A Unified Framework for Clustering Constrained Data without Locality Property

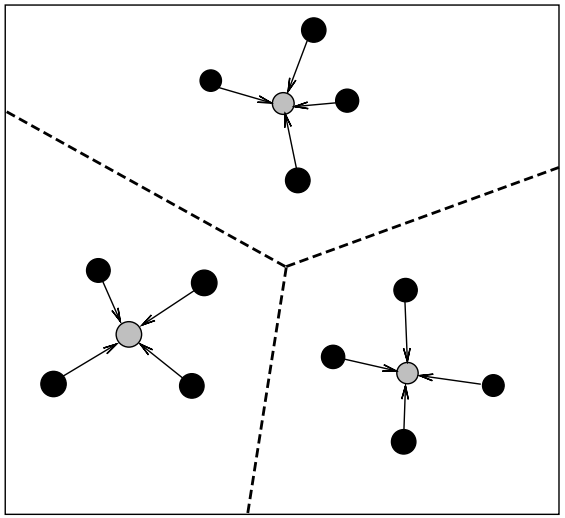
Hu Ding Jinhui Xu

Department of Computer Science and Engineering
State University of New York at Buffalo

January 6, 2015

Ordinary clustering: k-means/median

- Given $P \subset \mathbb{R}^d$, group it into k clusters and minimize the average (squared) distance from each point to its closest mean/median point.

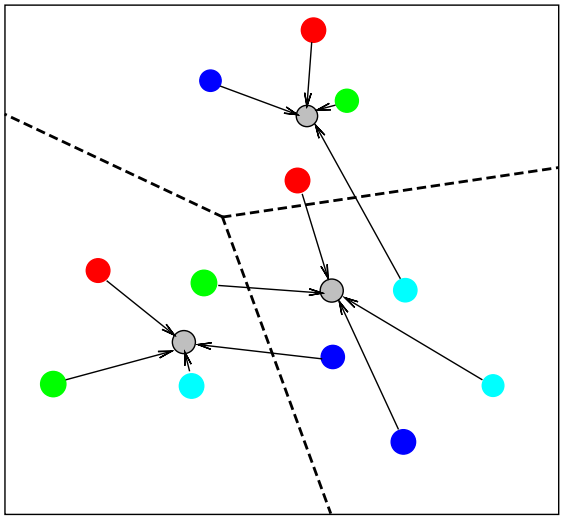


- **Locality property**, *i.e.*, each cluster lies entirely inside the Voronoi cell of its center.
- Many existing algorithms: [Badŏiu *et al.* 2002], [Ostrovsky *et al.* 2006], [Kumar *et al.* 2010], *etc.*

Motivations

Constrained clustering

- Locality property may no longer exist $\Longrightarrow$ most existing algorithms fail



Problem Description: Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

- Partition $P \subset \mathbb{R}^d$ into k clusters satisfying some additional constraint $\mathbb{C}$ and minimizing the objective function of the ordinary k-means (or k-median) problem.
- d could be quite high, and k is a constant.

Problem Description: Constrained Clustering

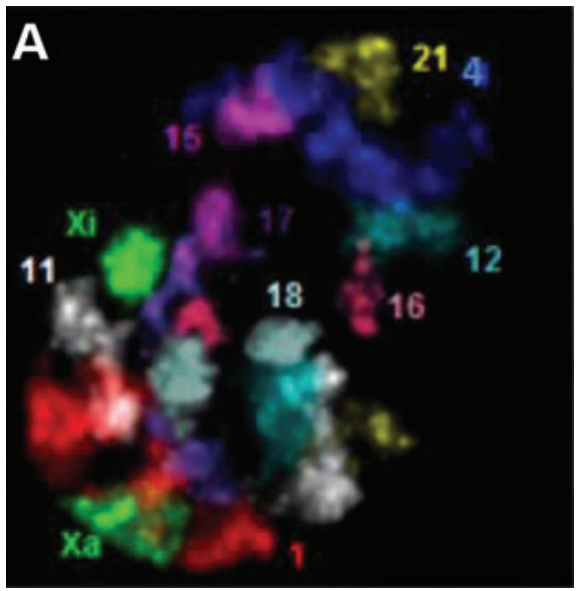
Constrained k-means/median (k-CMeans and k-CMedian)

- Partition $P \subset \mathbb{R}^d$ into k clusters satisfying some additional constraint $\mathbb{C}$ and minimizing the objective function of the ordinary k-means (or k-median) problem.
- d could be quite high, and k is a constant.
- **Hardness:** NP-hard even for the ordinary k-means/median clustering in high dimensions with $k = 2$ [Guruswami and Indyk 2003] and [Dasgupta 2008].

Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

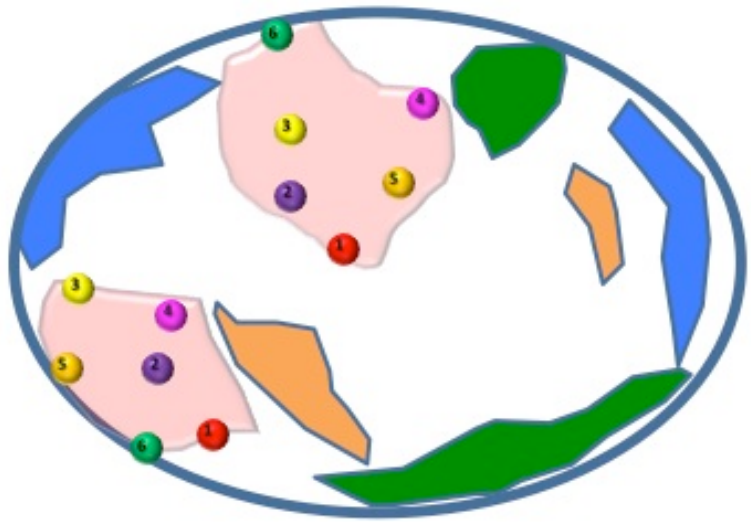
- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering



Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering



Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering

Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering

	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	< 30	*	AIDS
2	130**	< 30	*	Heart Disease
3	130**	< 30	*	Viral Infection
4	130**	< 30	*	Viral Infection
5	1485*	≥ 40	*	Cancer
6	1485*	≥ 40	*	Heart Disease
7	1485*	≥ 40	*	Viral Infection
8	1485*	≥ 40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

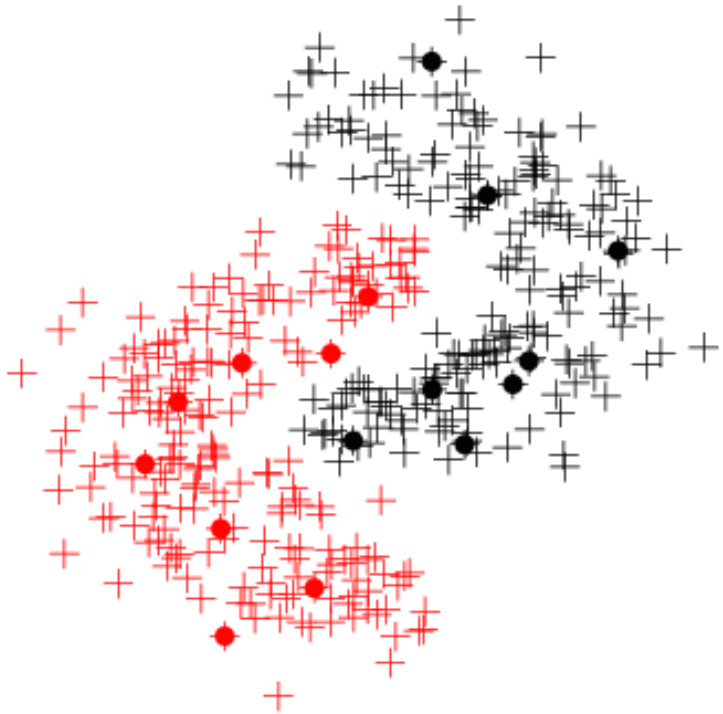
- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	≤ 40	*	Heart Disease
4	1305*	≤ 40	*	Viral Infection
9	1305*	≤ 40	*	Cancer
10	1305*	≤ 40	*	Cancer
5	1485*	> 40	*	Cancer
6	1485*	> 40	*	Heart Disease
7	1485*	> 40	*	Viral Infection
8	1485*	> 40	*	Viral Infection
2	1306*	≤ 40	*	Heart Disease
3	1306*	≤ 40	*	Viral Infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

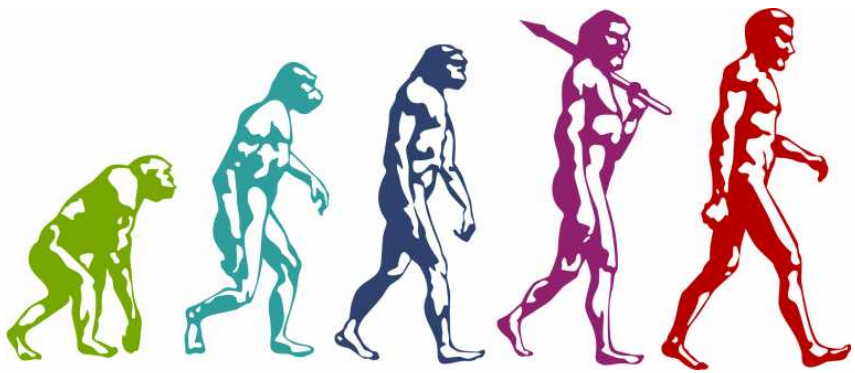
- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering



Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

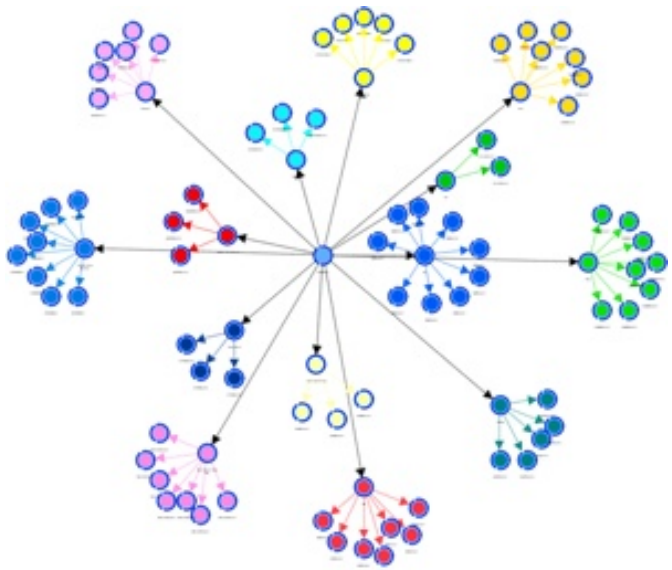
- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering



Some Examples of Constrained Clustering

Constrained k-means/median (k-CMeans and k-CMedian)

- Chromatic clustering
- r-gather clustering
- l-diversity clustering
- Semi-supervised clustering
- Evolutionary clustering
- Probabilistic clustering



Our Main Result: A Unified Framwork

Two steps for clustering in Euclidean space:

1. Identify the set of cluster centers, *i.e.*, the k mean or median points.
2. Partition the input points into k clusters based on these mean or median points. The partition step may be non-trivial for some problems.

Our Main Result: A Unified Framework

Theorem: Let P be a set of n points in $\mathbb{R}^d$, and $\mathbb{C}$ be some constraint.

Our Main Result: A Unified Framework

Theorem: Let P be a set of n points in $\mathbb{R}^d$, and $\mathbb{C}$ be some constraint. There exists an algorithm outputting $O(\log^{k+1} n)$ k-tuple candidates for the mean/median points in $O(n(\log^{k+2} n)d)$ time,

Our Main Result: A Unified Framework

Theorem: Let P be a set of n points in $\mathbb{R}^d$ and $\mathbb{C}$ be some constraint. There exists an algorithm outputting $O(\log^{k+1} n)$ k-tuple candidates for the mean/median points in $O(n(\log^{k+2} n)d)$ time, and with probability $1 - \frac{1}{n}$ at least one candidate yields $(1+\epsilon)$-approximation of k-CMeans/CMedian, if the corresponding partition step can be solved.

Our Results for a Class of k-CMeans/CMedian

Problems	Existing Results
l-diversity clustering	2-approx. for metric k-centers (only for a restricted version of l-diversity clustering)
chromatic clustering	$(1+\epsilon)$-approx. for chromatic k-cones clustering in $\mathbb{R}^d$; $(1+\epsilon)$-approx. for 2-center in $\mathbb{R}^2$
fault tolerant clustering	4 and 93-approx. for uniform and non-uniform metric k-median; 2-approx. for metric k-centers;
r-gather clustering	2-approx. for metric k-centers and 4-approx. for metric k-cellulars; $(4+\epsilon)$-approx. for k-centers in constant dimensional space
capacitated clustering	6 and 7-approx. for metric k-centers with uniform and non-uniform capacities
semi-supervised clustering	Heuristic algorithms
uncertain data clustering	$(1+\epsilon)$-approx. for k-means and unassigned k-median; $(3+\epsilon)$-approx. for assigned k-median; $(1+\epsilon)$-approx. for assigned k-median in constant dimensional space; $O(1)$-approx. for k-centers
Our results: $(1+\epsilon)$-approx. of k-means and k-median for all 7 problems in any (*i.e.*, both low and high) dimensional space	

Table 1: Existing and our new results for the class of constrained clustering problems.

The running time depends on the partition step for each constraint $\mathbb{C}$.

A Representative Approach for Ordinary Clustering

Used in [Badŏiu *et al.* 2002], [Kumar *et al.* 2010], *etc.*

Peeling + Sampling

A Representative Approach for Ordinary Clustering

Peeling + Sampling

- **Approximate the mean point**: Uniform random sampling [Inaba *et al.* 1994].

$$||m(S) - m(T)||^2 \leq O(\frac{1}{|S|})\delta^2, S \subset T.$$

- **Approximate the median point**: Similar but more complicated approach, and similar result [Badŏiu *et al.* 2002].

A Representative Approach for Ordinary Clustering

Peeling + Sampling

- **Approximate the mean point**: Uniform random sampling [Inaba *et al.* 1994].

