Effective and Scalable Clustering on Massive Attributed Graphs

Renchi Yang, Jieming Shi, Yin Yang,
Keke Huang, Shiqi Zhang, Xiaokui Xiao

April 2021

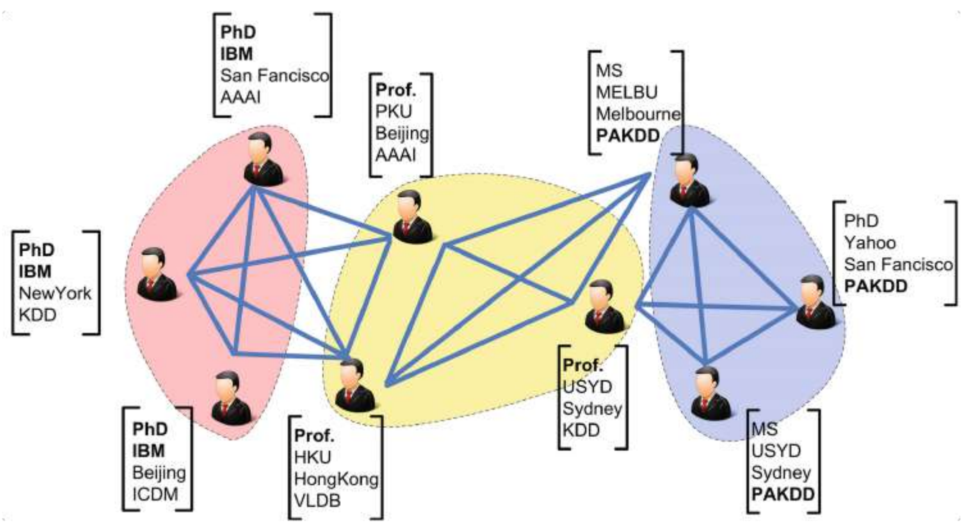
The Web Conference 2021

Outline

- Problem Definition
- Existing Work
- Challenges
- Objective
- Proposed ACMin
- Experiments

Problem Definition

- Given an attributed graph $G(V, R, E_V, E_R)$ and k, ***k*-AGC** (k-Attributed Graph Clustering) aims to partition the node set V of G into *disjoint* subsets: $C_1, C_2, \cdots, C_k$ such that
  - nodes in the *same* cluster C_i are *close*, otherwise distant
  - nodes in the *same* cluster C_i have *similar* attributes



V: node set, $|V|=n$

R: attribute set

E_V: edge set

E_R: node-attribute association set

Existing Work

- Edge-weighted-based clustering
  - Build a weighted graph $\hat{G}$ (weight is attribute similarity)
  - Apply classic graph clustering on $\hat{G}$
  - **No multi-hop information** → inferior clustering quality!

- Distance-based clustering
  - Build a distance matrix **M** based on attributes/topology
  - Apply classic data clustering (e.g., k-means) on **M**
  - $\boldsymbol{O(n^2)}$ **time & space** → inefficient & not scalable!

Existing Work

- Probabilistic-model-based clustering
  - Assume structure, attributes, & clusters $\backsim$ a distribution
  - Infer a probabilistic model
  - **Costly optimization process** → inefficient & not scalable!

- Embedding-based methods.
  - Learn an embedding per node
  - Apply *k*-means on the embeddings
  - Are **not specially designed for clustering** & rely on embedding quality → suboptimal clustering quality!

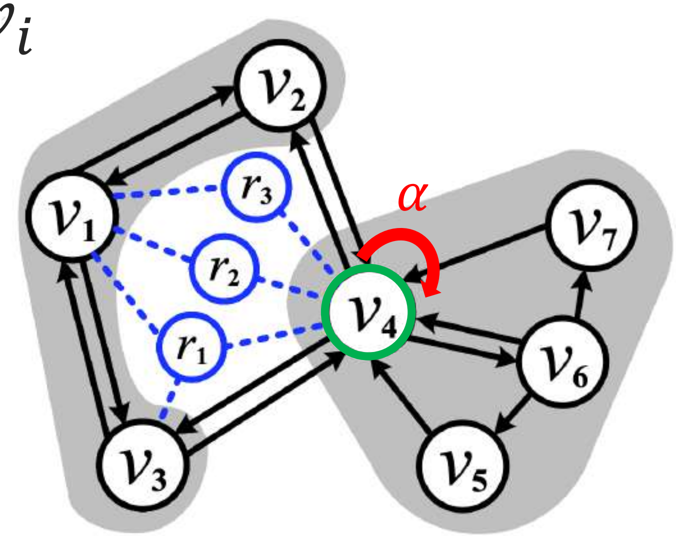
Challenges

? Formulate a quantitative objective to k-AGC which
- aims to optimize the clustering quality
- considers *multi-hop* (topology & attribute) relationships between nodes

