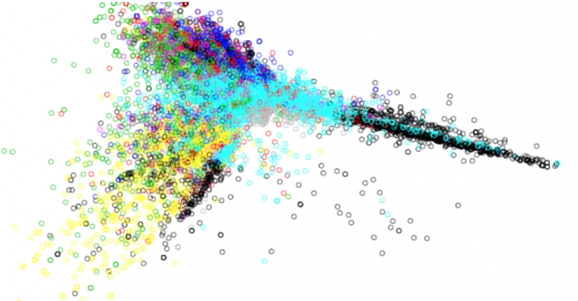
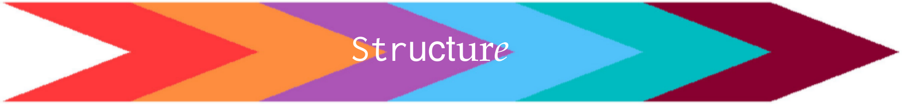
Hardness of Approximation for Metric Clustering

Karthik C. S.
(Rutgers University)

March 5th 2022



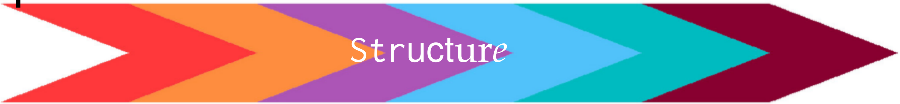
Spectrum of Computational Problems



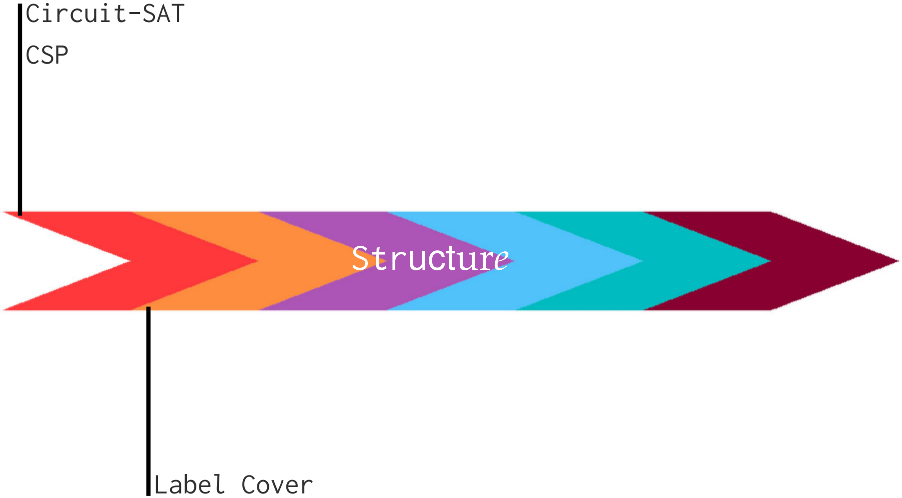
Spectrum of Computational Problems

Circuit-SAT

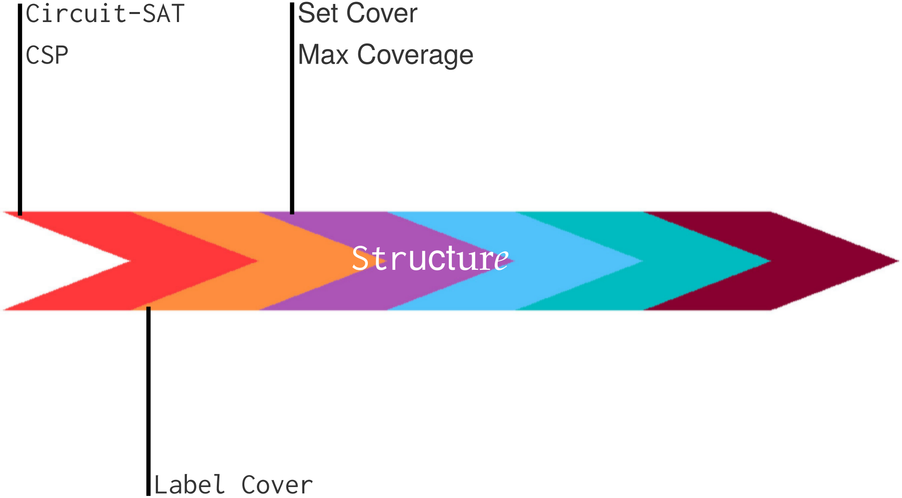
CSP



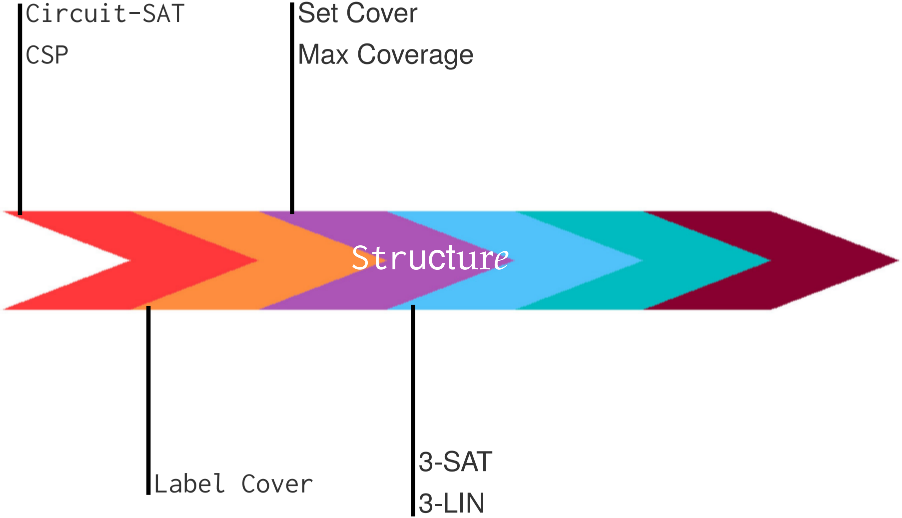
Spectrum of Computational Problems



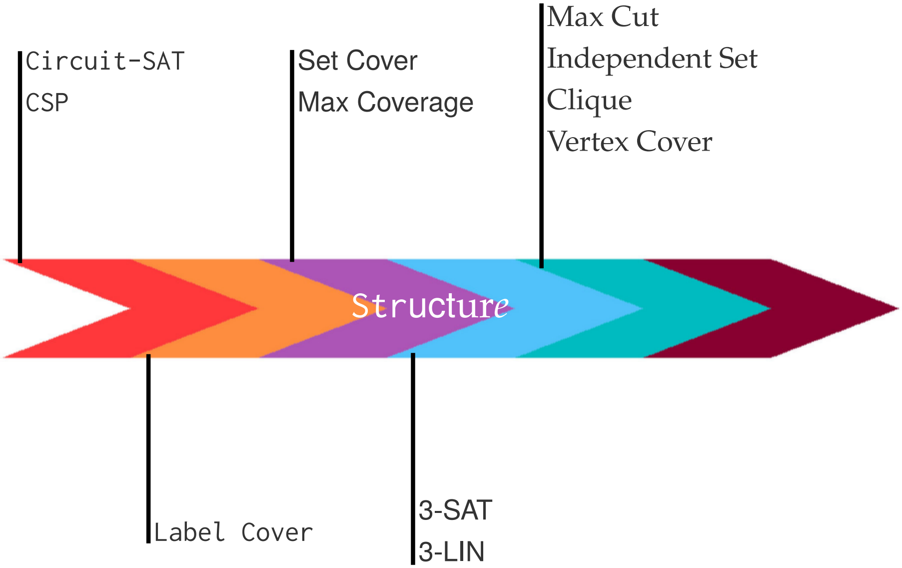
Spectrum of Computational Problems



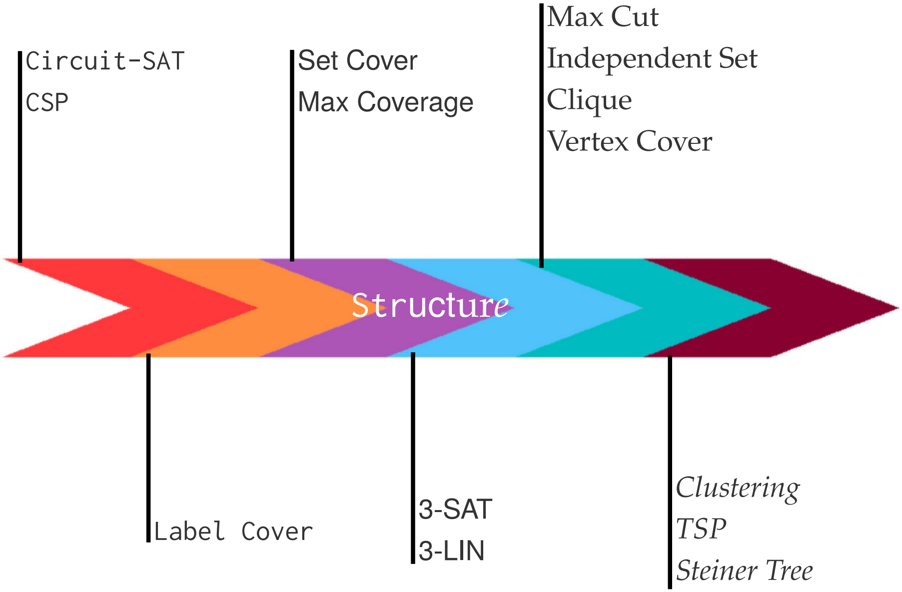
Spectrum of Computational Problems



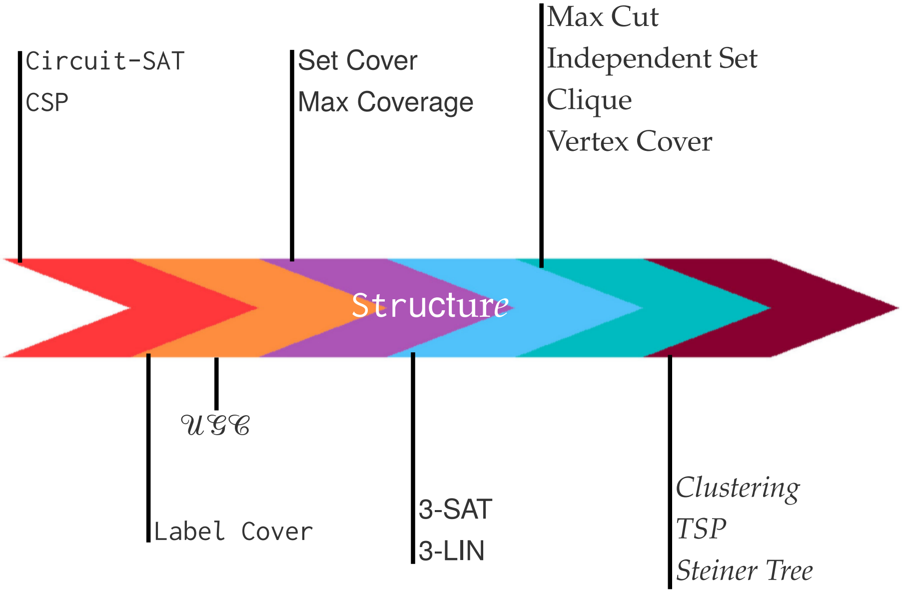
Spectrum of Computational Problems



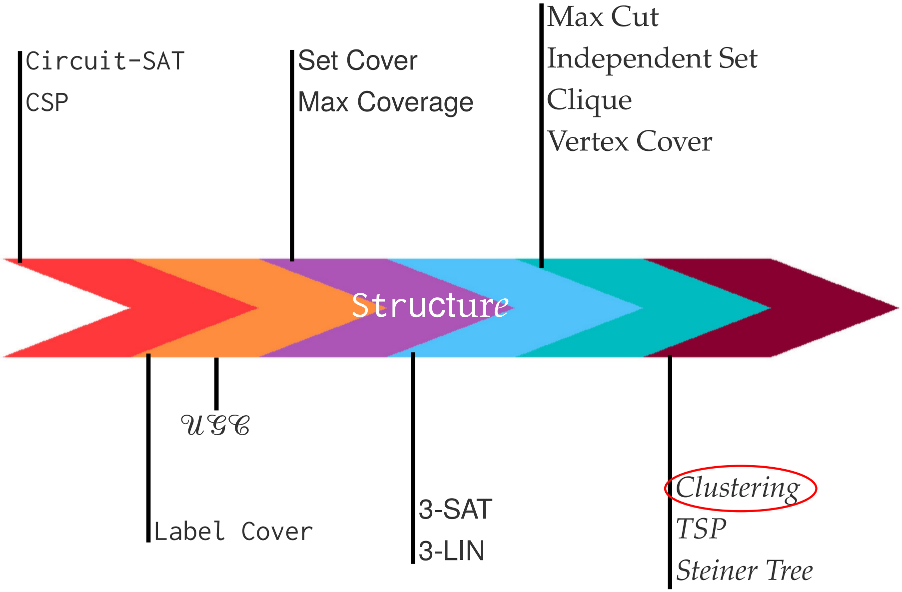
Spectrum of Computational Problems



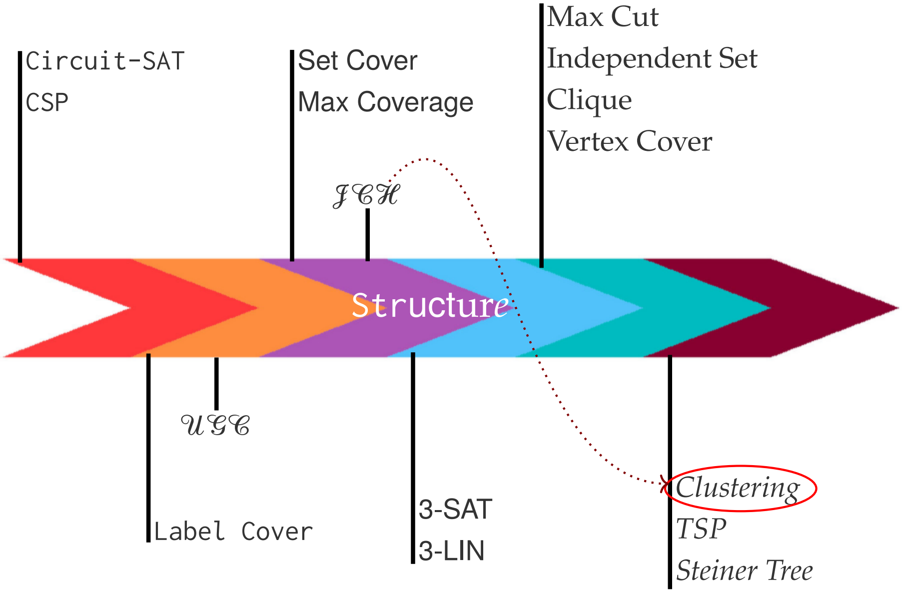
Spectrum of Computational Problems



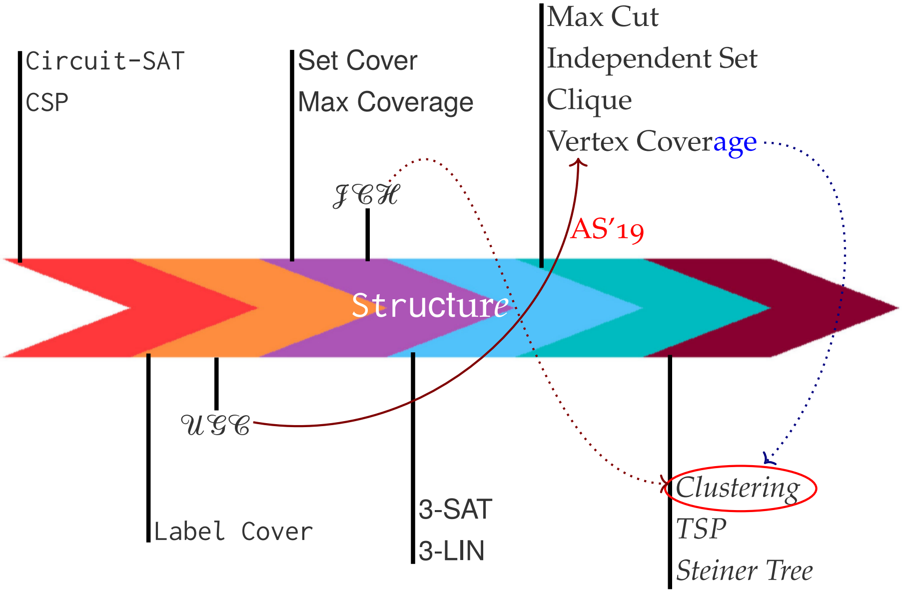