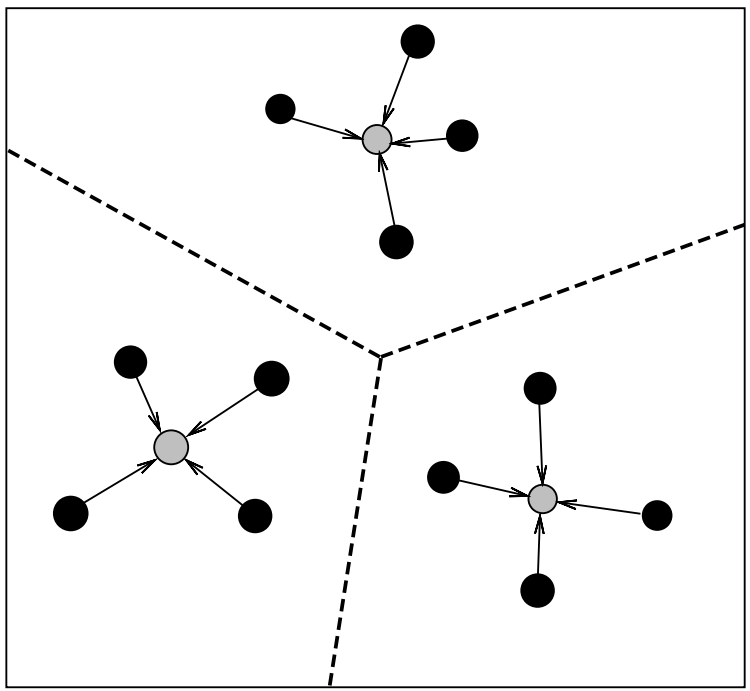
$$||m(S) - m(T)||^2 \leq O(\frac{1}{|S|})\delta^2, S \subset T.$$

- **Approximate the median point**: Similar but more complicated approach, and similar result [Badŏiu *et al.* 2002].

A Representative Approach for Ordinary Clustering

Peeling + Sampling

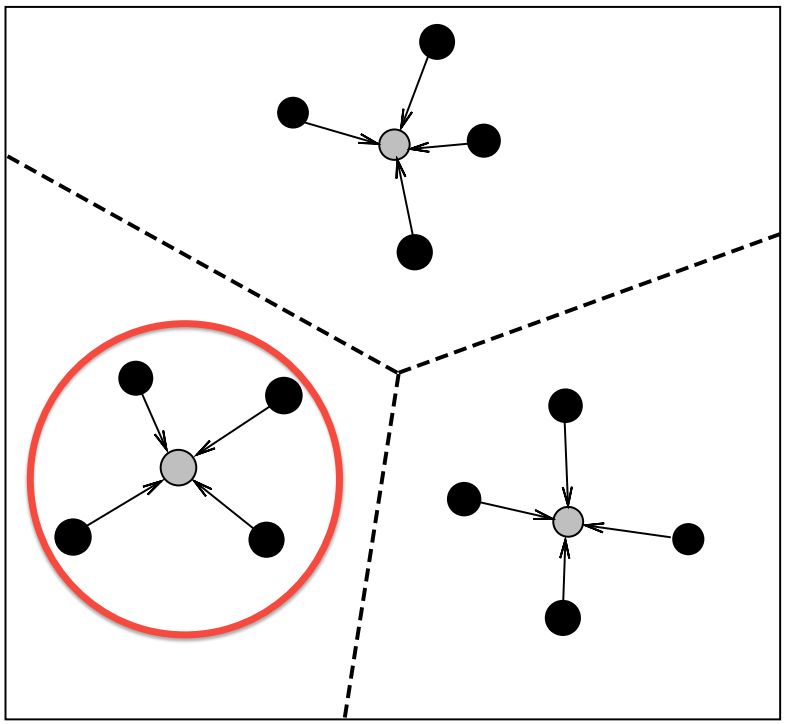
- Iteratively draw k peeling spheres
- Each time find the j-th mean/median point via random sampling



A Representative Approach for Ordinary Clustering

Peeling + Sampling

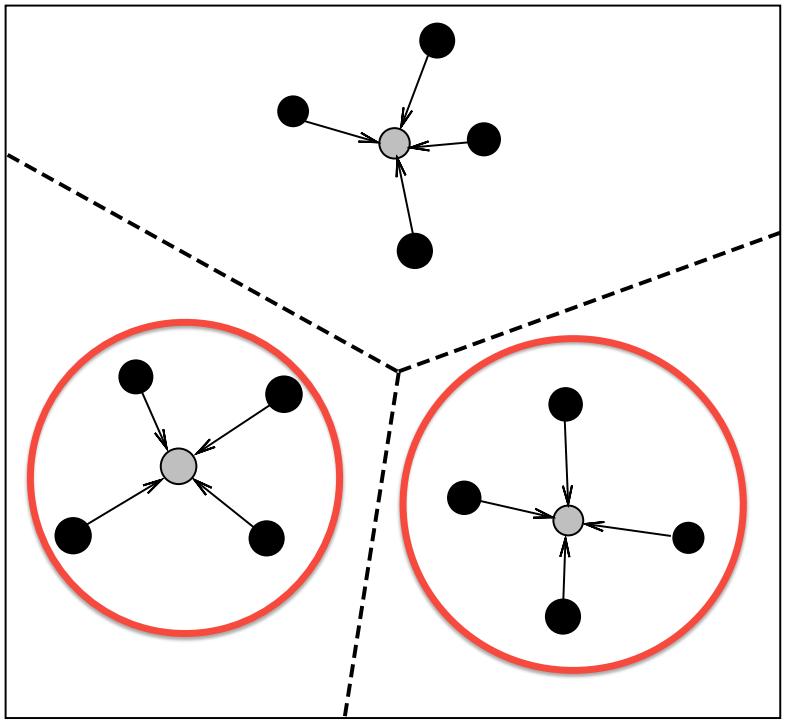
- Iteratively draw k peeling spheres
- Each time find the j-th mean/median point via random sampling



A Representative Approach for Ordinary Clustering

Peeling + Sampling

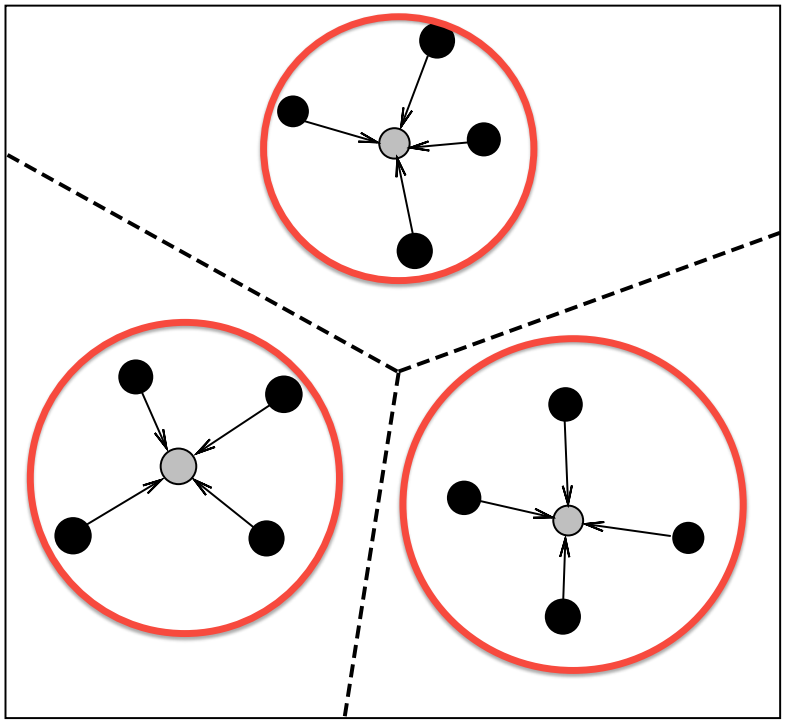
- Iteratively draw k peeling spheres
- Each time find the j-th mean/median point via random sampling



A Representative Approach for Ordinary Clustering

Peeling + Sampling

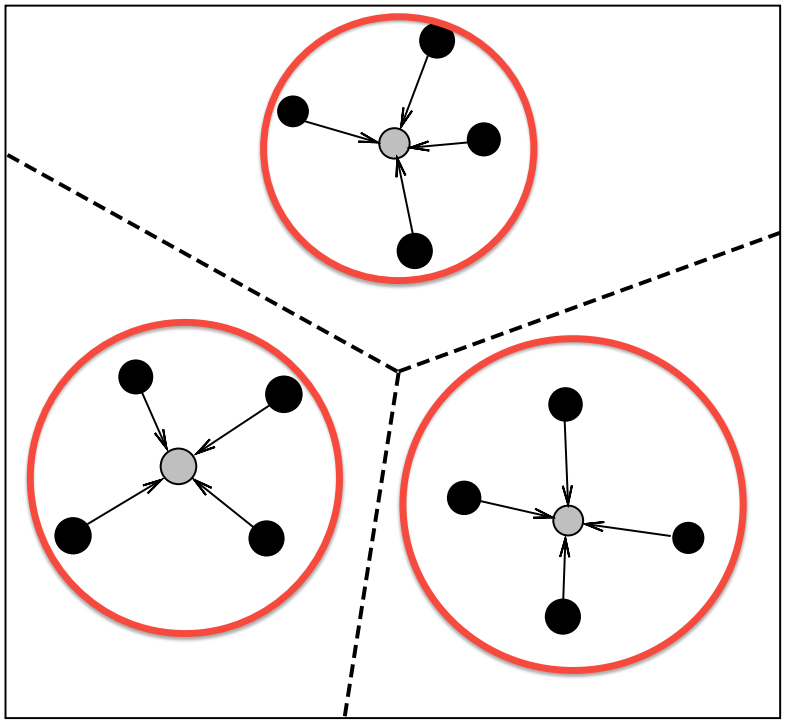
- Iteratively draw k peeling spheres
- Each time find the j-th mean/median point via random sampling



A Representative Approach for Ordinary Clustering

Peeling + Sampling

- Iteratively draw k peeling spheres
- Each time find the j-th mean/median point via random sampling
- Locality property ensures the correctness



Roadmap of the Remaining Part

1. Two novel techniques:
   - Simplex Lemma to find the mean/median points for the constrained clustering, instead of simply random sampling.
   - A Unified Constant Approximation to determine the radii of peeling spheres.
2. A Peeling + Enclosing algorithm which outputs a set of candidates for k mean/median points.
3. Algorithms for selecting the best candidate based on each individual constraint.

Roadmap of the Remaining Part

1. Two novel techniques:
   - Simplex Lemma to find the mean/median points for the constrained clustering, instead of simply random sampling.
   - A Unified Constant Approximation to determine the radii of peeling spheres.
2. A Peeling + Enclosing algorithm which outputs a set of candidates for k mean/median points.
3. Algorithms for selecting the best candidate based on each individual constraint.

Roadmap of the Remaining Part

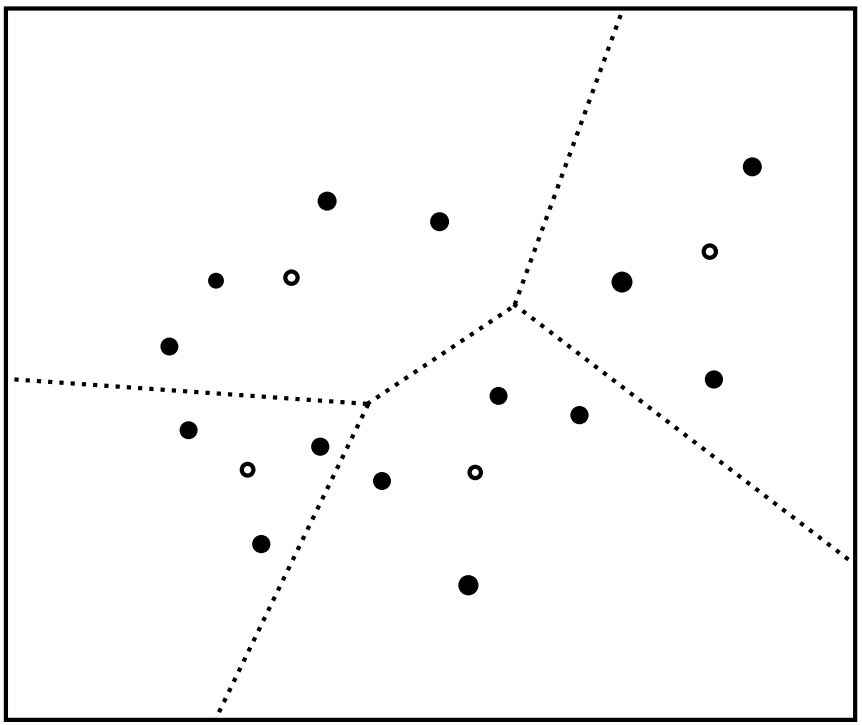
1. Two novel techniques:
   - Simplex Lemma to find the mean/median points for the constrained clustering, instead of simply random sampling.
   - A Unified Constant Approximation to determine the radii of peeling spheres.
2. A Peeling + Enclosing algorithm which outputs a set of candidates for k mean/median points.
3. Algorithms for selecting the best candidate based on each individual constraint.

Novel Technique I: Simplex Lemma

Generally speaking, **Simplex Lemma** enables us to approximate the mean point of some unknown point set with only partial knowledge.

Novel Technique I: Simplex Lemma

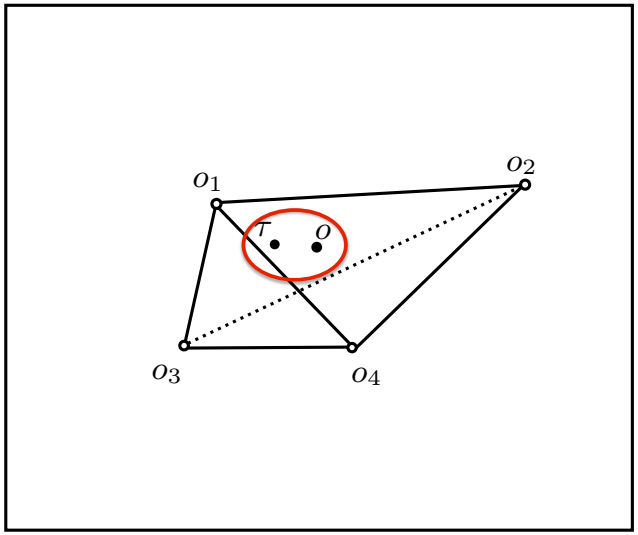
- Let P be an unknown set of points in $\mathbb{R}^d$.
- Let $P = \cup_{l=1}^{j} P_l$ be an unknown partition of P into j subsets.
- Let o_l be the known mean point of P_l for $1 \leq l \leq j$.



Note: The partition can be arbitrary, not necessary forming a Voronoi diagram.

Novel Technique I: Simplex Lemma

- Let V be the simplex formed by $\{o_1, \cdots, o_j\}$.
- Let o be the mean of P, and δ^2 be the variance (i.e., $\delta^2 = \frac{1}{|P|}\sum_{p \in P} ||p - o||^2$).



