Clustering

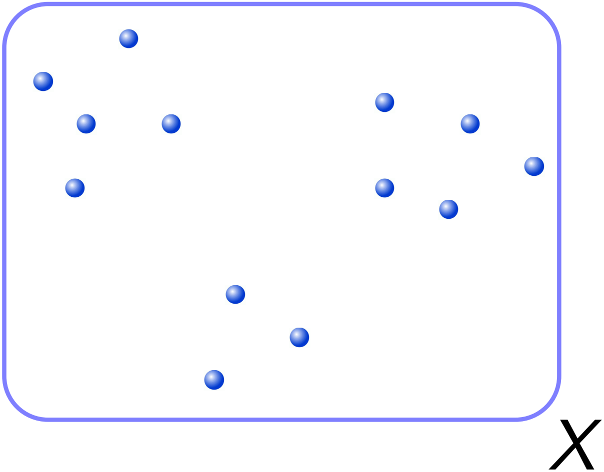
How Bad Is The k-Means++ Method?

Tobias Brunsch   Heiko Röglin

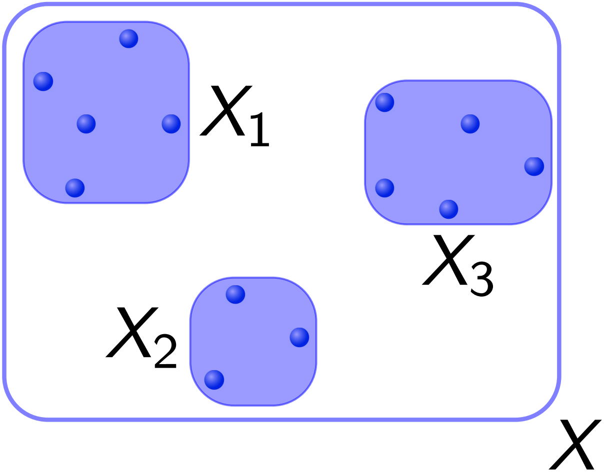
Department of Quantitative Economics
Maastricht University
The Netherlands

July 15, 2010

What Is Clustering?



What Is Clustering?



The k-Means Problem (k-Means)

Measuring the quality of a clustering:

The k-Means Problem (k-Means)

Measuring the quality of a clustering:

- Assign one *center* c_i to each cluster X_i

The k-Means Problem (k-Means)

Measuring the quality of a clustering:

- Assign one *center* c_i to each cluster X_i
- *Cluster potential*: $\Phi(X_i) := \sum_{x \in X_i} \|x - c_i\|^2$

The k-Means Problem (k-Means)

Measuring the quality of a clustering:

- Assign one *center* c_i to each cluster X_i
- *Cluster potential*: $\Phi(X_i) := \sum_{x \in X_i} \|x - c_i\|^2$
- *Clustering potential*: $\Phi(X) := \sum_i \Phi(X_i)$

The k-Means Problem (k-Means)

Measuring the quality of a clustering:

- Assign one *center* c_i to each cluster X_i
- *Cluster potential*: $\Phi(X_i) := \sum_{x \in X_i} \|x - c_i\|^2$
- *Clustering potential*: $\Phi(X) := \sum_i \Phi(X_i)$

Objective: Find clustering $(X_1, \ldots, X_k)$ and centers $c_1, \ldots, c_k$ with minimal potential $\Phi(X)$

The Challenge

k-means is $\mathcal{NP}$-hard,

- even in the plane and
- even for $k = 2$

The Challenge

k-means is $\mathcal{NP}$-hard,

- even in the plane and
- even for $k = 2$

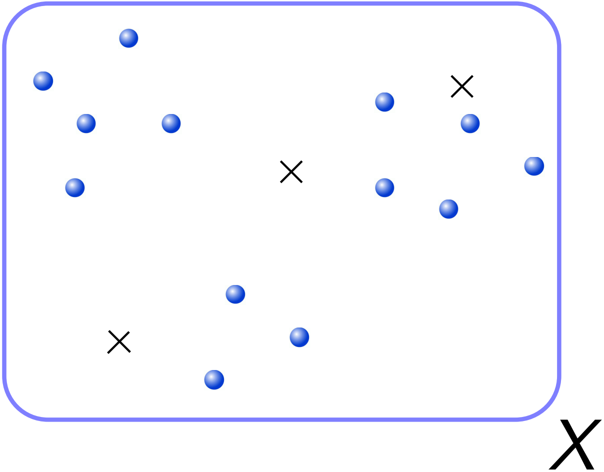
$\Longrightarrow$ Approximation algorithms, heuristics

