Fully Dynamic Clustering and Diversity Maximization in Doubling Metrics

Andrea Pietracaprina

University of Padova, Italy

Joint work with:

Paolo Pellizzoni (Max-Planck Inst. of Biochemistry, Germany)

Geppino Pucci (University of Padova, Italy)

Outline

- Problems and key notions
- Our contribution and comparison with previous work
- Augmented Cover Trees
- ACT-based fully-dynamic clustering and diversity
- Conclusions

k-center: Given a pointset S from metric space $(M, d(\cdot,\cdot))$ and $k \leq |S|$, k centers (set $C \subseteq S$) minimizing the *radius*

$$r_C(S) = \max_{x \in S} d(x, C)$$

Variants:

- **k-center with z outliers:** Disregard the z largest center-point distances in the max computation (z outliers).

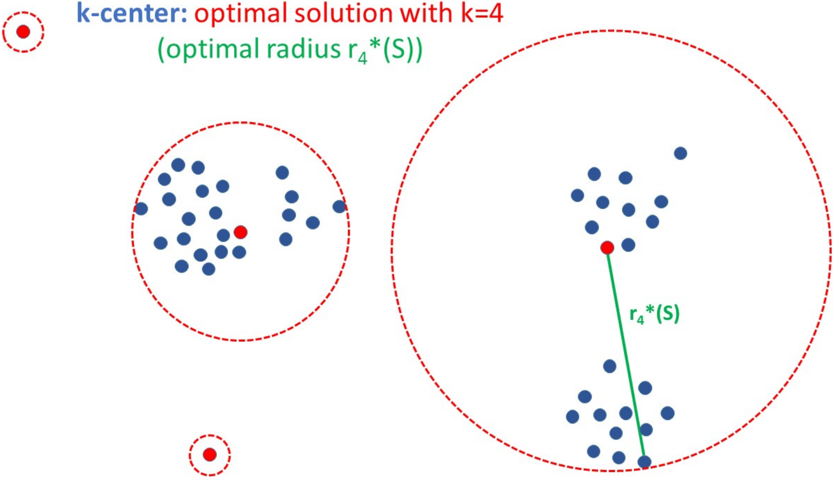
$$r_C(S, z) = \min_{S' \subset S, |S'| = z} \max_{x \in S - S'} d(x, C)$$

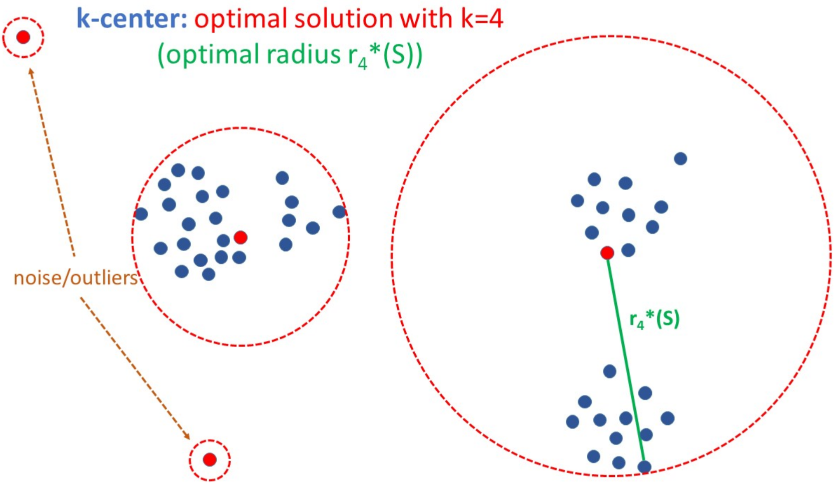
- **Matroid center:** Set C must be an independent set of a matroid (S, I). In this case, $k =$ rank of matroid.

  Used to model *fairness constraints*.

k-center

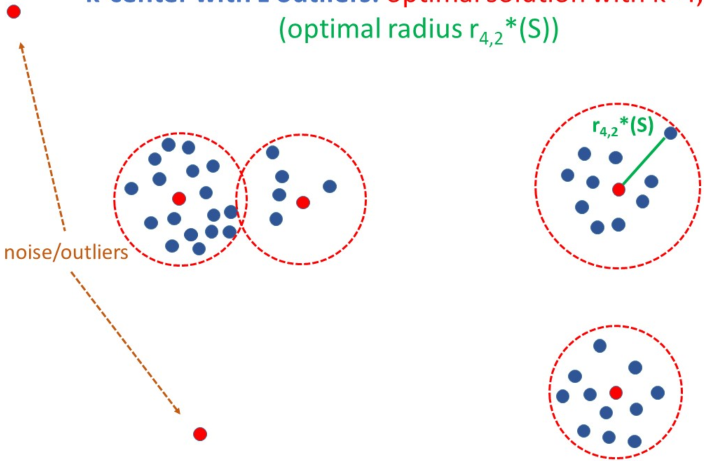
k-center: optimal solution with k=4
(optimal radius $r_4^*(S)$)




k-center: optimal solution with k=4
(optimal radius r4*(S))
noise/outliers
r4*(S)