Spectrum of Computational Problems



Spectrum of Computational Problems



Spectrum of Computational Problems



k-center

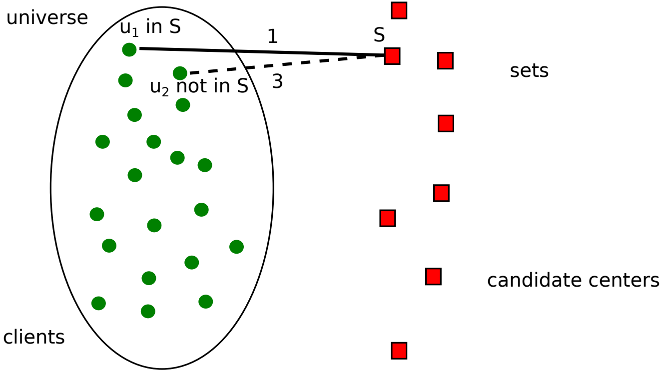
k-center modeling

- Input: $X, S \subseteq \mathbb{R}^d, k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq S$ and $|C| = k$
  - $\sigma : X \to C$
  - (C, σ) minimizes $\max_{x \in X} \|x - \sigma(x)\|_p$

State-of-the-art: General Metrics

- NP-hard [FPT81]
- Poly Time 3-approximation (Gonzalez Algorithm)
- NP-Hard to approximate to $3 - o(1)$ factor! [FPT81]

Proof Overview: General Metrics



Proof Overview: General Metrics

Theorem (Fowler-Paterson-Tanimoto'81)

Given input (X, S, k). It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\max_{x \in X} \Delta(x, \sigma^*(x)) \leq 1$

NO: For all (C, σ) we have $\max_{x \in X} \Delta(x, \sigma(x)) \geq 3$

State-of-the-art: ℓ_p Metrics

- ℓ_1 and ℓ_∞ metrics