? Design techniques to solve the objective such that
- $O(n^2)$ materialization cost is *not needed*
- optimization process can be done *efficiently*

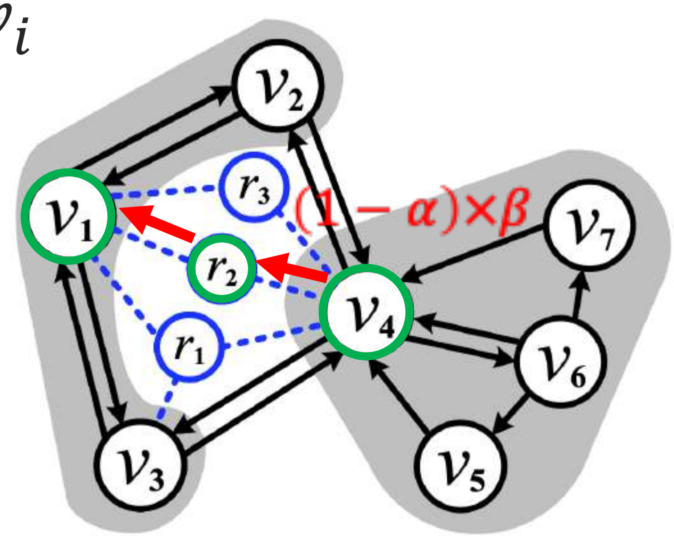
Objective: Attributed Random Walk (ARW) Model

- At each step, an ARW from node v_i
  - w.p. α, stops at current node v_j



Objective: Attributed Random Walk Model

- At each step, an ARW from node v_i
  - w.p. α, stops at current node v_j
  - w.p. $1-\alpha$, jumps to
    - w.p. β, a node v_l via an attribute w.p. $\mathrm{P}_R[v_j, v_l]$

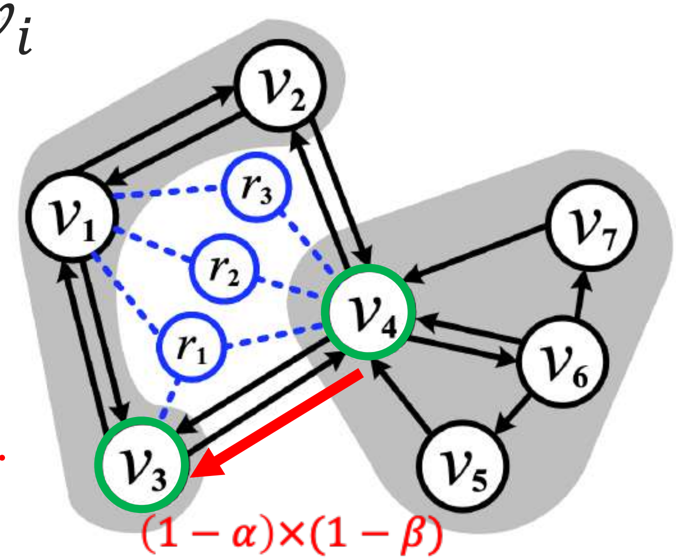


? $$\mathrm{P}_R[v_i, v_j] = \frac{\mathbf{R}[v_i]\cdot\mathbf{R}[v_j]^\top}{\sum_{v_l\in V}\mathbf{R}[v_i]\cdot\mathbf{R}[v_l]^\top}$$

