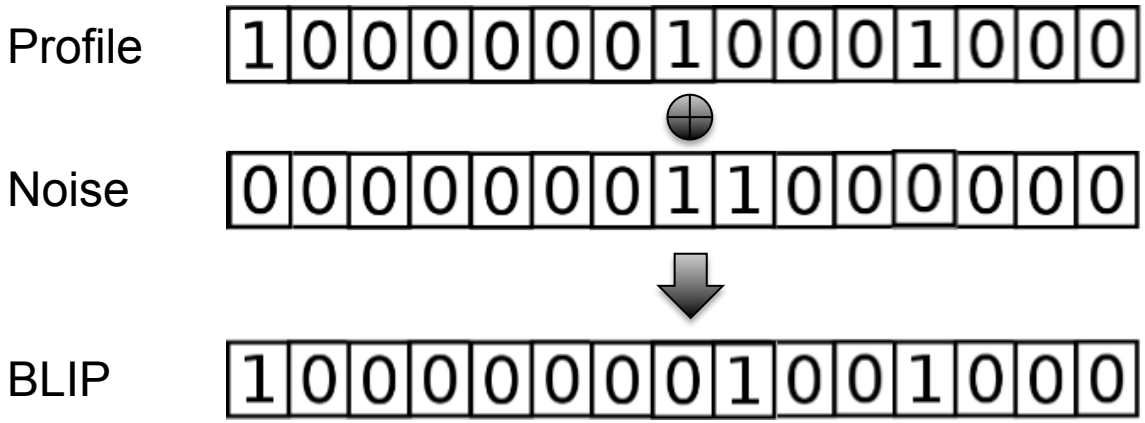
- Achieved by randomized perturbation of data
  - Interactive: perturbed for each query
  - Non-interactive: perturbed and published

- In personalization:
  - Altering one item will not change the probability of a user profile representation **t**

[1] Dwork, C.: Differential privacy, ICALP 2006

BLIP [2]

1. Profile representation: Bloom filter B_P
2. ε-differentially private representation: BLIP: $\tilde{\mathbf{B}}_P = \mathbf{B}_P \oplus \mathbf{Noise}$
   - Randomization by binary noise, i.i.d. Bernoulli distribution
   - Random flipping with probability $p_\epsilon = 1/(1+e^{\epsilon/K})$.



[2] Alaggan, M., Gambs, S., Kermarrec, A.-M.: BLIP: Non-interactive Differentially-Private Similarity Computation on Bloom Filters, SSS 2012

JLT [3]

1. Profile representation: Johnson-Lindenstrauss Transform

$$\mathbf{Y}_P = \sum_{j \in P} \mathbf{X}_j$$

   – Codeword $\mathbf{X}_j$ is real vector of length *L*: $X_j(i) \overset{i.i.d.}{\sim} \mathcal{N}(0, 1/L)$

2. (**ε,δ**)-differentially private representation: $\tilde{\mathbf{Y}}_P = \mathbf{Y}_P + \mathbf{Noise}$

   – Randomization by noise: $Noise(i) \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

with $L \geq 2(\log(N) + \log(2/\delta)), \quad \sigma \geq \frac{4}{\epsilon}\sqrt{\log(1/\delta)}$ and $\epsilon < \log(1/\delta)$

[3] Kenthapadi, K., Korolova, A., Mironov, I., Mishra, N.: Privacy via the Johnson Lindenstrauss transform. *Journal of Privacy and Confidentiality*, *5*(1), 39-71.

Our Contribution

THEORETICAL ANALYSIS

Single decoder for BLIP

- The adversary infers the presence of a single item
  - Hypothesis test:
    - Item *j* is not in the profile $\mathbb{P}[\tilde{B}_P(\ell), X_j(\ell)] = \mathbb{P}[\tilde{B}_P(\ell)]\mathbb{P}[X_j(\ell)]$
    - Item *j* is in the profile $\mathbb{P}[\tilde{B}_P(\ell), X_j(\ell)] = \mathbb{P}[\tilde{B}_P(\ell)|X_i(\ell)]\mathbb{P}[X_i(\ell)]$
  - Mutual information $I(\tilde{\mathbf{B}}_P; \mathbf{X}_j)$ measures the amount of information the BLIP is disclosing about the presence of item *j*.

