# A Sublinear-Time Spectral Clustering Oracle with Improved Preprocessing Time

Ranran Shen Pan Peng

University of Science and Technology of China

Background ●000000		
Graph Clustering		

- Input: G = (V, E) and  $k \ (k \ge 2)$
- Goal: partition V into k disjoint clusters  $C_1, \ldots, C_k$ , such that each cluster exhibits
  - tight connections inside
  - loose connections outside



Example (k = 3)

Background ●000000		
Graph Clustering		

- Input: G = (V, E) and  $k \ (k \ge 2)$
- Goal: partition V into k disjoint clusters  $C_1, \ldots, C_k$ , such that each cluster exhibits
  - tight connections inside
  - loose connections outside



Example (k = 3)

2 / 13

Global algorithms run in poly(n) time: n increases  $\Rightarrow$  impractical (n = |V|)

Background ●000000		
Graph Clustering		

- Input: G = (V, E) and  $k \ (k \ge 2)$
- Goal: partition V into k disjoint clusters  $C_1, \ldots, C_k$ , such that each cluster exhibits
  - tight connections inside
  - loose connections outside



Example (k = 3)

Global algorithms run in poly(n) time: n increases  $\Rightarrow$  impractical (n = |V|)We focus on sublinear-time spectral clustering oracles.

Ranran Shen, Pan Peng

A Sublinear-Time Spectral Clustering Oracle with Improved Preprocessing Time

Background 0●00000		
Query Access		

► Has query access to the adjacency list of the input graph G



Background 00●0000		

#### Two Phases

- Preprocessing phase (sublinear-time)
  - build a data structure



Preprocessing phase

Background 00●0000		

## Two Phases

- Preprocessing phase (sublinear-time)
  - build a data structure
- Query phase (sublinear-time)
  - ▶ answer WHICHCLUSTER(v) queries



Background 000●000		

# Requirements

- Consistent
- Close to the ground-truth clustering

Background 0000●00		

• *d*-bounded graphs: maximum degree  $\leq d$ 

• *d*-bounded graphs: maximum degree  $\leq d$ 



Example: a clusterable graph

•  $(k, \varphi, \varepsilon)$ -clusterable graphs  $(\varepsilon \ll \varphi)$ 

6 / 13

• *d*-bounded graphs: maximum degree  $\leq d$ 



Example: a clusterable graph

►  $(k, \varphi, \varepsilon)$ -clusterable graphs  $(\varepsilon \ll \varphi)$ ► has a *k*-partition of *V*, denoted by  $C_1, \ldots, C_k$ ,  $\frac{|C_i|}{|C_i|} \in O(1)$ 

• *d*-bounded graphs: maximum degree  $\leq d$ 



Example: a clusterable graph

- $(k, \varphi, \varepsilon)$ -clusterable graphs  $(\varepsilon \ll \varphi)$ 
  - ▶ has a *k*-partition of *V*, denoted by  $C_1, \ldots, C_k$ ,  $\frac{|C_i|}{|C_i|} \in O(1)$
  - ▶ tight connections inside: inner conductance  $\phi_{in}(C_i) \ge \varphi$

• *d*-bounded graphs: maximum degree  $\leq d$ 



Example: a clusterable graph

- $(k, \varphi, \varepsilon)$ -clusterable graphs  $(\varepsilon \ll \varphi)$ 
  - ▶ has a *k*-partition of *V*, denoted by  $C_1, \ldots, C_k$ ,  $\frac{|C_i|}{|C_j|} \in O(1)$
  - tight connections inside: inner conductance  $\phi_{in}(C_i) \ge \varphi$
  - loose connections outside: outer conductance  $\phi_{out}(C_i, V) \leq \varepsilon$

Background 00000€0		

# Previous Work

	[Pen20]	[GKL <sup>+</sup> 21]
conductance	$\varphi^2$	$\varphi^3$
gap	$c \sim \operatorname{poly}(k) \cdot \log n$	$c \gg \log k$
preprocessing	$O\left(\sqrt{n} \cdot \operatorname{poly}\left(\frac{k \log n}{n}\right)\right)$	$\operatorname{poly}(\frac{k}{\varepsilon})$ , $n^{1/2+O(\varepsilon)}$ , $\operatorname{poly}(\log n)$
time	$O\left(\sqrt{n \cdot \operatorname{poly}(\varepsilon)}\right)$	2 poly(log $n$ )
query	$O\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k \log n}{n})\right)$	$m^{1/2+O(\varepsilon)}$ , $poly(k \log n)$
time	$O\left(\sqrt{n \cdot \operatorname{poly}(-\varepsilon)}\right)$	$n \mapsto \operatorname{poly}(\underline{-\varepsilon})$
misclassification	$O(k, \sqrt{2})$	$O(\log k \cdot \epsilon)$ per cluster
error (fraction)	$\int O(h\sqrt{\varepsilon})$	