Normalized attribute similarity of two modes

Objective: Attributed Random Walk Model

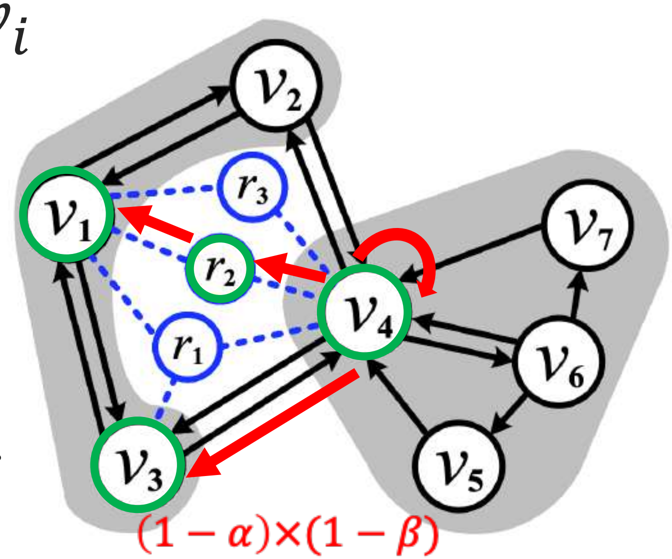
- At each step, an ARW from node v_i
  - w.p. α, stops at current node v_j
  - w.p. $1-\alpha$, jumps to
    - w.p. β, a node v_l via an attribute w.p. $\mathrm{P}_R[v_j, v_l]$
    - w.p. $1-\beta$, an out-neighbor v_l of v_j w.p. $\mathrm{P}_V[v_j, v_l]$



$$\mathrm{P}_V[v_j, v_l] = \frac{edge\ weight\ of\ (v_j, v_l)}{\sum_{v_x} edge\ weight\ of\ (v_j, v_x)}$$

Objective: Attributed Random Walk Model

- At each step, an ARW from node v_i
  - w.p. α, stops at current node v_j
  - w.p. $1-\alpha$, jumps to
    - w.p. β, a node v_l via an attribute w.p. $P_R[v_j, v_l]$
    - w.p. $1-\beta$, an out-neighbor v_l of v_j w.p. $P_V[v_j, v_l]$



The probability that a ARW from v_i stopping at v_j:

$$\mathbf{S}[v_i, v_j] = \alpha \sum_{\ell=0}^{\infty} (1-\alpha)^{\ell} \cdot ((1-\beta) \cdot \mathbf{P}_V + \beta \cdot \mathbf{P}_R)^{\ell} [v_i, v_j]$$

Objective: Average Attributed Multi-Hop Conductance (AAMC)

- Conductance

|cut(C)|: #edges crossing C and other clusters

|vol(C)|: #edges of nodes within C

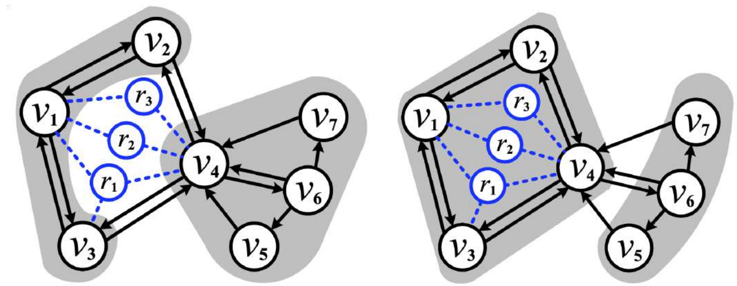
$$\widehat{\Phi}(C) = \frac{|\mathrm{cut}(C)|}{\min\{\mathrm{vol}(C), \mathrm{vol}(V \backslash C)\}}$$

The ratio of edges crossing C

- AAMC

$$\Phi(C) = \sum_{v_i \in C, v_j \in V \backslash C} \frac{\mathrm{S}[v_i, v_j]}{|C|}$$

the expected portion of attributed random walks escaping from C



(a) Clusters C_1, C_2 **(b) Clusters** C_1', C_2'

v4 is mutually connected to & shares 3 attributes with v2, v3

v2, v3 & v4 should be in the same cluster

Avg. conductance:
(a) 4/12; < (b) 4/10. ✗

AAMC:
(a) 0.123; > (b) 0.105. ✓

Objective: Objective Function

- Find k clusters $C_1, \cdots, C_k$ s.t. AAMC is minimized

$$\phi^* = \min_{C_1, C_2, \cdots, C_k} \frac{\sum_{i=1}^{k} \Phi(C_i)}{k}$$

$$\Downarrow$$

$$\phi^* = \min_{\mathbf{Y} \in \mathbb{1}^{k \times n}} \frac{2}{k} \cdot \mathrm{trace}\left(\left((\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}}\mathbf{Y}\right) \cdot (\mathbf{I} - \mathbf{S}) \cdot \left((\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}}\mathbf{Y}\right)^\top\right)$$

$$\mathbf{Y}[C_i, v_j] = \begin{cases} 1 & v_j \in C_i, \\ 0 & v_j \in V \setminus C_i, \end{cases}$$