- By testing sequentially all *N* items, the adversary reconstructs the user items set *P*
  - α Probability of missing a user item
  - η Probability of including at least one wrong item

$$\log \eta \geq \log N - \frac{I(\tilde{\mathbf{B}}_P; \mathbf{X}_j)}{1-\alpha}$$

Shift of paradigm

- We are now interested by testing if a given subset of *c* items is the true user items set *P*.
  - We call it a *joint* decoder
- By testing all *c*-items subsets, the adversary finds the true *P* with probability

$$\log \eta \geq \log N - \frac{I(\tilde{\mathbf{B}}_P; (\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_c}))}{c(1-\alpha)}$$

Theoretical performances of a joint decoder

- From [4]: $I(\tilde{\mathbf{B}}_P; (\mathbf{X}_{j_1}, \ldots, \mathbf{X}_{j_c}))/c \geq I(\tilde{\mathbf{B}}_P; \mathbf{X}_j)$
- This shows that:

  Joint decoding is always more efficient than single decoding
- Our paper gives these theoretical performances for
  - Single and Joint decoding, applied to
  - BLIP and JLT approach
- Depending on the setup, there is a substantial difference between performances of single and joint decoding

[4] Universal Fingerprinting: Capacity and Random-coding exponents, P. Moulin, ISIT 2008

A nice theoretical result but…

It does NOT work in practice because

- For $c \ll N$, the number of *c*-items subset is

$$\binom{N}{c} = O(N^c)$$

It is not tractable to test all *c*-items subsets!!!

Our Contribution

PRACTICAL DECODERS

Markov chain

- Instead of testing all subsets:
  - We do a guided random walk in the space of *c*-items subsets
  - This random walk leads to the most likely *c*-items subsets
- Input: The adversary observes one BLIP $\tilde{\mathbf{b}}$
- Starting point: a random subset $P^{(0)}$
- New state $P^{(t+1)}$ sampled with transition probability

$$\mathbb{P}[P^{(t+1)} = P | P^{(t)}] = \frac{\mathbb{P}[\tilde{\mathbf{B}}_P = \tilde{\mathbf{b}} | P]\mathbb{P}[P]}{\sum_{P' \in \mathcal{V}(P^{(t)}, i)} \mathbb{P}[\tilde{\mathbf{B}}_{P'} = \tilde{\mathbf{b}} | P']\mathbb{P}[P']}$$

Likelihood of *P*

Prior of *P*

Neighborhood of $P^{(t)}$

- Converges to $\mathbb{P}[P|\tilde{\mathbf{b}}]$ as $t \to \infty$, in practice at $t > T$ (burn-in period)

Monte Carlo

- Once the Markov chain has converged,
  - We sample subsets according to posteriori probability $\mathbb{P}[P|\tilde{\mathbf{b}}]$
  - We let the chain running for M more iterations.

- Possible outputs of the Monte Carlo Markov Chain:
  - Marginal a posteriori probability per item estimated by empirical frequency

$$\hat{\mathbb{P}}[j \in P|\tilde{\mathbf{b}}] = |\{t \in [T+1, T+M]|j \in P^{(t)}\}|/M$$

  - Maximum a posteriori estimator of the profile

$$\hat{P} = \arg \max_{T+1 \leq t \leq T+M} \mathbb{P}[P^{(t)}|\tilde{\mathbf{b}}]$$

EXPERIMENTS

Setup

- Datasets
  - Digg: Social bookmarking dataset
  - MovieLens: Movie rating dataset

	Nb of users	Training set size	Testing set size	N	c_{avg}	Sparsity %
Digg	531	331	200	1237	317	25.63%
MovieLens	943	600	343	1682	106	6.30%

- Attack algorithms:
  - Single decoder
  - Joint decoder with uniform prior
  - Joint decoder with prior estimated from the training set
  - Popularity-based attack (baseline)
    - The *c* most popular items