Simplex Lemma: For any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V s.t. at least one grid point τ satisfies

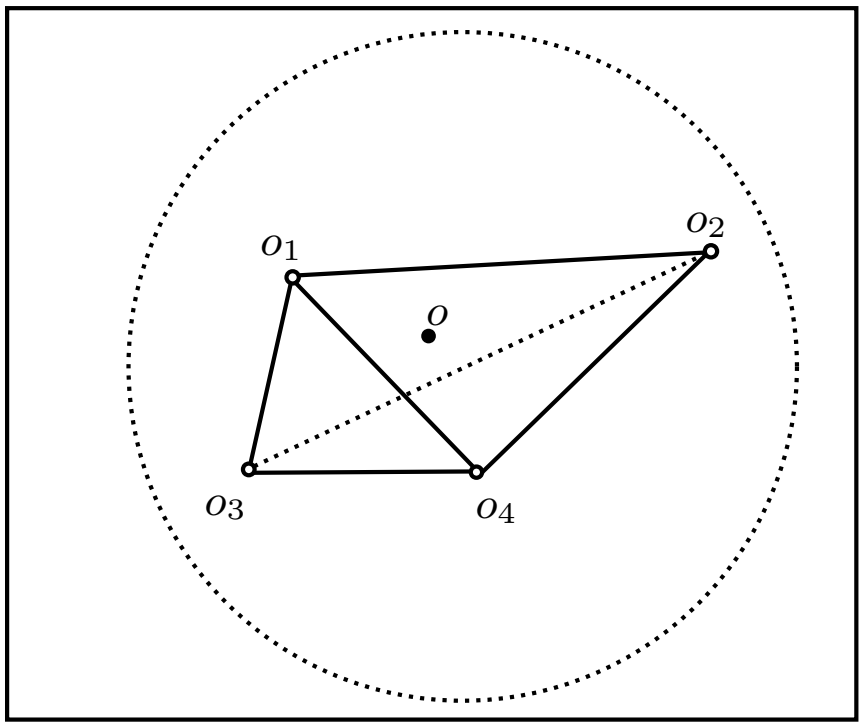
$$||\tau - o|| \leq \sqrt{\epsilon}\delta.$$

Note: The grid size is independent of the dimensionality d, and thus can be used in high dimensions.

Novel Technique I: Simplex Lemma

Proof (sketch)

1. o lies inside V.
2. Consider two cases: (1) Every P_l contains a large enough fraction of P (i.e., $\geq \frac{\epsilon}{4j}$), and (2) otherwise.
   - For case (1), the whole V is bounded by a $(j-1)$-dimensional ball with radius $4\sqrt{j/\epsilon}\delta$.
   - For case (2), it can be reduced to a case with a smaller j following an induction argument.

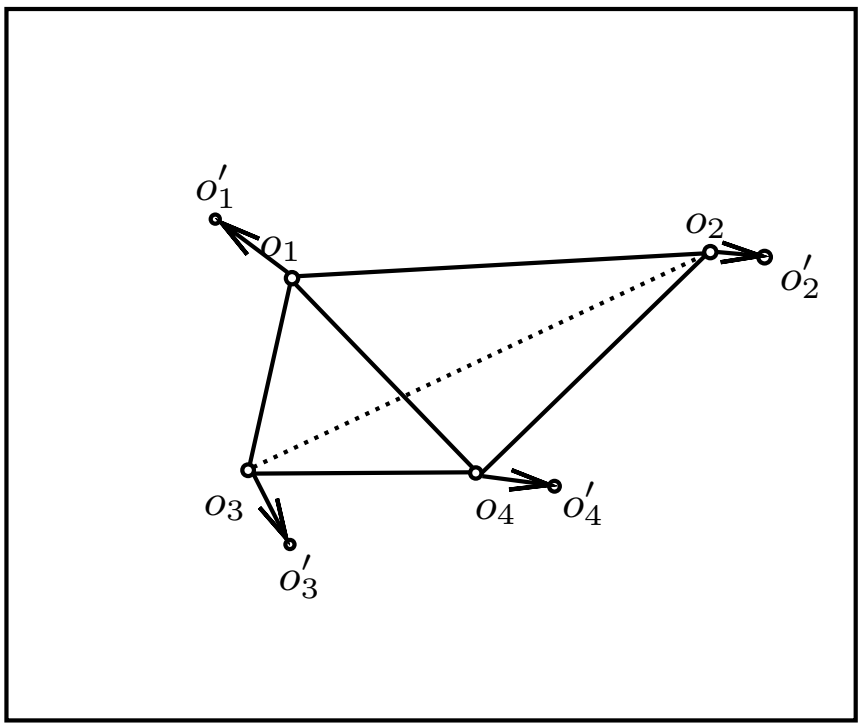


Novel Technique I: Simplex Lemma

A more general result:

- The exact position of o_l for each P_l is unknown.
- Instead, only an approximate position o'_l is known for each o_l, s.t.

$$||o'_l - o_l|| \leq L.$$



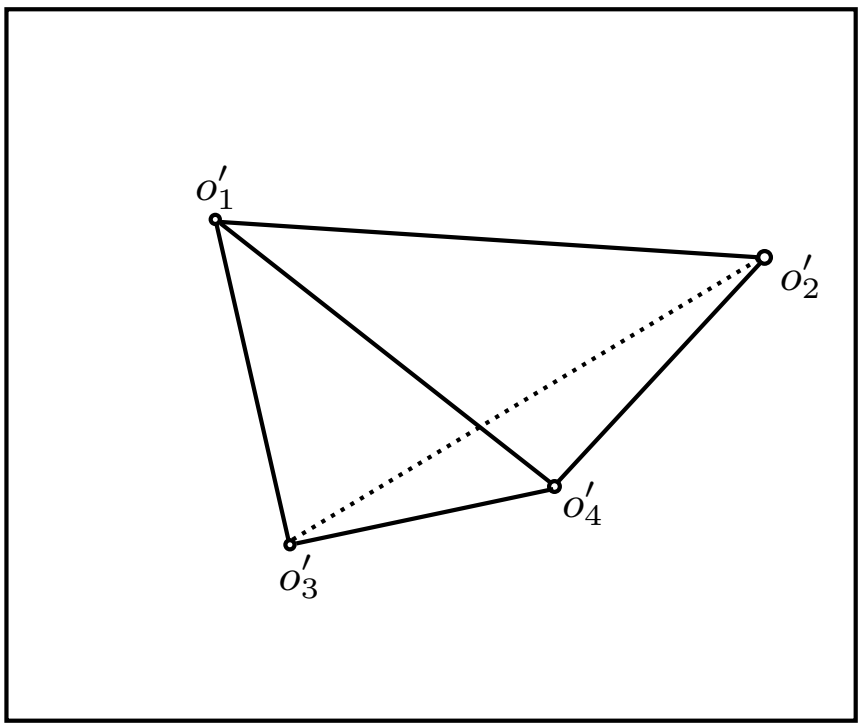
Novel Technique I: Simplex Lemma

A more general result:

- The exact position of o_l for each P_l is unknown.
- Instead, only an approximate position o'_l is known for each o_l, s.t.

$$||o'_l - o_l|| \leq L.$$

- Let V' be the simplex of $\{o'_1, \cdots, o'_j\}$.



Novel Technique I: Simplex Lemma

A more general result:

- Then, for any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V' such that at least one grid point τ satisfies the inequality

$$||\tau - o|| \leq \sqrt{\epsilon}\delta + (1 + \epsilon)L.$$

- **Note:** The right hand side contains two parts; one is from the variance, and the other is from the upper bound of $||o_l - o'_l||$.

Novel Technique I: Simplex Lemma

A more general result:

- Then, for any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V' such that at least one grid point τ satisfies the inequality

$$||\tau - o|| \leq \boxed{\sqrt{\epsilon}\delta} + (1+\epsilon)L.$$

- **Note:** The right hand side contains two parts; one is from the variance, and the other is from the upper bound of $||o_l - o'_l||$.

Novel Technique I: Simplex Lemma

A more general result:

- Then, for any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V' such that at least one grid point τ satisfies the inequality

$$||\tau - o|| \leq \sqrt{\epsilon}\delta + \boxed{(1+\epsilon)L}.$$

- **Note:** The right hand side contains two parts; one is from the variance, and the other is from the upper bound of $||o_l - o'_l||$.

Novel Technique I: Simplex Lemma

A more general result:

- Then, for any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V' such that at least one grid point τ satisfies the inequality

$$||\tau - o|| \leq \sqrt{\epsilon}\delta + (1 + \epsilon)L.$$

- **Note:** The right hand side contains two parts; one is from the variance, and the other is from the upper bound of $||o_l - o'_l||$.

Novel Technique I: Simplex Lemma

A more general result:

- Then, for any $0 < \epsilon \leq 1$, it is possible to construct a grid of size $O((8j/\epsilon)^j)$ inside V' such that at least one grid point τ satisfies the inequality

$$||\tau - o|| \leq \sqrt{\epsilon}\delta + (1 + \epsilon)L.$$

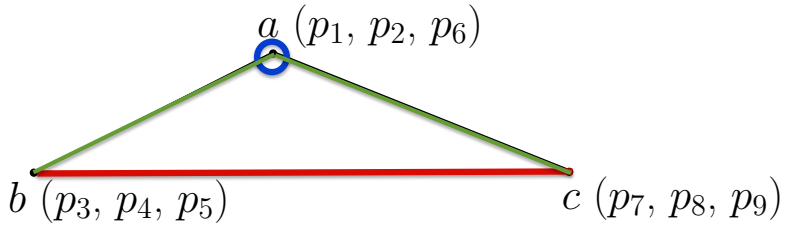
- **Note:** The right side contains two parts; one is from the variance, and the other is from the upper bound of $||o_l - o'_l||$
$\Longrightarrow$ A key for proving the correctness.

Novel Technique I: Simplex Lemma

Extension to median point:

- **Bad news:** A median point could lie outside the simplex.

- $P_1 = \{p_i \mid 1 \leq i \leq 5\}$,
  $P_2 = \{p_i \mid 6 \leq i \leq 9\}$.



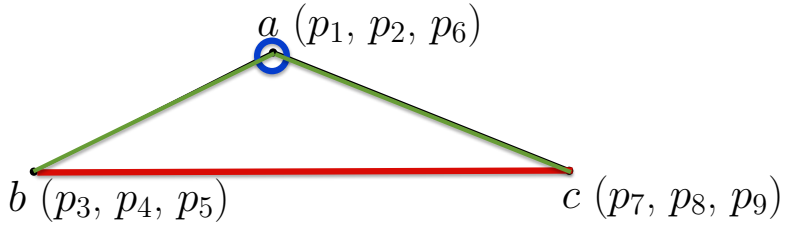
- **Good news:** Although a median point may lie outside the simplex, it is always in the surrounding region of the simplex, and thus a similar result can be obtained by building a grid in the surrounding region.

Novel Technique I: Simplex Lemma

Extension to median point:

- **Bad news:** A median point could lie outside the simplex.

- $P_1 = \{p_i \mid 1 \leq i \leq 5\}$,
  $P_2 = \{p_i \mid 6 \leq i \leq 9\}$.



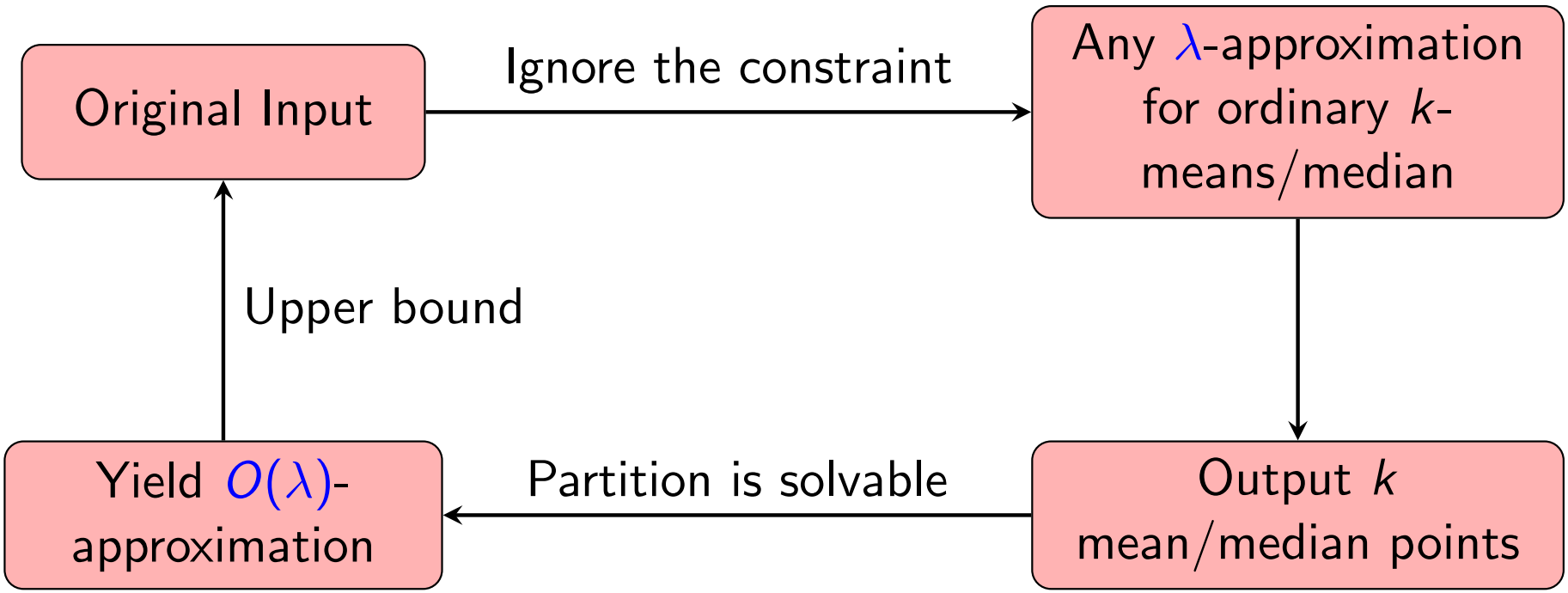
- **Good news:** Although a median point may lie outside the simplex, it is always in the surrounding region of the simplex, and thus a similar result can be obtained by building a grid in the surrounding region.

Novel Technique II: A Unified Constant Approximation

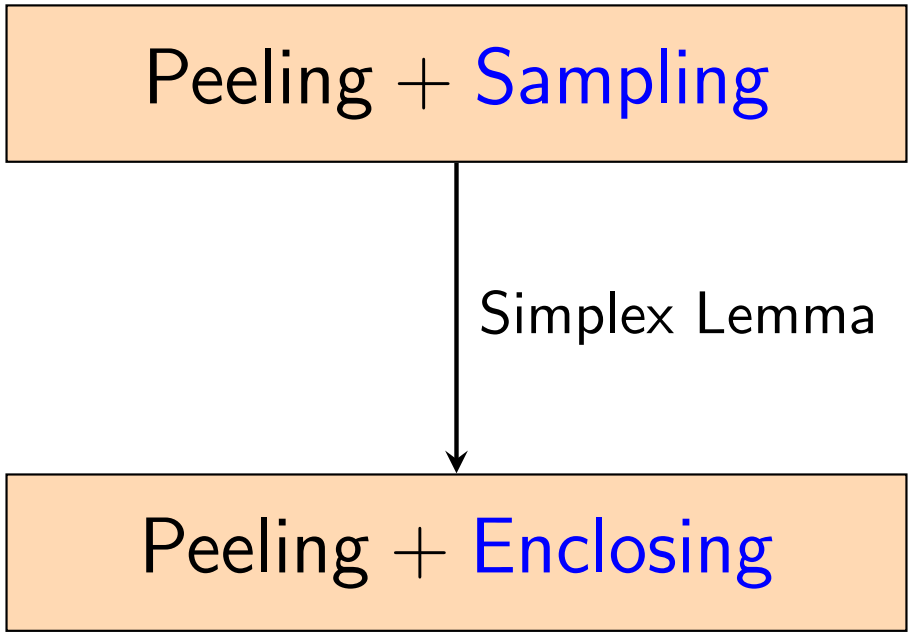
A natural question: For the same set of input points, how large is the difference between the solutions to the constrained (*i.e.*, k-CMeans/CMedian) and unconstrained (*i.e.*, the ordinary k-means/median) clusterings?

