Non-Exhaustive, Overlapping Clustering

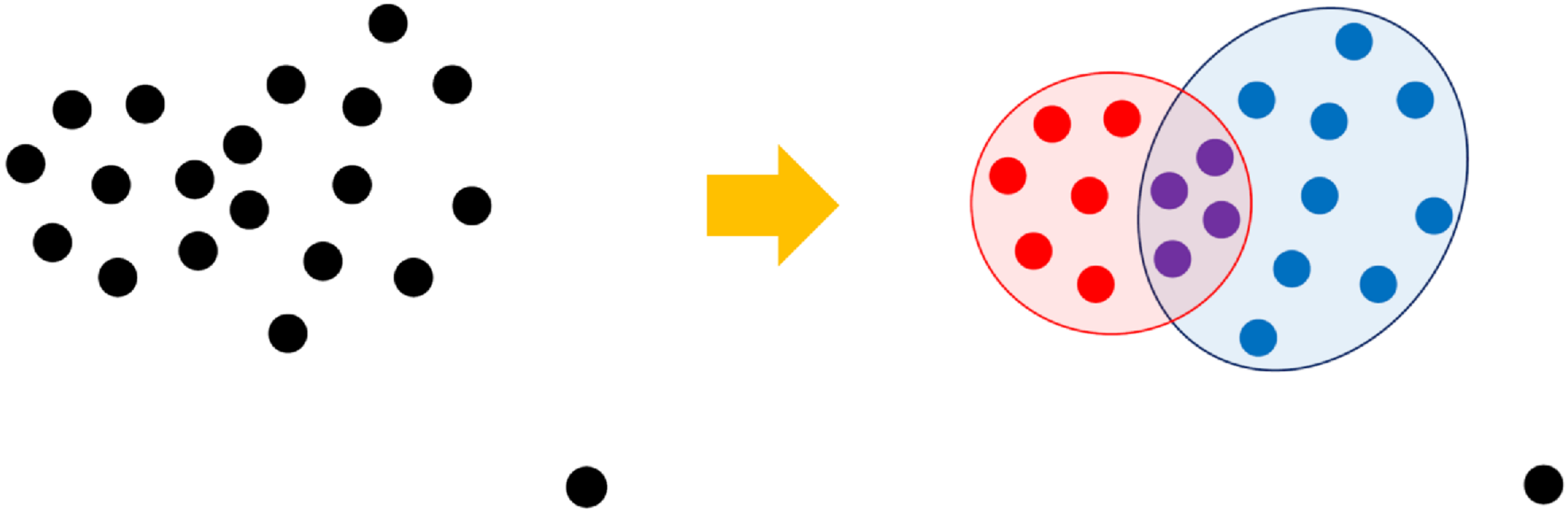
J. J. Whang, Y. Hou, D. F. Gleich, and I. S. Dhillon

IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

November 2019

Clustering

- Traditional disjoint, exhaustive clustering
  - Every single data point is assigned to exactly one cluster.
- **Non-exhaustive, overlapping clustering**
  - A data point is allowed to be outside of any cluster, and clusters can overlap.



K-Means Clustering

- K-Means seeks k clusters $C_1, C_2, \cdots, C_k$ in $X = \{x_1, x_2, \cdots, x_n\}$
  - $C_i \cap C_j = \emptyset \; \forall i \neq j$ **(disjoint)** and $C_1 \cup C_2 \cup \cdots \cup C_k = X$ **(exhaustive)**

- K-Means objective function

$$\min_{\{\mathcal{C}_j\}_{j=1}^k} \sum_{j=1}^{k} \sum_{\mathbf{x}_i \in \mathcal{C}_j} \|\mathbf{x}_i - \mathbf{m}_j\|^2, \text{ where } \mathbf{m}_j = \frac{\sum_{\mathbf{x}_i \in \mathcal{C}_j} \mathbf{x}_i}{|\mathcal{C}_j|}$$

- K-Means algorithm
  - Repeatedly assigning data points to their closest clusters and recomputing centers.

NEO-K-Means Objective

- NEO-K-Means (Non-Exhaustive, Overlapping K-Means)
- Assignment matrix $U = [u_{ij}]_{n\times k}$
  - $u_{ij} = 1$ if x_i belongs to cluster j
  - $u_{ij} = 0$ if x_i does not belong to cluster j

$$U = \begin{array}{cc} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} & \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} & \begin{array}{c} x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \end{array} \quad U^T U = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad U\mathbf{1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \boxed{\begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \end{bmatrix}}$$

cluster sizes

no. of clusters a data point belongs to

NEO-K-Means Objective

$$\min_{U} \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \underline{\|\mathbf{x}_i - \mathbf{m}_j\|^2}, \text{ where } \mathbf{m}_j = \frac{\sum_{i=1}^{n} u_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}}$$

$$\text{s.t. } \underline{trace(U^T U) = (1+\alpha)n}, \ \underline{\sum_{i=1}^{n} \mathbb{I}\{(U\mathbf{1})_i = 0\} \leq \beta n}.$$

Minimize the distance between a data point and its cluster center.
Add two constraints to control overlap and non-exhaustiveness.

