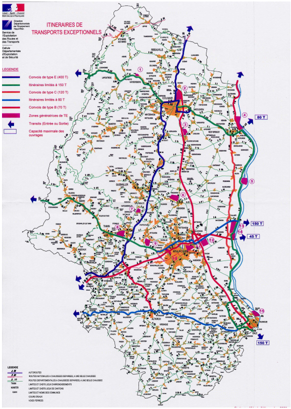
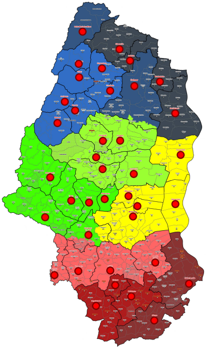
The Power of Local Search for Clustering in “Separable Instances”

Vincent Cohen-Addad Sorbonne Université & CNRS
Joint work with:
Philip N. Klein Brown University
Claire Mathieu Ecole normale supérieure & CNRS

What is Clustering?

Partition data points according to distances.

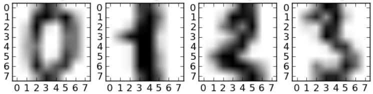




Group buildings to locate firestations

Underlying data: Road networks.

Partition data according to *similarity*.



Underlying data: Points in $\mathbb{R}^2$.

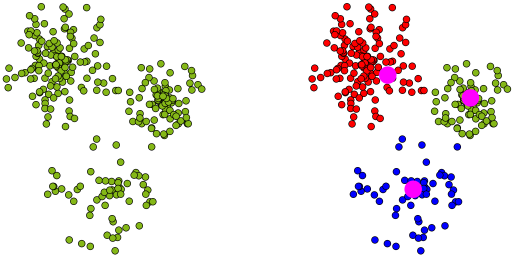
How to model clustering?

k-Clustering

Input: *data points* A in a metric space
Output: set C of k *centers* that minimizes

$$\sum_{a \in A} \min_{c \in C} \mathrm{d}(a, c)^p.$$

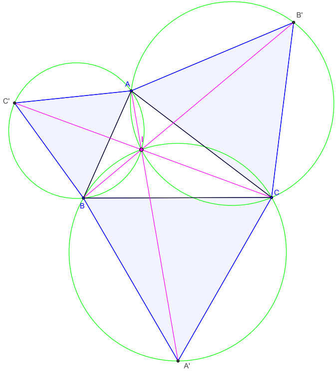
k-median is when $p = 1$, k-means is when $p = 2$.



The 1-median problem dates back to Fermat (1636).

Given three points $a, b, c \in \mathbb{R}^2$, find a point d that minimizes

$$d(a,d) + d(b,d) + d(c,d).$$



If more than 3 points, it is hard to compute exactly!

Algorithms for Clustering: History

- k-median:

1964	Introduction of the Problem	[Hakimi]
1979	NP-Hardness	[Kariv and Hakimi]
2002	623-approx	[Charikar et al.]
2004	$3+\varepsilon$-approx	[Arya et al.]
2013	$1+\sqrt{3}\approx 2.732+\varepsilon$-approx	[Li and Svensson]
2015	(current best) $2.675+\varepsilon$	[Byrka et al.]

- k-means:

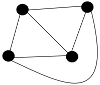
1967	Introduction of the Problem	[MacQueen]
2004	(current best) $16+\varepsilon$	[Kanungo et al.]

NP-Hard

To obtain better than $1+2/e\approx 1.735$ approx for k-median in polynomial time.

Focus on real-world:

- Road Networks
  planar graphs
- Machine learning and image compression
  low-dimensional Euclidean space



Previous Work on Restricted Metrics

Planar graphs

Nothing Better than General Case

$\mathbb{R}^{O(1)}$

k-median	$(1+\varepsilon)$	[Arora et al. '98]
k-means	9	[Kanungo et al. '04]

Recent Results for $\mathbb{R}^{O(1)}$

[C.-A. and Mathieu, SoCG '15]

Local search achieves a $(1+\varepsilon)$-approximation using $(1+\varepsilon)k$ centers **for k-median**.