Lloyd's Algorithm (k-Means Method, k-Means)

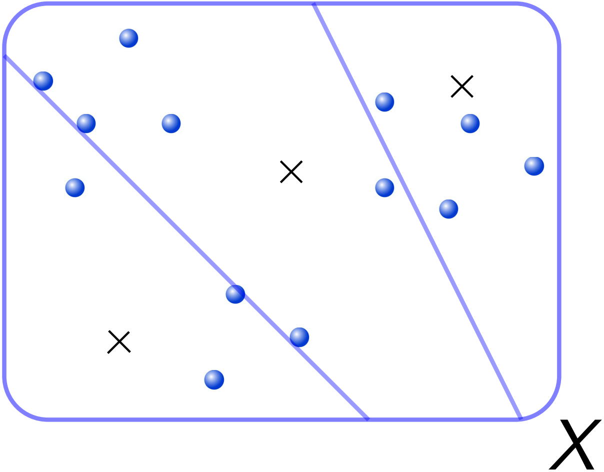
Observations:

- The optimal centers c_i for given clusters X_i are their centers of mass
- The optimal clusters X_i for given centers c_i are the points nearest to c_i

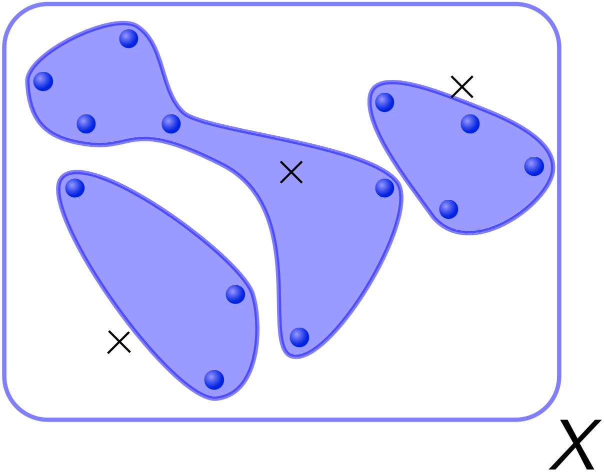
Lloyd's Algorithm



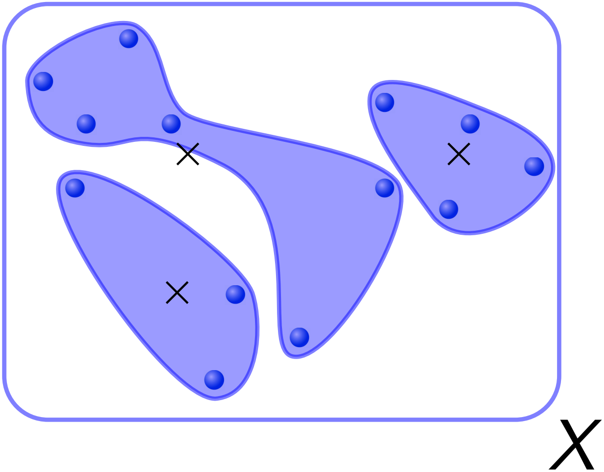
Lloyd's Algorithm



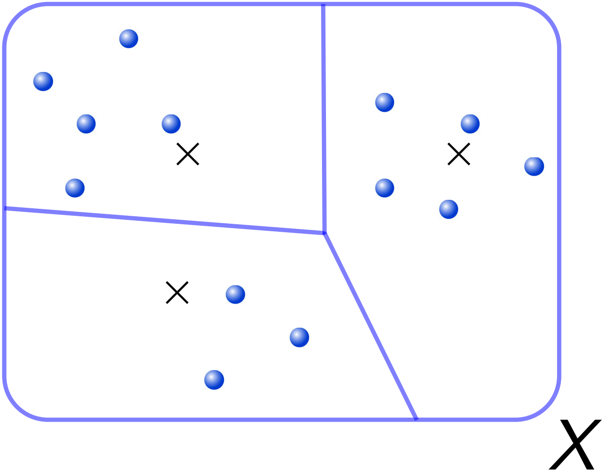
Lloyd's Algorithm



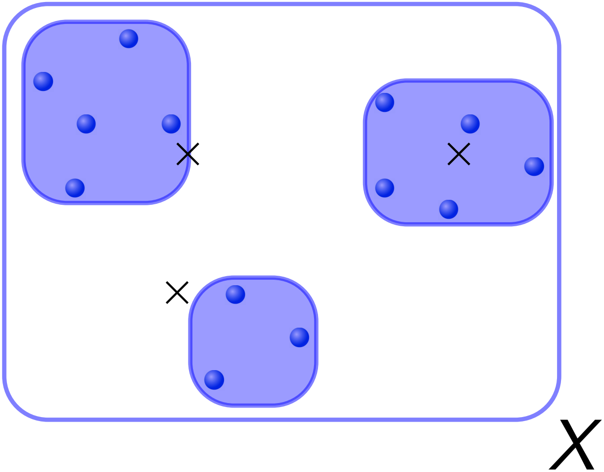
Lloyd's Algorithm



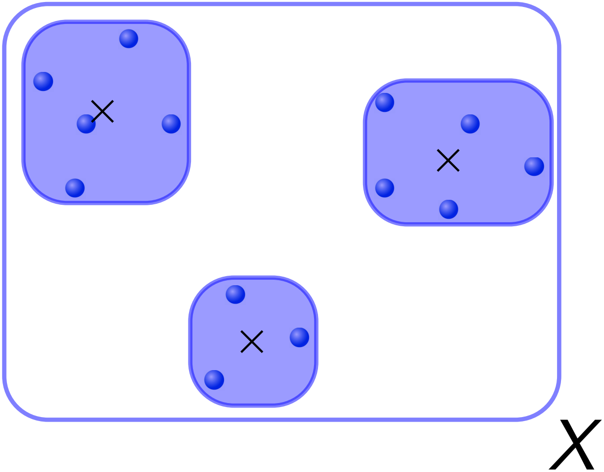
Lloyd's Algorithm



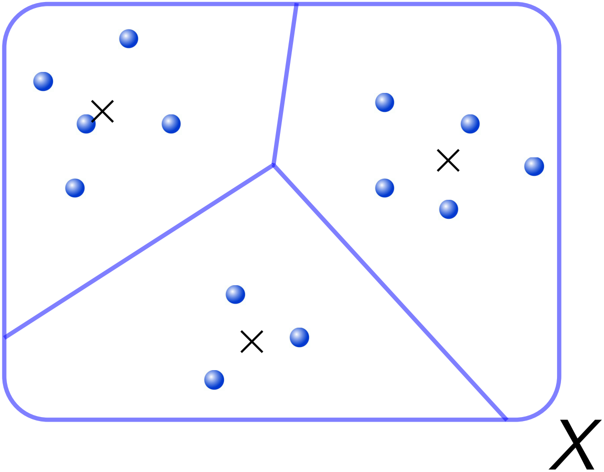
Lloyd's Algorithm



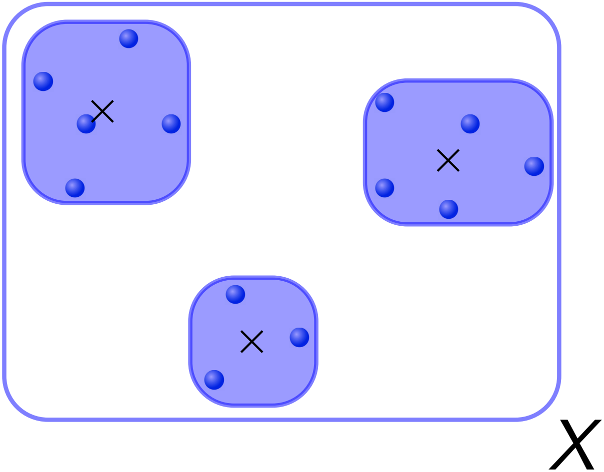
Lloyd's Algorithm



Lloyd's Algorithm



Lloyd's Algorithm



Advantages And Disadvantages

	Practitioners	**Theoreticians**
Advantages		
Simple to implement		

Advantages And Disadvantages

	Practitioners	**Theoreticians**
Advantages		
Simple to implement		
Fast in practice		

Advantages And Disadvantages

	Practitioners	**Theoreticians**
Advantages		
Simple to implement		
Fast in practice		
Disadvantages		
Exponential worst-case time		

Advantages And Disadvantages

	Practitioners	**Theoreticians**
Advantages		
Simple to implement		
Fast in practice		
Disadvantages		
Exponential worst-case time		
Requires good initialization		

Tackling The Disadvantages

- Polynomial time in the framework of *smoothed analysis*
- Approximation guarantee with *k-means++* seeding technique

k-Means++ Seeding

Centers chosen **from the input set** step-by-step

k-Means++ Seeding

Centers chosen **from the input set** step-by-step

1. Choose the first center uniformly at random

k-Means++ Seeding

Centers chosen **from the input set** step-by-step

1. Choose the first center uniformly at random
2. Choose point $x \in X$ with probability $\frac{D^2(x)}{\Phi(X)}$ as next center

$$\left(D^2(x) = \min_{c_i} \|x - c_i\|^2 \right)$$

Asymptotic Bounds

Theorem (Arthur and Vassilvitskii, 2007)

The expected approximation ratio of k-means++ is $O(\log k)$.