- α: overlap, β: non-exhaustiveness
- $\alpha = 0, \beta = 0$: equivalent to the traditional K-Means objective

NEO-K-Means Objective

$$\min_{U} \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|^2, \text{ where } \mathbf{m}_j = \frac{\sum_{i=1}^{n} u_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}}$$

$$\text{s.t.} \quad \underline{trace(U^T U) = (1+\alpha)n}, \quad \underline{\textstyle\sum_{i=1}^{n} \mathbb{I}\{(U\mathbf{1})_i = 0\} \le \beta n}.$$

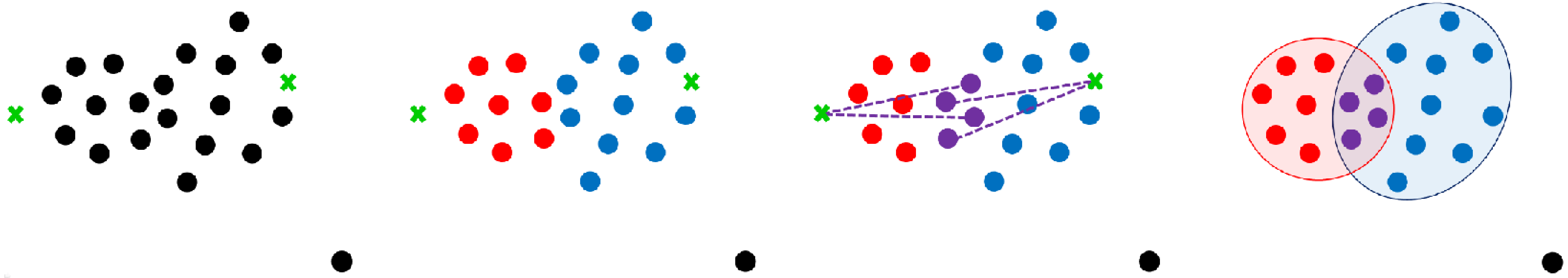
$(1+\alpha)n$ assignments are made.

At most βn data points can have no membership in any cluster.

- α: overlap, β: non-exhaustiveness
- $\alpha = 0, \beta = 0$: equivalent to the traditional K-Means objective

NEO-K-Means Algorithm

- A simple iterative algorithm that **monotonically decreases the NEO-K-Means objective**.
- Example ($n = 20, \alpha = 0.15, \beta = 0.05$)
  - Assign $n - \beta n$ $(= 19)$ data points to their closest clusters.
  - Make $\beta n + \alpha n$ $(= 4)$ assignments by taking minimum distances.



NEO-K-Means via LRSDP

- The NEO-K-Means algorithm
  - Fast iterative algorithm, but susceptible to initialization
- LRSDP initialization
  - Make the NEO-K-Means get **more accurate and reliable solutions**

Fast and scalable
Trapped in local optima

Slow and not scalable
Globally optimized

Faster than CVX
Locally optimized

Semidefinite Programs (SDPs)

- Semidefinite Programming (SDP)
  - Convex problem → globally optimized via off-the-shelf SDP solvers
  - **Problems with fewer than 100 data points**
- Low-rank SDP
  - Non-convex → locally optimized via an augmented Lagrangian method
  - **Problems with tens of thousands of data points**

Canonical SDP

$$\begin{aligned} \text{maximize} \quad & \operatorname{trace}(\boldsymbol{C}\boldsymbol{X}) \\ \text{subject to} \quad & \boldsymbol{X} \succeq 0, \boldsymbol{X} = \boldsymbol{X}^T, \\ & \operatorname{trace}(\boldsymbol{A}_i \boldsymbol{X}) = b_i \\ & i = 1, \ldots, m \end{aligned}$$

Low-rank SDP

$$\begin{aligned} \text{maximize} \quad & \operatorname{trace}(\boldsymbol{C}\boldsymbol{Y}\boldsymbol{Y}^T) \\ \text{subject to} \quad & \boldsymbol{Y} : n \times k \\ & \operatorname{trace}(\boldsymbol{A}_i \boldsymbol{Y}\boldsymbol{Y}^T) = b_i \\ & i = 1, \ldots, m \end{aligned}$$