  - Poly Time 3-approximation
  - NP-Hard to approximate to $3 - o(1)$ factor! [FG88]
- Euclidean metric
  - Poly Time 2.74-approximation! [NSS13]
  - NP-Hard to approximate to 2.65 factor [FG88]

Proof Overview: ℓ_p Metrics

Vertex Coverage:

- Input: $G(V, E),\ k$

Proof Overview: ℓ_p Metrics

Vertex Coverage:

- Input: $G(V, E),\ k$
- Objective: Max Fraction of E covered by k vertices in V

Vertex Coverage:

- Input: $G(V, E)$, k
- Objective: Max Fraction of E covered by k vertices in V

Theorem (Karp'72)

It is NP-hard to distinguish:

Vertex Coverage:

- Input: $G(V, E),\ k$
- Objective: Max Fraction of E covered by k vertices in V

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Vertex Coverage is 1

Vertex Coverage:

- Input: $G(V, E),\ k$
- Objective: Max Fraction of E covered by k vertices in V

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Vertex Coverage is 1

NO: Vertex Coverage is < 1

Proof Overview: ℓ_p Metrics

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Vertex Coverage is 1

NO: Vertex Coverage is < 1

$\Downarrow$

Proof Overview: ℓ_p Metrics

Theorem (Karp'72)

It is NP-hard to distinguish:

YES: Vertex Coverage is 1

NO: Vertex Coverage is < 1

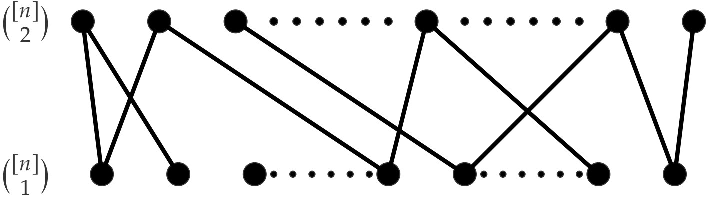
$$\Downarrow$$

Theorem (Fowler-Paterson-Tanimoto'81)

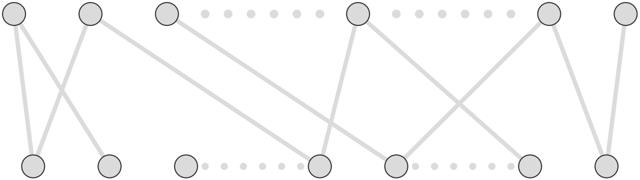
Fix $\varepsilon > 0$. Given input (X, S, k) in $\mathbb{R}^n$. It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\max_{x \in X} \|x - \sigma^*(x)\|_1 \leq 1$

NO: For all (C, σ) we have $\max_{x \in X} \|x - \sigma(x)\|_1 \geq 3$

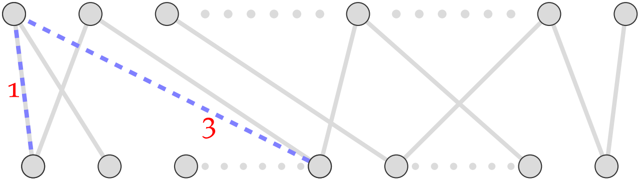

$\binom{[n]}{2}$
$\binom{[n]}{1}$

Points in $\{0,1\}^n$



Graph Embedding

Points in $\{0, 1\}^n$



Proof Overview: ℓ_p Metrics

Theorem (Fowler-Paterson-Tanimoto'81)

