Differentially-Private Clustering of Easy Instances

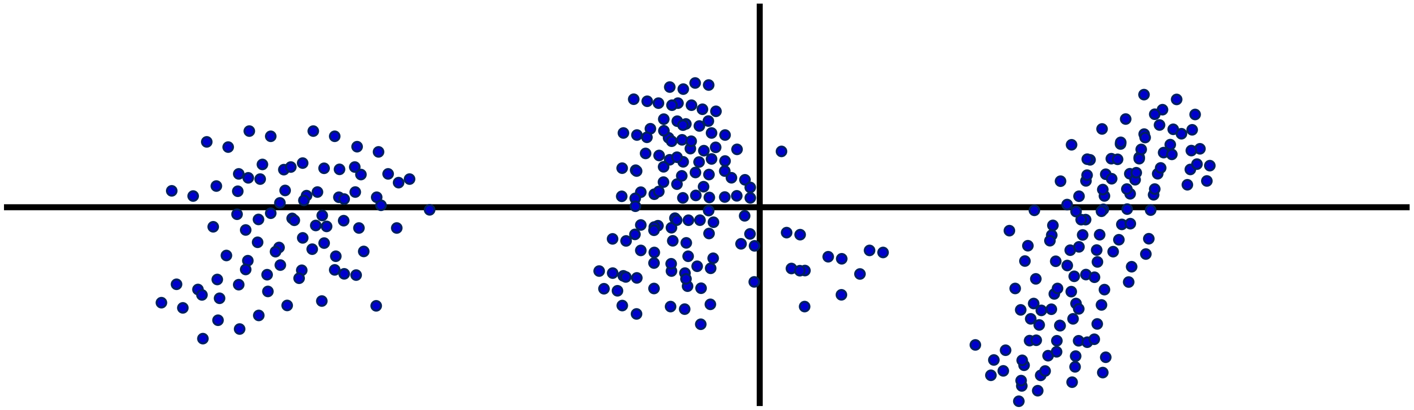
Eliad Tsfadia

Joint work with

Edith Cohen, Haim Kaplan, Yishay Mansour, Uri Stemmer

What is Clustering?

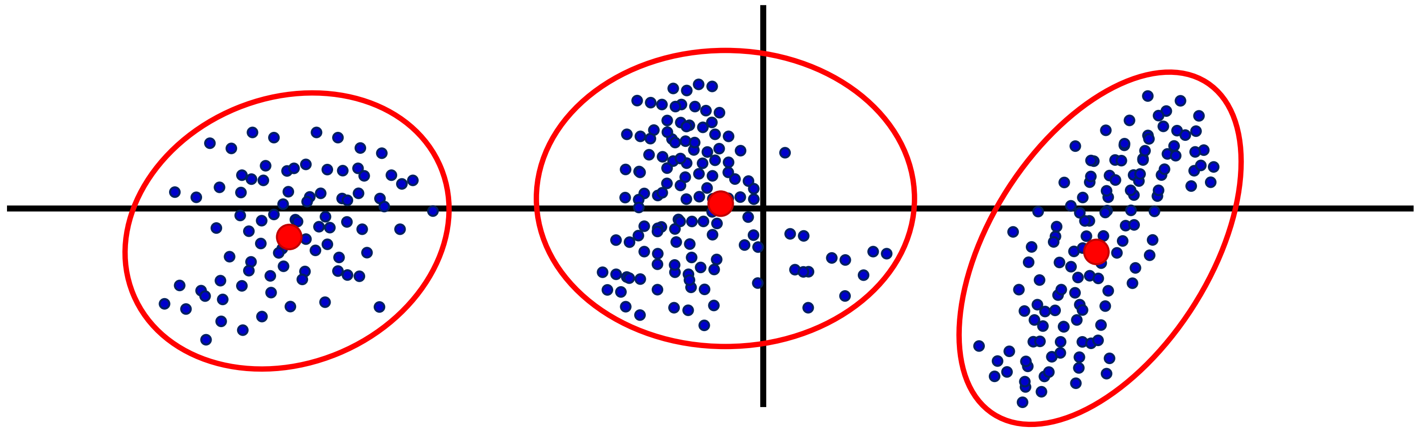
Input: Data points $\boldsymbol{S} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\} \in \left(\mathbb{R}^{\boldsymbol{d}}\right)^{\boldsymbol{n}}$



What is Clustering?