NEO-K-Means as an SDP

$$U = \begin{array}{cc} & \begin{array}{cc} c_1 & c_2 \end{array} \\ \begin{array}{c} \\ \\ \\ \\ \end{array} & \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \end{array} \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array}$$

$$\mathrm{f} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 0 \end{bmatrix}$$

Overlap

$$g = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Non-exhaustiveness

Co-occurrence matrix

$$Z = \begin{bmatrix} \frac{w_1{}^2}{w_1 + w_2} & \frac{w_1 w_2}{w_1 + w_2} & 0 & 0 \\ \frac{w_2 w_1}{w_1 + w_2} & \frac{w_2{}^2}{w_1 + w_2} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{w_2{}^2}{w_2 + w_3} & \frac{w_2 w_3}{w_2 + w_3} & 0 \\ 0 & \frac{w_3 w_2}{w_2 + w_3} & \frac{w_3{}^2}{w_2 + w_3} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

SDP-like Formulation for NEO-K-Means

- NEO-K-Means with a **discrete assignment** matrix (Non-convex, combinatorial problem)

$$
\begin{aligned}
\underset{\mathbf{Z},\mathbf{f},\mathbf{g}}{\text{maximize}} \quad & \operatorname{trace}(\mathbf{KZ}) - \mathbf{f}^T\mathbf{d} \\
\text{subject to} \quad & \boxed{\begin{aligned} & \operatorname{trace}(\mathbf{W}^{-1}\mathbf{Z}) = k, && (a) \\ & Z_{ij} \geq 0, && (b) \\ & \mathbf{Z} \succeq 0, \mathbf{Z} = \mathbf{Z}^T && (c) \end{aligned}} \\
& \boxed{\begin{aligned} & \mathbf{Ze} = \mathbf{Wf}, && (d) \\ & \mathbf{e}^T\mathbf{f} = (1+\alpha)n, && (e) \\ & \mathbf{e}^T\mathbf{g} \geq (1-\beta)n, && (f) \\ & \mathbf{f} \geq \mathbf{g}, && (g) \end{aligned}} \\
& \boxed{\begin{aligned} & \operatorname{rank}(\mathbf{Z}) = k, && (h) \\ & \mathbf{f} \in \mathcal{Z}^n_{\geq 0}, \mathbf{g} \in \{0,1\}^n. && (i) \end{aligned}}
\end{aligned}
$$

Z must arise from an assignment matrix

Overlap & non-exhaustiveness constraints

Combinatorial problem

SDP for NEO-K-Means

- **Convex relaxation** of NEO-K-Means

$$
\begin{aligned}
\underset{\mathbf{Z},\mathbf{f},\mathbf{g}}{\text{maximize}} \quad & \text{trace}(\mathbf{KZ}) - \mathbf{f}^T\mathbf{d} & \\
\text{subject to} \quad & \text{trace}(\mathbf{W}^{-1}\mathbf{Z}) = k, & (a) \\
& Z_{ij} \geq 0, & (b) \\
& \mathbf{Z} \succeq 0, \mathbf{Z} = \mathbf{Z}^T & (c) \\
& \mathbf{Ze} = \mathbf{Wf}, & (d) \\
& \mathbf{e}^T\mathbf{f} = (1+\alpha)n, & (e) \\
& \mathbf{e}^T\mathbf{g} \geq (1-\beta)n, & (f) \\
& \mathbf{f} \geq \mathbf{g}, & (g) \\
& 0 \leq \mathbf{g} \leq 1 & (h)
\end{aligned}
$$

Z must arise from an assignment matrix

Overlap & non-exhaustiveness constraints

Relaxation

Low-Rank SDP for NEO-K-Means

- Low-rank factorization of $Z = YY^T$(Y: $n \times k$, non-negative)
  - Lose convexity but **only requires linear memory**

$$
\begin{array}{ll}
\underset{\boldsymbol{Y},\mathbf{f},\mathbf{g},\mathbf{s},r}{\text{minimize}} & \mathbf{f}^T\mathbf{d} - \operatorname{trace}(\boldsymbol{Y}^T\boldsymbol{K}\boldsymbol{Y}) \\
\text{subject to} & k = \operatorname{trace}(\boldsymbol{Y}^T\boldsymbol{W}^{-1}\boldsymbol{Y}) \\
& 0 = \boldsymbol{Y}\boldsymbol{Y}^T\mathbf{e} - \boldsymbol{W}\mathbf{f} \\
& 0 = \mathbf{e}^T\mathbf{f} - (1+\alpha)n \\
& 0 = \mathbf{f} - \mathbf{g} - \mathbf{s} \\
& 0 = \mathbf{e}^T\mathbf{g} - (1-\beta)n - r \\
& Y_{ij} \geq 0, \mathbf{s} \geq 0, r \geq 0 \\
& 0 \leq \mathbf{f} \leq k\mathbf{e}, 0 \leq \mathbf{g} \leq 1
\end{array}
$$

