k-means and k-medians under dimension reduction

Yury Makarychev, TTIC

Konstantin Makarychev, Northwestern
Ilya Razenshteyn, Microsoft Research

Simons Institute, November 2, 2018

Euclidean k-means and k-medians

Given a set of points X in $\mathbb{R}^m$

Partition X into k clusters $C_1, \dots, C_k$ and find a "center" c_i for each C_i so as to minimize the cost

$$\sum_{i=1}^{k} \sum_{u \in C_i} d(u, c_i) \qquad (k\text{-median})$$

$$\sum_{i=1}^{k} \sum_{u \in C_i} d(u, c_i)^2 \qquad (k\text{-means})$$

Dimension Reduction

Dimension reduction $\varphi: \mathbb{R}^m \to \mathbb{R}^d$ is a random map that preserves distances within a factor of $(1+\varepsilon)$ with probability at least $1-\delta$:

$$\frac{1}{1+\varepsilon}\|u-v\| \le \|\varphi(u)-\varphi(v)\| \le (1+\varepsilon)\|u-v\|$$

[Johnson-Lindenstrauss '84] There exists a random linear dimension reduction with $d = O\left(\frac{\log 1/\delta}{\varepsilon^2}\right)$.

[Larsen, Nelson '17] The dependence of d on ε and δ is optimal.

Dimension Reduction

JL preserves all distances between points in X whp when $d = \Omega(\log |X| / \varepsilon^2)$.

Numerous applications in computer science.

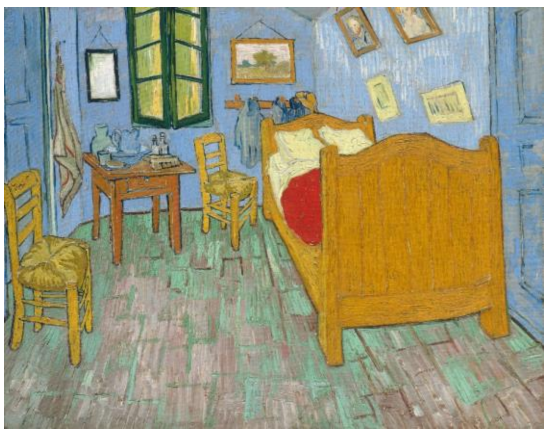
Dimension Reduction Constructions:

- [JL '84] Project on a random d-dimensional subspace
- [Indyk, Motwani '98] Apply a random Gaussian matrix
- [Achlioptas '03] Apply a random matrix with ± 1 entries
- [Ailon, Chazelle '06] Fast JL-transform

k-means under dimension reduction

[Boutsidis, Zouzias, Drineas '10]

Apply a dimension reduction φ to our dataset X



dimension reduction

Cluster $\varphi(X)$ in dimension d.

k-means under dimension reduction

want

Optimal clusterings of X and $\varphi(X)$ have approximately the same cost.

even better

The cost of every clustering is approximately preserved.

For what dimension d can we get this?

k-means under dimension reduction

	d	distortion
Folklore	$\sim \log n / \varepsilon^2$	$1 + \varepsilon$
Boutsidis, Zouzias, Drineas '10	$\sim k / \varepsilon^2$	$2 + \varepsilon$
Cohen, Elder, Musco, Musco, Persu '15	$\sim k / \varepsilon^2$	$1 + \varepsilon$
	$\sim \log k / \varepsilon^2$	$9 + \varepsilon$
MMR '18	$\sim \log(k/\varepsilon) / \varepsilon^2$	$1 + \varepsilon$
Lower bound	$\sim \log k / \varepsilon^2$	$1 + \varepsilon$

k-medians under dimension reduction

	d	distortion
Prior work	—	—
Kirszsbraun Thm $\Rightarrow$	$\sim \log n / \varepsilon^2$	$1+\varepsilon$
MMR '18	$\sim \log(k/\varepsilon) / \varepsilon^2$	$1+\varepsilon$
Lower bound	$\sim \log k / \varepsilon^2$	$1+\varepsilon$