Asymptotic Bounds

Theorem (Arthur and Vassilvitskii, 2007)

The expected approximation ratio of k-means++ is $O(\log k)$.

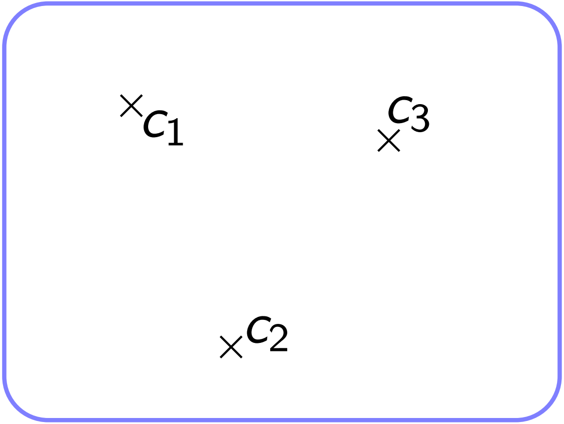
Observation

There is a family of instances on which the expected approximation ratio of k-means++ is $\Omega(\log k)$.

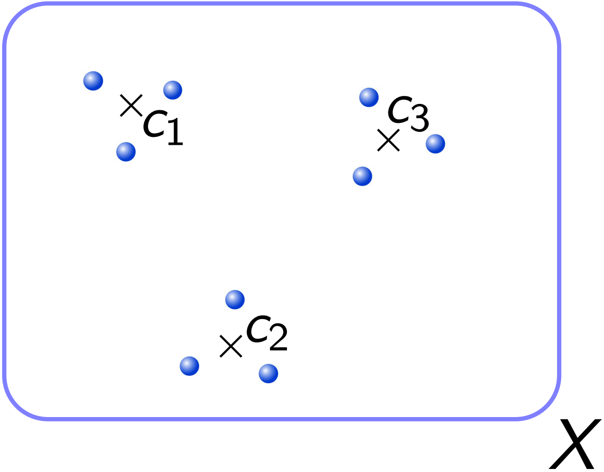
Open Question

Does k-means++ yield an $O(1)$-approximation with constant probability?