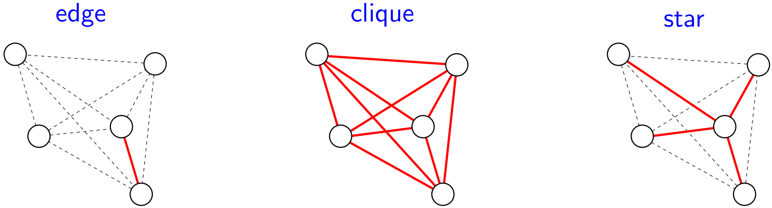
k-center with z outliers

k-center with z outliers: optimal solution with k=4, z=2
(optimal radius $r_{4,2}^*(S)$)



Diversity maximization: Given a pointset S from metric space $(M, d(\cdot,\cdot))$ and $k \leq |S|$, determine k points (set $C \subseteq S$) maximizing a given *diversity function* div(C)

Typical instantiations: div(C) = min (aggregate) distance of a specific subgraph induced by C, e.g.,



Applications

← News/document aggregators

↑ e-commerce ↑

← Facility location

Dynamic setting

Observations. Above problems are NP-hard and best approximations are often costly. Also, practical scenarios entail dynamically evolving data.

FULLY-DYNAMIC SETTING: at each time step t support

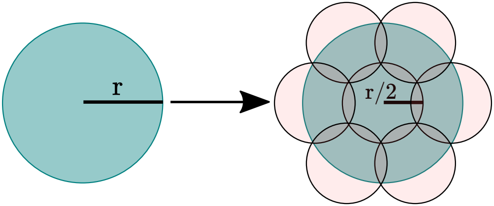
- update: insert/delete a point p in/from S
- query: return a good solution for the current S

GOALs (w.r.t. full recomputation):

- significantly smaller update/query times
- comparable accuracy
- (quasi-)linear space

Doubling dimension

Doubling dimension of a metric space (M, D): minimum D such that every *ball* of radius r can be covered by 2^D balls of radius $r/2$



- Generalizes notion of Euclidean dimension
- E.g., related to expansion for networks under shortest-path distances

Our Contribution

1. Augmented Cover Tree: enhancement of Cover Tree supporting fully-dynamic k-center (with outliers), matroid-center, diversity maximization
   - update time: $O\left(c^D \log \Delta\right)$, with $c = O(1)$ and $\Delta =$ aspect ratio.
   - linear space

   (Extra factor k for matroid-center).

2. Coreset-based fully-dynamic algorithms for the above problems
   - $(\alpha_{\text{static}} + \epsilon)$-approx. with $\alpha_{\text{static}} =$ best approx. in static setting. ($+1$ for k-center with outliers)
   - Query time $O\left(\text{poly}(k, (c/\epsilon)^D) \log \Delta\right)$ (independent of $|S|$!).

Remarks: data structure oblivious to k, ϵ, D, Δ – algorithms oblivious to D, Δ.

Comparison with previous work

- **Cover Tree:**

  [BeygelzimerEtAl-ICML06]: original structure, more complex implementation and analysis, no support for extra info.

- **k-center**

  - [ChanEtAl-WWWW2018], [BateniEtAl-SODA23]: randomized algorithms, update time dependent on k, superlinear space, data structure dependent on k and ϵ.
  - [GoranciEtAl-ALENEX21]: randomized algorithms, superlinear space, data structure dependent on ϵ.

- **k-center with z outliers.**

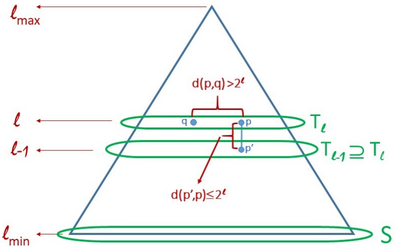
  - [ChanEtAl-COCOON22]: randomized algorithm, bicriteria $(14 + \epsilon)$-approximation ratio (ours is $(3 + \epsilon)$ and non-bicriteria!) update time dependent on k, superlinear space, data structure dependent on k and ϵ.

