Clustering Perturbation Resilient Instances

Maria-Florina Balcan

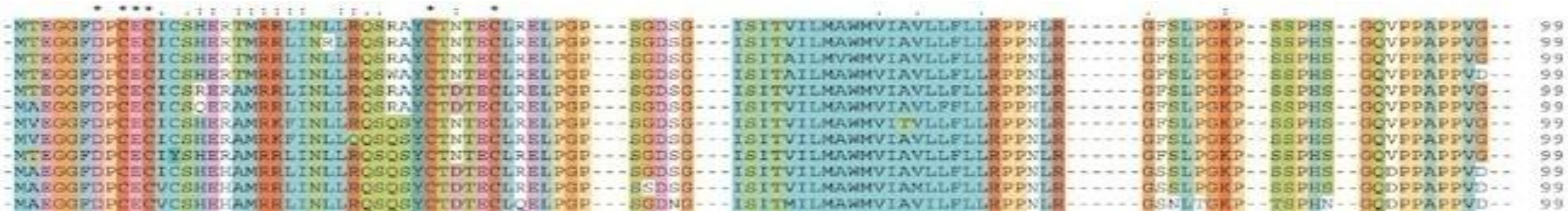
Carnegie Mellon University

Clustering Comes Up Everywhere

- Clustering news articles or web pages or search results by topic.



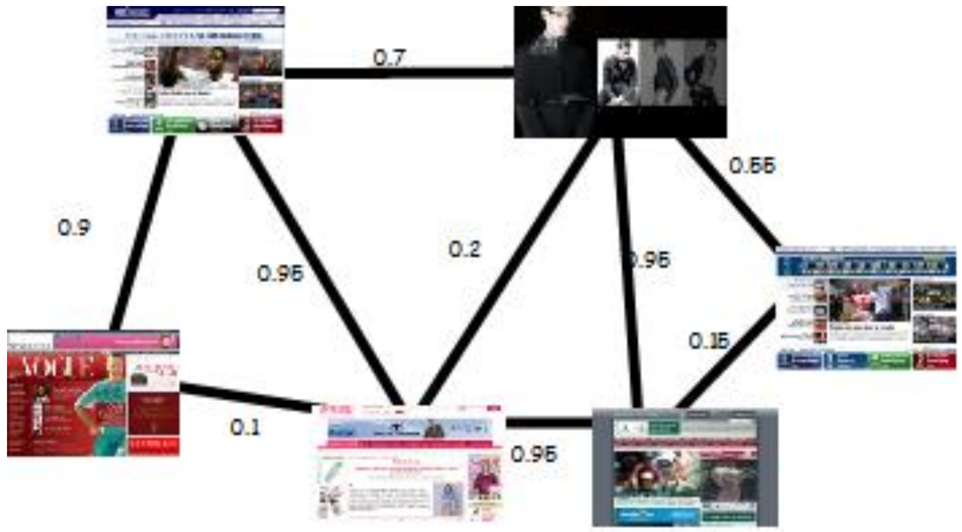
- Clustering protein sequences by function or genes according to expression profile.

- Clustering images by whom is in them.

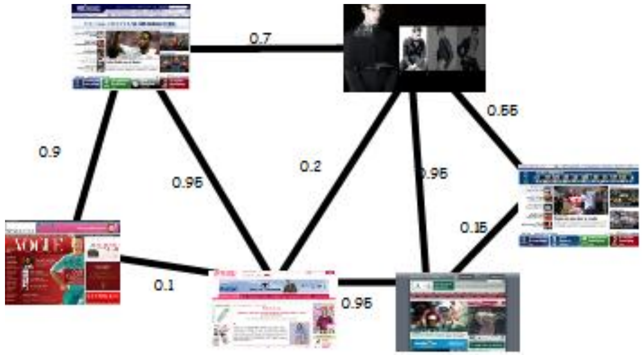
Classic Approach: Objective Based Clustering

- S set of n objects. [documents, web pages]
- Also have a distance/dissimilarity measure.
- View objects as nodes in weighted graph based on distances.

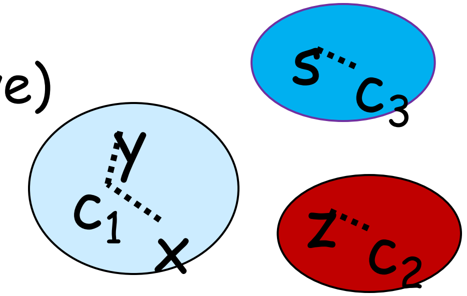


Classic Approach: Objective Based Clustering

- View objects as nodes in weighted graph based on distances.



