Simplified Inapproximability of k-means

Melanie Schmidt

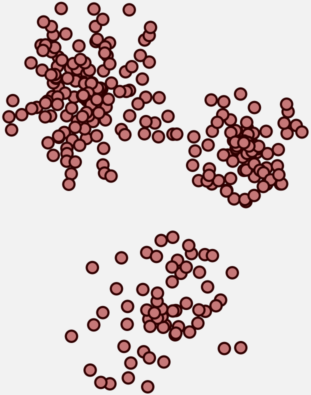
Joint work with Euiwoong Lee and John Wright

07.11.2015

Definition

The k-means problem

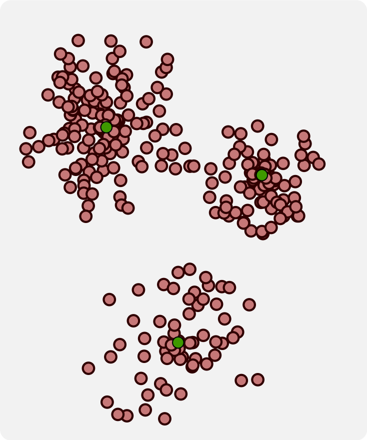
Definition



The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,

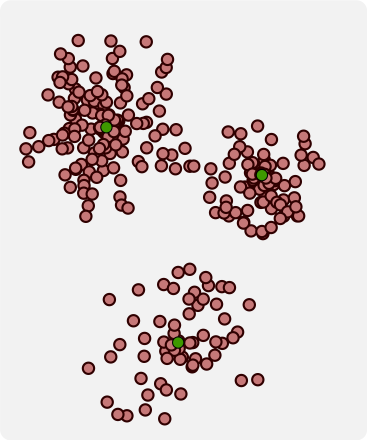
Definition



The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,
- compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers

Definition



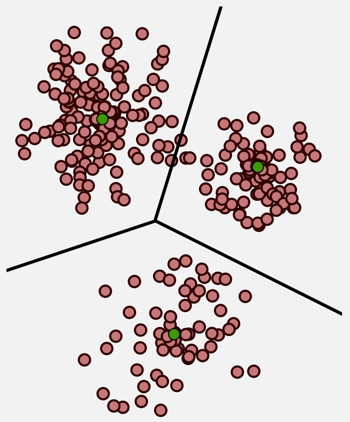
The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,
- compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers
- which minimizes

$$\mathrm{cost}(P, C) = \sum_{p \in P} \min_{c \in C} ||p - c||^2,$$

the sum of the squared distances.

Definition



The k-means problem

- Given a point set $P \subseteq \mathbb{R}^d$,
- compute a set $C \subseteq \mathbb{R}^d$ with $|C| = k$ centers
- which minimizes

$$\mathrm{cost}(P, C) = \sum_{p \in P} \min_{c \in C} ||p - c||^2,$$

the sum of the squared distances.

- induces a partitioning of P

Complexity

	Small dimension d	Large dimension d
Small k		
Large k		

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[ADHP09] Aloise, Deshpande, Hansen, Popat: NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, 2009.
[FL11] Feldman, Langberg, A unified framework for approximating and clustering data, STOC 2011.
[IKI94] Inaba, Katoh, Imai: Appl. of Weighted Voronoi Diagrams and Rand. to Variance-Based k-Clustering, SoCG 1994.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Y. Wu, A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright: Improved and Simplified Inapproximability for k-means, CORR 2015.
[M00] Matoušek: On approximate geometric k-clustering
[MNV09] Mahajan, Nimbhorkar, Varadarajan, The Planar k-means Problem is NP-Hard, WALCOM 2009.

Complexity

	Small dimension d	Large dimension d
Small k		NP-hard for $k = 2$ [ADHP09], but PTAS, best running time $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]
Large k		

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[ADHP09] Aloise, Deshpande, Hansen, Popat: NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, 2009.
[FL11] Feldman, Langberg, A unified framework for approximating and clustering data, STOC 2011.
[IKI94] Inaba, Katoh, Imai: Appl. of Weighted Voronoi Diagrams and Rand. to Variance-Based k-Clustering, SoCG 1994.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Y. Wu, A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright: Improved and Simplified Inapproximability for k-means, CORR 2015.
[M00] Matoušek: On approximate geometric k-clustering
[MNV09] Mahajan, Nimbhorkar, Varadarajan, The Planar k-means Problem is NP-Hard, WALCOM 2009.