Plan

k-means

- Challenges
- Warm up: $d \sim \log n / \varepsilon^2$
- Special case: “distortions” are everywhere sparse
- Remove outliers: the general case → the special case
- Outliers

k-medians

- Overview of our approach

Out result for k-means

Let $X \subset \mathbb{R}^m$

$\varphi: \mathbb{R}^m \to \mathbb{R}^d$ be a random dimension reduction.

$$d \geq c \log \frac{k}{\varepsilon\delta} / \varepsilon^2$$

With probability at least $1 - \delta$:

$$(1-\varepsilon)\operatorname{cost}\mathcal{C} \leq \operatorname{cost}\varphi(\mathcal{C}) \leq (1+\varepsilon)\operatorname{cost}\mathcal{C}$$

for every clustering $\mathcal{C} = (C_1, \dots, C_k)$ of X

Challenges

Let $\mathcal{C}^*$ be the optimal k-means clustering.

Easy:

$$\operatorname{cost} \mathcal{C}^* \approx \operatorname{cost} \varphi(\mathcal{C}^*)$$

with probability $1 - \delta$

Hard: Prove that there is no other clustering $\mathcal{C}'$ s.t.

$$\operatorname{cost} \varphi(\mathcal{C}') < (1 - \varepsilon) \operatorname{cost} \mathcal{C}^*$$

since there are exponentially many clusterings $\mathcal{C}'$

(can't use the union bound)

Warm-up

Consider a clustering $\mathcal{C} = (C_1, \dots, C_k)$.

Write the cost in terms of pair-wise distances:

$$\operatorname{cost} \mathcal{C} = \sum_{i=1}^{k} \frac{1}{2|C_i|} \sum_{u,v \in C_i} \|u - v\|^2$$

all distances $\|u - v\|$ are preserved within $1 + \varepsilon$

$\Downarrow$

$\operatorname{cost} \mathcal{C}$ is preserved within $1 + \varepsilon$

Sufficient to have $d \sim \log n / \varepsilon^2$

Problem & Notation

Assume that $\mathcal{C} = (C_1, \dots, C_k)$ is a random clustering that depends on φ.

Want to prove: $\operatorname{cost} \mathcal{C} \approx \operatorname{cost} \varphi(\mathcal{C})$ whp.

The distance between u and v is $(1+\varepsilon)$-preserved or distorted depending on whether

$$\|\varphi(u) - \varphi(v)\| \approx_{1+\varepsilon} \|u - v\|$$

Think $\delta = \operatorname{poly}(1/k, \varepsilon)$ is sufficiently small.

Distortion graph

Connect u and v with an edge if the distance between them is distorted.

+ Every edge is present with probability at most δ.
– Edges are not independent.
– $\mathcal{C}$ depends on the set of edges.
– May have high-degree vertices.
– All distances in a cluster may be distorted.