Such information can be used
(1) For determining the radii of peeling spheres; (2) Of interest in its own right.

Novel Technique II: A Unified Constant Approximation



A Unified Algorithm



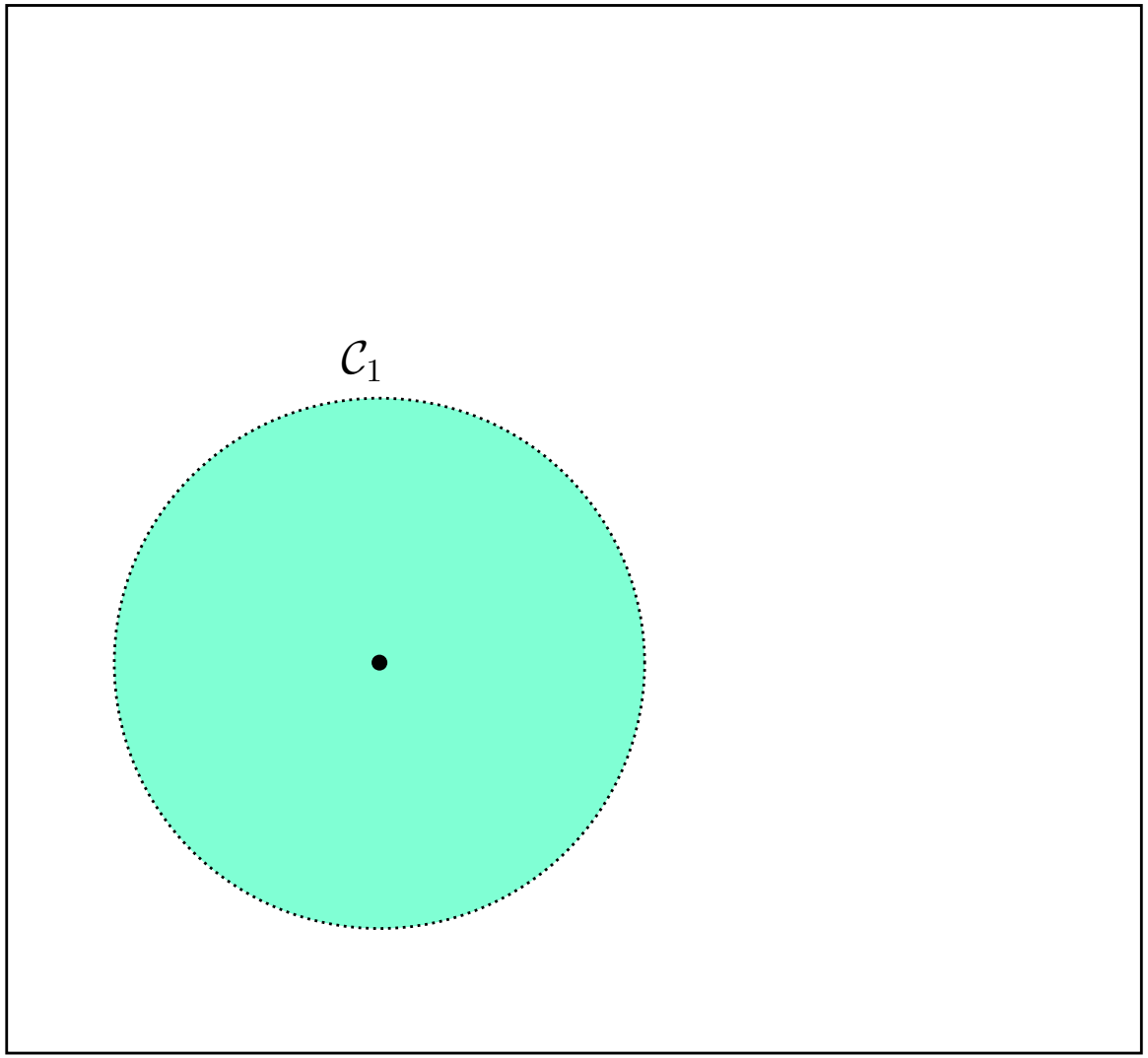
A Unified Algorithm

Main Steps:

1. Run the constant approximation algorithm to obtain an upper bound (assuming the partition step is solvable)
2. Iteratively find the k mean/median points; for each $0 \leq j \leq k-1$:
   1. Draw j peeling spheres centered at the j previously obtained mean/median points, each radus is based on the upper bound
   2. Find an extra point π via random sampling after the peeling
   3. Build a j-dimensional simplex
   4. Find the $(j+1)$-th mean/median point via Simplex Lemma
3. Output a set of k-tuple candidates for the k mean/median points

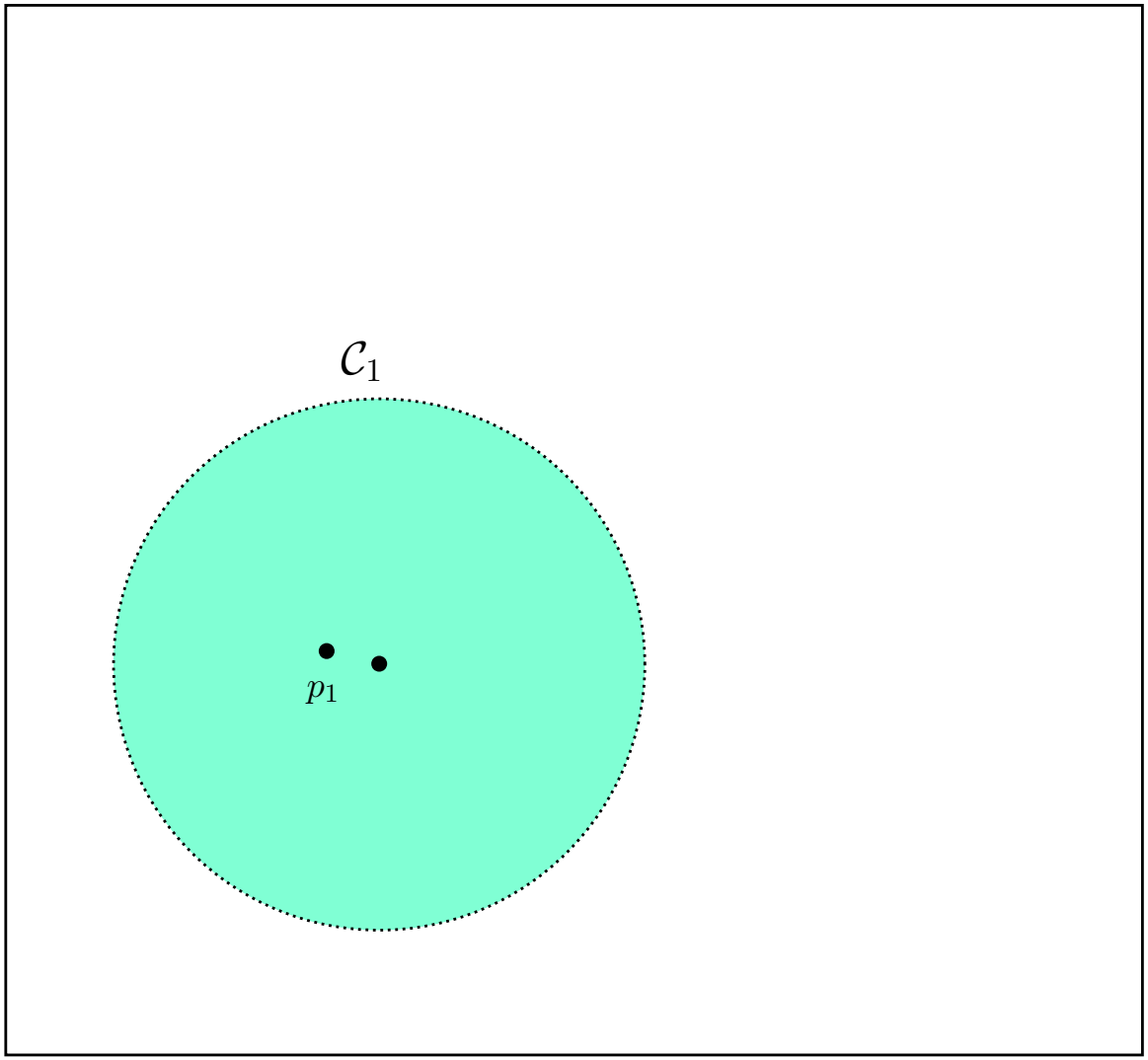
A Unified Algorithm

An illustration for the case of $k = 4$.



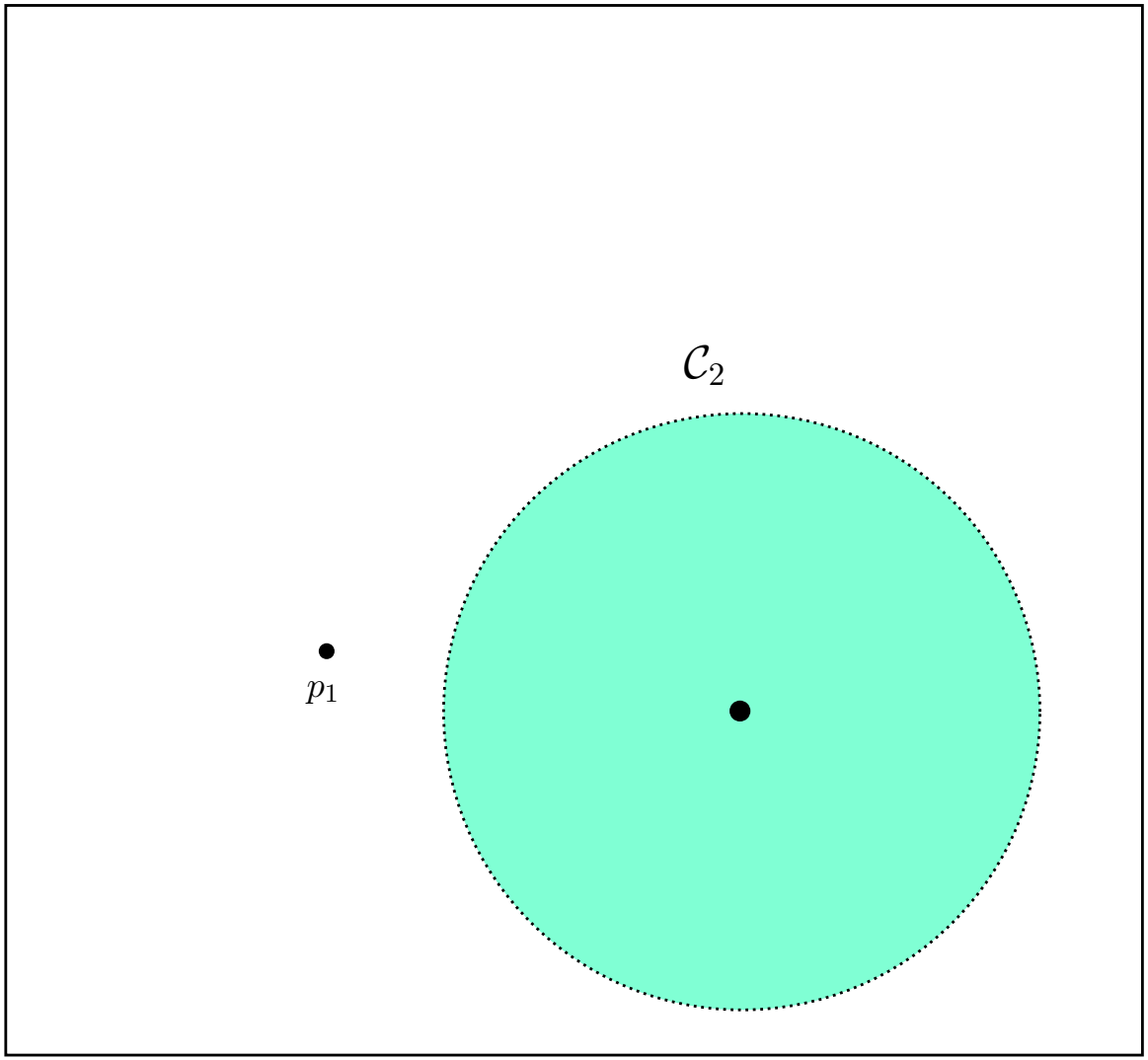
A Unified Algorithm

An illustration for the case of $k = 4$.