Input: Data points $\boldsymbol{S} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\} \in \left(\mathbb{R}^d\right)^n$

"Task": Identify groups of data points, and assign each point to one of the groups



But what is a "good" clustering?

k-means: Identify $\mathrm{C} = \{\boldsymbol{c_1}, \dots, \boldsymbol{c_k}\}$ that minimize $\mathbf{cost}(\boldsymbol{C}) = \sum_{i\in[n]} \min_{j\in[k]} \left\|\boldsymbol{x_i} - \boldsymbol{c_j}\right\|^2$

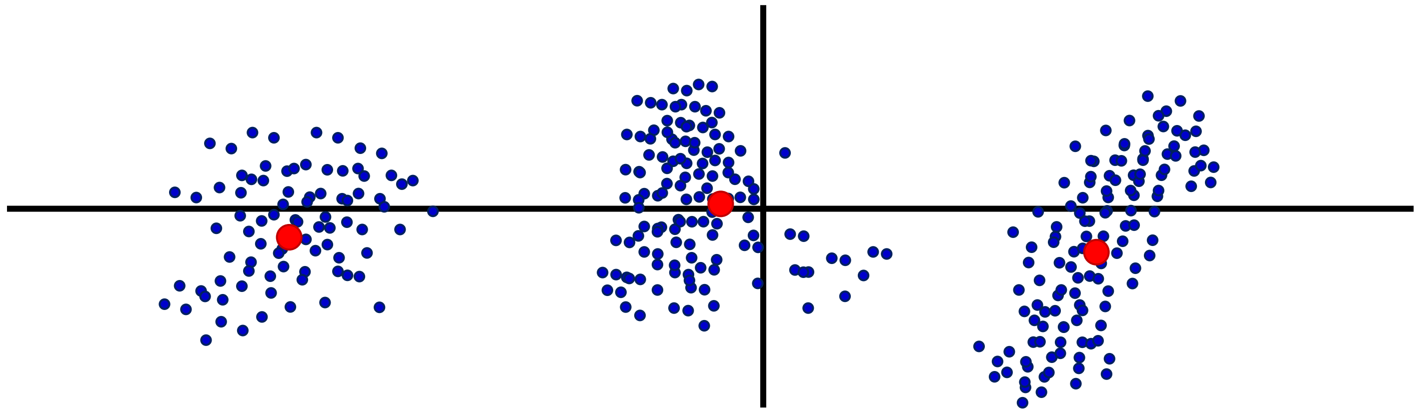
k-GMM: The points are samples from an (unknown) mixture of k Gaussians, and the goal is to estimate the parameters of the mixture (e.g., the means).

Other Problems: k-medians, k-centers, mixtures of other distributions….

When the instances are **well-separated**, clustering task is (essentially) the same

Differentially Private Clustering

<u>Input</u>: Data points $\boldsymbol{S} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\} \in \left(\mathbb{R}^{\boldsymbol{d}}\right)^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$