Complexity

	Small dimension d	Large dimension d
Small k	Optimal solution by enumerating Voronoi diagrams [IKI94]	NP-hard for $k = 2$ [ADHP09], but PTAS, best running time $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon, k)})$ [FL11]
Large k		

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[ADHP09] Aloise, Deshpande, Hansen, Popat: NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, 2009.
[FL11] Feldman, Langberg, A unified framework for approximating and clustering data, STOC 2011.
[IKI94] Inaba, Katoh, Imai: Appl. of Weighted Voronoi Diagrams and Rand. to Variance-Based k-Clustering, SoCG 1994.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Y. Wu, A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright: Improved and Simplified Inapproximability for k-means, CORR 2015.
[M00] Matoušek: On approximate geometric k-clustering
[MNV09] Mahajan, Nimbhorkar, Varadarajan, The Planar k-means Problem is NP-Hard, WALCOM 2009.

Complexity

	Small dimension d	Large dimension d
Small k	Optimal solution by enumerating Voronoi diagrams [IKI94]	NP-hard for $k = 2$ [ADHP09], but PTAS, best running time $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]
Large k		APX-hard [ACKS15], factor is ≥ 1.0013 [LSW15] and $\leq 9 + \varepsilon$ [KMN+02]

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[ADHP09] Aloise, Deshpande, Hansen, Popat: NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, 2009.
[FL11] Feldman, Langberg, A unified framework for approximating and clustering data, STOC 2011.
[IKI94] Inaba, Katoh, Imai: Appl. of Weighted Voronoi Diagrams and Rand. to Variance-Based k-Clustering, SoCG 1994.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Y. Wu, A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright: Improved and Simplified Inapproximability for k-means, CORR 2015.
[M00] Matoušek: On approximate geometric k-clustering
[MNV09] Mahajan, Nimbhorkar, Varadarajan, The Planar k-means Problem is NP-Hard, WALCOM 2009.

Complexity

	Small dimension d	Large dimension d
Small k	Optimal solution by enumerating Voronoi diagrams [IKI94]	NP-hard for $k = 2$ [ADHP09], but PTAS, best running time $\mathcal{O}(nd + 2^{\mathrm{poly}(1/\varepsilon, k)})$ [FL11]
Large k	NP-hard for $d = 2$ [MNV09] no PTAS known, but no APX-hardness proof either	APX-hard [ACKS15], factor is ≥ 1.0013 [LSW15] and $\leq 9 + \varepsilon$ [KMN+02]

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[ADHP09] Aloise, Deshpande, Hansen, Popat: NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, 2009.
[FL11] Feldman, Langberg, A unified framework for approximating and clustering data, STOC 2011.
[IKI94] Inaba, Katoh, Imai: Appl. of Weighted Voronoi Diagrams and Rand. to Variance-Based k-Clustering, SoCG 1994.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Y. Wu, A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright: Improved and Simplified Inapproximability for k-means, CORR 2015.
[M00] Matoušek: On approximate geometric k-clustering
[MNV09] Mahajan, Nimbhorkar, Varadarajan, The Planar k-means Problem is NP-Hard, WALCOM 2009.

Complexity

	Small dimension d	Large dimension d
Small k	Optimal solution by enumerating Voronoi diagrams [IKI94]	NP-hard for $k = 2$ [ADHP09], but PTAS, best running time $\mathcal{O}(nd + 2^{\text{poly}(1/\varepsilon,k)})$ [FL11]
Large k	NP-hard for $d = 2$ [MNV09] no PTAS known, but no APX-hardness proof either	APX-hard [ACKS15], factor is ≥ 1.0013 [LSW15] and $\leq 9 + \varepsilon$ [KMN+02]

[ACKS15] Awasthi, Charikar, Krishnaswamy, Sinop. The hardness of approximation of euclidean k-means, SoCG 2015.
[ADHP09] Aloise, Deshpande, Hansen, Popat: NP-hardness of Euclidean sum-of-squares clustering, Machine Learning, 2009.
[FL11] Feldman, Langberg, A unified framework for approximating and clustering data, STOC 2011.
[IKI94] Inaba, Katoh, Imai: Appl. of Weighted Voronoi Diagrams and Rand. to Variance-Based k-Clustering, SoCG 1994.
[KMN+02] Kanungo, Mount, Netanyahu, Piatko, Silverman, Y. Wu, A local search approx. alg. for k-means clustering, SoCG 2002.
[LSW13] Lee, S. Wright: Improved and Simplified Inapproximability for k-means, CORR 2015.
[M00] Matoušek: On approximate geometric k-clustering
[MNV09] Mahajan, Nimbhorkar, Varadarajan, The Planar k-means Problem is NP-Hard, WALCOM 2009.

- What is the best possible approximation factor?
- PTAS for $d = 2$, constant d?