Fix $\varepsilon > 0$. Given input (X, S, k) in $\mathbb{R}^n$. It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\max_{x \in X} \|x - \sigma^*(x)\|_1 \leq 1$

NO: For all (C, σ) we have $\max_{x \in X} \|x - \sigma(x)\|_1 \geq 3$

k-means & k-median

k-means and k-median modeling

- Input: $X, S \subseteq \mathbb{R}^d$, $k \in \mathbb{N}$
- Output: A classification (C, σ):
  - $C \subseteq S$ and $|C| = k$
  - $\sigma : X \to C$
  - k-means: (C, σ) minimizes $\sum_{x \in X} \|x - \sigma(x)\|_p^2$
  - k-median: (C, σ) minimizes $\sum_{x \in X} \|x - \sigma(x)\|_p$

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee'22)

Discrete Version

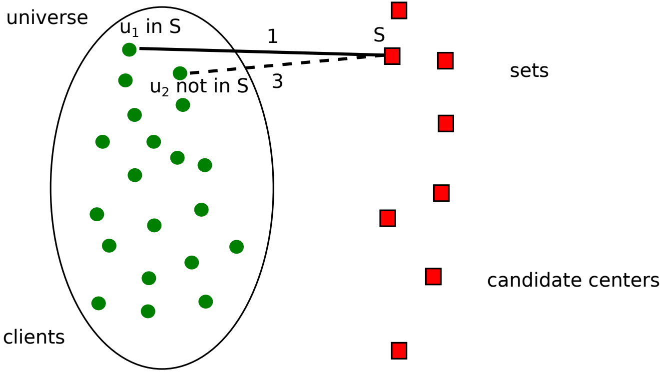
	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

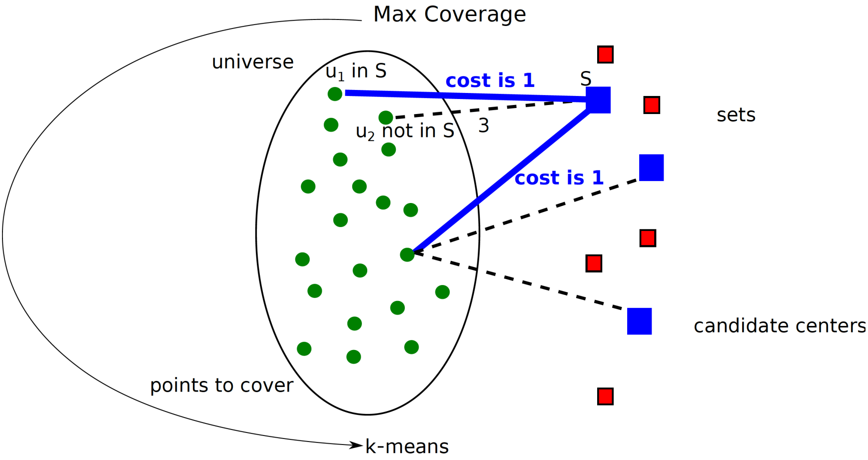
Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

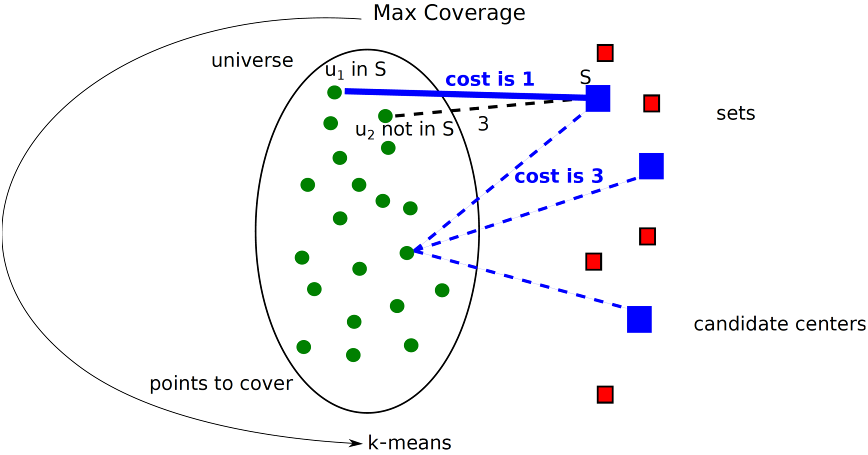
k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Proof Overview: General Metrics




Max Coverage
universe
u_1 in S
cost is 1
S
u_2 not in S
3
sets
cost is 1
candidate centers
points to cover
k-means

Proof Overview: General Metrics



Proof Overview: General Metrics

Theorem (Guha-Khuller'99)

Fix $\varepsilon > 0$. Given input (X, S, k). It is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that $\sum_{x \in X} \Delta(x, \sigma^*(x))^2 \leq |X|$

NO: For all (C, σ) we have $\sum_{x \in X} \Delta(x, \sigma(x))^2 \geq (1 + 8/e - \varepsilon) \cdot |X|$

Johnson Coverage Hypothesis

(α, t)-Johnson Coverage Problem

Given $E \subseteq \binom{[n]}{t}$, and k as input, distinguish between:

Completeness: There exists $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ such that

$$\forall T \in E,\ \exists S_i \in \mathscr{C},\ S_i \subset T.$$

Soundness: For every $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ we have

$$\Pr_{T \sim E}[\exists S_i,\ S_i \subset T] \leq \alpha.$$

Johnson Coverage Hypothesis

(α, t)-Johnson Coverage Problem

Given $E \subseteq \binom{[n]}{t}$, and k as input, distinguish between:

Completeness: There exists $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ such that

$$\forall T \in E,\ \exists S_i \in \mathscr{C},\ S_i \subset T.$$

Soundness: For every $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ we have

$$\Pr_{T \sim E}[\exists S_i,\ S_i \subset T] \leq \alpha.$$

Johnson Coverage Hypothesis (Cohen-Addad–K–Lee'22)

$\forall \varepsilon > 0$, $\exists t_\varepsilon \in \mathbb{N}$ such that $(1 - \frac{1}{e} + \varepsilon, t_\varepsilon)$-Johnson Coverage problem is NP-hard.

Theorem (Cohen-Addad–K–Lee'22)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee'22)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|x - \sigma^*(x)\|_0^2 \leq n',$$

Theorem (Cohen-Addad–K–Lee'22)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee'22)

Assuming $(1\text{-}\frac{1}{e}, t)$ Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee'22)

Assuming $(1\text{-}\frac{1}{e}, t)$ Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|x - \sigma(x)\|_0^2 \geq \left(1 + \frac{8}{e}\right) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee'22)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee'22)

Assuming (0.93,2) Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee'22)

Assuming (0.93,2) Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

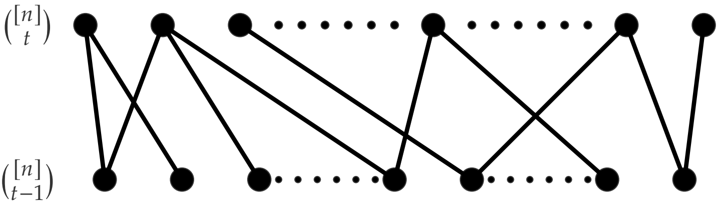
YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

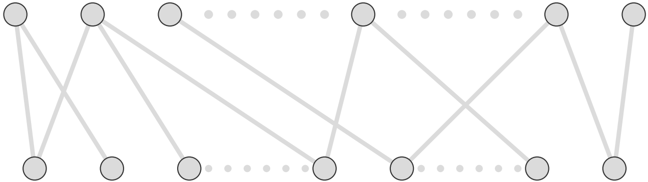
$$\sum_{x \in X} \|x - \sigma(x)\|_0^2 \geq 1.56 \cdot n'.$$

Johnson Graph Embedding

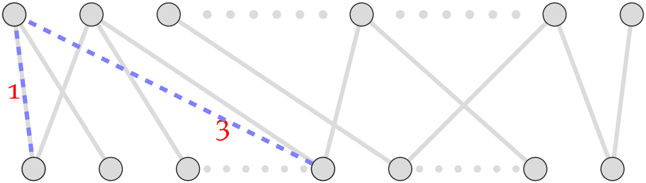


Johnson Graph Embedding

Points in $\{0, 1\}^n$



Points in $\{0,1\}^n$

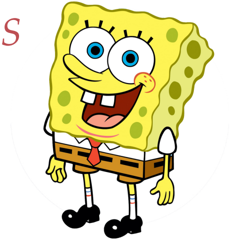


Containment Game

Containment Game

$T \in \binom{[n]}{t}$

$\binom{[n]}{t-1} \ni S$



Containment Game

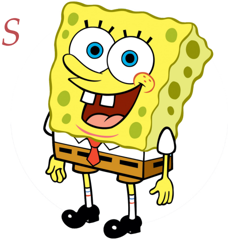
$T \in \binom{[n]}{t}$

$\binom{[n]}{t-1} \ni S$

Containment Game

$T \in \binom{[n]}{t}$

$\binom{[n]}{t-1} \ni S$



Public Randomness

Containment Game

$T \in \binom{[n]}{t}$

$\binom{[n]}{t-1} \ni S$

Public Randomness

GOAL

Determine if $S \subset T$

Containment Game: Protocols

- Deterministic Protocol:
  - Message length: $O(t \log n)$ bits
  - Completeness: 1, Soundness: 0

Containment Game: Protocols

- Deterministic Protocol:
  - Message length: $O(t \log n)$ bits
  - Completeness: 1, Soundness: 0
- Randomized Protocol:
  - Message length: $O_{\varepsilon,t}(1)$ bits

Containment Game: Protocols

- Deterministic Protocol:
  - Message length: $O(t \log n)$ bits
  - Completeness: 1, Soundness: 0
- Randomized Protocol:
  - Message length: $O_{\varepsilon,t}(1)$ bits
  - Completeness: 1, Soundness: ε

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$
- Message length: $(t - 1) \cdot \log_2 q$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \to \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$
- Message length: $(t-1) \cdot \log_2 q$
- Soundness: $t \cdot (1 - \Delta(\mathscr{C}))$

Containment Game: Randomized Protocol

- Let $\mathscr{C} : \mathbb{F}_q^{\log n} \longrightarrow \mathbb{F}_q^{c \cdot \log n}$
- Alice and Bob pick randomly $i \in [c \cdot \log n]$
- Bob sends to Alice $S_i := \{\mathscr{C}(u)_i \mid u \in S\}$
- Alice checks if $S_i \subseteq T_i := \{\mathscr{C}(u)_i \mid u \in T\}$
- Message length: $(t-1) \cdot \log_2 q$
- Soundness: $t \cdot (1 - \Delta(\mathscr{C})) \approx O_t(1/\sqrt{q})$ (for AG codes)