Cost of a cluster

The cost of C_i is

$$\frac{1}{2|C_i|}\sum_{u,v\in C_i}\|u-v\|^2$$

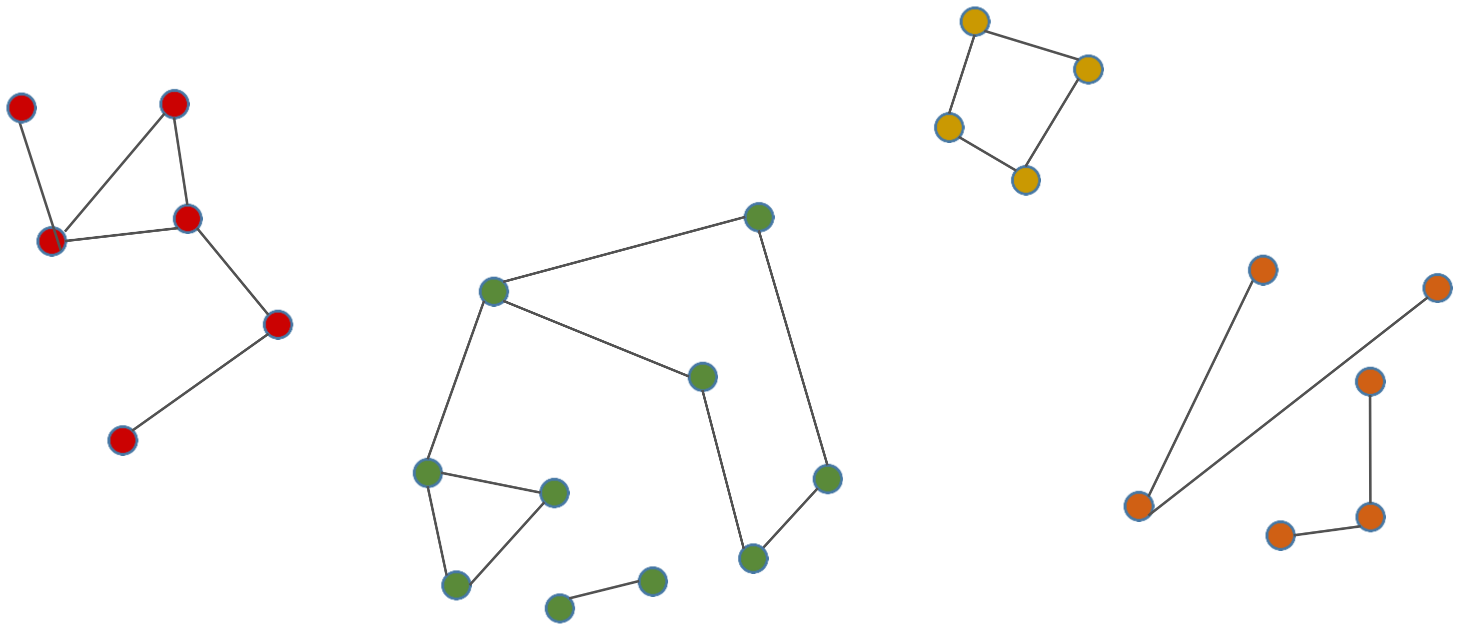
+ Terms for non-edges (u,v) are $(1+\varepsilon)$ preserved.

$$\|u-v\| \approx \|\varphi(u)-\varphi(v)\|$$

– Need to prove that

$$\sum_{\substack{u,v\in C_i\\(u,v)\in E}}\|u-v\|^2 = \sum_{\substack{u,v\in C_i\\(u,v)\in E}}\|\varphi(u)-\varphi(v)\|^2 \pm \varepsilon' \mathrm{cost}\,\mathcal{C}$$

Everywhere-sparse edges



Assume every $u \in C_i$ is connected to at most a θ fraction of all v in C_i (where $\theta \ll \varepsilon$).

Everywhere-sparse edges

+ Terms for non-edges (u, v) are $(1 + \varepsilon)$ preserved.

+ The contribution of terms for edges is small:
for an edge (u, v) and any $w \in C_i$

$$\|u - v\| \leq \|u - w\| + \|w - v\|$$

$$\|u - v\|^2 \leq 2\left(\|u - w\|^2 + \|w - v\|^2\right)$$

Everywhere-sparse edges

$$\|u-v\|^2 \le 2\left(\|u-w\|^2 + \|w-v\|^2\right)$$

- Replace the term for every edge with two terms $\|u-w\|^2, \|w-v\|^2$ for random $w \in C_i$.
- Each term is used at most 2θ times, in expectation.

$$\sum_{\substack{(u,v)\in E \\ u,v\in C_i}} \|u-v\|^2 \le 4\theta \sum_{u,v\in C_i} \|u-v\|^2$$

Everywhere-sparse edges

$$\sum_{u,v \in C_i} \|u - v\|^2 \approx \sum_{(u,v) \notin E} \|u - v\|^2$$

$$\approx$$

$$\sum_{(u,v) \notin E} \|\varphi(u) - \varphi(v)\|^2 \approx \sum_{u,v \in C_i} \|\varphi(u) - \varphi(v)\|^2$$