[Bandyapadhyay and Varadarajan, SoCG '16]

Local search achieves a $(1+\varepsilon)$-approximation using $(1+\varepsilon)k$ centers **for k-means**.

Main open problems:

- Obtain better than general case in planar graphs
- Obtain $(1+\varepsilon)$ for $\mathbb{R}^{O(1)}$ for k-means using k centers
- Design a unified approach for well-clusterable instances

Our Results

Local search is a PTAS for uniform facility location in edge-weighted planar graphs.

Local search is a PTAS for k-median in edge-weighted planar graphs.

Local search is a PTAS for k-means in $\mathbb{R}^d$.

Techniques: **Separators**

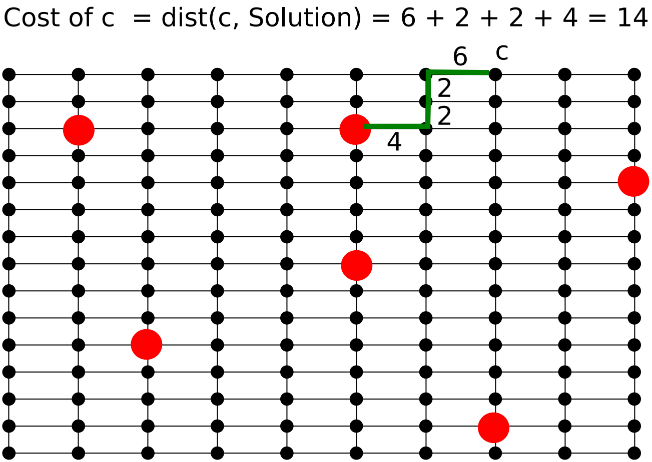
Planar graphs

Planar separator [Lipton and Tarjan, SIAM J. App. Math. '79]:

$\mathbb{R}^{O(1)}$

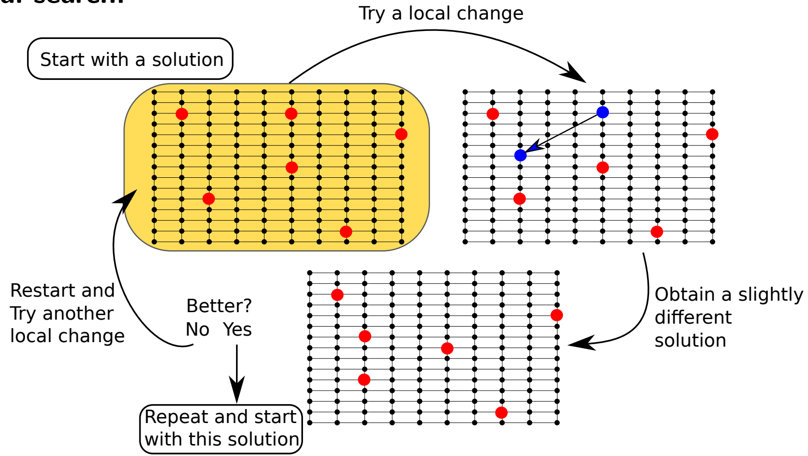
Isoperimetric inequality through [Bhattiprolu and Har-Peled, SoCG '16].

Local search is a PTAS for uniform facility location in edge-weighted planar graphs.



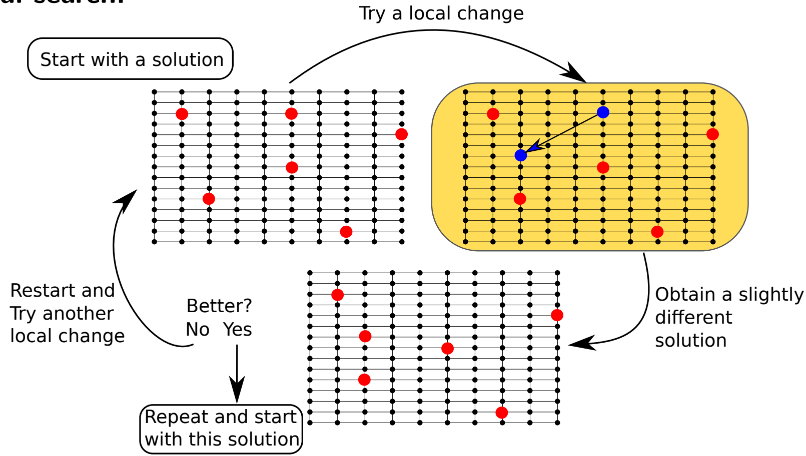
Cost of the solution: 6 (opening cost) + $\sum_c$ (cost of c)