- Pick an objective to optimize (e.g., a classic center based objective)
  - k-median: find centers $\{c_1, c_2, \dots, c_k\}$ to min $\sum_p \min d(p, c_i)$
  - k-means: find centers $\{c_1, c_2, \dots, c_k\}$ to min $\sum_p \min d^2(p, c_i)$
  - k-center: find centers to minimize the maximum radius.



Standard Theoretical Approach in ML, TCS

However many of these problems are provably NP-hard to optimize in poly time in the worst case….

- k-median: NP-hard to approximate within $\left(1+\frac{1}{e}\right)$

  can be approximated within a $(1+\sqrt{3}+\epsilon)$ factor

- K-center: NP-hard to optimize within a factor of 2

  can be approximated within a factor of 2

- Asymmetric K-center: NP-hard to optimize within a factor of $\log^* k$

  can be approximated within a factor of $\log^* k$

Cool new direction: exploit additional properties of the data to circumvent lower bounds.

α-Perturbation Resilient Instances

Definition

α-Perturbation: (S, d), d distance function, an α-perturbation is any function d' s.t. $\forall p, q \in S, d(p,q) \leq d'(p,q) \leq \alpha d(p,q)$.

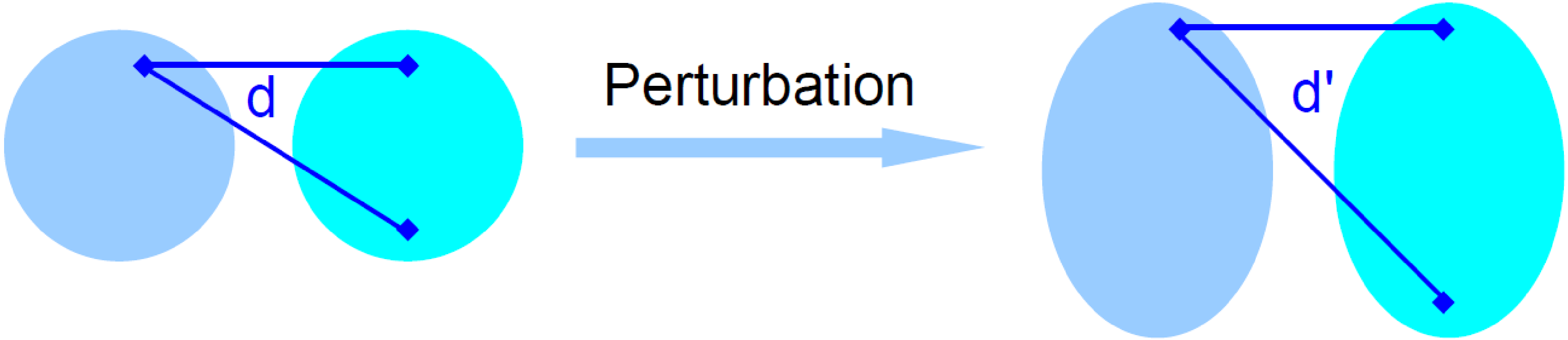
Definition [Bilu and Linial, 2010]:

A clustering instance (S, d) is α-perturbation resilient (α-PR) for objective Φ if for any α-Perturbation d', $OPT_{d'} = OPT_d$ (i.e., the optimal clustering for Φ under d', $OPT_{d'}$ is equal to the optimal clustering OPT for Φ under d).

α-Perturbation Resilient Instances

Definition

A clustering instance (S, d) is α-perturbation resilient for objective Φ if for any α-Perturbation d', $OPT_{d'} = OPT_d$ (i.e., the optimal clustering for Φ under d', $OPT_{d'}$ is equal to the optimal clustering OPT for Φ under d).



Related Work: Positive Results Exploiting PR

- Poly time algo for finding OPT for α-PR instances of max cut when $\alpha > \sqrt{n}$ [Bilu-Linial, 2010]
- Poly time algo for finding OPT for α-PR for max cut when $\alpha = \tilde{\Theta}(\sqrt{\log n})$ [Markarychev et al 2013]
- Poly time algo for finding OPT for α-PR for any center based objective when $\alpha > 3$ (e.g., k-median, k-means, k-center)

Our Results: Positive Results Exploiting PR

Center based objectives & Min-sum [Balcan-Liang'12] [Balcan-Liang'14]

- Poly time algo for finding OPT for α-PR for any center based objective when $\alpha > 1 + \sqrt{2}$ (e.g., k-median, k-means, k-center)
- Poly time algo for a generalization (α, ϵ)-PR for k-median.
- Poly time algo for finding OPT for α-PR min-sum instances when $\alpha > 3 \frac{\max_i |C_i|}{\min_i |C_i|}$

K-center [Balcan-Haghtalab-White'15]

- Tight poly time algo for finding OPT for α-PR for k-center when $\alpha > 2$. This is tight!!!!
- Poly time algo for finding OPT for α-PR for asymmetric k-center, $\alpha > 3$.