Everywhere-sparse edges

$$\sum_{u,v \in C_i} \|u - v\|^2 \approx \sum_{(u,v) \notin E} \|u - v\|^2$$

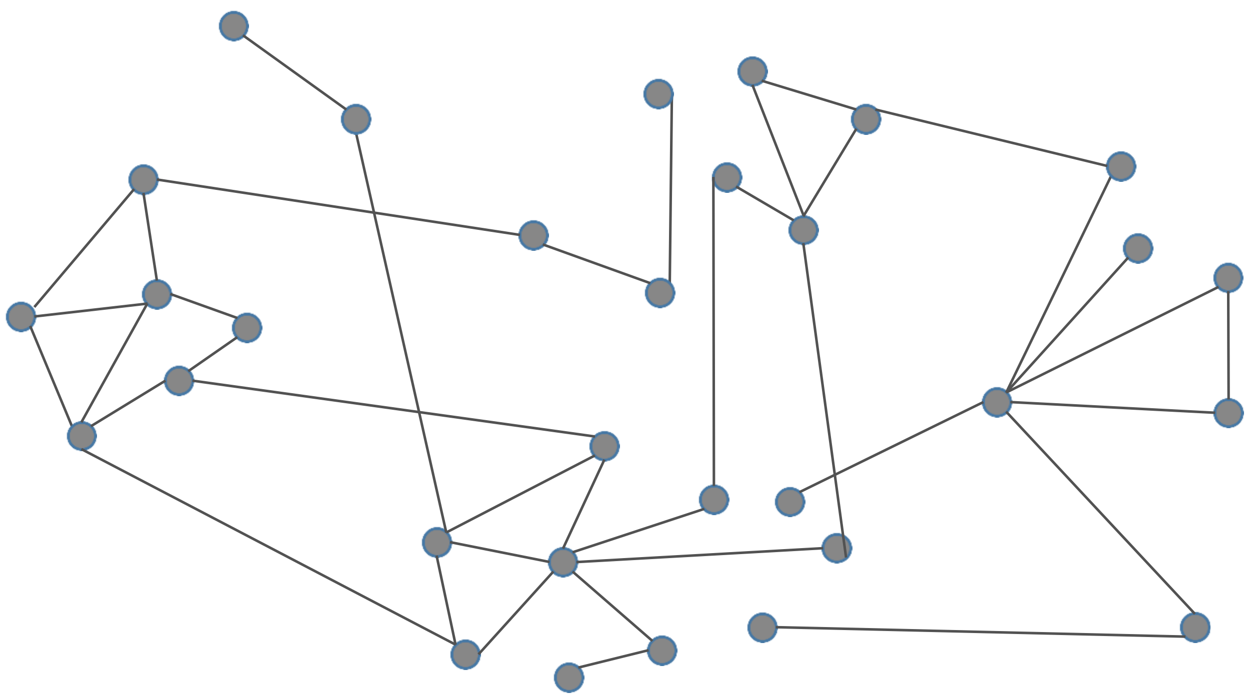
$$\approx$$

$$\sum_{(u,v) \notin E} \|\varphi(u) - \varphi(v)\|^2 \approx \sum_{u,v \in C_i} \|\varphi(u) - \varphi(v)\|^2$$

Edges are not necessarily everywhere sparse!