Local search:



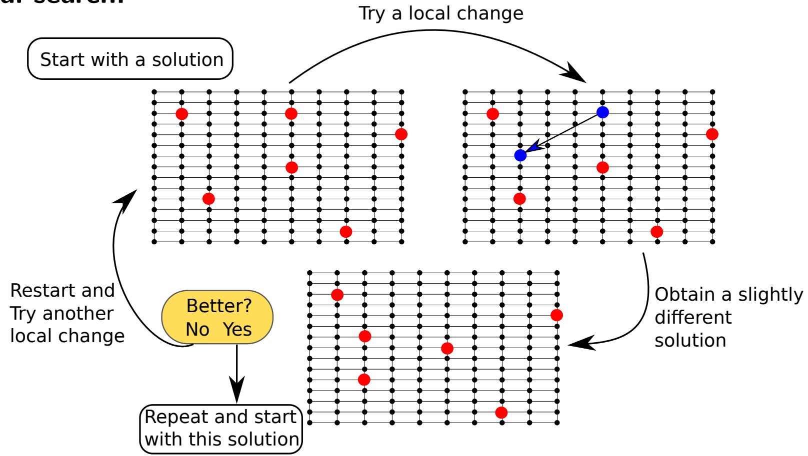
Repeat
Find better solution S among sets that differ from S in at most $1/\varepsilon^2$ centers
Replace S by S
Until: local optimum

Local search:



Repeat
Find better solution S among sets that differ from S in at most $1/\varepsilon^2$ centers
Replace S by S
Until: local optimum

Local search:



Repeat
Find better solution S among sets that differ from S in at most $1/\varepsilon^2$ centers
Replace S by S
Until: local optimum

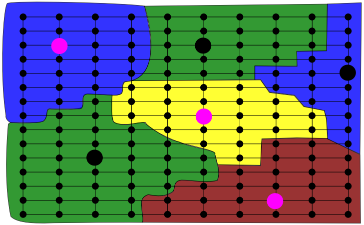
Why does any $1/\varepsilon^2$-locally-optimal solution have value $(1+\varepsilon)\text{OPT}$?

Proof structure:

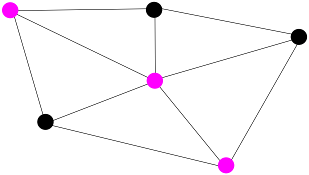
1. Define a structured near-optimal solution OPT'
2. Compare the local solution $\mathcal{L}$ to OPT'

- Local optimum
- Global optimum

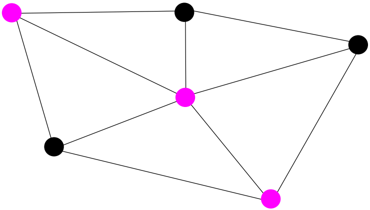
Contract the clusters of the clustering $\mathcal{L} \cup \mathrm{OPT}$.



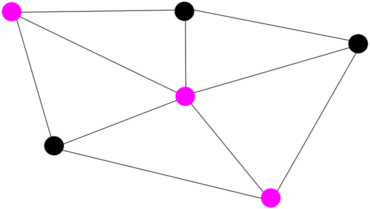
Contraction



Obtain a planar graph $\tilde{G}$



What do we know about planar graphs?



What do we know about planar graphs?