Proposed ACMin: Basic Idea

$$\phi^* = \min_{\mathbf{Y}\in\mathbb{1}^{k\times n}} \frac{2}{k}\cdot \mathrm{trace}\left(\left(\boxed{(\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}}\mathbf{Y}}\right)\cdot(\mathbf{I}-\mathbf{S})\cdot\left(\boxed{(\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}}\mathbf{Y}}\right)^\top\right)$$

$$\cancel{\mathbf{Y}[C_i, v_j] = \begin{cases} 1 & v_j \in C_i, \\ 0 & v_j \in V \setminus C_i, \end{cases}}$$

1

$$\phi^* = \min \frac{2}{k}\cdot \mathrm{trace}(\boxed{\mathbf{F}}\cdot(\mathbf{I}-\mathbf{S})\cdot\boxed{\mathbf{F}}^\top)$$

optimal when F is the top-*k* eigenvectors of S

2

find Y such that $(\mathbf{Y}\mathbf{Y}^\top)^{-1/2}\mathbf{Y}$ approximates F

$$\min \|\mathbf{X}\mathbf{F} - (\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}}\mathbf{Y}\|_F^2 \quad s.t.\ \mathbf{Y}\in\mathbb{1}^{k\times n},\ \mathbf{X}^\top\mathbf{X}=\mathbf{I}$$

Proposed ACMin: Find F

- Compute $\mathbf{S} = \alpha \sum_{\ell=0}^{\infty} (1-\alpha)^{\ell} \cdot ((1-\beta) \cdot \mathbf{P}_V + \beta \cdot \mathbf{P}_R)^{\ell}$
- Compute the top-k eigenvectors $\mathbf{F}$ of $\mathbf{S}$

$$O(|E_V| \cdot |V| + |R| \cdot |V|^2)!$$

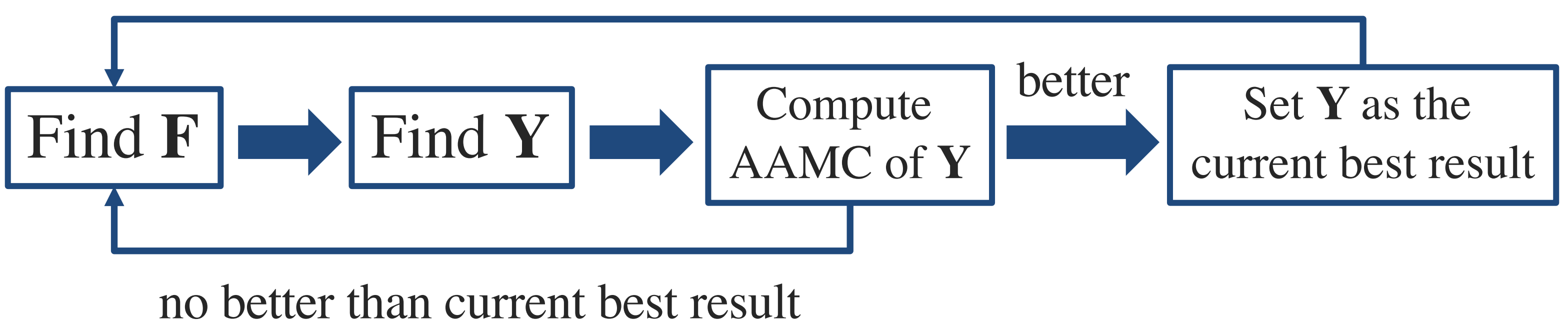
- $\mathbf{F}$ is the top-k eigenvectors of $(1-\beta) \cdot \mathbf{P}_V + \beta \cdot \mathbf{P}_R$!
- $\mathbf{P}_R = \widehat{\mathbf{R}}\mathbf{R}^\top$ $\quad \widehat{\mathbf{R}}[v_i] = \frac{\mathbf{R}[v_i]}{\mathbf{R}[v_i] \cdot \mathbf{r}^\top}\ \forall v_i \in V$, where $\mathbf{r} = \sum_{v_j \in V} \mathbf{R}[v_j]$
- **for** $\ell \leftarrow 1$ *to* t_e **do**
  - $\mathbf{Z}_\ell \leftarrow (1-\beta) \cdot \mathbf{P}_V \mathbf{F}_{\ell-1}^\top + \beta \cdot \widehat{\mathbf{R}}(\mathbf{R}^\top \mathbf{F}_{\ell-1}^\top)$;
  - $\mathbf{F}_\ell \leftarrow \mathrm{QR}(\mathbf{Z}_\ell)$;