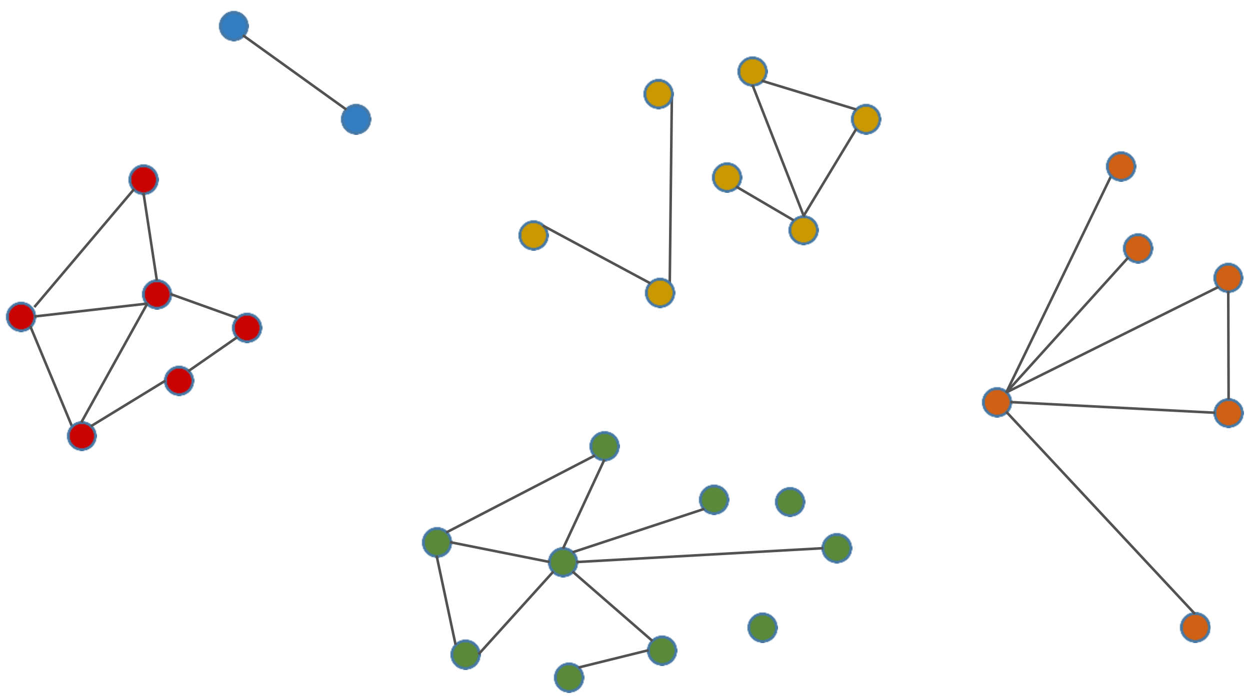
Outliers

Want: remove "outliers" so that in the remaining set X' edges are everywhere sparse in every cluster.



$(1-\theta)$ non-distorted core

Want: remove "outliers" so that in the remaining set X' edges are everywhere sparse in every cluster.



$(1-\theta)$ non-distorted core

Want: remove "outliers" so that in the remaining set X' edges are everywhere sparse in every cluster.

Find a subset $X' \subset X$ (which depends on $\mathcal{C}$) s.t.

- Edges are sparse in the obtained clusters:

Every $u \in C_i \cap X'$ is connected to at most a θ fraction of all v in $C_i \cap X'$.

- Outliers are rare:

For every u,

$$\Pr(u \notin X') \leq \theta$$

All clusters are large

Assume all clusters are of size $\sim n/k$. Let $\theta = \delta^{1/4}$.

outliers = all vertices of degree at least $\sim \theta n/k$

Every vertex has degree at most δn in expectation.

By Markov,

$$\Pr(u \text{ is an outlier}) \leq \frac{\delta k}{\theta} \leq \theta$$

Remove $\theta n \ll n/k$ vertices in total, so all clusters still have size $\sim n/k$.

Crucially use that all clusters are large!

Main Combinatorial Lemma

Idea: assign "weights" to vertices so that all clusters have a large weight.

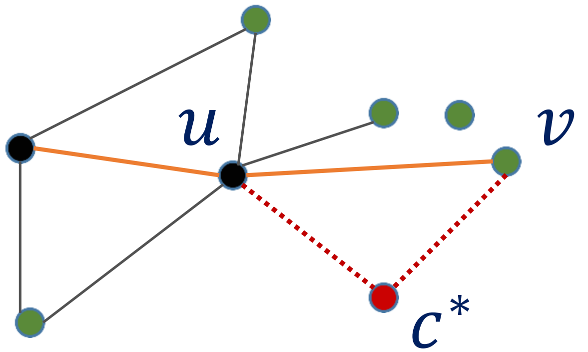
- There is a measure μ on X and random set R s.t. $\mu(x) \geq \frac{1}{|C_i \setminus R|}$ for $x \in C_i \setminus R$ (always)
- $\mu(X) \leq 4k^3/\theta^2$
- $\Pr(x \in R) \leq \theta$

All clusters $C_i \setminus R$ are "large" w.r.t. measure μ.

Can apply a variant of the previous argument.

Edges Incident on Outliers

Need to take care of edges incident on outliers.

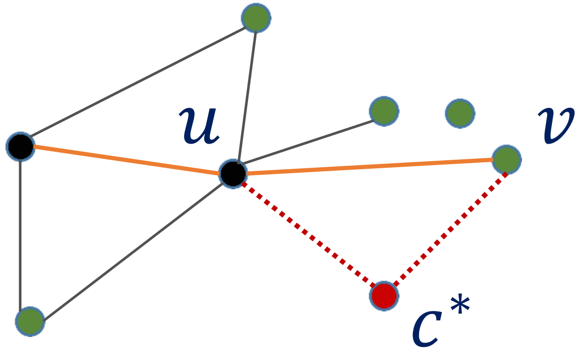


Say, u is an outlier and v is not.