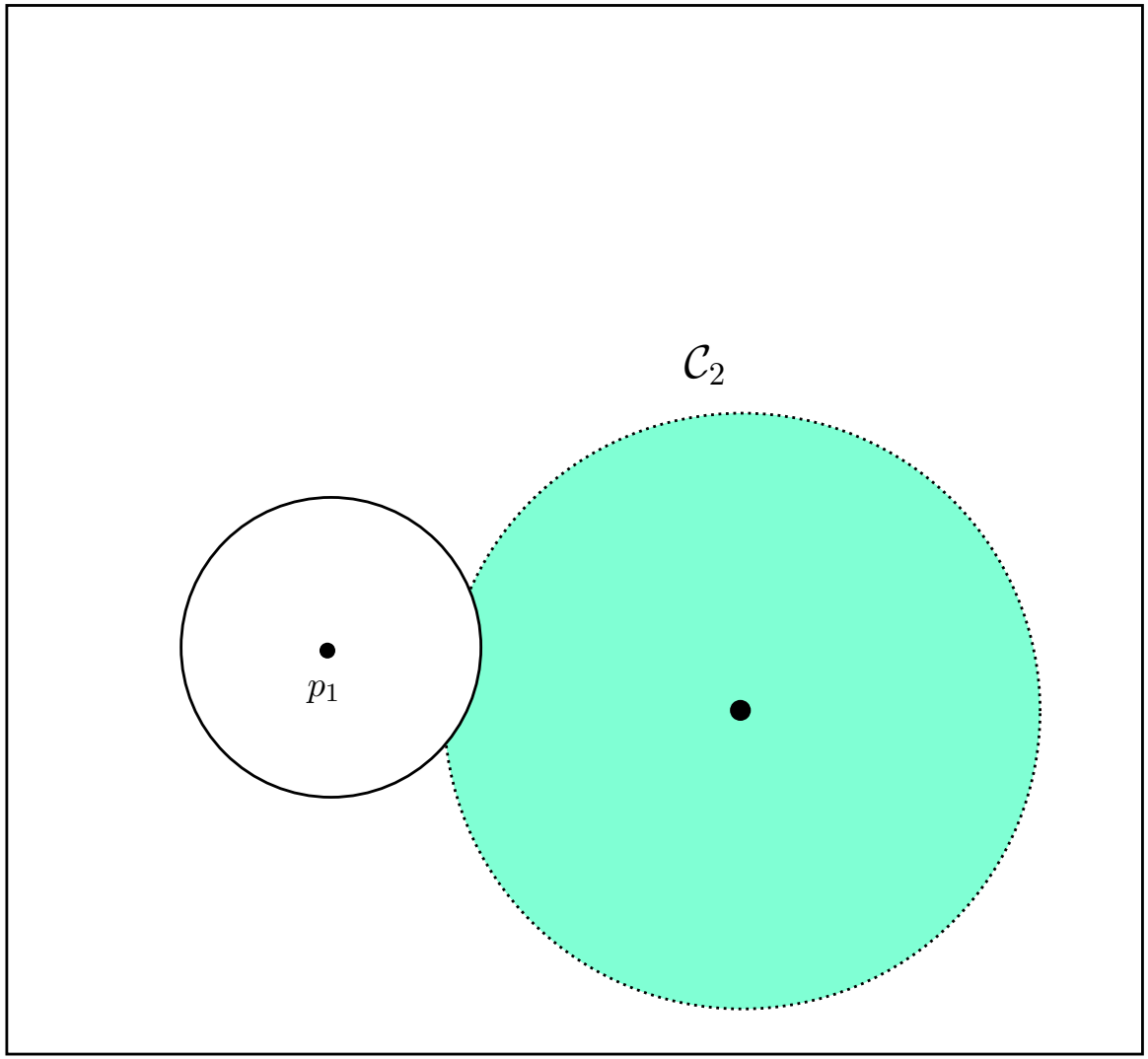
A Unified Algorithm

An illustration for the case of $k = 4$.



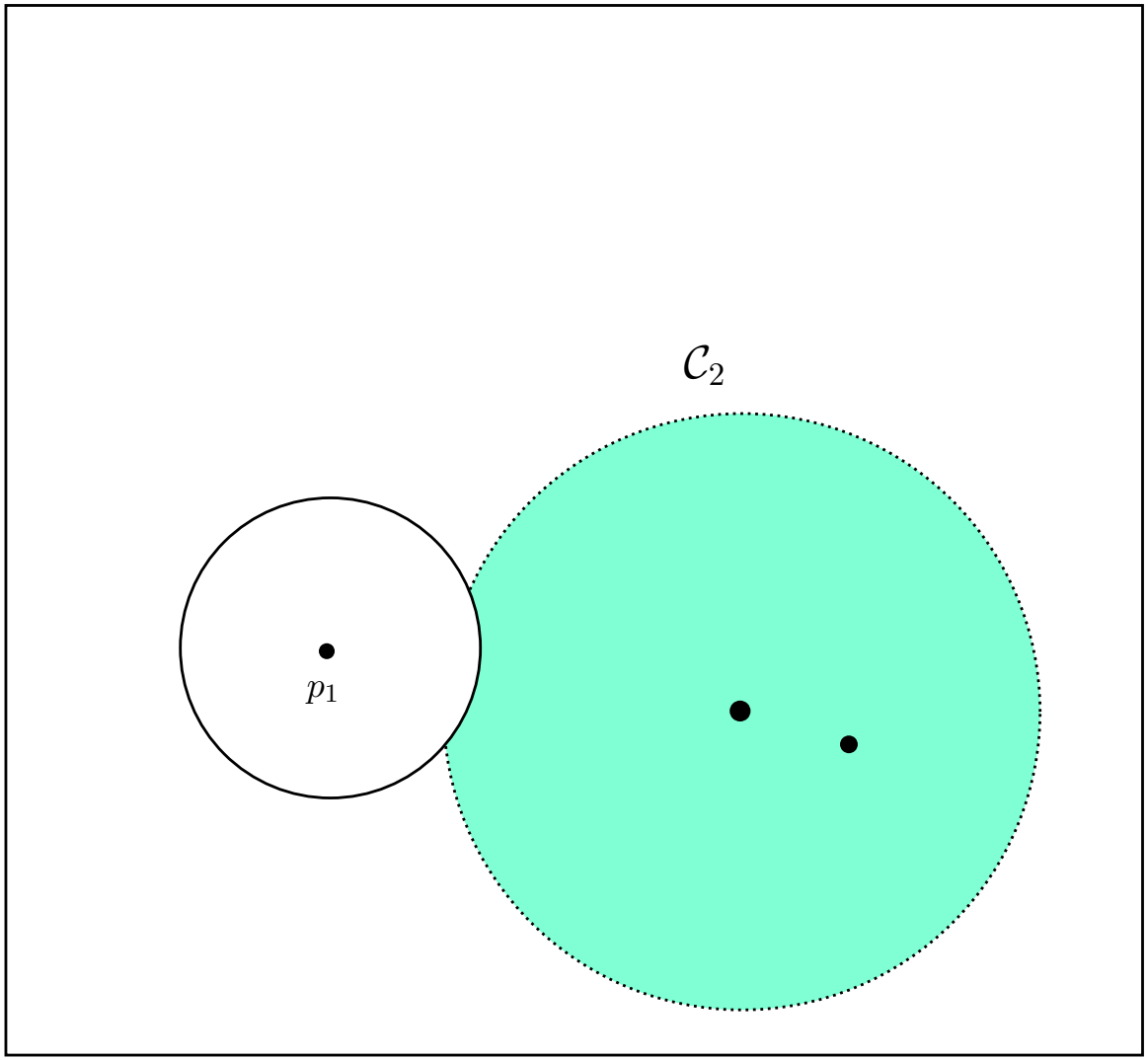
A Unified Algorithm

An illustration for the case of $k = 4$.



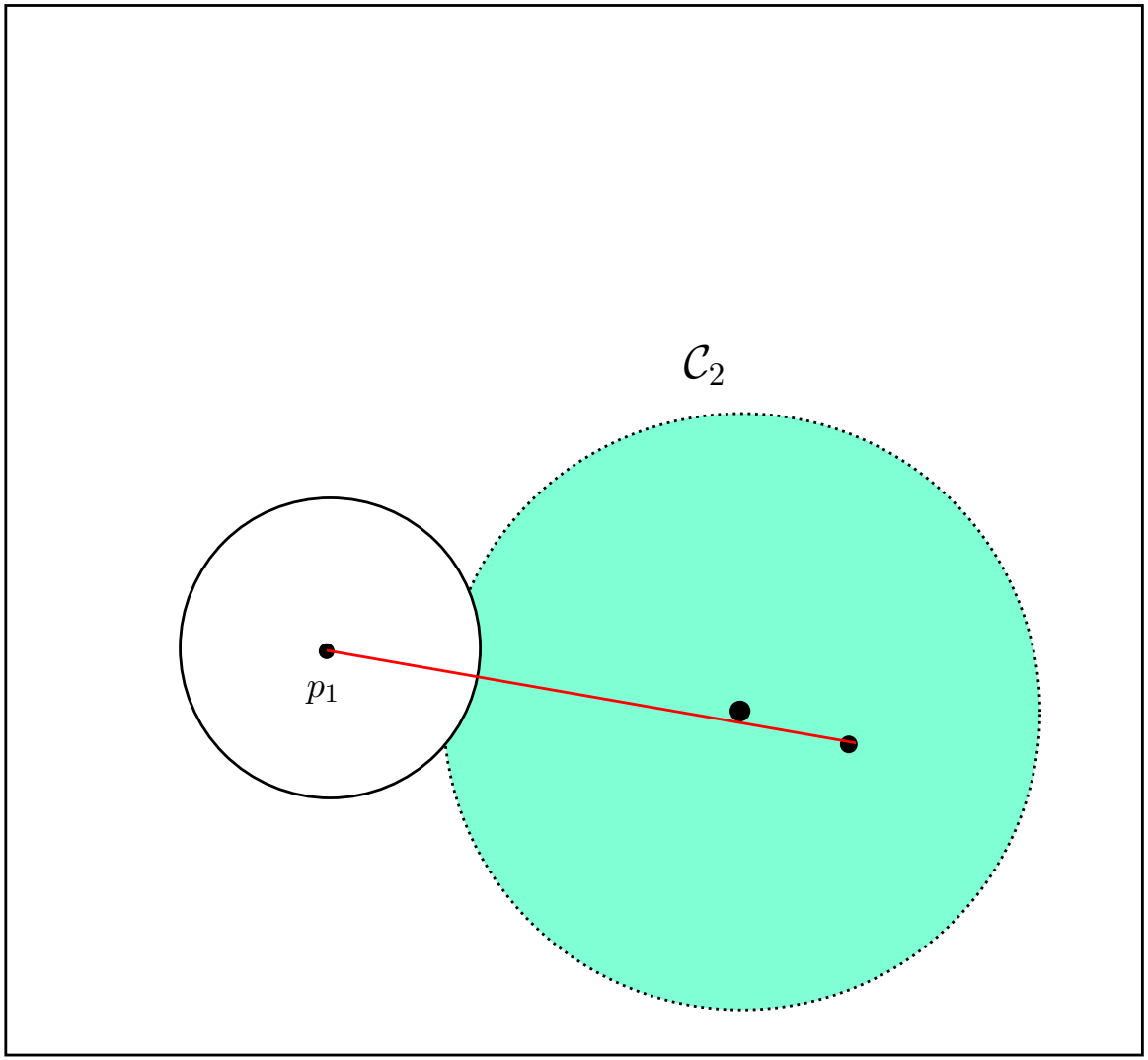
A Unified Algorithm

An illustration for the case of $k = 4$.



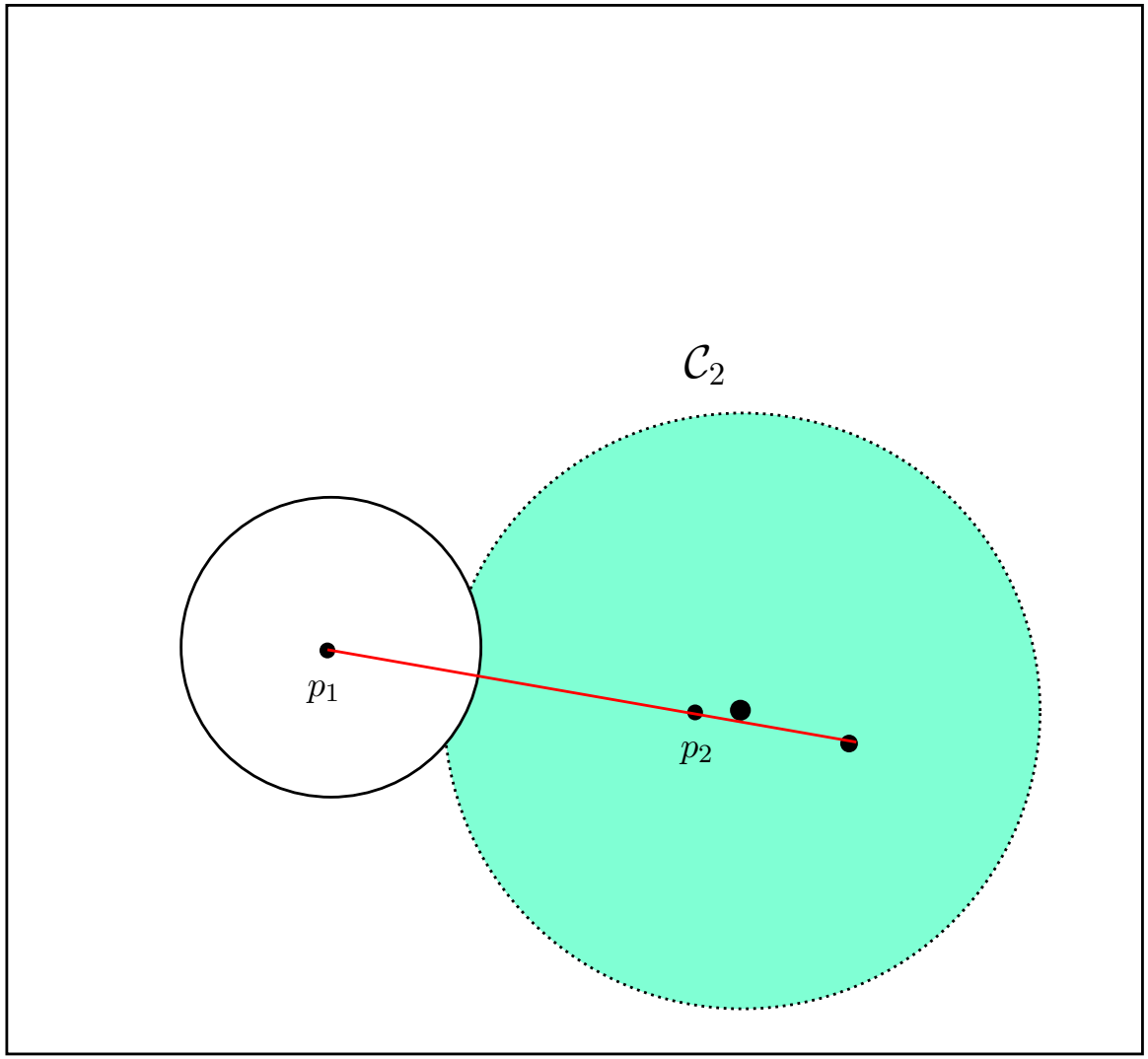
A Unified Algorithm

An illustration for the case of $k = 4$.