Background 000000●		

#### Motivation

	[Pen20]	[GKL <sup>+</sup> 21]
conductance	$\varphi^2$	$\varphi^{3}$
gap	$\mathcal{E} \ll \overline{\operatorname{poly}(k) \cdot \log n}$	$\mathcal{E} \swarrow \overline{\log k}$
preprocessing	$O\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k \log n}{n})\right)$	$\operatorname{poly}(\frac{k}{\varepsilon})$ , $n^{1/2+O(\varepsilon)}$ , $\operatorname{poly}(\log n)$
time	$O\left(\sqrt{n \cdot \operatorname{poly}(-\varepsilon)}\right)$	$2^{-1}$ (log $n$ )
query	$O\left(\sqrt{n} \cdot \operatorname{poly}\left(\frac{k \log n}{n}\right)\right)$	$n^{1/2+O(\varepsilon)}$ , $noly(k \log n)$
time	$O\left(\sqrt{n \cdot \operatorname{poly}(-\varepsilon)}\right)$	$n \to poly(-\varepsilon)$
misclassification	$O(k \sqrt{c})$	$O(\log k \cdot s)$ per cluster
error (fraction)	$O(h\sqrt{c})$	$O(\log \kappa \cdot \varepsilon)$ per cluster

Can we get a spectral clustering oracle with

- ▶ better conductance gap than [Pen20] and
- ▶ better preprocessing time than [GKL<sup>+</sup>21]?

	Results ●O	
<b>A B</b>		

#### Our Results

	[Pen20]	[GKL+21]	this work
conductance	$\varphi^2$	$\varphi^3$	$\varphi^2$
gap	$c \propto \overline{\operatorname{poly}(k) \cdot \log n}$	$c \sim \log k$	$c \ll poly(k)$
preprocessing	$\widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{2})\right)$	$\widetilde{O}(2^{\operatorname{poly}(\frac{k}{\varepsilon})}).$	$\widetilde{O}(\operatorname{poly}(k) \cdot$
time	$O((v^n \operatorname{poly}(\varepsilon)))$	$n^{1/2+O(\varepsilon)})$	$n^{1/2+O(\varepsilon)})$
query	$\widetilde{O}\left(\sqrt{m} - n \operatorname{sl}(k)\right)$	$\widetilde{O}(n^{1/2+O(\varepsilon)}.$	$\widetilde{O}(n^{1/2+O(\varepsilon)}.$
time	$O\left(\sqrt{n} \cdot \operatorname{poly}(\frac{-}{\varepsilon})\right)$	$\operatorname{poly}(\frac{k}{\varepsilon}))$	$\operatorname{poly}(k))$
misclassification error (fraction)	$O\left(k\sqrt{\varepsilon} ight)$	$O\left(\log k \cdot \varepsilon\right)$ per cluster	$O\left(\mathrm{poly}(k)\cdotarepsilon^{1/3} ight)$ per cluster

	Results ●O	
<b>•</b> •		

#### Our Results

	[Pen20]	[GKL+21]	this work
conductance	$\varphi^2$	$\varphi^3$	$\varphi^2$
gap	$c \propto \overline{\operatorname{poly}(k) \cdot \log n}$	$\mathcal{E} = \log k$	$c \ll \overline{\operatorname{poly}(k)}$
preprocessing time	$\widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{\varepsilon})\right)$	$\widetilde{O}(\frac{2^{\operatorname{poly}(rac{k}{\varepsilon})}}{n^{1/2+O(\varepsilon)}})$	$\widetilde{O}( rac{ ext{poly}(k)}{n^{1/2+O(arepsilon)}})$
		$\widetilde{\alpha}(-1/2+O(c))$	$\widetilde{O}(-1/2+O(c))$
query	$\widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{2})\right)$	$O(n^{1/2+O(\varepsilon)})$	$O(n^{1/2+O(\varepsilon)})$
time		$\operatorname{poly}(\frac{\kappa}{\varepsilon}))$	$\operatorname{poly}(k))$
misclassification error (fraction)	$O\left(k\sqrt{arepsilon} ight)$	$O\left(\log k \cdot arepsilon ight)$ per cluster	$O\left(\mathrm{poly}(k)\cdot arepsilon^{1/3} ight)$ per cluster

- ► Conductance gap: poly(k)
  - better than  $poly(k) \cdot \log n$  in [Pen20]
  - a slightly worse than  $\log k$  in [GKL+21]