Reducing vertex cover ($\triangle$-free) to k-means

(Awasthi et. al., SoCG 2015)

Reducing vertex cover ($\triangle$-free) to k-means

(Awasthi et. al., SoCG 2015)

Vertex Cover instance

Graph $G = (V, E)$

Reducing vertex cover ($\triangle$-free) to k-means
(Awasthi et. al., SoCG 2015)

Vertex Cover instance

Graph $G = (V, E)$

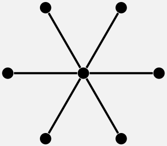
k-means instance

For $e = (i, j)$, define $x_e \in \mathbb{R}^{|V|}$ by

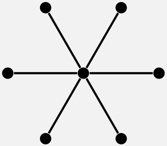
$$(x_e)_i = (x_e)_i = 1 \text{ and } (x_e)_\ell = 0 \text{ for } \ell \neq i, j$$

$$x_e = (0, \ldots, 0, \underbrace{1}_{i}, 0, \ldots, 0, \underbrace{1}_{j}, 0 \ldots, 0)$$

Example

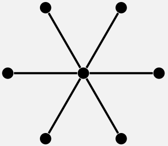


Example



$$
\begin{array}{lcccccccccccc}
 & v_1 & v_2 & v_3 & \dots & & & & & & & & \\
e_1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
e_2 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
e_3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
e_3 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
e_3 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
e_3 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
 & \frac{1}{6} & 0 & 1 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 & \frac{1}{6} & 0
\end{array}
$$

Example

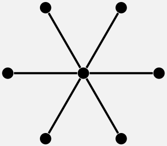


$$
\begin{array}{ll}
 & v_1 v_2 v_3 \ldots \\
e_1 & 101000000000 \\
e_2 & 001010000000 \\
e_3 & 001000000010 \\
e_3 & 001100000000 \\
e_3 & 001001000000 \\
e_3 & 001000100000 \\
 & \tfrac{1}{6}01\tfrac{1}{6}\tfrac{1}{6}\tfrac{1}{6}\tfrac{1}{6}000\tfrac{1}{6}0
\end{array}
$$

Cluster cost with (001000000000): $|E'|$

With centroid: $|E'| - 1$

Example

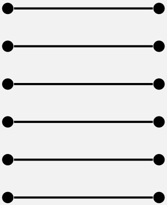


$v_1 v_2 v_3 \ldots$

e_1 101000000000
e_2 001010000000
e_3 001000000010
$\dot{e}_3$ 001100000000
e_3 001001000000
e_3 001000100000
$\frac{1}{6}\,0\,1\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,0\,0\,0\,\frac{1}{6}\,0$

Cluster cost with (001000000000): $|E'|$

With centroid: $|E'| - 1$



$v_1 v_2 v_3 \ldots$

e_1 101000000000
e_2 010000100000
e_3 000000011000
e_4 000000000101
e_5 000010000010
e_6 000101000000
$\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}\,\frac{1}{6}$

Cluster cost with (000000000000): $2|E'|$

With centroid: $2|E'| - 2$

Idea

Idea

- star cluster E' costs $|E'| - 1$

Idea

- star cluster E' costs $|E'| - 1$
- small vertex cover implies k star clusters $\rightsquigarrow$ small cost $(m - k)$

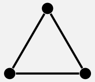
Idea

- star cluster E' costs $|E'| - 1$
- small vertex cover implies k star clusters $\rightsquigarrow$ small cost $(m - k)$
- hope: small cost implies many stars and small enough vertex cover

Idea

- star cluster E' costs $|E'| - 1$
- small vertex cover implies k star clusters $\rightsquigarrow$ small cost $(m - k)$
- hope: small cost implies many stars and small enough vertex cover

Problem: Triangles



$$\begin{array}{cccc} & v_1 & v_2 & v_3 \\ e_1 & 1 & 1 & 0 \\ e_2 & 0 & 1 & 1 \\ e_3 & 1 & 0 & 1 \\ & \frac{2}{3} & \frac{2}{3} & \frac{2}{3} \end{array}$$

Cluster cost:
$3 \cdot (2 \cdot \frac{1}{3^2} + (\frac{2}{3})^2) = 3 - 1$

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$$\Downarrow$$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$$\Downarrow$$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

Awasthi et. al., Part II

APX-hardness for VC in graphs with max. degree D

$$\Downarrow$$

APX-hardness for VC in $\triangle$-free graphs with max. degree poly(D, ε^{-1})

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$$\Downarrow$$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

Awasthi et. al., Part II

APX-hardness for VC in graphs with max. degree D

$$\Downarrow$$

APX-hardness for VC in $\triangle$-free graphs with max. degree poly(D, ε^{-1})