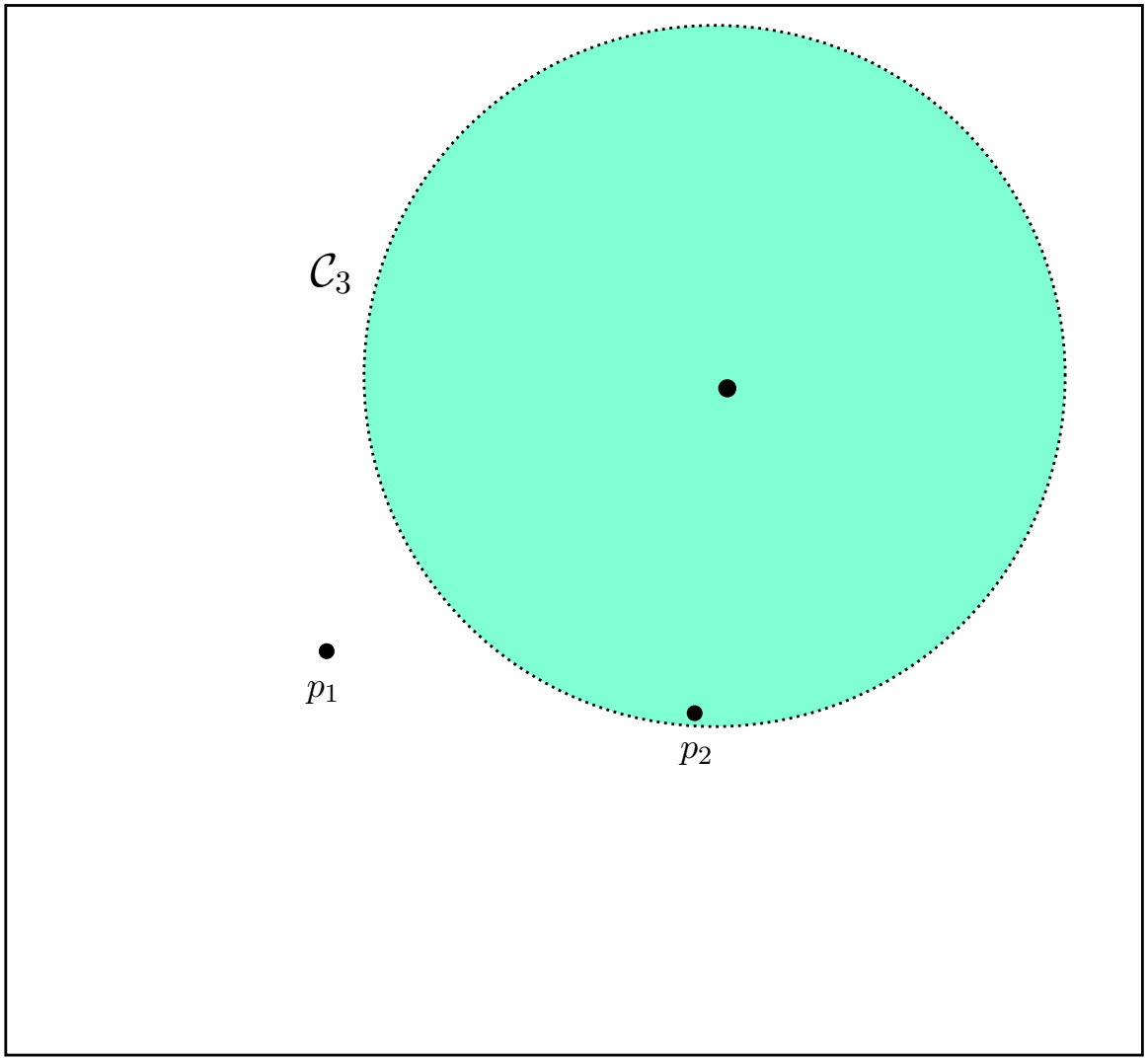
A Unified Algorithm

An illustration for the case of $k = 4$.



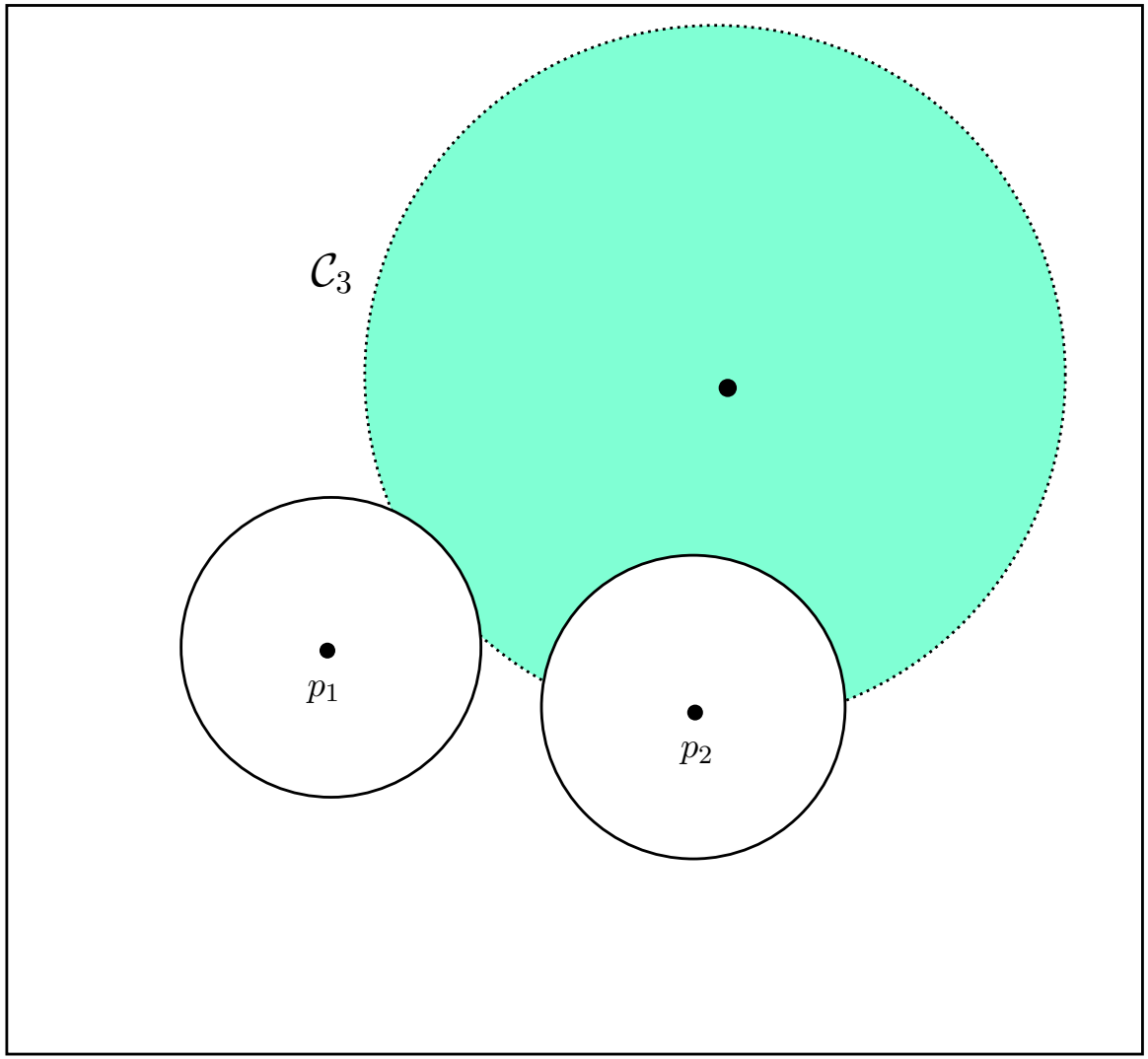
A Unified Algorithm

An illustration for the case of $k = 4$.



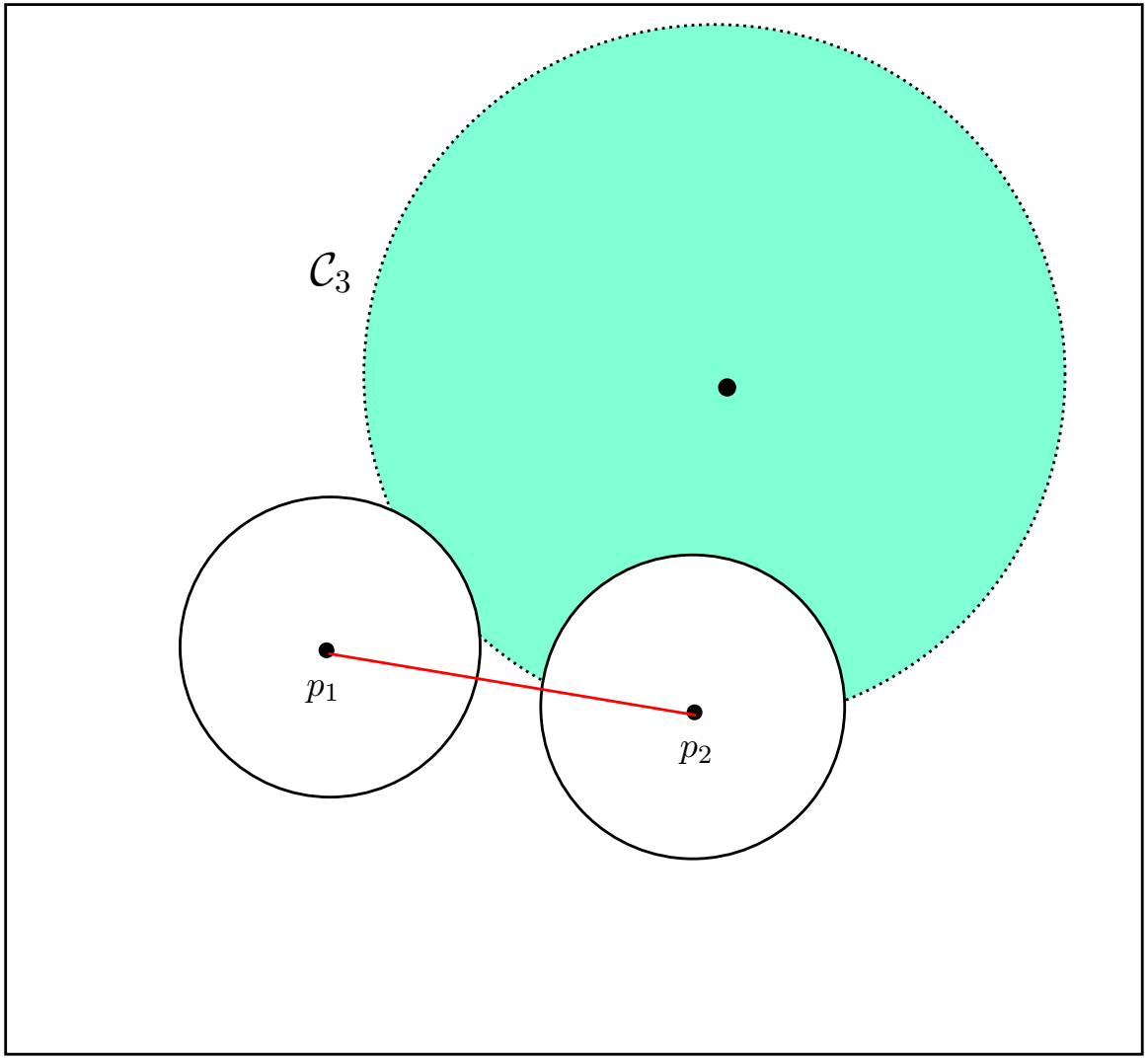
A Unified Algorithm

An illustration for the case of $k = 4$.



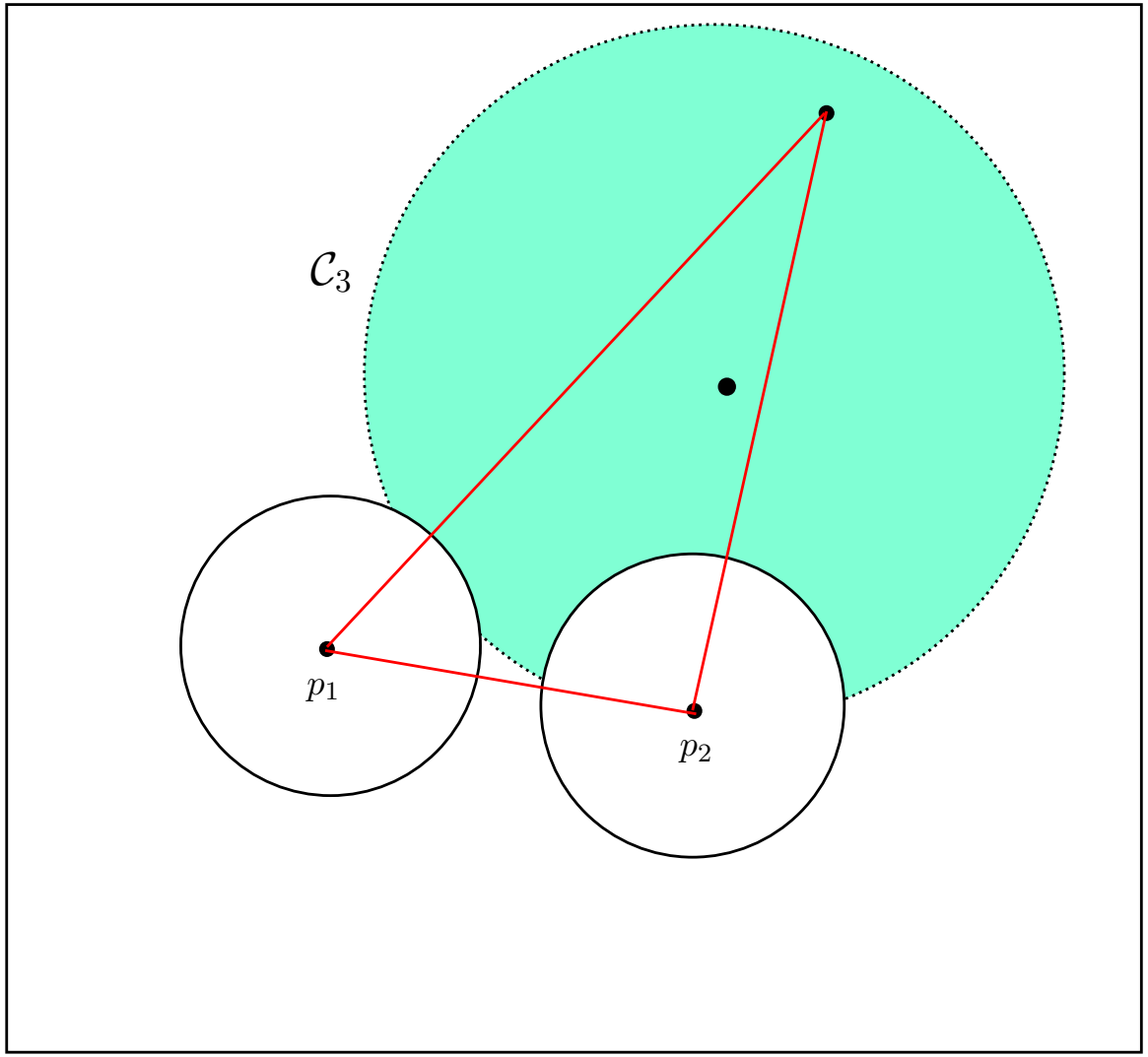
A Unified Algorithm

An illustration for the case of $k = 4$.