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$

$$\vdots$$

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

Embedding Transcript into Hamming metric

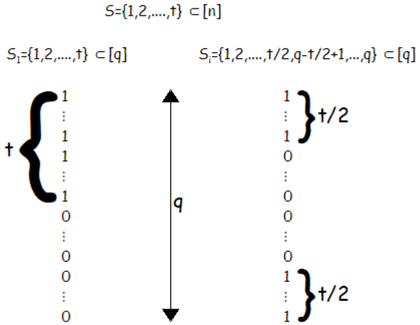
- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$



Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \to \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$

- $X = \{\tau(T) \mid T \in E\}$

Embedding Transcript into Hamming metric

- Construct $\tau : 2^{[n]} \longrightarrow \{0,1\}^{q \cdot c \cdot \log n}$
- Fix $i \in [c \cdot \log n]$ and $S \in 2^{[n]}$:

$$\tau(S)_i = e_{S_i}, \quad \text{where } S_i = \{\mathscr{C}(u)_i \mid u \in S\} \subseteq [q]$$

- $X = \{\tau(T) \mid T \in E\}$
- $\mathcal{S} = \left\{\tau(S) \,\middle|\, S \in \binom{[n]}{t-1}\right\}$

Structural Observations

Suppose $S \subset T$

For every block i, we have $S_i \subset T_i$

$S_i = \{1,2,\ldots,t/2, q-t/2+1,\ldots,q\} \subset [q]$ $\qquad$ $T_i = S_i \cup \{t+1\} \subset [q]$

$$\begin{array}{ccc} 1 & \quad & 1 \\ \vdots & & \vdots \\ 1 & & 1 \\ 0 & & 0 \\ \vdots & & \vdots \\ 0 & & 0 \\ 0 & & 1 \\ 0 & & 0 \\ \vdots & & \vdots \\ 0 & & 0 \\ 1 & & 1 \\ \vdots & & \vdots \\ 1 & & 1 \end{array}$$

$$\boxed{|\tau(T\_i)-\tau(S\_i)|=1}$$

Structural Observations

Suppose $S \not\subset T$

For most blocks i, we have $S_i \not\subset T_i$

$S_i \setminus T_i = \{q\}$	$T_i \setminus S_i = \{t+1,t+2\}$
1	1
⋮	⋮
1	1
0	0
⋮	⋮
0	0
0	1
0	1
0	0
⋮	⋮
0	0
1	1
⋮	⋮
1	1
1	0

$$\boxed{|\tau(T\_i)-\tau(S\_i)| \geq 3}$$

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

$$\sigma(\tau(T)) = \tau(S_i), \quad \text{where } S_i \subset T$$

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

$$\sigma(\tau(T)) = \tau(S_i), \quad \text{where } S_i \subset T$$

- Fix $T \in E$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is 1

Completeness of Reduction

- $\mathcal{S}' := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{t-1}$ be a cover of $E \subseteq \binom{[n]}{t}$
- Build $\sigma : X \to C \subseteq \mathcal{S}$:

$$\sigma(\tau(T)) = \tau(S_i), \quad \text{where } S_i \subset T$$

- Fix $T \in E$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is 1

- k-means objective is:

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 = (c \cdot \log n)^2 \cdot |X|$$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

- $\exists E' \subseteq E$, s.t. $\forall T \in E'$, T does not contain any subset in $\mathcal{S}'$

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

- $\exists E' \subseteq E$, s.t. $\forall T \in E'$, T does not contain any subset in $\mathcal{S}'$
- Fix $\tau(T) \in X_{E'}$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is mostly 3

Soundness of Reduction

- $\sigma : X \to C \subseteq \mathcal{S}$ is some classification
- Build $\mathcal{S}' \subseteq \binom{[n]}{t-1}$ of size k:

$$S \in \mathcal{S}' \iff \tau(S) \in C$$

- $\exists E' \subseteq E$, s.t. $\forall T \in E'$, T does not contain any subset in $\mathcal{S}'$
- Fix $\tau(T) \in X_{E'}$ and $i \in [c \cdot \log n]$

Distance between $\tau(T)$ and $\sigma(\tau(T))$ on block i is mostly 3

- k-means objective is:

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 = (c \cdot \log n)^2 \cdot |X \setminus X_{E'}| + 9 \cdot (c \cdot \log n)^2 \cdot |X_{E'}|$$

Our Embedding in Hamming metric

Theorem (Cohen-Addad–K–Lee'22)

Assuming (α, t)-Johnson coverage problem is NP-hard,

given input $X, \mathcal{S} \subseteq \{0,1\}^{O(\log n)}$, it is NP-hard to distinguish:

YES: There exists (C^*, σ^*) such that

$$\sum_{x \in X} \|(x - \sigma^*(x)\|_0^2 \leq n',$$

NO: For all (C, σ) we have

$$\sum_{x \in X} \|(x - \sigma(x)\|_0^2 \geq (1 + 8 \cdot (1 - \alpha)) \cdot n'.$$

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee'22)

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

Continuous Version

k-means in ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric ≈ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee'22)

Johnson graph Embedding

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19, Cohen-Addad–K–Lee'22)

Johnson graph Embedding

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

Use Feige's Instance

Continuous Version

k-means in ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric ≈ 1.36 (JCH), 1.07 (UGC)

Our Results (Cohen-Addad–K'19,Cohen-Addad–K–Lee'22)

Johnson graph Embedding

Discrete Version

	k-means (JCH)	k-median (JCH)	k-means (UGC)	k-median (UGC)
ℓ_1-metric	3.94	1.73	1.56	1.14
ℓ_2-metric	1.73	1.27	1.17	1.06
ℓ_∞-metric	3.94	1.73	3.94*	1.73*

Use Feige's Instance

Continuous Version

k-means in ℓ_2-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

k-median in ℓ_1-metric $\approx$ 1.36 (JCH), 1.07 (UGC)

Decoding Vertex Cover

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - $\approx$0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

  - Hypergraph Turán number: Open since 1940s!