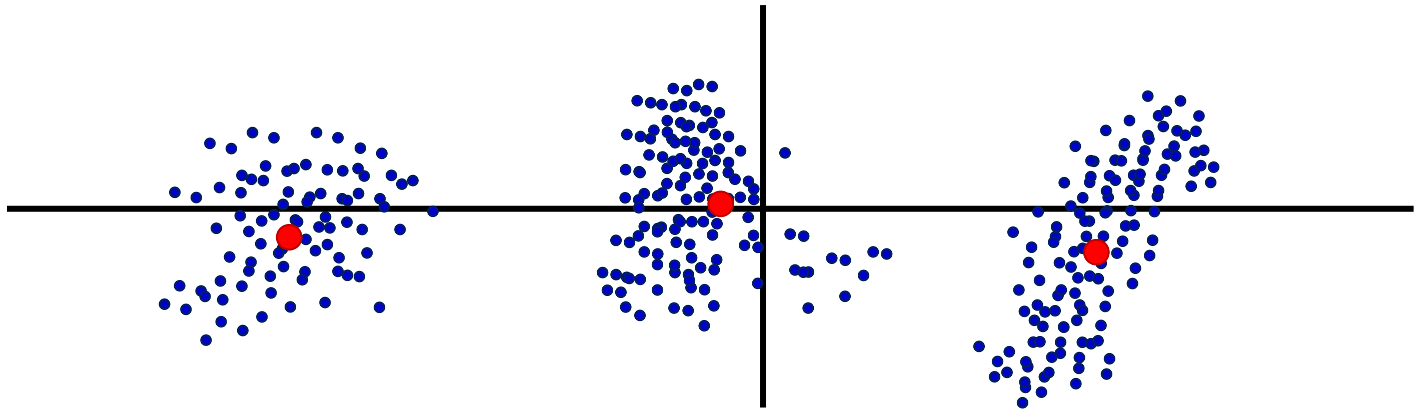


What is <u>Differentially Private</u> Clustering?

[Dwork, McSherry, Nissim, Smith 06] (informal, for now)

Differentially Private Clustering

<u>Input</u>: Data points $\boldsymbol{S} = \{\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\} \in (\mathbb{R}^{\boldsymbol{d}})^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$



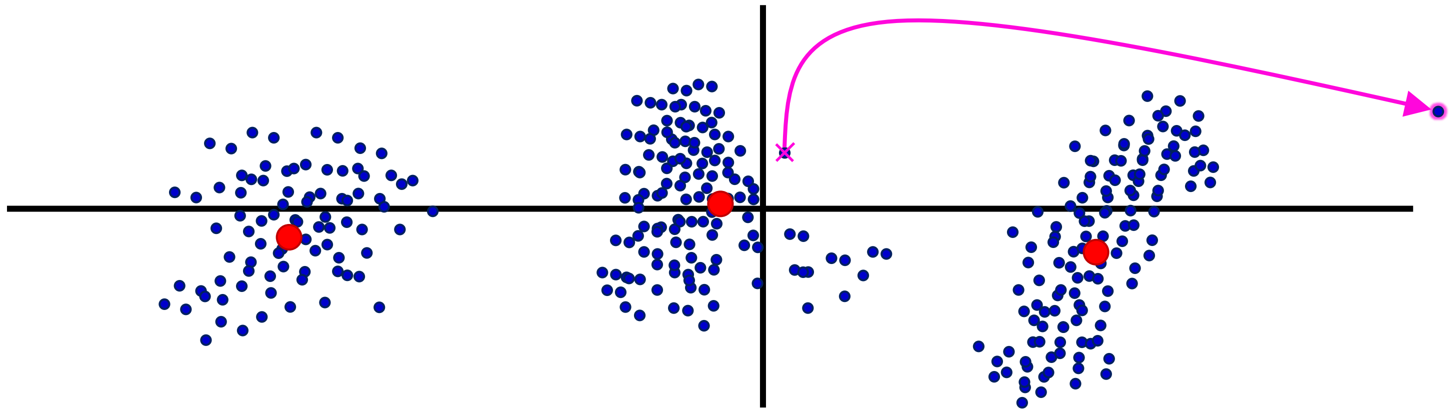
What is <u>Differentially Private</u> Clustering?

[Dwork, McSherry, Nissim, Smith 06] (informal, for now)

- ✓ Every data point $\boldsymbol{x_i}$ represents the (sensitive) information of one individual
- ✓ **Goal:** the output (the set of centers) does not reveal information that is specific to any single individual

Differentially Private Clustering

Input: Data points $\boldsymbol{S} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\} \in (\mathbb{R}^d)^n$ and parameter $\boldsymbol{k}$



What is Differentially Private Clustering?