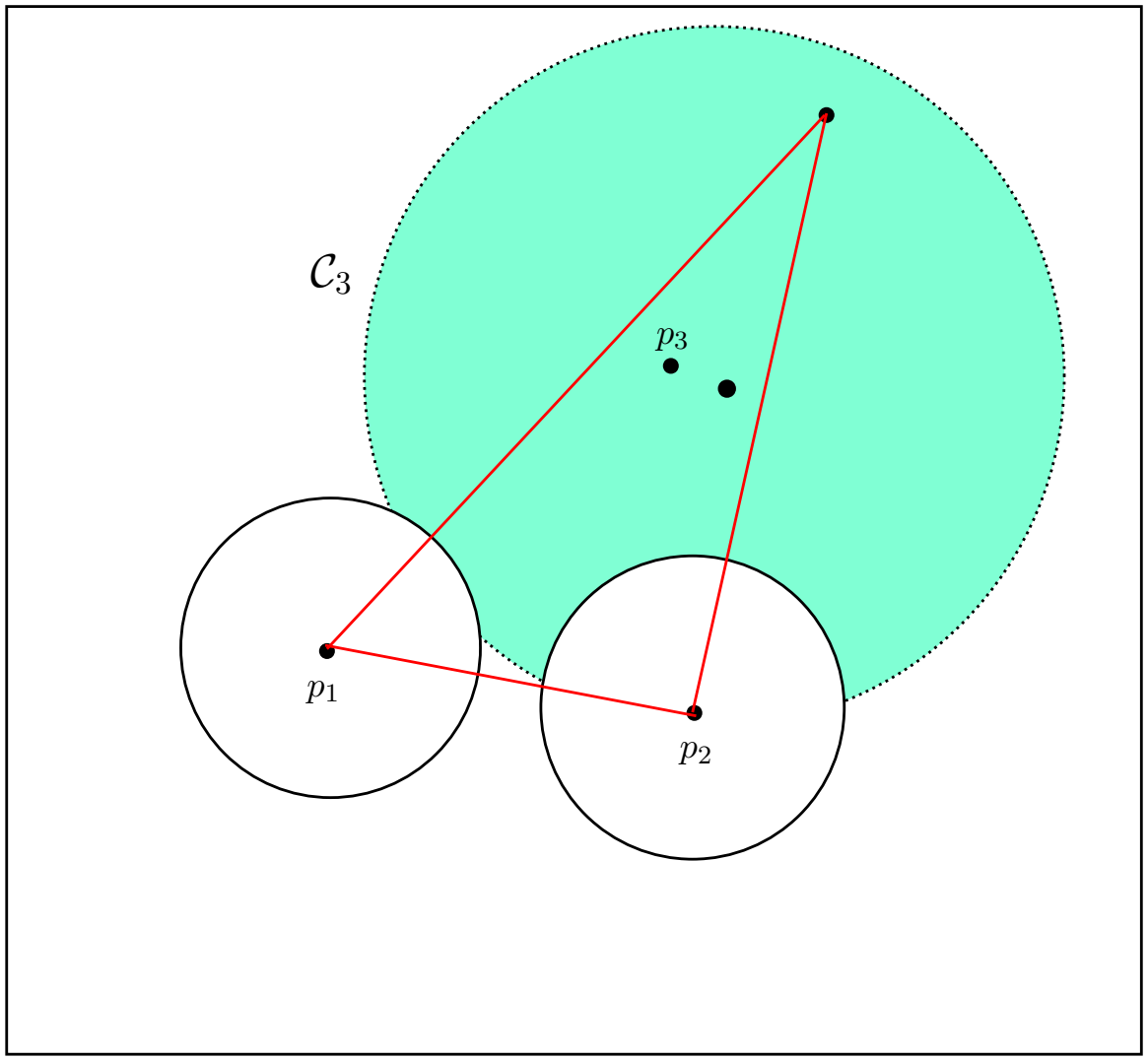
A Unified Algorithm

An illustration for the case of $k = 4$.



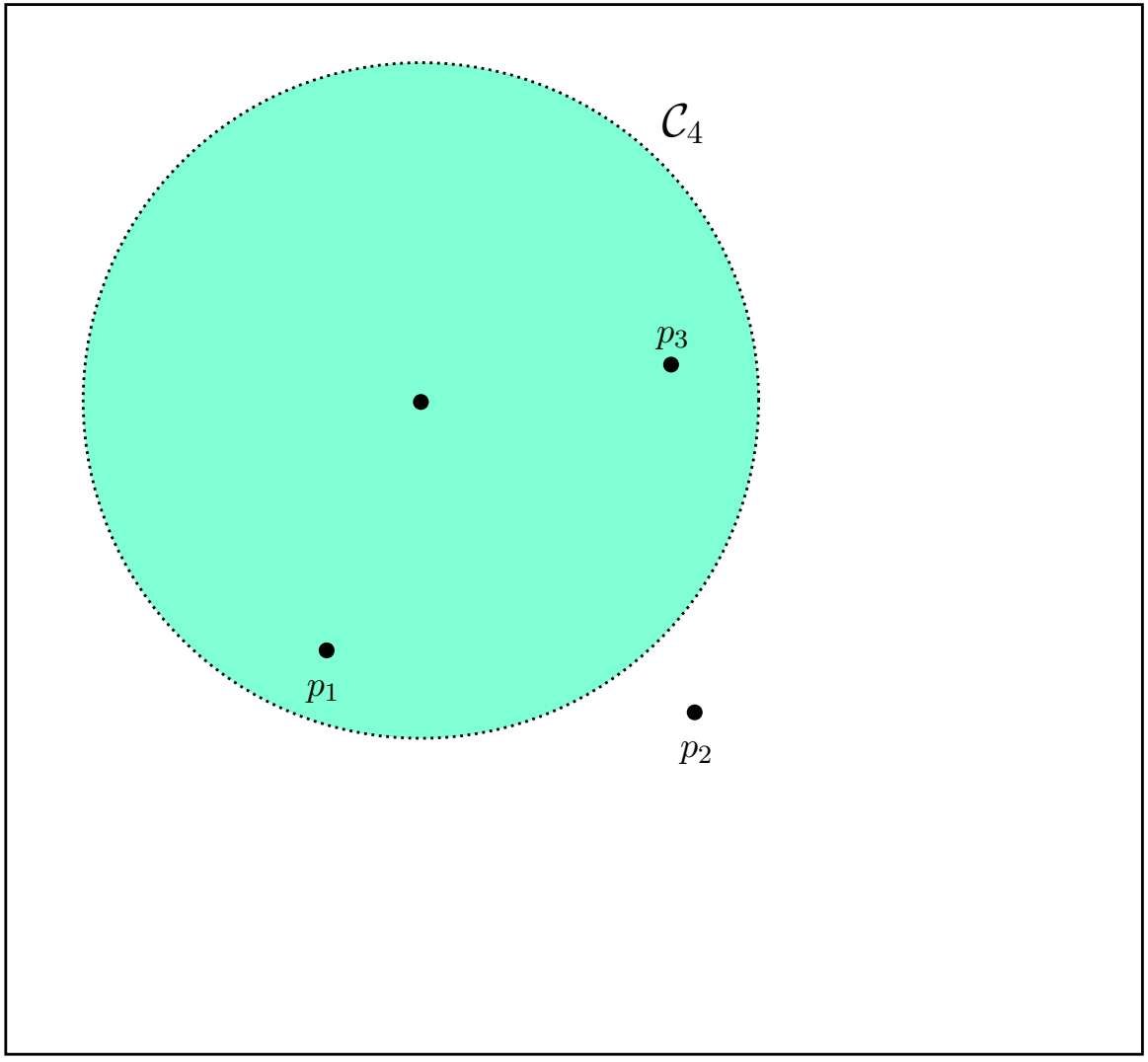
A Unified Algorithm

An illustration for the case of $k = 4$.



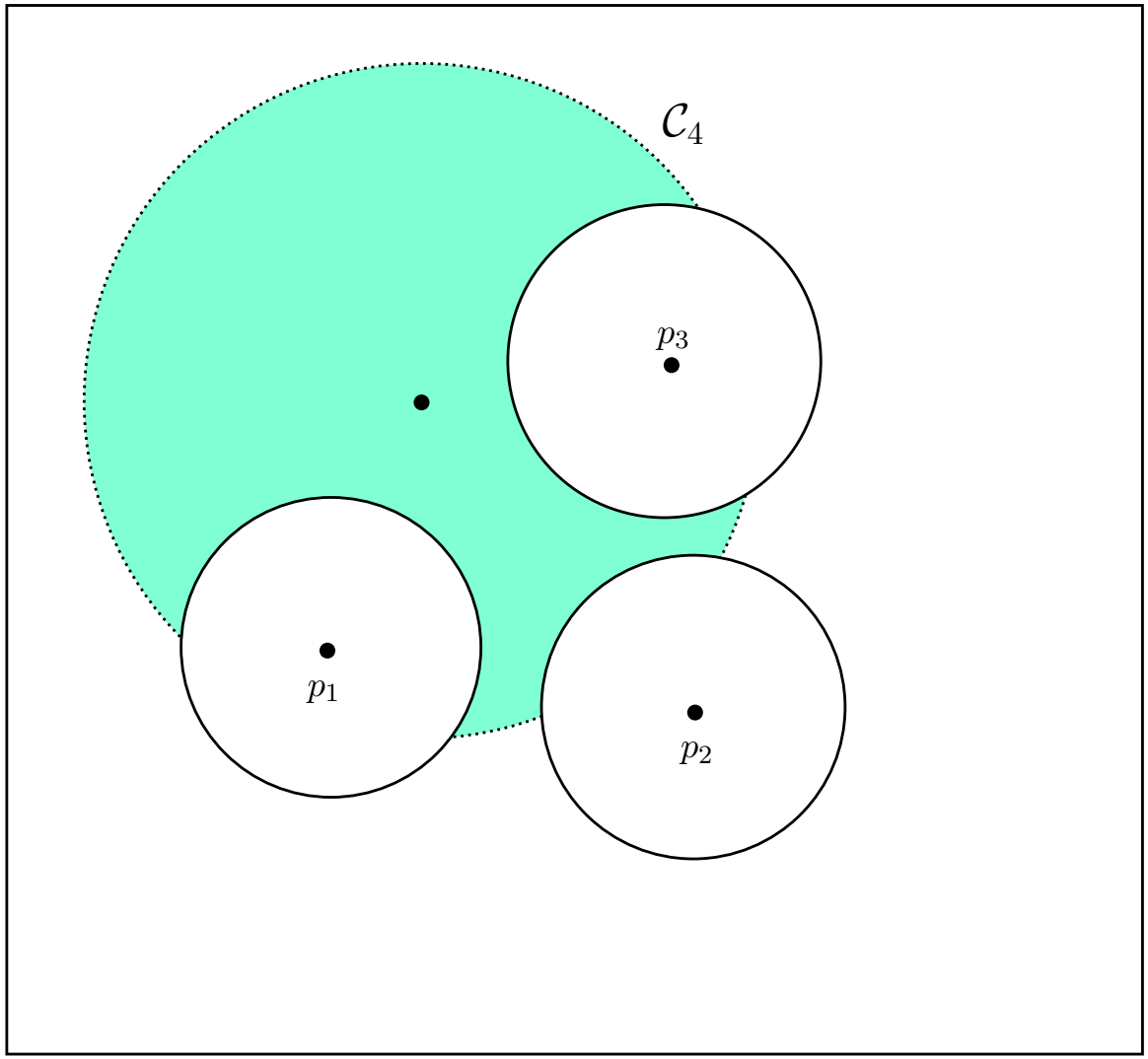
A Unified Algorithm

An illustration for the case of $k = 4$.



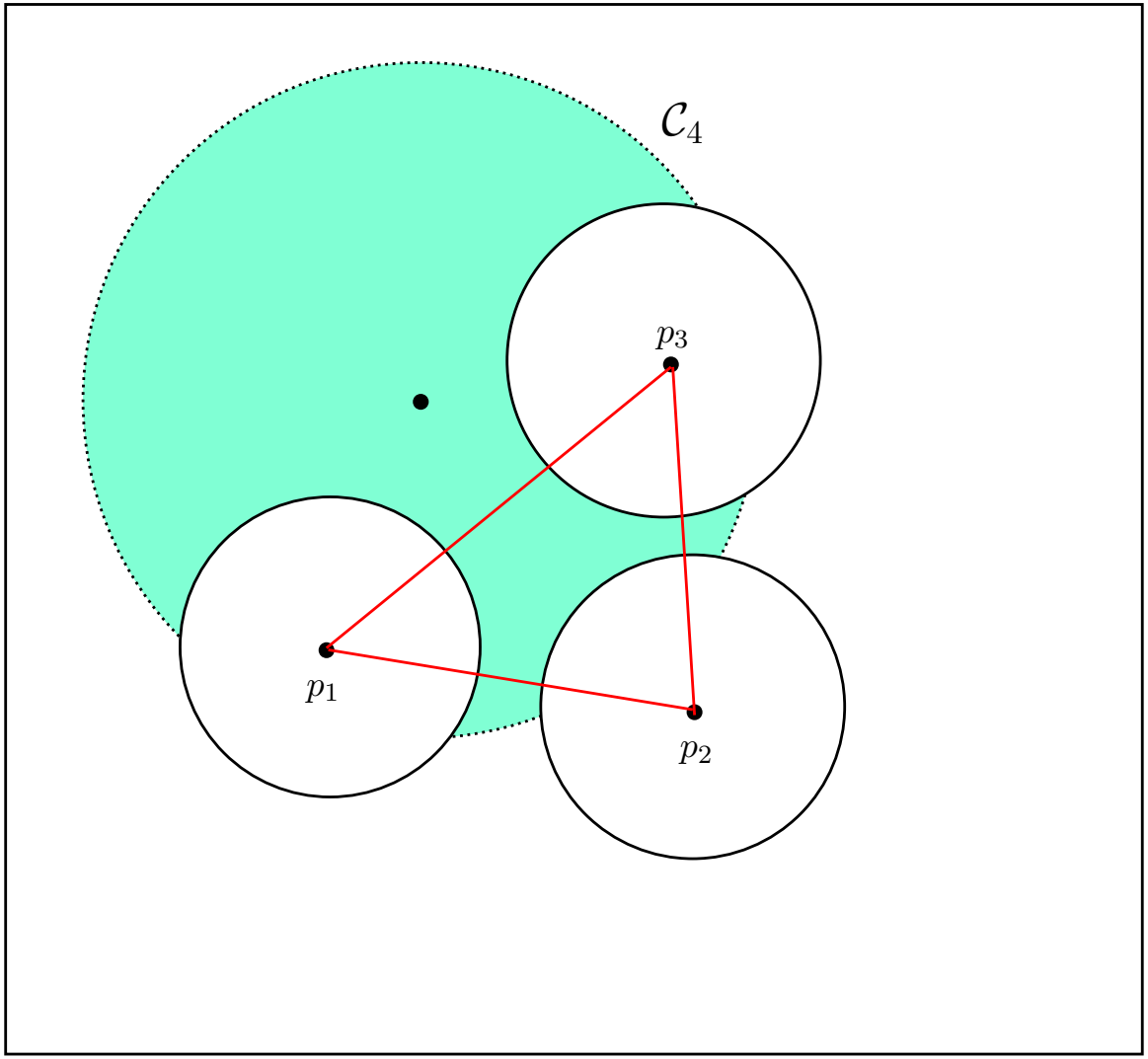
A Unified Algorithm

An illustration for the case of $k = 4$.