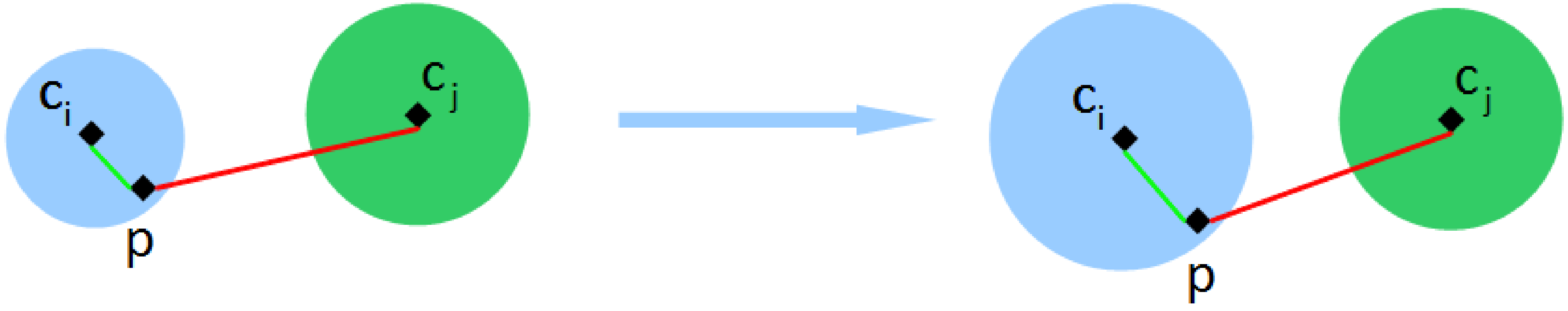
Structural Properties Induced by α-PR

Claim For any center based objective, α-PR implies α-center proximity.

I.e., α-PR implies that $\forall p \in C_i, \alpha d(p, c_i) < d(p, c_j)$.

Proof

- d': blow up all pairwise distances within the optimal cluster by α
- OPT and centers do not change, so $\forall p \in C_i, d'(p, c_i) < d'(p, c_j)$.
- $\alpha d(p, c_i) = d'(p, c_i) < d'(p, c_j) = d(p, c_j)$



Structural Properties Induced by α-PR

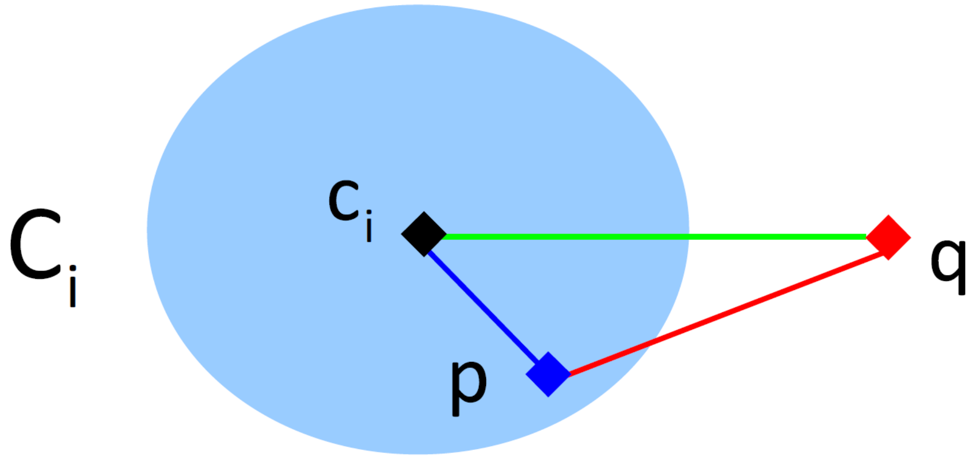
Claim For any center based objective, α-PR implies α-center proximity.

I.e., α-PR implies that $\forall p \in C_i, \alpha d(p, c_i) < d(p, c_j)$.