Johnson Coverage Hypothesis: Discussion

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- Pick $\mathscr{C} := \{S_1, \ldots, S_k\} \subseteq \binom{[n]}{1}$: Max Coverage problem
  - As t increases, gap approaches $1 - \frac{1}{e}$
- LP Integrality gap:

> Determine smallest collection in $\binom{[n]}{t-1}$ that hits all of $\binom{[n]}{t}$

  - Hypergraph Turán number: Open since 1940s!
  - Recently resolved for $t = 3$
  - Improved SDP gaps for Clustering

Johnson Coverage Hypothesis: What can we show?

- $t = 2$: Vertex Coverage problem

Johnson Coverage Hypothesis: What can we show?

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!

Johnson Coverage Hypothesis: What can we show?

- $t = 2$: Vertex Coverage problem
  - ≈ 0.9292 gap is tight!
- 3-Hypergraph Vertex Coverage problem is NP-Hard to approximate to a factor of $7/8$

Key Takeaways

- Improved Inapproximability of

Key Takeaways

- Improved Inapproximability of
- k-means and k-median

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Containment Protocol

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Containment Protocol
- And Geometric Realization of Johnson Graphs

Key Takeaways

- Improved Inapproximability of
- k-means and k-median
- In ℓ_p-metrics
- Using Transcript of Containment Protocol
- And Geometric Realization of Johnson Graphs

Open: Is JCH true?

O
PROBLEMS
E
N

State-of-the-art for k-means

Discrete Version

	JCH	UGC	NP≠P
ℓ_1-metric	3.94	1.56	1.38
ℓ_2-metric	1.73	1.17	1.17
ℓ_∞-metric	3.94	3.94	3.94

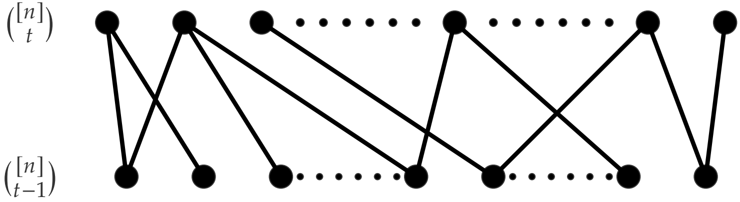
Continuous Version

General metric ≈ 4 (NP≠P)

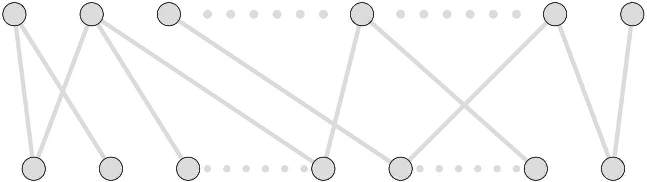
ℓ_2-metric ≈ 1.36 (JCH), 1.07 (UGC), 1.06 (NP≠P)

ℓ_1-metric ≈ 2.10 (JCH), 1.16 (NP≠P)

ℓ_∞-metric $\approx$???

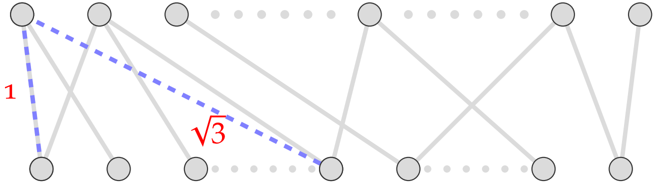

$\binom{[n]}{t}$
$\binom{[n]}{t-1}$

Points in $\{0, 1\}^d$



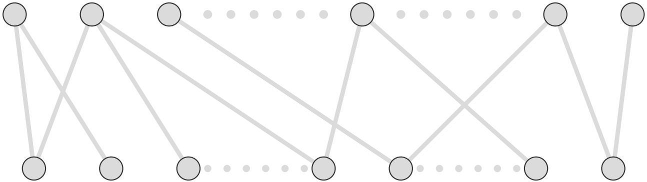
Inapproximability of Clustering in Euclidean metrics

Points in $\{0,1\}^d$



Inapproximability of Clustering in Euclidean metrics

Points in $\{0, 1\}^d$



Is there a better embedding of the Johnson Graph into the Euclidean metric?

Tight inapproximability of k-center in Euclidean metrics?

Other Open Problems

Can we prove strong inapproximability results for:

Other Open Problems

Can we prove strong inapproximability results for:

- k-minsum in ℓ_p-metrics

Other Open Problems

Can we prove strong inapproximability results for:

- k-minsum in ℓ_p-metrics
- Capacitated Clustering

Other Open Problems

Can we prove strong inapproximability results for:

- k-minsum in ℓ_p-metrics
- Capacitated Clustering
- Fair Clustering

THANK
YOU!