VC in $\triangle$-free graphs is 1.36-hard

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$$\Downarrow$$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

Awasthi et. al., Part II

APX-hardness for VC in graphs with max. degree D

$$\Downarrow$$

APX-hardness for VC in $\triangle$-free graphs with max. degree poly(D, ε^{-1})

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$\Downarrow$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

New Part II

APX-hardness for VC in graphs that are 4-regular

$\Downarrow$

APX-hardness for VC in $\triangle$-free graphs and maximum degree 4

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$$\Downarrow$$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

New Part II

APX-hardness for VC in graphs that are 4-regular

$$\Downarrow$$

APX-hardness for VC in $\triangle$-free graphs and maximum degree 4

Chlebík, Clebíková, 2006

Given a 4-regular graph G, it is NP-hard to distinguish

- G has a vertex cover of size $\leq \alpha_{\text{min}}|V(A)|$
- every vertex cover in G has size $\geq \alpha_{\text{max}}|V(A)|$

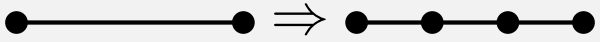
Here, $\alpha_{\text{max}}/\alpha_{\text{min}} \geq 1.0192$.

Idea I

Replace every edge by three edges

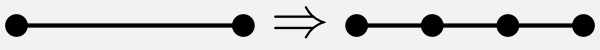
Idea I

Replace every edge by three edges



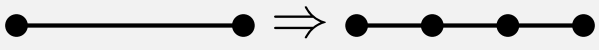
Idea I

Replace every edge by three edges ($\rightsquigarrow +4n$ vertices and $+4n$ edges)



Idea I

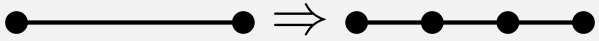
Replace every edge by three edges ($\rightsquigarrow +4n$ vertices and $+4n$ edges)



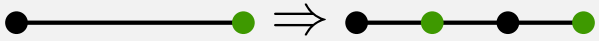
Minimum vertex cover size increases by $2n$:

Idea I

Replace every edge by three edges ($\rightsquigarrow +4n$ vertices and $+4n$ edges)

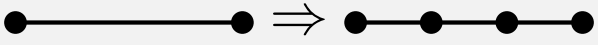


Minimum vertex cover size increases by $2n$:

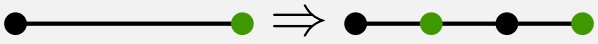


Idea I

Replace every edge by three edges ($\rightsquigarrow +4n$ vertices and $+4n$ edges)

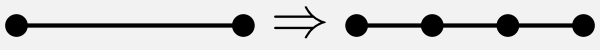


Minimum vertex cover size increases by $2n$:

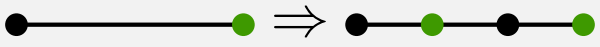


Idea I

Replace every edge by three edges ($\rightsquigarrow +4n$ vertices and $+4n$ edges)

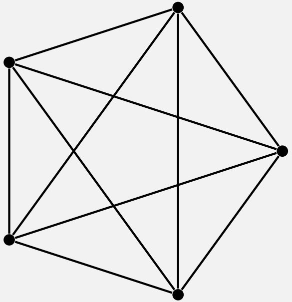


Minimum vertex cover size increases by $2n$:

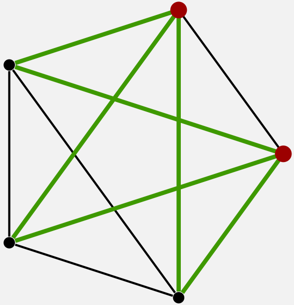


$\rightsquigarrow$ NP-hard to decide between $\leq (\alpha_{\text{min}} + 2)n$ and $\geq (\alpha_{\text{max}} + 2)n$

Idea II

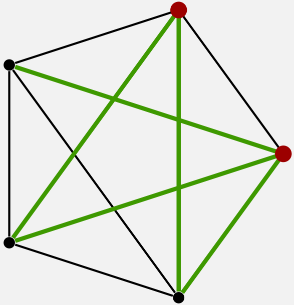


Idea II



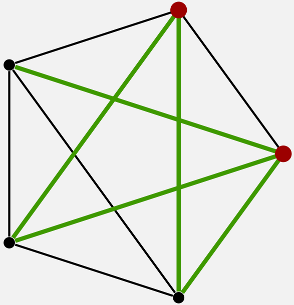
- Let E' with $|E'| \geq m/2$ be the edges of a large cut