Implication For any center based objective

If $\alpha > 1 + \sqrt{2}$, then for any $p \in C_i, q \notin C_i$,

- $d(c_i, p) < d(c_i, q)$
- $d(p, c_i) < d(p, q)$

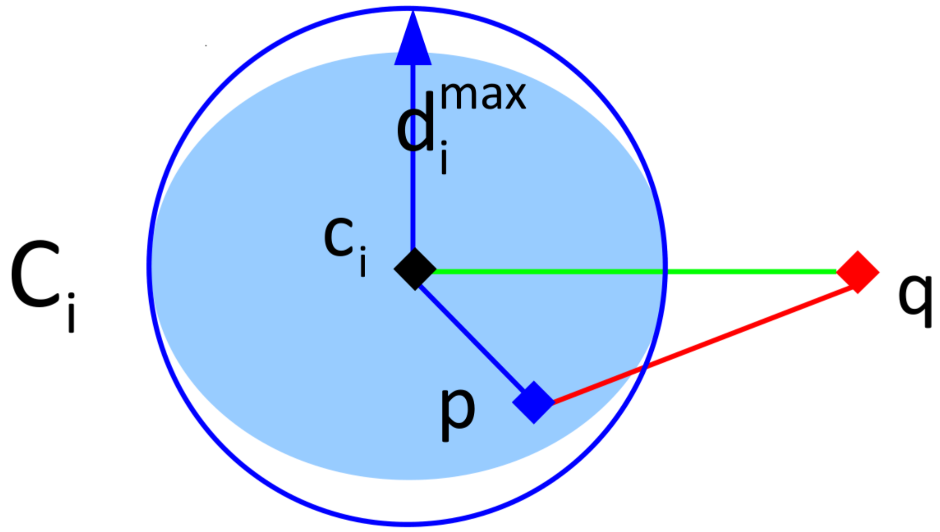


Structural Properties Induced by α-PR

Implication: For any center based objective

If $\alpha > 1 + \sqrt{2}$, then for any $p \in C_i, q \notin C_i$,

- $d(c_i, p) < d(c_i, q)$
- $d(p, c_i) < d(p, q)$



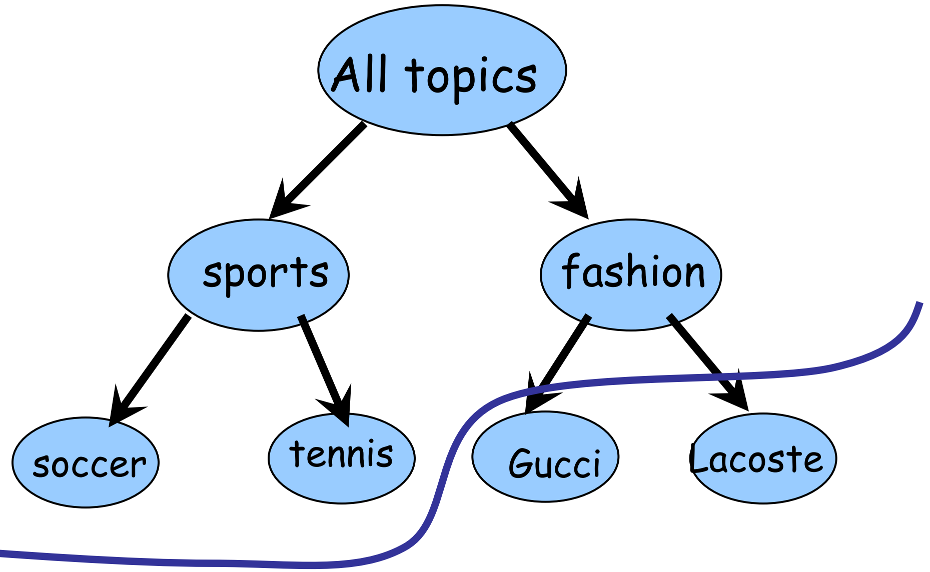
Let $d_i^{max} = \max_{p \in C_i} d(p, c_i)$. Construct ball $B = B(c_i, d_i^{max})$.

- The ball covers exactly C_i
- Points inside are closer to the center than to points outside: for any points $p \in B, q \notin B, d(p, c_i) < d(p, q)$

Algorithm for Clustering $1+\sqrt{2}$-PR instances

Step 1: Closure Linkage

- Begin with each point being a cluster
- Repeat until one cluster remains: merge the two clusters with minimum closure distance



Step 2: Apply **dynamic programming** to extract the minimum k-cost clustering.