Privacy measures

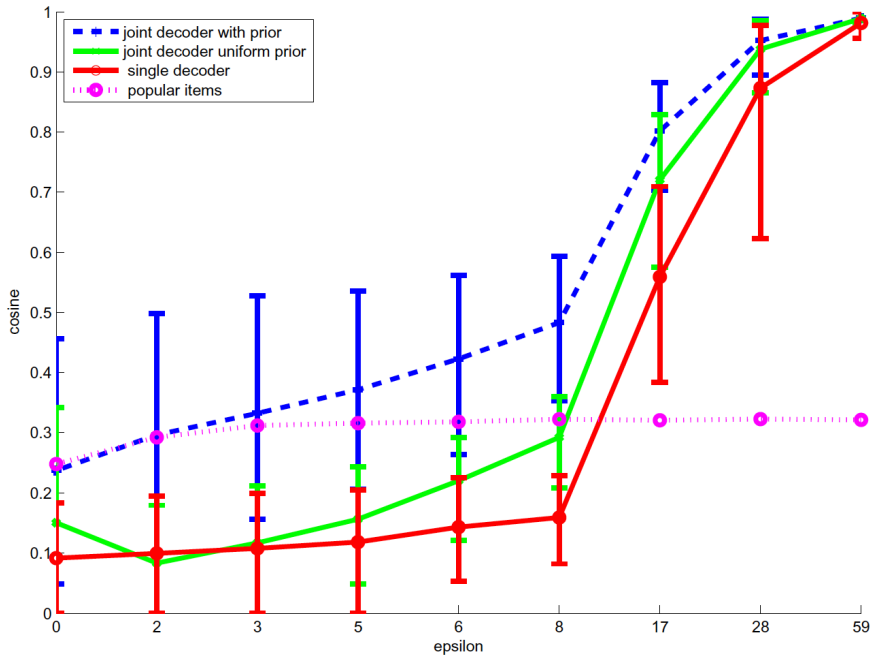
- Profile reconstruction
  - Cosine between original and reconstructed profiles

$$\cos(P, \hat{P}) = \frac{|P.\hat{P}|}{|P||\hat{P}|}$$

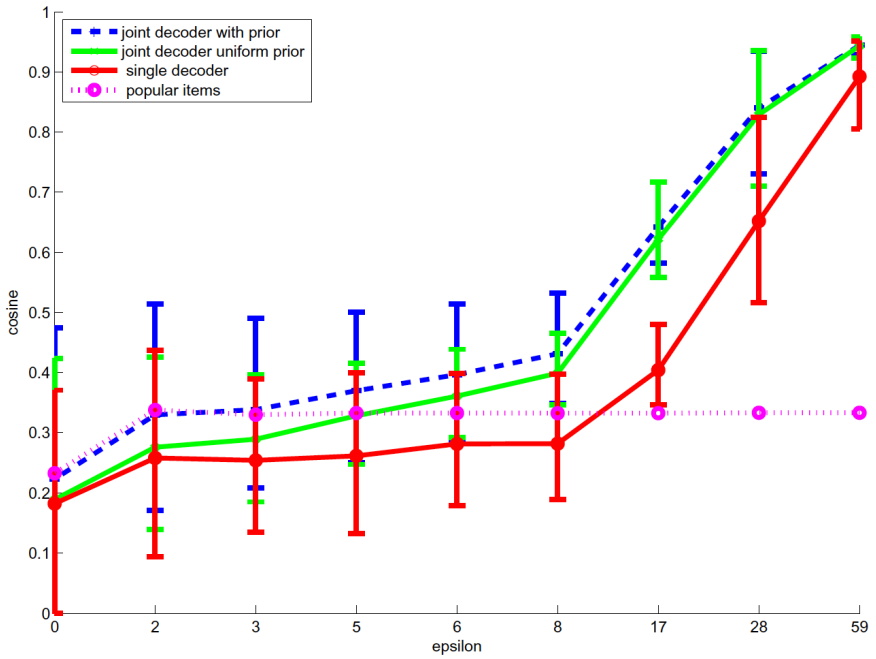
- Presence of individual item
  - Mean average precision of top-*R* ranked items

$$\text{mAP@}K = \frac{1}{Q} \sum_{q=1}^{Q} \left(\frac{1}{R} \sum_{r=1}^{R} \text{precision}_q(r) \right)$$