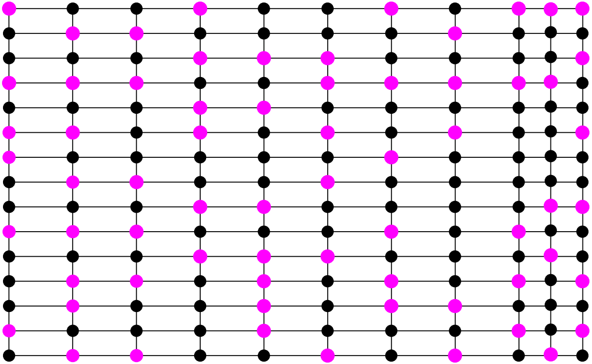
Planar separator [Lipton and Tarjan, SIAM J. App. Math. '79]

For any planar graph with n vertices, there exists a balanced separator with $O(\sqrt{n})$ vertices.

$1/\varepsilon^2$-division – Corollary of Lipton and Tarjan

If $\tilde{G}$ planar then $\exists$ a partition into **regions** such that:

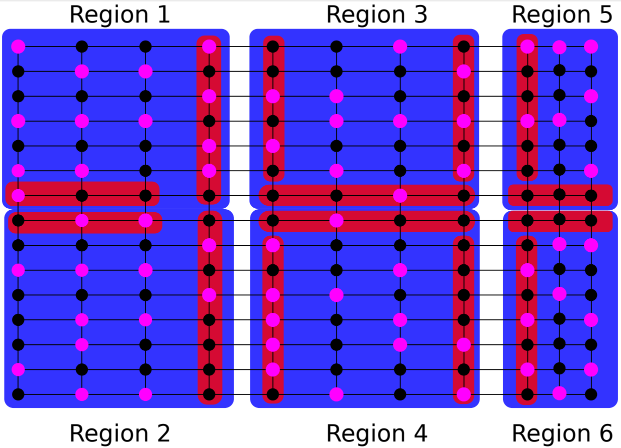
- at most $1/\varepsilon^2$ vertices in each
- at most $\varepsilon V(\tilde{G})$ boundary vertices



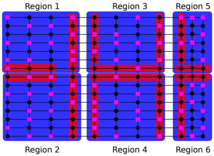
$1/\varepsilon^2$-division – Corollary of Lipton and Tarjan

If $\tilde{G}$ planar then $\exists$ a partition into **regions** such that:

- at most $1/\varepsilon^2$ vertices in each
- at most $\varepsilon V(\tilde{G})$ boundary vertices



Consider the boundary vertices of a $1/\varepsilon^2$-division of $\tilde{G}$



New solution $\mathrm{OPT}' \leftarrow \mathrm{OPT} \cup$ boundary vertices

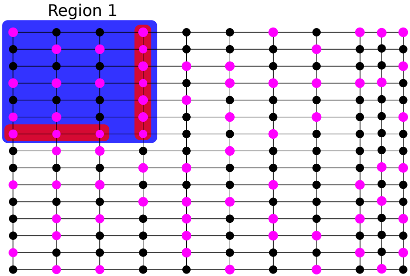
Facility opening cost is ok: $f(|\mathrm{OPT}| + \varepsilon(|\mathrm{OPT}| + |\mathcal{L}|))$

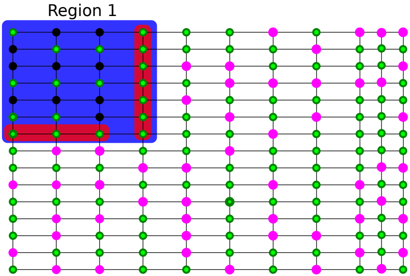
Client cost is optimal: $\mathrm{OPT} \subseteq \mathrm{OPT}' \implies \mathrm{d}(c, \text{closest facility})$ can only decrease

Comparing $\mathcal{L}$ to OPT$'$

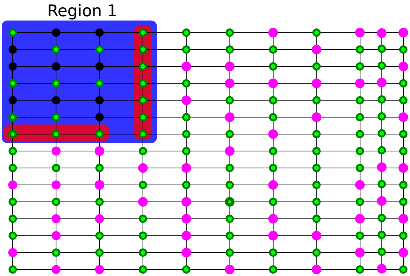
For each region, define a mixed solution M:

$\{$ Facilities of OPT$' \in$ Region $\} \cup \{$ Facilities of $\mathcal{L} \notin$ Region$\}$





Compare $\mathcal{L}$ to M.