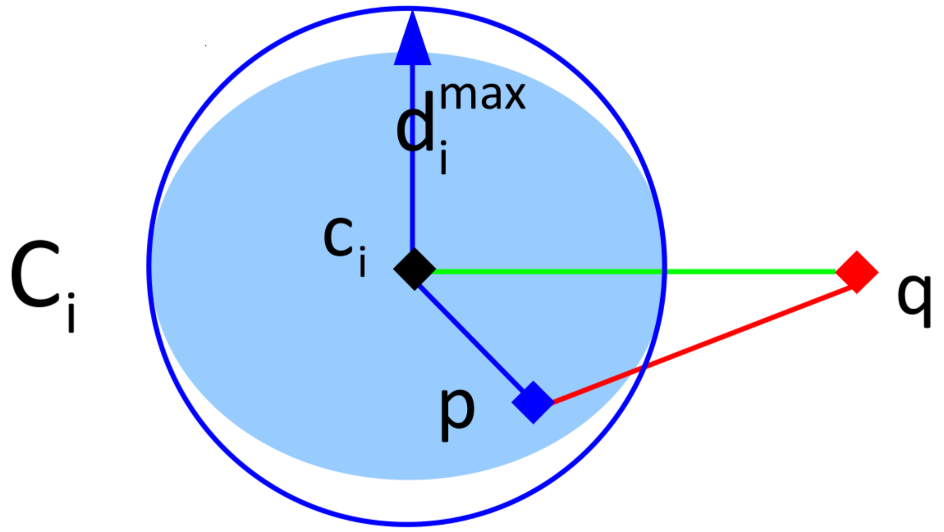
Fact: If $\alpha \geq 1+\sqrt{2}$, the tree output contains OPT as a pruning.

Structural Properties Induced by α-PR

Implication For any center based objective

If $\alpha > 1 + \sqrt{2}$, then for any $p \in C_i, q \notin C_i$,

- $d(c_i, p) < d(c_i, q)$
- $d(p, c_i) < d(p, q)$



Let $d_i^{max} = \max_{p \in C_i} d(p, c_i)$. Construct a ball $B = B(c_i, d_i^{max})$.

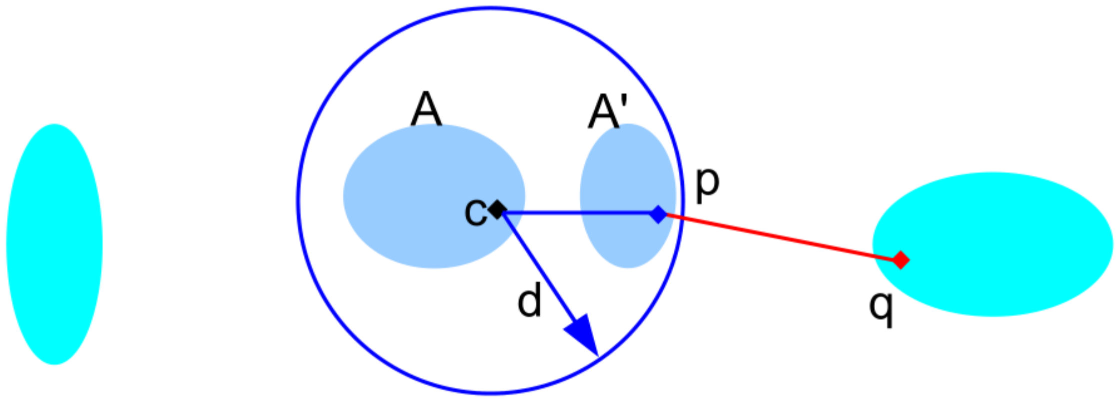
- The ball covers exactly C_i
- Points inside are closer to the center than to points outside: for any points $p \in B, q \notin B, d(p, c_i) < d(p, q)$

Closure Distance

Closure distance between 2 sets: radius of the minimum ball that covers the sets and has some margin outside the sets.

Definition: The closure distance $d_S(A, A')$ between A and A' is the minimum d, s. t. $\exists c \in A \cup A'$ satisfying:

- Coverage: the ball $B(c, d)$ covers $A \cup A'$
- Margin : pts inside are closer to the center than to pts outside, i.e., $\forall p \in B(c, d), q \notin B(c, d), d(p, c) < d(p, q)$