[Dwork, McSherry, Nissim, Smith 06] (informal, for now)

- ✓ Every data point $\boldsymbol{x}_i$ represents the (sensitive) information of one individual
- ✓ **Goal:** the output (the set of centers) does not reveal information that is specific to any single individual
- ✓ **Requirement:** the output distribution remains roughly the same for every arbitrarily change of a single input point

Previous Results

The construction of differentially private clustering algorithms has attracted a lot of attention over the last decade, and many different algorithms have been suggested.

However, none of these algorithms have been implemented: They are not particularly simple and suffer from large hidden constants that translate to a significant loss in utility, compared to non-private implementations.

Can we construction **practical** differentially private algorithms for clustering problems?

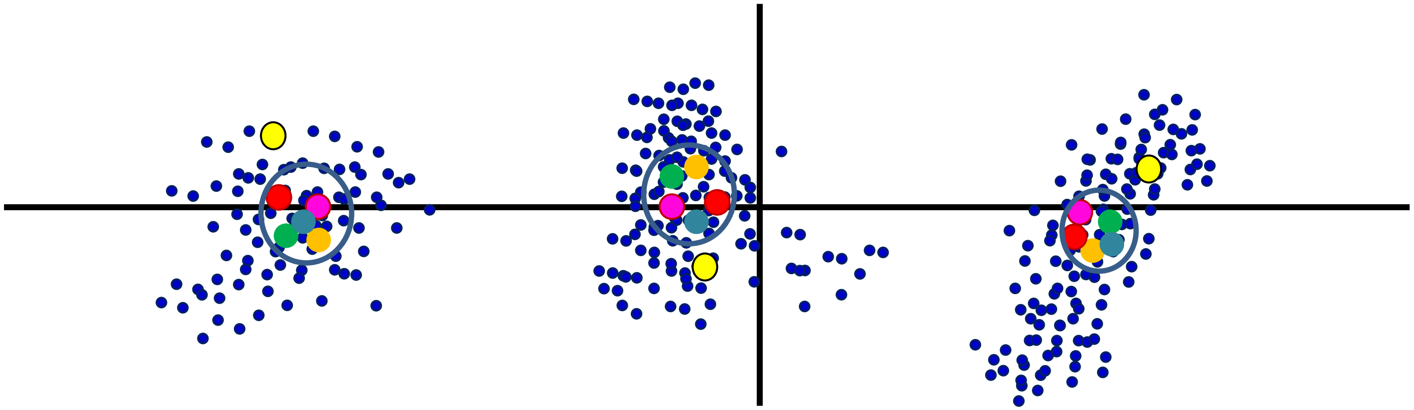
Our Approach

Input: Data points $\boldsymbol{S} = \{\boldsymbol{x_1}, \dots, \boldsymbol{x_n}\} \in \left(\mathbb{R}^d\right)^{\boldsymbol{n}}$ and parameter $\boldsymbol{k}$

Sample and Aggregate: (1) Randomly split $\boldsymbol{S}$ into m subsets

(2) Execute some non-private algorithm in each subset.

=> each execution gives a *k-tuple* of points over $\mathbb{R}^{\boldsymbol{d}}$



k-tuple Clustering

Input: k-tuples $\boldsymbol{T} = \{\boldsymbol{Y_1}, \dots, \boldsymbol{Y_m}\} \in \left(\left(\mathbb{R}^d\right)^{\boldsymbol{k}}\right)^{\boldsymbol{m}}$ with the (utility) promise that the tuples are partitioned by ``far balls''.

Goal: Privately identify a new k-tuple that is ``close'' to them (e.g., the yellow tuple).

Our Results

1. Two **simple** algorithms for solving the k-tuple clustering problem.
2. Solving private k-means and k-GMM (under common separation assumptions) by a reduction.
   I. Much simpler (and implementable) algorithms than existing ones.
   II. Private k-GMM: reduce sample complexity, weaker separation assumption and modularity, compared to [KSSU19].
3. We present empirical results over synthetic data.
   - First ``practical'' differentially private algorithm for clustering very separated instances.
   - First approach to bridge the gap between the theory and practice of differentially private clustering methods.