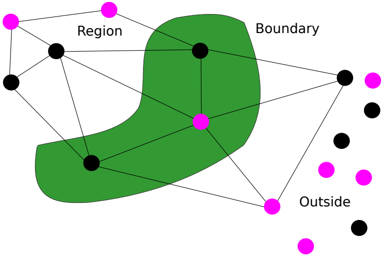
M and $\mathcal{L}$ differ by at most $1/\varepsilon^2$ facilities.

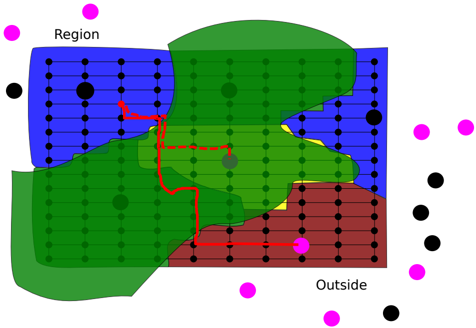
Local optimality implies that $\text{cost}(M) \geq \text{cost}(\mathcal{L})$.

What is the cost of M w.r.t to OPT and $\mathcal{L}$?

Connection cost in M:

Claim: $\forall x \in$ cluster of the region: its closest facility in OPT' is in M



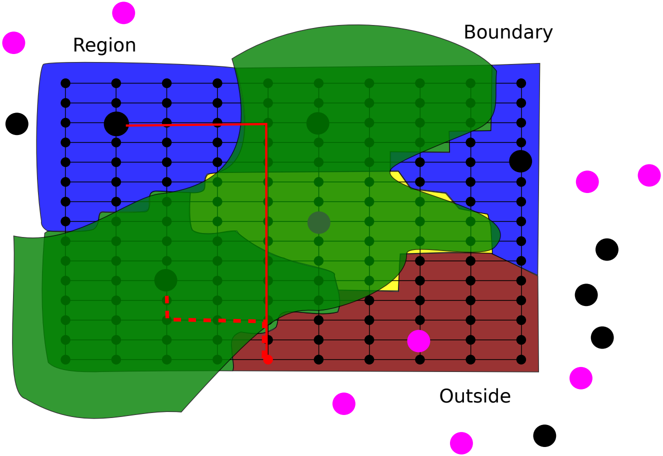


If x is internal then

$$\mathrm{d}(x, M) \leq \mathrm{d}(x, \mathrm{OPT}')$$

Claim: $\forall y \notin$ region: $d(x, M) \leq d(x, \mathcal{L})$

Exact same reasoning w.r.t to $\mathcal{L}$:



Cost of M:

Facility opening cost:

$f \cdot (|\{\mathrm{OPT}' \in \text{region}\}| + |\{\mathcal{L} \notin \text{region}\}|)$

Client service cost: at most

$\sum_{x \text{ internal}} \mathrm{d}(x, \mathrm{OPT}') + \sum_{y \text{ external}} \mathrm{d}(y, \mathcal{L})$

Local optimality: $\text{cost}(M) \geq \text{cost}(\mathcal{L})$

$$\text{cost}(M) \leq \sum_{x \text{ internal}} \mathrm{d}(x, \mathrm{OPT}') + \sum_{y \text{ external}} \mathrm{d}(y, \mathcal{L}) +$$
$$f \cdot |\{\mathrm{OPT}' \in \text{Region}\}| + f \cdot |\{\mathcal{L} \notin \text{Region}\}|$$

$$\text{cost}(\mathcal{L}) = \sum_{x \text{ internal}} \mathrm{d}(x, \mathcal{L}) + \sum_{y \text{ external}} \mathrm{d}(y, \mathcal{L}) +$$
$$f \cdot |\{\mathcal{L} \in \text{Region}\}| + f \cdot |\{\mathcal{L} \notin \text{Region}\}|$$

$$\sum_{x \text{ internal}} \mathrm{d}(x, \mathcal{L}) + f|\{\mathcal{L} \in \text{Reg.}\}| \leq \sum_{x \text{ internal}} \mathrm{d}(x, \mathrm{OPT}') + f|\{\mathrm{OPT}' \in \text{Reg.}\}|$$