Profile reconstruction

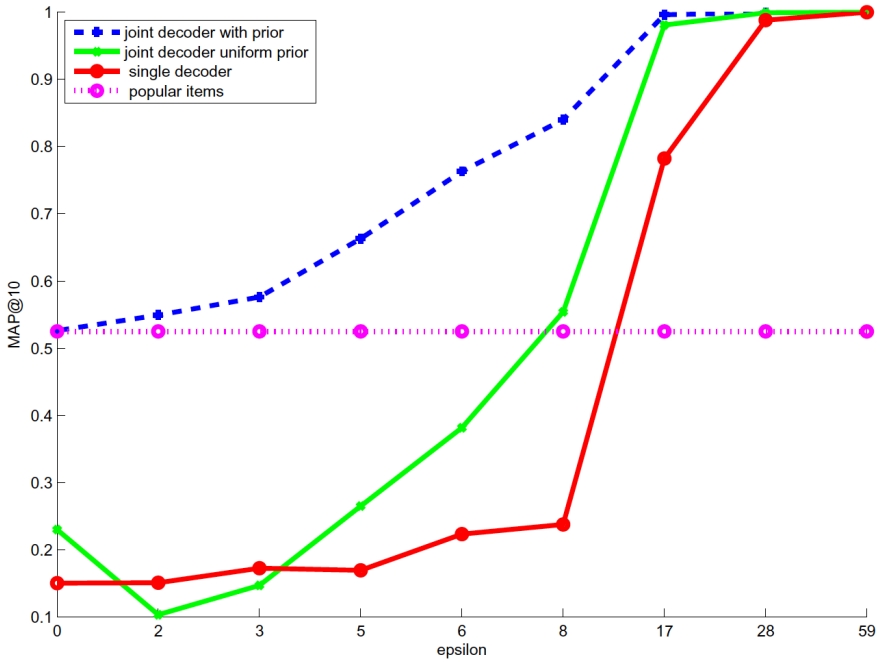


MovieLens

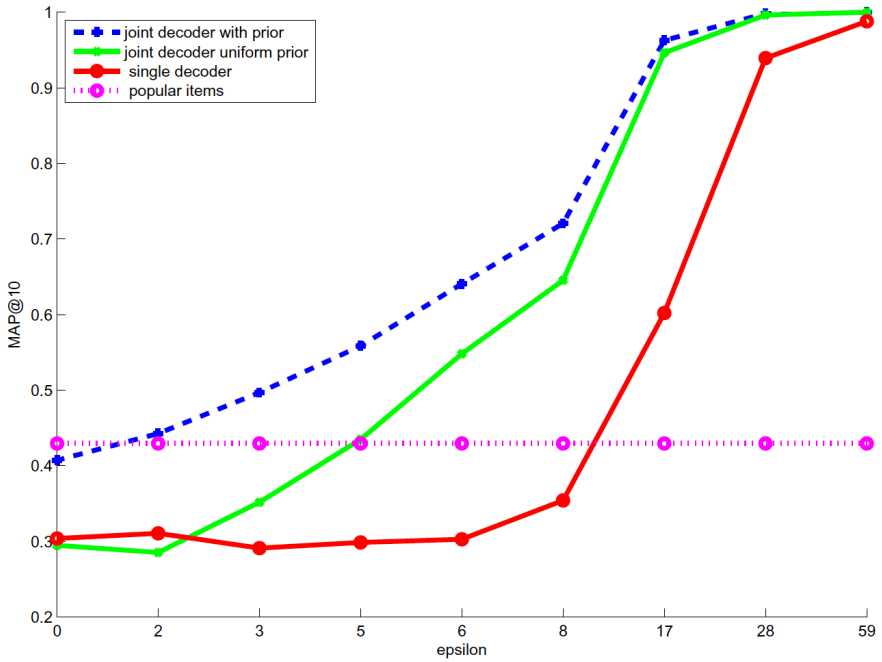


Digg

Presence of individual item



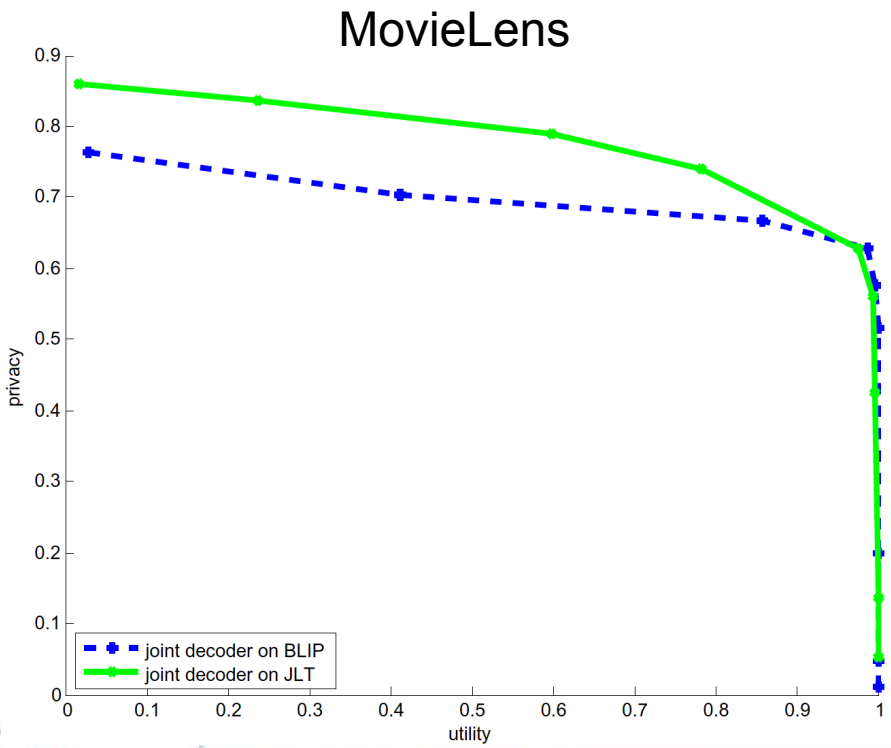
MovieLens

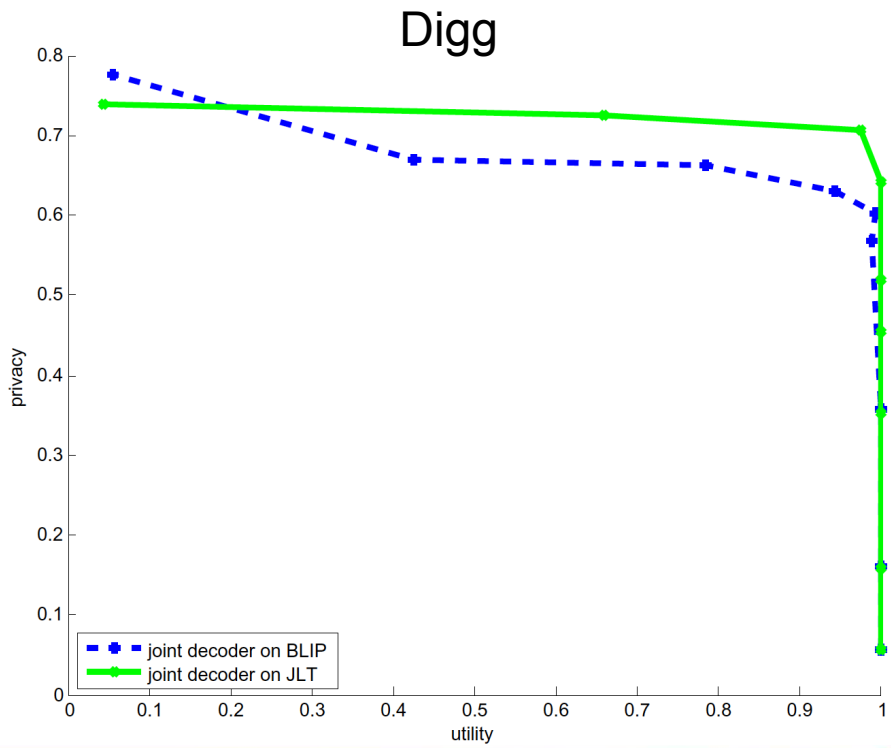


Digg

Utility vs privacy

- Privacy-Utility tradeoff
  - Privacy: $1 - \cos(P, \hat{P})$
  - Utility: $recall@10$
- *recall*@k: probability that the most similar profile is among the top-*k* ranked profiles





Conclusion

- Two attacks
  - Single and joint decoding
- Evaluated against two differentially private schemes:
  - BLIP and JLT

- Theoretical analysis and experimental results shows:

  Joint decoding is more powerful than single decoding

- Practical implementations of a joint decoder
  - We use a Monte Carlo Markov Chain (MCMC)
  - There are alternatives (Belief propagation, Iterative joint decoders)

Conclusion

- Our attacks help:
  - Understand the privacy guarantees of differentially-private mechanisms
  - Experimentally tune the parameter **ε**
  - Compare different non-interactive mechanisms
- Open Question
  - Towards a new definition of differential privacy?

$$\frac{I(\mathcal{F}(\mathbf{X}); \mathbf{X})}{|\mathbf{X}|} < \epsilon$$

where $\mathbf{X}$ is a collection of $|\mathbf{X}|$ items.

Questions?

THANK YOU!