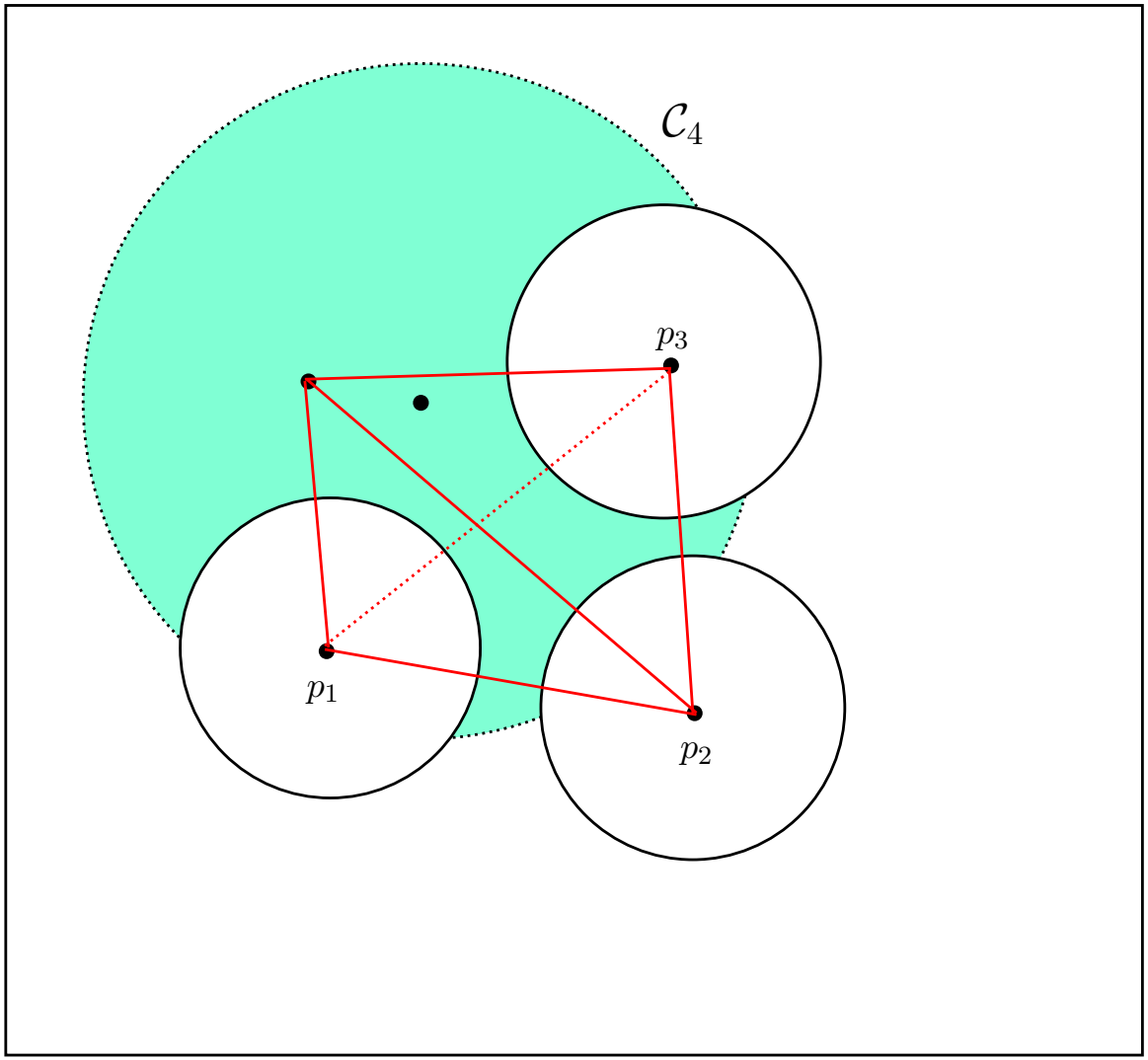
A Unified Algorithm

An illustration for the case of $k = 4$.



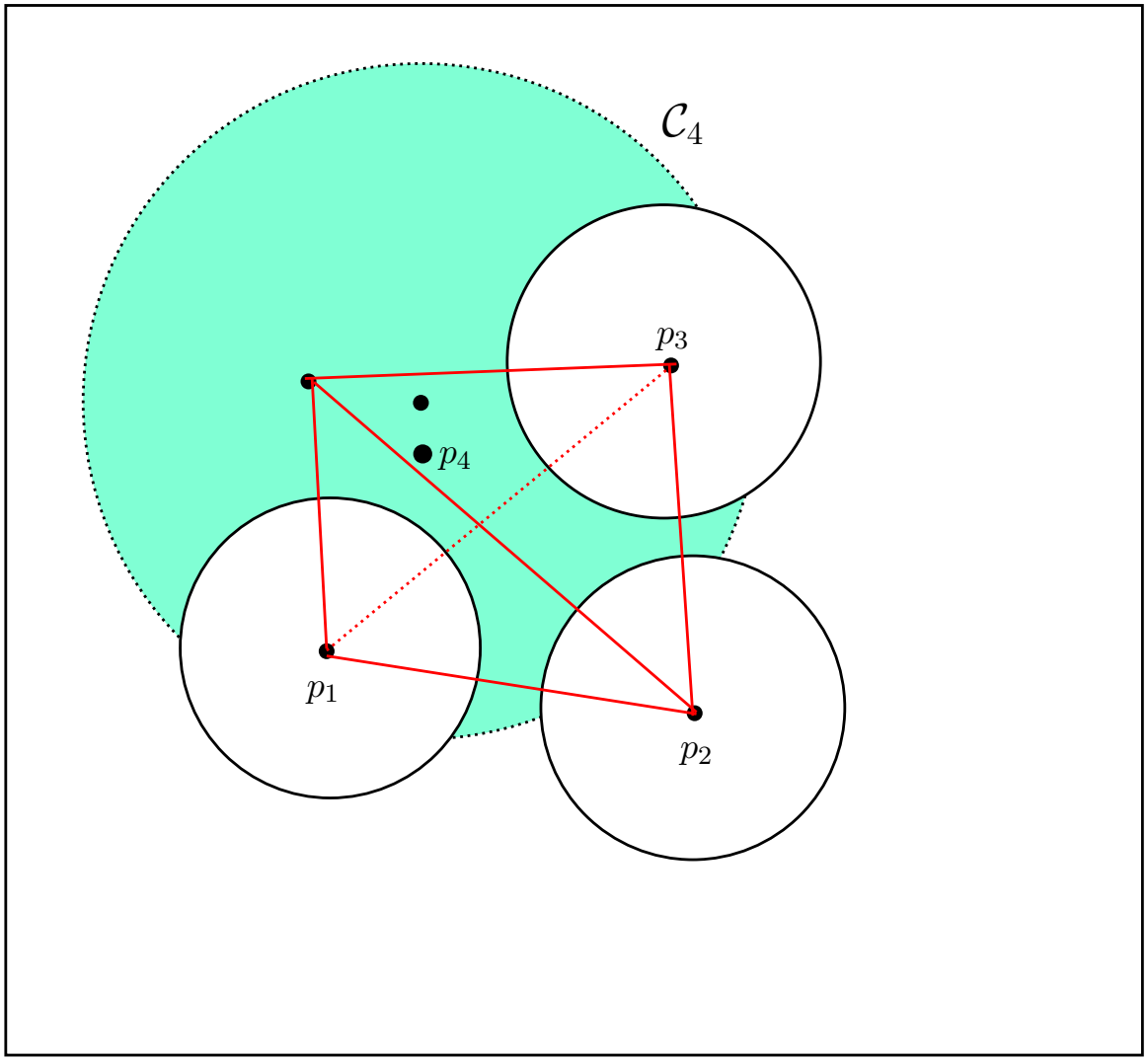
A Unified Algorithm

An illustration for the case of $k = 4$.



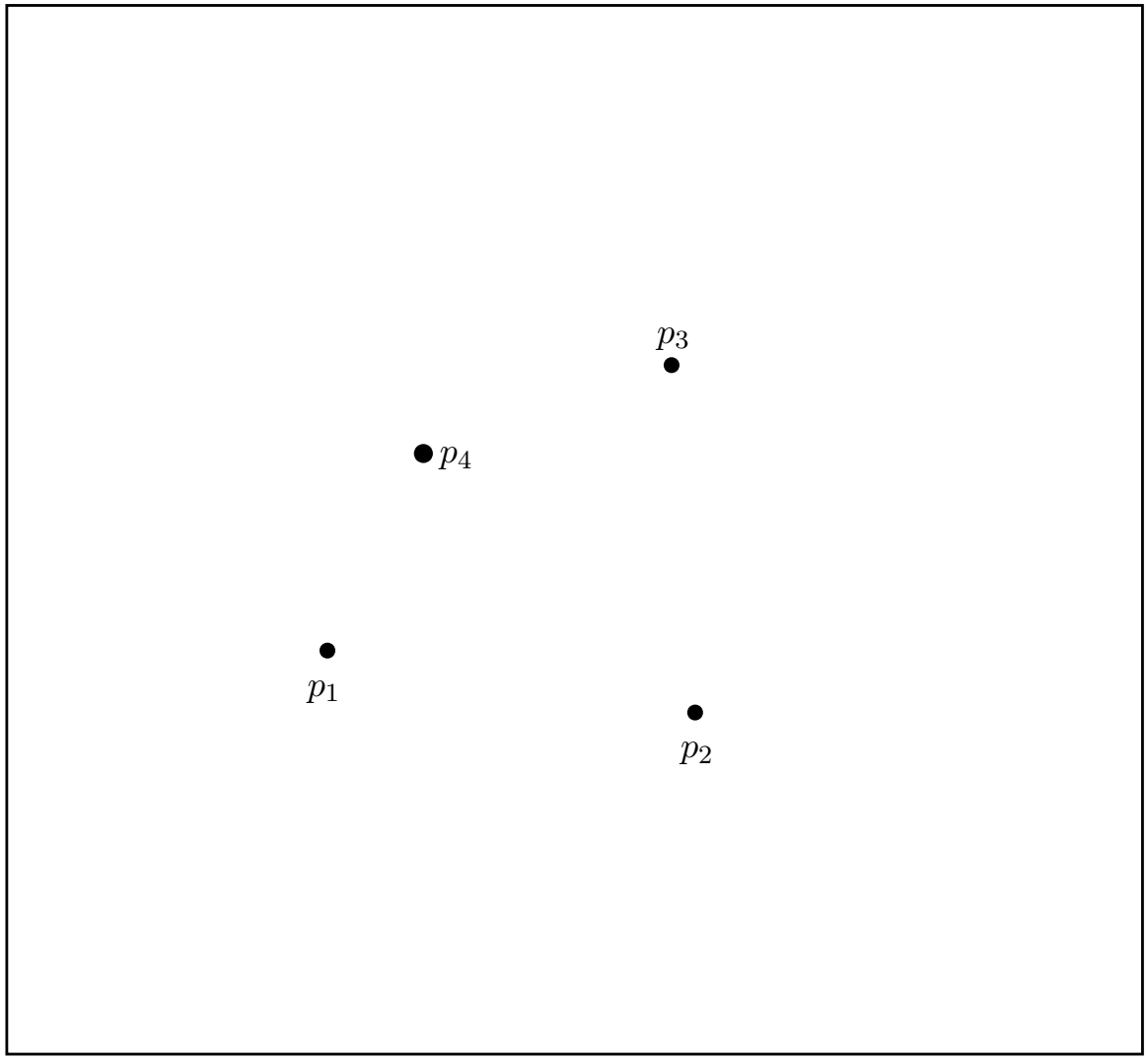
A Unified Algorithm

An illustration for the case of $k = 4$.



A Unified Algorithm

An illustration for the case of $k = 4$.



A Unified Algorithm

A key lemma for the correctness:

$$||p_i - m_i|| \leq \boxed{\epsilon\delta_i + O(\sqrt{\epsilon})\delta_{opt}},$$

where m_i is the optimal mean/median, δ_i is the cost of the i-th cluster, and δ_{opt} is the overall cost.

A Unified Algorithm

Theorem: Let P be a set of n points in $\mathbb{R}^d$ and $\mathbb{C}$ be some constraint. There exists an algorithm outputting $O(\log^{k+1} n)$ k-tuple candidates for the mean/median points in $O(n(\log^{k+2} n)d)$ time, and with probability $1 - \frac{1}{n}$ at least one candidate yields a $(1+\epsilon)$-approximation of k-CMeans/CMedian, if the corresponding partition step can be solved.

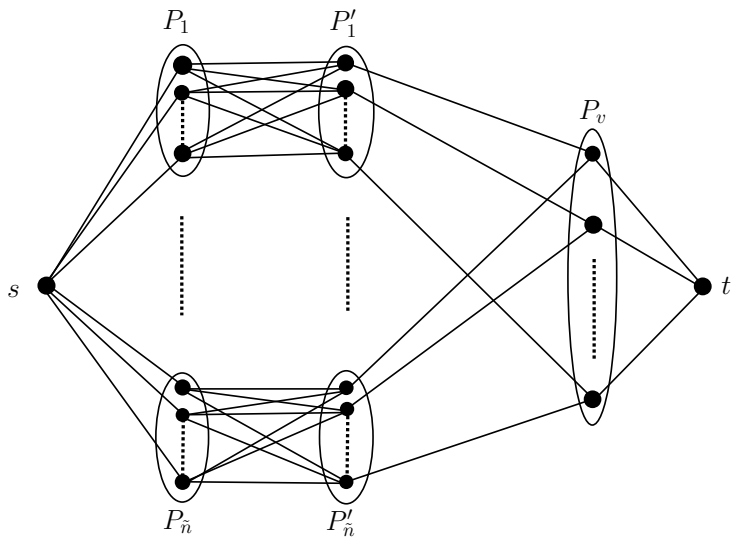
Selection Algorithms

1. For each k-tuple candidate, solve the partition step, *i.e.*, generate the k clusters satisfying the constraint $\mathbb{C}$.

2. Select the one with the smallest objective value.

Selection Algorithms

1. For each k-tuple candidate, solve the partition step, *i.e.*, generate the k clusters satisfying the constraint $\mathbb{C}$.
   - Some problems are easy, *e.g.*, chromatic clustering, fault tolerant clustering, probabilistic clustering.
   - Some problems are harder, *e.g.*, r-gather clustering, diversity clustering, semi-supervised clustering, involving min-cost max-flow techniques (please refer to our paper for details).



2. Select the one with the smallest objective value.

Summary

1. A unified algorithm outputting k-tuple candidates for k-CMeans/CMedian
   - Simplex Lemma
   - A unified constant approximation
   - Peeling + Enclosing
2. Selection algorithm for each individual constraint
3. Open problems
   - Is it possible to improve the time complexity of the unified framework to be linear, such as using *core-set*?
   - Can the unified framework be applied to other types of k-CMeans/CMedian?
   - Can it be extended to some non-Euclidean space?
   - How to improve the selection algorithms?

Thank You!

Any Question?