	Results ●O	
<b>•</b> •		

#### Our Results

[Pen20]	[GKL+21]	this work
$\varepsilon \ll \frac{\varphi^2}{\varphi^2}$	$\varepsilon \ll \frac{\varphi^3}{2}$	$\varepsilon \ll \frac{\varphi^2}{\varphi^2}$
$\sim \qquad \operatorname{poly}(k) \cdot \log n$	$c \sim \log k$	$c \sim \operatorname{poly}(k)$
$\widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{2})\right)$	$\widetilde{O}(2^{\operatorname{poly}(\frac{k}{\varepsilon})}).$	$\widetilde{O}(\mathrm{poly}(k)\cdot$
$O\left(\sqrt{n} \operatorname{poly}(\varepsilon)\right)$	$n^{1/2+O(\varepsilon)})$	$n^{1/2+O(\varepsilon)})$
$\widetilde{O}\left(\sqrt{k}, \frac{1}{k}\right)$	$\widetilde{O}(n^{1/2+O(\varepsilon)}\cdot$	$\widetilde{O}(n^{1/2+O(\varepsilon)}.$
$O\left(\sqrt{n} \cdot \operatorname{poly}(\frac{1}{\varepsilon})\right)$	$\operatorname{poly}(\frac{k}{\varepsilon}))$	$\operatorname{poly}(k))$
$O\left(k\sqrt{\varepsilon} ight)$	$O\left(\log k \cdot \varepsilon\right)$ per cluster	$O\left(\mathrm{poly}(k)\cdot\varepsilon^{1/3} ight)$ per cluster
	$[Pen20]$ $\varepsilon \ll \frac{\varphi^2}{\operatorname{poly}(k) \cdot \log n}$ $\widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{\varepsilon})\right)$ $\widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{\varepsilon})\right)$ $O\left(k\sqrt{\varepsilon}\right)$	$\begin{array}{ c c c } \hline [Pen20] & [GKL^+21] \\ \hline \varepsilon \ll \frac{\varphi^2}{\operatorname{poly}(k) \cdot \log n} & \varepsilon \ll \frac{\varphi^3}{\log k} \\ \hline \widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{\varepsilon})\right) & \frac{\widetilde{O}(2^{\operatorname{poly}(\frac{k}{\varepsilon})} \cdot n^{1/2 + O(\varepsilon)})}{n^{1/2 + O(\varepsilon)}} \\ \hline \widetilde{O}\left(\sqrt{n} \cdot \operatorname{poly}(\frac{k}{\varepsilon})\right) & \frac{\widetilde{O}(n^{1/2 + O(\varepsilon)} \cdot n^{1/2 + O(\varepsilon)} \cdot n^{1/2 + O(\varepsilon)} \cdot n^{1/2 + O(\varepsilon)})}{\operatorname{poly}(\frac{k}{\varepsilon}))} \\ \hline O\left(k\sqrt{\varepsilon}\right) & \frac{O\left(\log k \cdot \varepsilon\right)}{\operatorname{per cluster}} \\ \end{array}$

- Conductance gap: poly(k)
  - better than  $poly(k) \cdot \log n$  in [Pen20]
  - a slightly worse than  $\log k$  in [GKL<sup>+</sup>21]

• Preprocessing time: polynomial in k, better than exponential in [GKL+21]

Ranran Shen, Pan Peng

	Results O●	
Our Results		

Our oracle is robust against a few edge deletions.

#### Theorem (Robust; Informal)

Let  $G_0 = (V, E)$  be a  $(k, \varphi, \varepsilon)$ -clusterable graph, where  $\frac{\varepsilon}{\varphi^4} \ll \frac{1}{\operatorname{poly}(k)}$ .

• G is obtained from  $G_0$  by deleting at most  $O(d\varphi^2)$  edges in each cluster, or

• G is obtained from  $G_0$  by randomly deleting at most  $O(\frac{kd^2}{d+\log k})$  edges in  $G_0$ Then w.h.p., there exists a clustering oracle for G with the same guarantees as presented in above table.

10 / 13

# Our Technique: A Nice Gap

 $f_x$ : spectral embedding of  $x \in V$ .  $\langle f_x, f_y \rangle$ : dot product of  $f_x$  and  $f_y$ .

#### Lemma 1 (Informal)

For most vertices in a  $(k,\varphi,\varepsilon)\text{-clusterable graph,}$ 

- if x and y are in the same cluster, then  $\langle f_x, f_y \rangle$  is close to  $O(\frac{k}{n})$
- if x and z are in the different clusters, then  $\langle f_x, f_z \rangle$  is close to 0.



#### Example: dot product gap

		Experiments ●0
Experiments		

Input graph: generated by SBM

► can handle graphs with a smaller conductance gap than [CPS15]

p	0.02	0.025	0.03	0.035	0.04	0.05	0.06	0.07
error([CPS15])	-	0.6208	0.4970	0.1996	0.0829	0.0168	0.0030	0.0003
error (this work)	0.3887	0.0030	0.0004	<mark>0</mark>	<mark>0</mark>	<mark>0</mark>	<mark>0</mark>	0

robust against a few adversarial edge deletions

delNum	0	25	32	40	45	50	55	60	65
error	0.0	0.00007	0.00007	0.00013	0.00047	0.00080	0.00080	0.00080	0.00087

	Experiments O●

# Thanks!

References:

[CPS15] Czumaj A, Peng P, Sohler C. Testing cluster structure of graphs. STOC 2015.

[Pen20] Peng P. Robust clustering oracle and local reconstructor of cluster structure of graphs. SODA 2020.

[GKL<sup>+</sup>21] Gluch G, Kapralov M, Lattanzi S, Mousavifar A and Sohler C. Spectral clustering oracles in sublinear time. SODA 2021.