Idea II



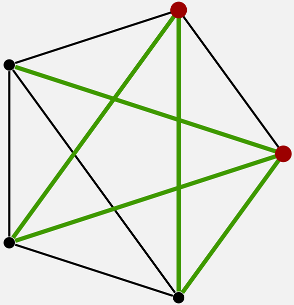
- Let E' with $|E'| \geq m/2$ be the edges of a large cut
- Pick $E_1 \subseteq E'$ with $|E_1| = m/2 = n$ (E_1 is bipartite)

Idea II



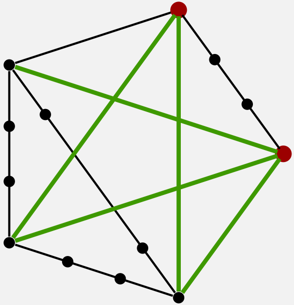
- Let E' with $|E'| \geq m/2$ be the edges of a large cut
- Pick $E_1 \subseteq E'$ with $|E_1| = m/2 = n$ (E_1 is bipartite)
- n remaining edges, E_2

Idea II



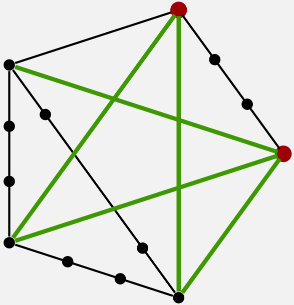
- Let E' with $|E'| \geq m/2$ be the edges of a large cut
- Pick $E_1 \subseteq E'$ with $|E_1| = m/2 = n$ (E_1 is bipartite)
- n remaining edges, E_2
- Only split edges in E_2

Idea II



- Let E' with $|E'| \geq m/2$ be the edges of a large cut
- Pick $E_1 \subseteq E'$ with $|E_1| = m/2 = n$ (E_1 is bipartite)
- n remaining edges, E_2
- Only split edges in E_2

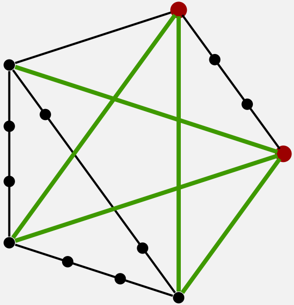
Idea II



- Let E' with $|E'| \geq m/2$ be the edges of a large cut
- Pick $E_1 \subseteq E'$ with $|E_1| = m/2 = n$ (E_1 is bipartite)
- n remaining edges, E_2
- Only split edges in E_2

$\rightsquigarrow$ $2n$ new edges and vertices, min. vertex cover size increases by n

Idea II



- Let E' with $|E'| \geq m/2$ be the edges of a large cut
- Pick $E_1 \subseteq E'$ with $|E_1| = m/2 = n$ (E_1 is bipartite)
- n remaining edges, E_2
- Only split edges in E_2

$\rightsquigarrow$ $2n$ new edges and vertices, min. vertex cover size increases by n
$\rightsquigarrow$ Gap between $(\alpha_{\min} + 1)n$ and $(\alpha_{\max} + 1)n$

Clebík, Chlebík, Clebíková, 2006

APX-hardness for VC in 4-regular graphs

Clebík, Chlebík, Clebíková, 2006

APX-hardness for VC in 4-regular graphs

New Part II

APX-hardness for VC in 4-regular graphs

$$\Downarrow$$

APX-hardness for VC in 4-regular $\triangle$-free graphs

Clebík, Chlebík, Clebíková, 2006

APX-hardness for VC in 4-regular graphs

New Part II

APX-hardness for VC in 4-regular graphs

$$\Downarrow$$

APX-hardness for VC in 4-regular $\triangle$-free graphs

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$$\Downarrow$$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

Clebík, Chlebík, Clebíková, 2006

APX-hardness for VC in 4-regular graphs

New Part II

APX-hardness for VC in 4-regular graphs

$\Downarrow$

APX-hardness for VC in 4-regular $\triangle$-free graphs

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$\Downarrow$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

Theorem

It is NP-hard to approximate k-means within a factor of 1.0013.

Clebík, Chlebík, Clebíková, 2006

APX-hardness for VC in 4-regular graphs

New Part II

APX-hardness for VC in 4-regular graphs

$\Downarrow$

APX-hardness for VC in 4-regular $\triangle$-free graphs

Awasthi et. al., Part I

$(1+\varepsilon)$-hardness for vertex cover in $\triangle$-free graphs with $D \cdot n$ edges

$\Downarrow$

$(1+\varepsilon')$-hardness for k-means with $\varepsilon' \in \Theta(\varepsilon/D)$

Theorem

It is NP-hard to approximate k-means within a factor of 1.0013.

Thanks!