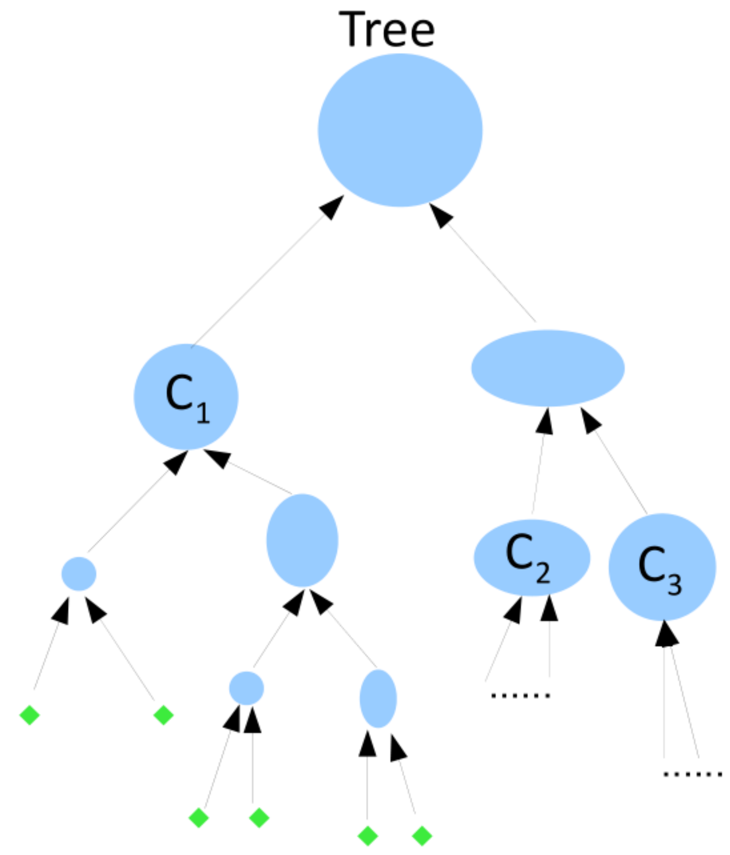


Clustering $1 + \sqrt{2}$ center based Resilient Instances

Closure Linkage Algorithm

- Begin with each point being a cluster
- Repeat until one cluster remains: merge the two clusters with minimum closure distance
- Output the tree obtained

Theorem: If $\alpha \geq 1 + \sqrt{2}$, the tree output contains OPT as a pruning.



Clustering $1+\sqrt{2}$ center based Resilient Instances

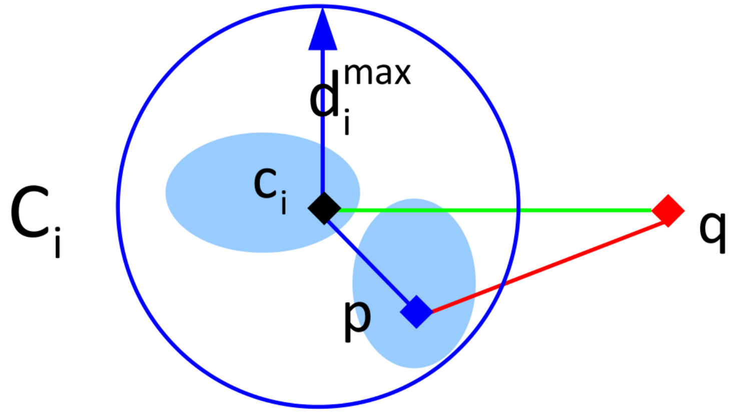
If $\alpha \geq 1+\sqrt{2}$, the tree output by closure linkage contains OPT as a pruning.

Proof idea: induction, show that current clustering is laminar w.r.t. OPT

Show that algo will not merge a strict subset A in C_i with a subset A' outside C_i.

- Pick $B \subset C_i \setminus A$ such that $c_i \in A \cup B$
- Then $d_S(A,B) \leq d_i^{max} = \max_{p \in C_i} d(p, c_i)$

  since the two conditions of closure distance are satisfied



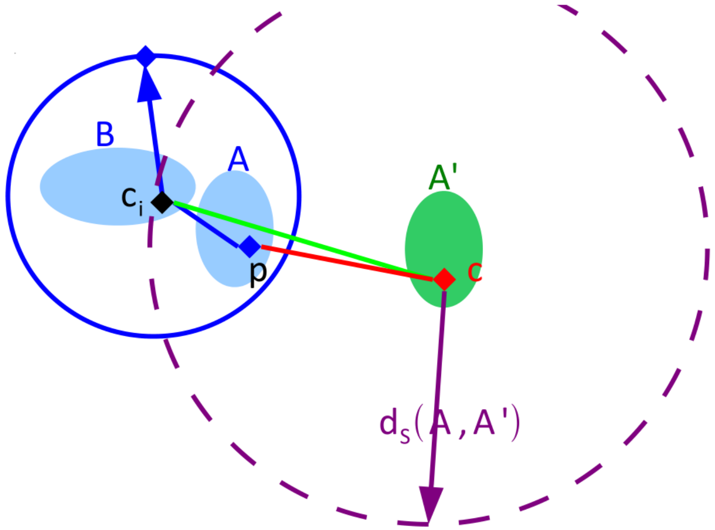
Clustering $1+\sqrt{2}$ center based Resilient Instances

If $\alpha \geq 1+\sqrt{2}$, the tree output by closure linkage contains OPT as a pruning.