No previous results for fully-dynamic matroid-center and diversity maximization

Augmented cover tree: definition

Original Cover tree for pointset S

- Levels indexed by $\ell \in [\ell_{\min}, \ell_{\max}]$
- $\ell_{\max} - \ell_{\min} = O(\log \Delta)$
- Each node associated with a $p \in S$
- Each $p \in S$ associated with ≥ 1 nodes
- **Define $T_\ell = \{$points at level $\ell\}$:**
  - $T_\ell \subseteq T_{\ell-1}$
  - $d(p,q) > 2^\ell$ for $p, q \in T_\ell$;
  - $d(p', p) \leq 2^\ell$ for $p' \in T_{\ell-1}$ child of p.



Compaction: to achieve linear space, chains of degree-1 nodes are coalesced.

Augmentation: each node v carries addidtional info:

- weight: number of points in subtree rooted at v
- mis: maximal independent set of points in the subtree rooted at v

Augmented cover tree: dynamic maintainance

DYNAMIC MAINTAINANCE

Key notion: cover set at each level of T for an arbitrary point q

- $Q^q_{\ell_{\max}} = \{\text{root of } T\}$
- $Q^q_\ell = \{p \in T_\ell : d(p,q) \le 2^{\ell+1} \wedge p.\text{parent} \in Q^q_{\ell+1}\}$, for $\ell_{\min} \le \ell < \ell_{\max}$

Lemma: for every q and ℓ

$$|Q^q_\ell| \le 4^D \quad \text{and } |\{\text{children of } Q^q_\ell\}| \le 12^D.$$

Insert/delete of a point q: essentially entails updating all cover sets Q^q_ℓ and the info associated with its nodes.

$\Rightarrow$ running time $= O\left(c^D \log \Delta\right)$

Remark: an extra factor proportional to the rank of the matroid is needed to maintain maximum independent sets, if needed.

ACT-based fully dynamic clustering/diversity

FULLY-DYNAMIC CLUSTERING/DIVERSITY

Main idea:

- Extract from T a *small* coreset $\bar{C} \subset S$ which represents S well for the problem at hand.
- Run best static approximation on $\bar{C}$

Definition: An (ϵ, k)-coreset for S is a subset $\bar{C} \subseteq S$ such that

$$r_{\bar{C}}(S) \leq \epsilon r^*_k(S)$$

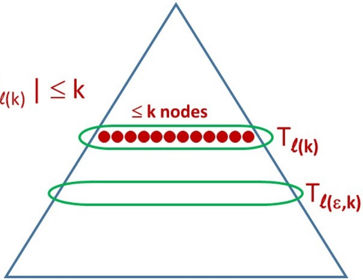
where $r^*_k(S)$ is the optimal radius for k-center.

ACT-based fully dynamic clustering/diversity

Define:

$\ell(k)$ = max index with $|T_{\ell(k)}| \leq k$

$\ell(\varepsilon,k) = \ell(k) - \lceil \log_2 (8/\varepsilon) \rceil$



Lemma: $T_{\ell(\epsilon,k)}$ is an (ϵ, k)-coreset for S

- $|T_{\ell(\epsilon,k)}| \leq k(64/\epsilon)^D$
- Construction time: $O\left(k((64/\epsilon)^D + \log \Delta)\right)$.

Problem-specific coresets

Problem	Coreset	Approximation ratio
k-center	$T_{\ell(\epsilon,k)}$	$2+\epsilon$
k-center with z outliers	$T_{\ell(\epsilon,k+z)}$	$3+O(\epsilon)$
Matroid-center	$\mathrm{MIS}(T_{\ell(\epsilon,k)})$	$3+O(\epsilon)$
Diversity maximization	$\mathrm{DM}(T_{\ell(\epsilon,k)})$	$\alpha_{\text{static}}+O(\epsilon)$

- $\mathrm{MIS}(T_{\ell(\epsilon,k)})$ = union of all maximal independent sets at $T_{\ell(\epsilon,k)}$ (k = rank of matroid)
- $\mathrm{DM}(T_{\ell(\epsilon,k)})$ depends on the diversity function:
  - *edge/cycle variants*: $\mathrm{DM}(T_{\ell(\epsilon,k)}) = (T_{\ell(\epsilon,k)})$
  - *other variants*: $\mathrm{DM}(T_{\ell(\epsilon,k)}) = \mathrm{MIS}(T_{\ell(\epsilon,k)})$ w.r.t. k-bounded cardinality matroid.

α_{static} = best static approximation.

Conclusions

Summary:

- Fully-Dynamic deterministic algorithms for k-center (with outliers), matroid center and diversity maximization.
- The algorithms feature:
  - Accuracy comparable to best static algorithms
  - For data of low doubling dimension, small update and query times (independent of dataset size)

Future work:

- Experimental analysis (under way)
- Lower dependency on doubling dimension
- Extension to other problems (e.g., diversity maximization with matroid constraint, other clustering problems).

Thank you