Consider a fixed optimal clustering $C_1^*, \ldots, C_k^*$ for X.

Let c^* be the optimal center for u.

Edges Incident on Outliers



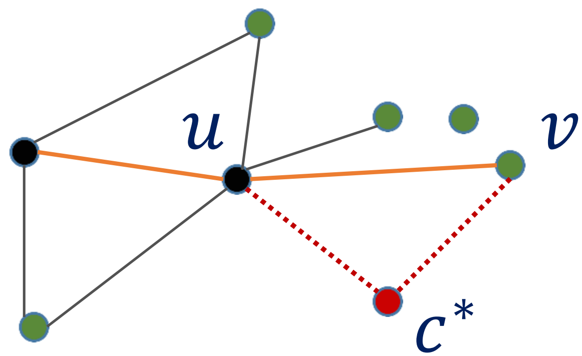
$$\|u - v\| = \|v - c^*\| \pm \|c^* - u\|$$

$$\wr\wr$$

$$\|\varphi(u) - \varphi(v)\| = \|\varphi(v) - \varphi(c^*)\| \pm \|\varphi(c^*) - \varphi(u)\|$$

May assume that the distances between non-outliers and the optimal centers are $(1 + \varepsilon)$-preserved.

Edges Incident on Outliers



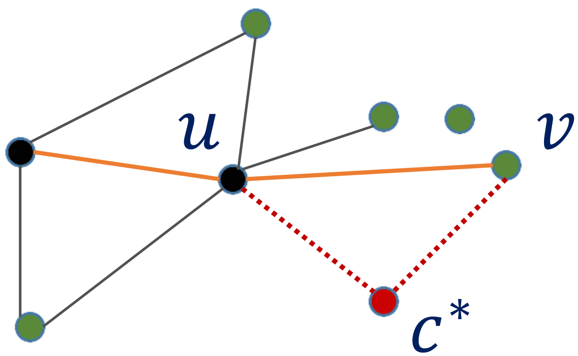
$$\|u - v\| = \|v - c^*\| \pm \|c^* - u\|$$

$$\approx$$

$$\|\varphi(u) - \varphi(v)\| = \|\varphi(v) - \varphi(c^*)\| \pm \|\varphi(c^*) - \varphi(u)\|$$

$$\mathbb{E}[\sum_{u \notin X'} \|c_u^* - u\|^2] \leq \theta \sum_{u \in X} \|c_u^* - u\|^2 = \theta \text{ OPT}$$

Edges Incident on Outliers



$$\|u - v\| = \|v - c^*\| \pm \|c^* - u\|$$

$$\approx$$

$$\|\varphi(u) - \varphi(v)\| = \|\varphi(v) - \varphi(c^*)\| \pm \|\varphi(c^*) - \varphi(u)\|$$

Taking care of $\|\varphi(c^*) - \varphi(u)\|$ is a bit more difficult.

QED

k-medians under dimension reduction

k-medians

– No formula for the cost of the clustering in terms of pairwise distances.

– Not obvious when $d \sim \log n$ (then all pairwise distances are approximately preserved).

[was asked by Ravi Kannan in a tutorial @ Simons]

+ Kirzsbraun Theorem ⇒ the $d \sim \log n$ case

+ Prove a Robust Kirzsbraun Theorem

Our methods for k-means + Robust Kirzsbraun ⇒ $d \sim \log k$ for k-medians

Summary

- Prove that the cost of every k-means and k-medians clustering is preserved up to $(1+\varepsilon)$ under dimension reduction, when $d \geq c \log \frac{k}{\varepsilon\delta} / \varepsilon^2$.
- The bound on d almost matches the lower bound.
- k-means: improves the bound $d \geq \frac{ck}{\varepsilon^2}$ by Cohen et al.
- k-medians: no results were known.
- Applies to k-clustering with the ℓ_p-objective when

$$d \geq c\, p^4 \log \frac{k}{\varepsilon\delta} / \varepsilon^2$$