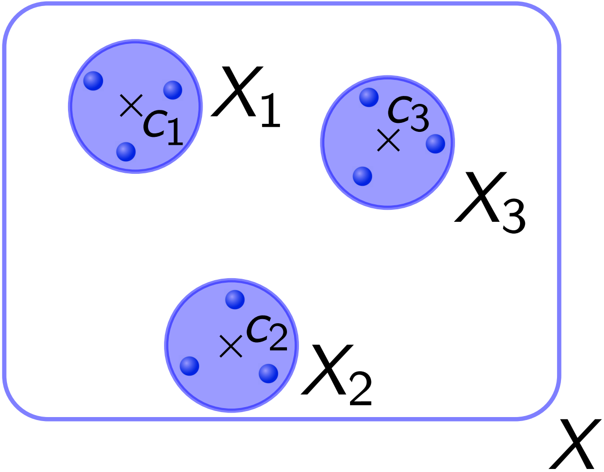
The Instance



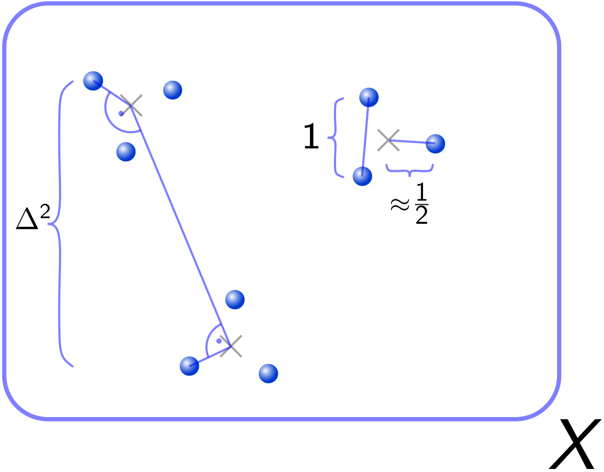
The Instance



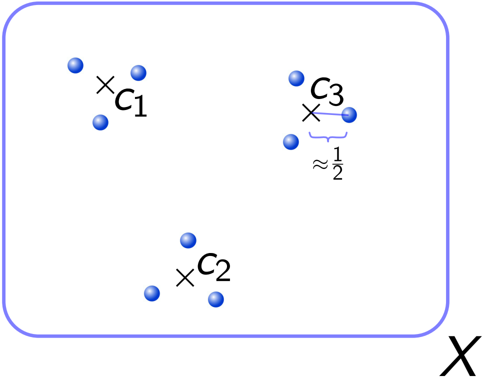
The Instance



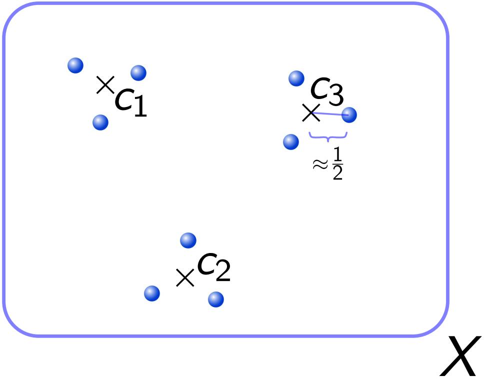
The Instance



Optimal Clustering C^*

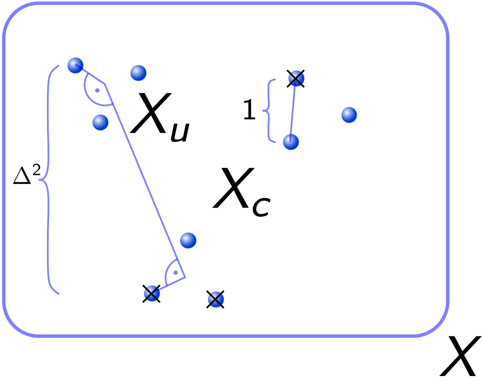


Optimal Clustering C^*

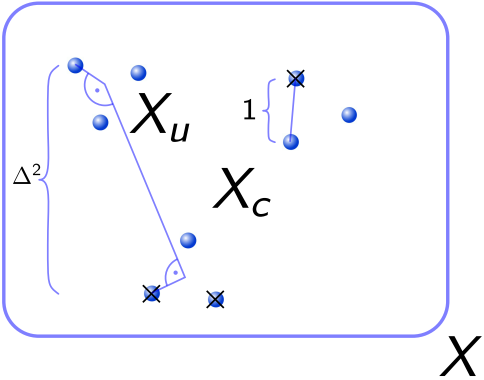


$$\Phi^*(X) \approx k \cdot k \cdot \frac{1}{2} = \frac{k^2}{2}$$

Discrete Clustering With s Covered Sets X_i

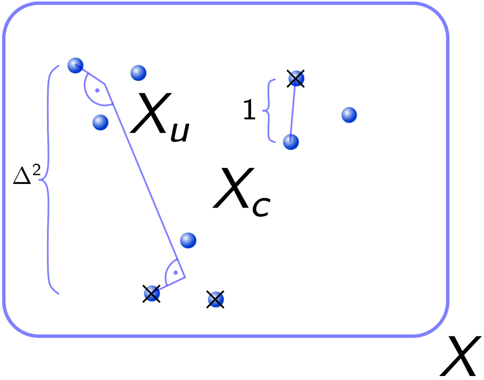


Discrete Clustering With s Covered Sets X_i



$$\Phi(X) = \Phi(X_c) + \Phi(X_u) \approx s \cdot k \cdot 1 + (k - s) \cdot k \cdot \Delta^2$$

Discrete Clustering With s Covered Sets X_i



Covering probability: $\frac{\Phi(X_u)}{\Phi(X)} \approx \frac{1}{1+\frac{s}{(k-s)\cdot\Delta^2}} =: p_s$

How Many Sets To Cover?