orthogonal iterations

$$O(k \cdot (|E_V| + |E_R|))!$$

Proposed ACMin: Find Y

- Alternative optimization $O(k^2 \cdot |V|)!$
  - Updating Y with X fixed: $\max_{\mathbf{Y} \in \mathbb{1}^{K \times n}} \operatorname{trace}((\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}} \mathbf{Y}\mathbf{F}^\top \mathbf{X}^\top)$
  - Updating X with Y fixed: $\max_{\mathbf{X}^\top \mathbf{X} = \mathbf{I}} \operatorname{trace}((\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}} \mathbf{Y}\mathbf{F}^\top \mathbf{X}^\top)$



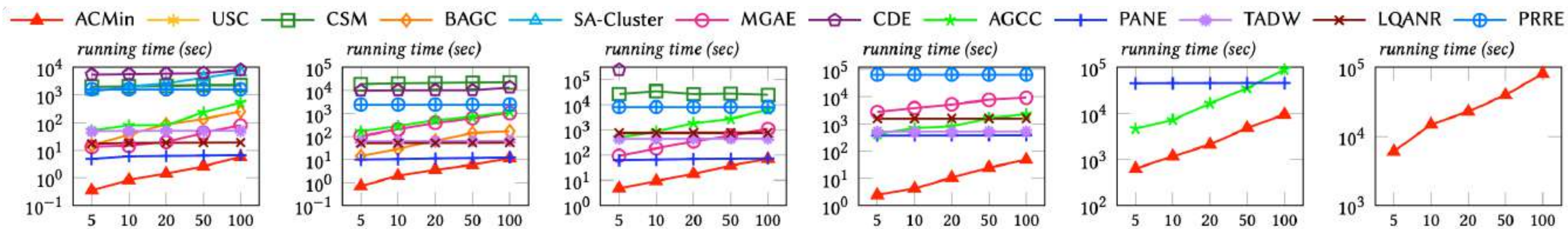
Experiments: Datasets and Setup

Table 2: Datasets. (K=10^3, M=10^6, B=10^9)

Name	$\|V\|$	$\|E_V\|$	$\|R\|$	$\|E_R\|$	$\|C\|$
Cora [29, 52, 53, 60, 64]	2.7K	5.4K	1.4K	49.2K	7
Citeseer [29, 52, 53, 60, 64]	3.3K	4.7K	3.7K	105.2K	6
Pubmed [52, 60, 64, 66]	19.7K	44.3K	0.5K	988K	3
Flickr [24, 29, 34, 58, 60]	7.6K	479.5K	12.1K	182.5K	9
TWeibo [60]	2.3M	50.7M	1.7K	16.8M	8
MAG-Scholar-C [3]	10.5M	265.2M	2.78M	1.1B	8

- $\alpha = 0.2, \beta = 0.35, \#iterations = 200$
- Competitors
  - Distance-based: *CSM*, *SA-Cluster*
  - Probabilistic-model-based: *BAGC*
  - embedding-based (dim=128):
    MGAE, *CDE*, *AGCC*, *TADW*, *PANE*, *LQANR*, *PRRE*

Experiments: Efficiency



(a) *Cora* (b) *Citeseer* (c) *Pubmed* (d) *Flickr* (e) *TWeibo* (f) *MAG-Scholar-C*

Figure 2: Running time with varying k (best viewed in color).

- $k = 5, 10, 20, 50, 100$
- y-axis is in log-scale
- ACMin is by up to orders of magnitude faster
- ACMin is the only method able to handle MAG-Scholar-C (1.68 hours when $k = 5$)

Experiments: Clustering Quality with Ground-truth

Table 3: CA, NMI and AAMC with ground-truth (Large CA, NMI, and small AAMC indicate high clustering quality).