Proof idea: induction, show that current clustering is laminar w.r.t. OPT

- $d_S(A, A') > d_i^{max}$
- Suppose center c for the ball defining $d_S(A, A')$ is from A'
- Since $c \notin C_i, d(c_i, p) < d(p, c)$ for any $p \in C_i$.

  By margin, $c_i \in B(c, d_S(A, A'))$, so $d_S(A, A') \geq d(c_i, c)$
- Since $c \notin C_i$, $d(c_i, c) > d_i^{max}$

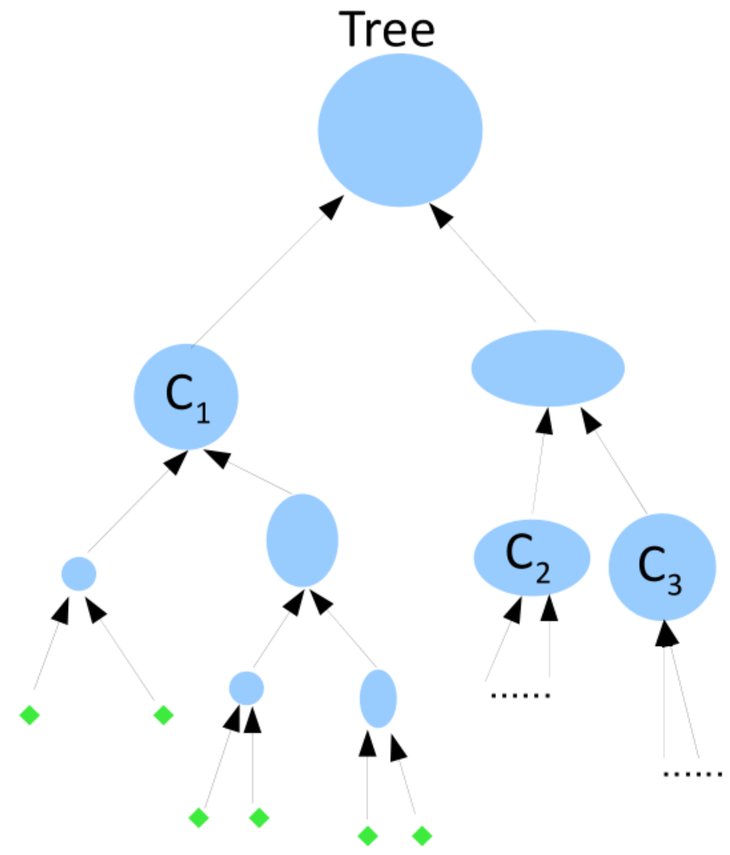


- A similar argument holds for the case $c \in A$

Clustering $1+\sqrt{2}$ center based Resilient Instances

Closure Linkage Algorithm

- Begin with each point being a cluster
- Repeat until one cluster remains: merge the two clusters with minimum closure distance
- Output the tree obtained



Theorem: If $\alpha \geq 1+\sqrt{2}$, the tree output contains OPT as a pruning.

(α, ϵ)-Perturbation Resilience

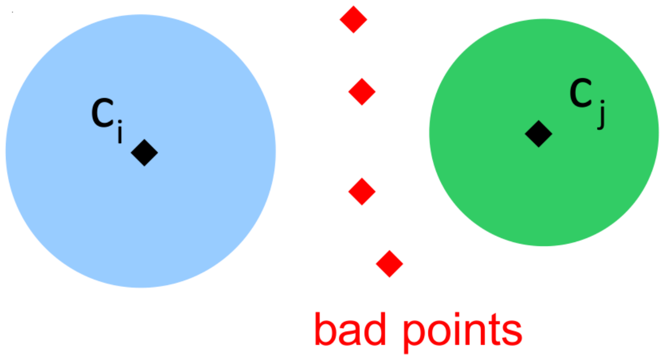
- α-PR imposes a strong restriction: OPT does not change after perturbation
- We propose a realistic relaxation of this condition.

Definition:

A clustering instance (S, d) is (α, ϵ)-perturbation resilient to a given objective function Φ if for any α-Perturbation d', the optimal clustering $OPT_{d'}$ is ϵ-close to the optimal clustering OPT_d.

Structural Property of (α, ϵ)-PR k-median

Theorem: Assume $\min_i |C_i| = O(\epsilon n)$. Except for at most ϵn bad points, any other point is α times closer to its own center than to other centers.