In the end:

$$r \geq \frac{\Phi(X)}{\Phi^*(X)} \geq \frac{\Phi(X_u)}{\Phi^*(X)} \approx 2\Delta^2 \cdot \left(1 - \frac{s}{k}\right) \quad (r \text{ - approximation factor})$$

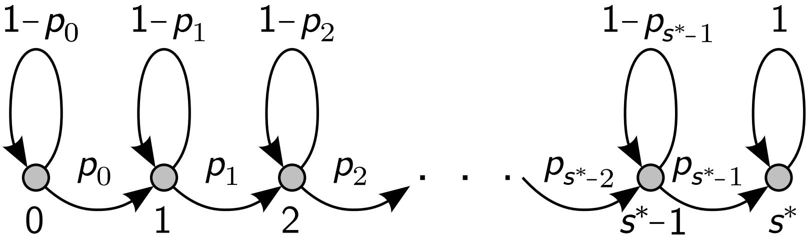
How Many Sets To Cover?

In the end:

$$r \geq \frac{\Phi(X)}{\Phi^*(X)} \geq \frac{\Phi(X_u)}{\Phi^*(X)} \approx 2\Delta^2 \cdot \left(1 - \frac{s}{k}\right) \quad (r \text{ - approximation factor})$$

$$\Longrightarrow s \gtrsim k \cdot \left(1 - \frac{r}{2\Delta^2}\right) =: s^*$$

Markov Chain



Expected Number Of Steps X

$$\mathbf{E}[X] = \sum_{s=0}^{s^*-1} \frac{1}{p_s} \gtrsim k + \frac{k}{\Delta^2} \cdot \left(\ln \frac{\Delta^2}{r} - \frac{r}{2} \right)$$

Expected Number Of Steps X

$$\mathbf{E}[X] = \sum_{s=0}^{s^*-1} \frac{1}{p_s} \gtrsim k + \frac{k}{\Delta^2} \cdot \left(\ln \frac{\Delta^2}{r} - \frac{r}{2} \right)$$

$\Longrightarrow$ Choose $\Delta^2 = r \cdot \exp\left(\frac{1+\epsilon}{2} r\right)$

Expected Number Of Steps X

$$\mathbf{E}[X] = \sum_{s=0}^{s^*-1} \frac{1}{p_s} \gtrsim k + \frac{k}{\Delta^2} \cdot \left(\ln \frac{\Delta^2}{r} - \frac{r}{2} \right)$$

$$\Longrightarrow \text{Choose } \Delta^2 = r \cdot \exp\left(\frac{1+\epsilon}{2} r\right)$$

$$\Longrightarrow \mathbf{E}[X] \gtrsim k + \frac{\epsilon}{2} \cdot \frac{k}{\exp\left(\frac{1+\epsilon}{2} r\right)}$$

Expected Number Of Steps X

$$\mathbf{E}[X] = \sum_{s=0}^{s^*-1} \frac{1}{p_s} \gtrsim k + \frac{k}{\Delta^2} \cdot \left(\ln \frac{\Delta^2}{r} - \frac{r}{2} \right)$$

$$\Longrightarrow \text{Choose } \Delta^2 = r \cdot \exp\left(\tfrac{1+\epsilon}{2} r\right)$$

$$\Longrightarrow \mathbf{E}[X] \gtrsim k + \frac{\epsilon}{2} \cdot \frac{k}{\exp\left(\frac{1+\epsilon}{2} r\right)}$$

If $r \in o(\log k)$, then $\mathbf{Pr}[X \leq k]$ is exponentially small in k

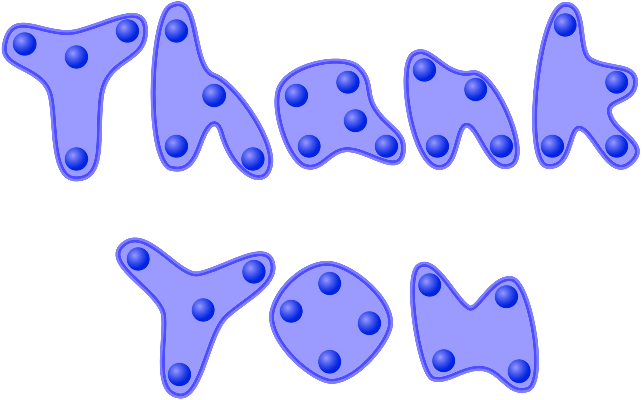
(Hoeffding Inequality + workaround)

Open Questions

1. Do k-means++ and k-means together yield an $O(1)$-approximation with constant probability?

Open Questions

1. Do k-means++ and k-means together yield an $O(1)$-approximation with constant probability?
2. Can we slightly modify k-means++ to guarantee better bounds?


Thank
You