Solution	*Cora*			*Citeseer*			*Pubmed*			*Flickr*			*TWeibo*			*MAG-Scholar-C*		
	CA	NMI	AAMC	CA	NMI	AAMC	CA	NMI	AAMC	CA	NMI	AAMC	CA	NMI	AAMC	CA	NMI	AAMC
Ground-truth	1.0	1.0	0.546	1.0	1.0	0.531	1.0	1.0	0.505	1.0	1.0	0.691	1.0	1.0	0.719	1.0	1.0	0.63
TADW	0.554	0.402	0.593	0.539	0.333	0.569	0.483	0.096	0.55	0.16	0.062	0.733	-	-	-	-	-	-
LQANR	0.64	0.492	0.559	0.587	0.374	0.549	0.403	0.022	0.612	0.127	0.002	0.739	-	-	-	-	-	-
PRRE	0.547	0.396	0.604	0.576	0.322	0.592	0.62	0.269	0.518	0.454	0.321	0.713	-	-	-	-	-	-
PANE	0.601	0.462	0.577	0.677	0.421	0.537	0.618	0.252	0.512	0.402	0.265	0.708	0.215	0.004	0.752	-	-	-
CSM	0.308	0.149	0.612	0.247	0.11	0.615	0.393	0.022	0.565	-	-	-	-	-	-	-	-	-
SA-Cluster	0.001	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BAGC	0.001	0.134	-	0.183	0	-	-	-	-	-	-	-	-	-	-	-	-	-
MGAE	0.633	0.456	0.571	0.661	0.408	0.545	0.419	0.076	0.556	0.266	0.109	0.729	-	-	-	-	-	-
CDE	0.473	0.332	0.581	0.535	0.318	0.571	0.663	0.259	0.547	0.254	0.11	0.714	-	-	-	-	-	-
AGCC	0.642	0.496	0.553	0.668	0.409	0.526	0.668	0.272	0.492	0.471	0.369	0.706	0.406	0.007	0.723	-	-	-
USC	0.635	0.455	0.706	0.495	0.326	0.682	0.548	0.212	0.614	-	-	-	-	-	-	-	-	-
ACMin	**0.656**	**0.498**	**0.544**	**0.68**	**0.422**	**0.525**	**0.691**	**0.308**	**0.487**	**0.757**	**0.608**	**0.698**	**0.408**	**0.01**	**0.686**	**0.659**	**0.497**	**0.57**

- CA: clustering accuracy w.r.t. ground truth labels
- NMI: normalized mutual information
- AAMC

Experiments: Clustering Quality without Ground-truth

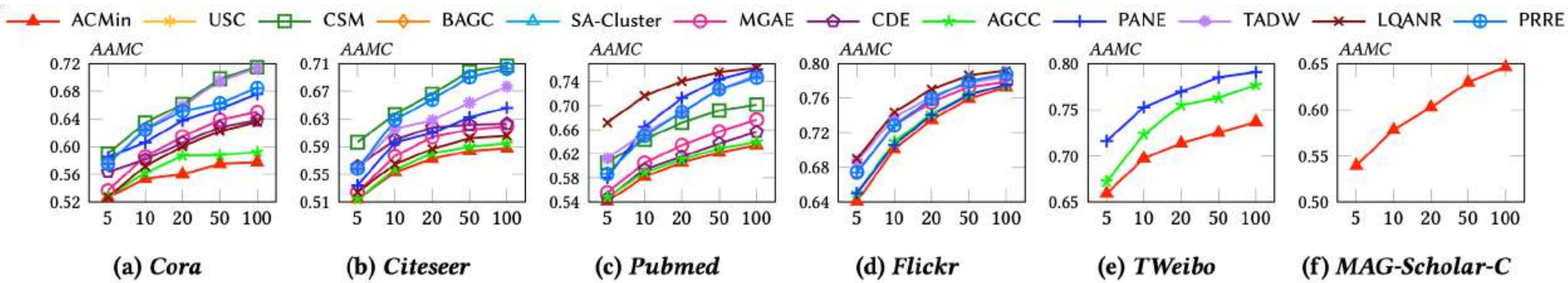


Figure 3: AAMC with varying k (best viewed in color).

