Clustering Perturbation Resilient k-Median Instances

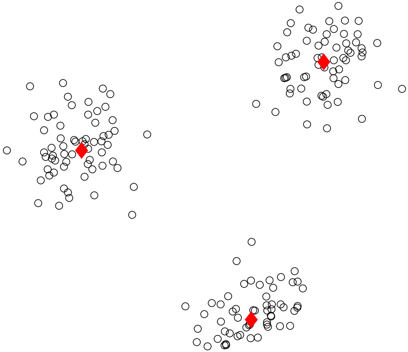
Yingyu Liang

Joint work with Maria Florina Balcan
Georgia Institute of Technology

k-Median Clustering

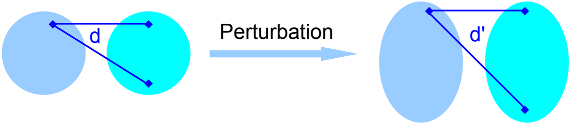
- Given the distances d on a set S of points
- Find centers $\{c_1, \ldots, c_k\}$ to minimize the k-median cost

$$\sum_{p \in P} \min_i d(p, c_i)$$



New Direction: Perturbation Resilience

α-perturbation of d: $d(p,q) \le d'(p,q) \le \alpha d(p,q)$, for any $p, q \in S$

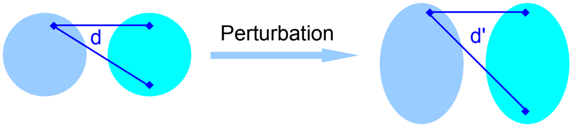


α-Perturbation Resilience [Bilu-Linial,ICS10]

The optimal clustering does not change after α-perturbation.

New Direction: Perturbation Resilience

α-perturbation of d: $d(p,q) \leq d'(p,q) \leq \alpha d(p,q)$, for any $p, q \in S$

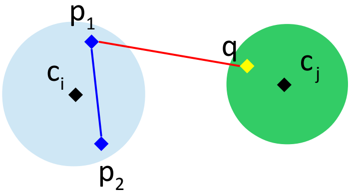


(α, ϵ)-Perturbation Resilience `[Balcan-Liang,ICALP12]`

The optimal clustering changes on at most ϵ fraction of points after α-perturbation.

Our Results

1. Structural property of α-PR for $\alpha > 4$:
   except for $\epsilon|S|$ bad points, all points satisfy strict separation.



2. Approximation algorithm:
   produces $1 + O(\epsilon/\rho)$-approx, where $\rho = \min_i |C_i^*|/n$

Faster Algorithm

Key: structural property preserved in random sample of small size

Sublinear algorithm:

- perform approximation algorithm on a sample of size $\tilde{\Theta}(\frac{k}{\epsilon^2})$
- produces $2(1 + O(\epsilon/\rho))$-approx
- runs in time logarithmic in $\#$points