Local optimality: $\text{cost}(M) \geq \text{cost}(\mathcal{L})$

$$\text{cost}(M) \leq \sum_{x \text{ internal}} \text{d}(x, \text{OPT}') + \sum_{y \text{ external}} \text{d}(y, \mathcal{L}) + \\ f \cdot |\{\text{OPT}' \in \text{Region}\}| + f \cdot |\{\mathcal{L} \notin \text{Region}\}|$$

$$\text{cost}(\mathcal{L}) = \sum_{x \text{ internal}} \text{d}(x, \mathcal{L}) + \sum_{y \text{ external}} \text{d}(y, \mathcal{L}) + \\ f \cdot |\{\mathcal{L} \in \text{Region}\}| + f \cdot |\{\mathcal{L} \notin \text{Region}\}|$$

$$\sum_{x \text{ internal}} \text{d}(x, \mathcal{L}) + f|\{\mathcal{L} \in \text{Reg.}\}| \leq \sum_{x \text{ internal}} \text{d}(x, \text{OPT}') + f|\{\text{OPT}' \in \text{Reg.}\}|$$

$$\sum_{x \text{ internal}} \mathrm{d}(x, \mathcal{L}) + f|\{\mathcal{L} \in \text{Reg.}\}| \leq \sum_{x \text{ internal}} \mathrm{d}(x, \mathrm{OPT}') + f|\{\mathrm{OPT}' \in \text{Reg.}\}|$$

Sum over all regions

$$\mathrm{cost}(L) \leq \mathrm{cost}(\mathrm{OPT}) + f|\text{boundary vertices}|$$

$$\mathrm{cost}(L) \leq \mathrm{cost}(\mathrm{OPT}) + \varepsilon \cdot f \cdot |\mathcal{L} \cup \mathrm{OPT}|$$

$$(1-\varepsilon)\mathrm{cost}(L) \leq (1+\varepsilon)\mathrm{cost}(\mathrm{OPT})$$

Polynomial-time:

Ensure that enough progress is made at each step $\implies$ lose additional εOPT.

Repeat
Find a solution S that improves the cost by a factor $(1 + \varepsilon/k)$ among sets that differ from S in at most $1/\varepsilon^2$ centers
Replace S by S
Until: local optimum

Proof for $\mathbb{R}^{O(1)}$

Building upon [Bhattiprolu and Har-Peled SoCG '16]

There exists $1/\varepsilon^{O(d)}$-division of the Voronoi partition of a set of points in $\mathbb{R}^d$.

Proof works directly.

Our Results

	Best known approx.	
	Previous	**New**
$\mathbb{R}^{O(1)}$	$1+\varepsilon$ (k-median) $9+\varepsilon$ (k-means)	$1+\varepsilon$ by Local Search
H-minor free graphs	2.675 (k-median, UFL) $25+\varepsilon$ (k-means)	

New result: Perform "local search" in time $n \cdot k \cdot (\log n)^{O(1/\varepsilon^d)}$ in d-dimensional Euclidean spaces.

Open: Perform "local search" in $f(\varepsilon)\text{poly}(n)$ in H-minor-free graphs?
PTAS for non-uniform facility location in H-minor-free graphs?

Our Results

	Best known approx.	
	Previous	**New**
$\mathbb{R}^{O(1)}$	$1+\varepsilon$ (*k*-median) $9+\varepsilon$ (*k*-means)	$1+\varepsilon$ by Local Search
H-minor free graphs	2.675 (*k*-median, UFL) $25+\varepsilon$ (*k*-means)	

New result: Perform "local search" in time $n \cdot k \cdot (\log n)^{O(1/\varepsilon^d)}$ in d-dimensional Euclidean spaces.

Open: Perform "local search" in $f(\varepsilon)\mathrm{poly}(n)$ in H-minor-free graphs?
PTAS for non-uniform facility location in H-minor-free graphs?
Thanks for your attention!