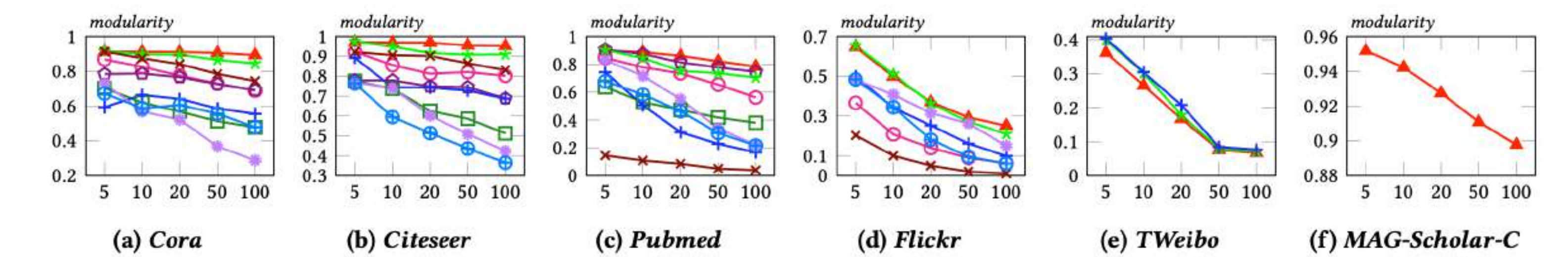


Figure 4: Modularity with varying k (best viewed in color).

- $k = 5, 10, 20, 50, 100$
- AAMC & modularity

Thank You!

Comparison with Spectral Clustering

- Spectral clustering applies *k*-means to generate **Y**
- Spectral clustering optimizes

$$\frac{2}{k} \cdot \mathrm{trace}(\mathbf{F}\mathbf{Y}^\top(\mathbf{Y}\mathbf{Y}^\top)^{-1}\mathbf{Y} \cdot (\mathbf{I}-\mathbf{S}) \cdot (\mathbf{F}\mathbf{Y}^\top(\mathbf{Y}\mathbf{Y}^\top)^{-1}\mathbf{Y})^\top)$$

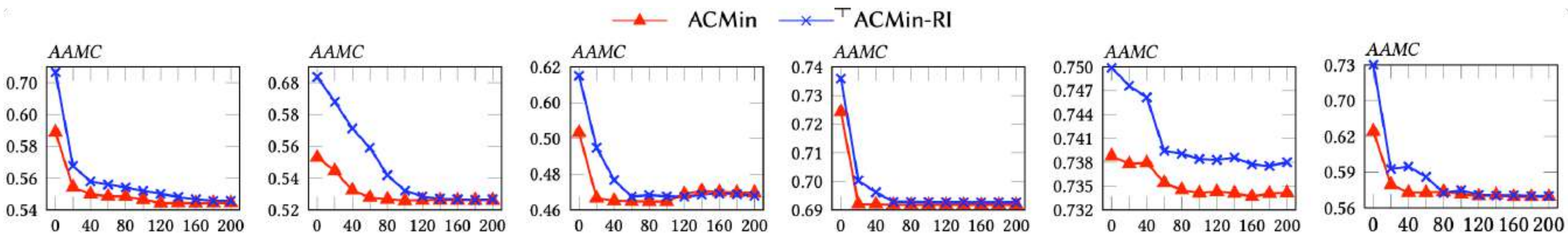
where **F** is the top-k eigenvectors of **S**,

- In contrast, ACMin optimizes

$$\phi^* = \min_{\mathbf{Y}\in\mathbb{1}^{k\times n}} \frac{2}{k} \cdot \mathrm{trace}(((\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}}\mathbf{Y}) \cdot (\mathbf{I}-\mathbf{S}) \cdot ((\mathbf{Y}\mathbf{Y}^\top)^{-\frac{1}{2}}\mathbf{Y})^\top)$$

$$\mathbf{Y}[C_i, v_j] = \begin{cases} 1 & v_j \in C_i, \\ 0 & v_j \in V \setminus C_i, \end{cases}$$

Experiments: Convergence Analysis



(a) *Cora* (b) *Citeseer* (c) *Pubmed* (d) *Flickr* (e) *TWeibo* (f) *MAG-Scholar-C*

Figure 5: AAMC with varying t_e (best viewed in color).

- #iterations = 0,20,40,60,80,100,120,140,160,180,200
- ACMin-RI: ACMin without effective initialization of **F**