Proof sketch:

- Carefully construct a perturbation that forces all the bad points move
- By (α, ϵ)-PR, there could be at most ϵn bad points

Structural Property of (α, ϵ)-PR k-median

- Assume more than $\epsilon n + 1$ bad points; select ϵn of them.

Perturbation: blow up all pairwise distances by α, except

- between selected bad points and their second nearest centers
- between the other points and their own centers

Intuition: ideally, after the perturbation,

- selected bad points assigned to their second nearest centers
- all the other points stay

(α, ϵ)-PR k-median

Theorem: If $\min_i |C_i| = \Omega(\epsilon n), \alpha > 4$, then the tree output contains a pruning that is ϵ-close to the optimal clustering. Moreover, the cost of this pruning is $1 + O(\epsilon/\rho)$-approximation where $\rho = \min_i |C_i| / n$.

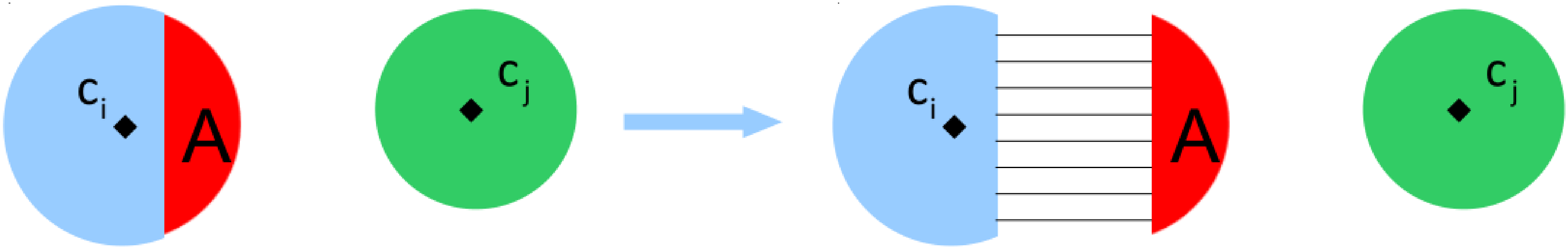
Idea:

- If $\alpha > 4$ except for the bad points we have strict separation, each good point is closer to points in its own cluster than to any other cluster.
- Run a robust version of single linkage. Guaranteed to have a pruning that is a good approximation.

Structural Property of α-PR Min-Sum

Claim: α-PR implies that for any $A \subseteq C_i, \alpha d(A, C_i \setminus A) < d(A, C_j)$.

Proof: blow up the distances between A and $C_i \setminus A$ by α.



Structural Property of α-PR Min-Sum

Claim: α-PR implies that for any $A \subseteq C_i$, $\alpha d(A, C_i \setminus A) < d(A, C_j)$.

Implications when $\alpha > 3\frac{\max_i |C_i|}{\min_i |C_i|}$

(1) For any point, its $\min_i |C_i|/2$ nearest neighbors are from the same optimal cluster

(2) Any strict subset of an optimal cluster has smaller average distance to the other points in the same cluster than to those in other clusters

Algorithm for α-PR Min-Sum

- Connect each point with its $\min_i |C_i|/2$ nearest neighbors
- Perform average linkage on the components

Theorem: If $\alpha > 3\frac{\max_i |C_i|}{\min_i |C_i|}$, then the tree contains OPT as a pruning.

- Implication (1) guarantees that the components are pure
- Implication (2) guarantees that no strict subset of an optimal cluster will be merged with a subset outside the cluster

Our Results: Positive Results Exploiting PR

Center based objectives & Min-sum [Balcan-Liang'12] [Balcan-Liang'14]

- Poly time algo for finding OPT for α-PR for any center based objective when $\alpha > 1 + \sqrt{2}$ (e.g., k-median, k-means, k-center)
- Poly time algo for a generalization (α, ϵ)-PR for k-median.
- Poly time algo for finding OPT for α-PR min-sum instances when $\alpha > 3\frac{\max_i |C_i|}{\min_i |C_i|}$

K-center [Balcan-Haghtalab-White'15]

- Tight poly time algo for finding OPT for α-PR for k-center when $\alpha > 2$. This is tight!!!!
- Poly time algo for finding OPT for α-PR for asymmetric k-center, $\alpha > 3$.

Great Research Direction

Exploit additional properties of the data to circumvent computational hardness lower bounds.

- Polynomial time algorithm for finding (nearly) optimal solutions for perturbation resilient instances.
- Also consider a more realistic relaxation (α, ϵ)-PR