Z is replaced by YY^T

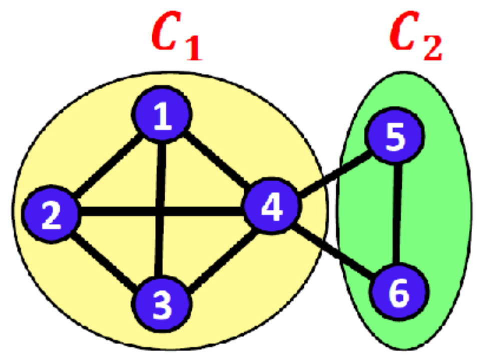
s, r : slack variables

Solving the NEO-K-Means Low-Rank SDP

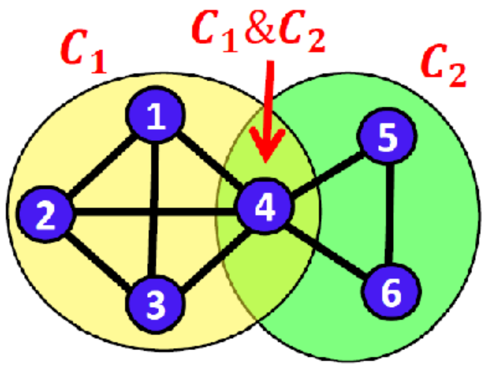
- Augmented Lagrangian method to optimize **the NEO-K-Means Low-Rank SDP**
  - minimizing an **augmented Lagrangian** of the problem

$$
\begin{aligned}
\mathcal{L}_{\mathcal{A}}(Y, \mathbf{f}, \mathbf{g}, \mathbf{s}, r; \boldsymbol{\lambda}, \boldsymbol{\mu}, \gamma, \sigma) = {} & \underbrace{\mathbf{f}^T\mathbf{d} - \operatorname{trace}(\mathbf{Y}^T\mathbf{K}\mathbf{Y})}_{\text{the objective}} \\
& - \lambda_1(\operatorname{trace}(\mathbf{Y}^T\mathbf{W}^{-1}\mathbf{Y}) - k) + \frac{\sigma}{2}(\operatorname{trace}(\mathbf{Y}^T\mathbf{W}^{-1}\mathbf{Y}) - k)^2 \\
& - \boldsymbol{\mu}^T(\mathbf{Y}\mathbf{Y}^T\mathbf{e} - \mathbf{W}\mathbf{f}) + \frac{\sigma}{2}(\mathbf{Y}\mathbf{Y}^T\mathbf{e} - \mathbf{W}\mathbf{f})^T(\mathbf{Y}\mathbf{Y}^T\mathbf{e} - \mathbf{W}\mathbf{f}) \\
& - \lambda_2(\mathbf{e}^T\mathbf{f} - (1+\alpha)n) + \frac{\sigma}{2}(\mathbf{e}^T\mathbf{f} - (1+\alpha)n)^2 \\
& - \gamma^T(\mathbf{f} - \mathbf{g} - \mathbf{s}) + \frac{\sigma}{2}(\mathbf{f} - \mathbf{g} - \mathbf{s})^T(\mathbf{f} - \mathbf{g} - \mathbf{s}) \\
& - \lambda_3(\mathbf{e}^T\mathbf{g} - (1-\beta)n - r) + \frac{\sigma}{2}(\mathbf{e}^T\mathbf{g} - (1-\beta)n - r)^2
\end{aligned}
$$

Extending NEO-K-Means to Graph Clustering



- Graph Clustering
  - Find tightly connected groups
  - Traditional setting: assign a node to exactly one cluster.

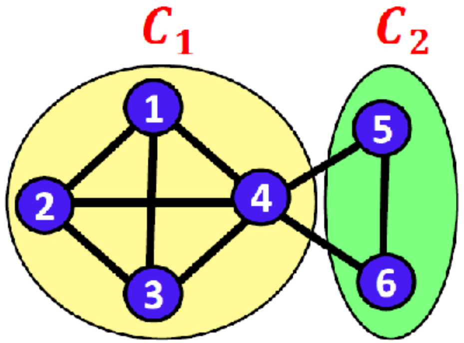


- **NEO-K-Means** can be extended to **graph clustering**
  - **Overlapping community detection**
  - Weighted kernel NEO-K-Means objective is mathematically equivalent to an overlapping community detection objective.

Community Detection Using NEO-K-Means

- **Normalized Cut**: a traditional graph clustering objective
  - Assumes a **disjoint, exhaustive** clustering

$$\texttt{ncut}(G) = \min_{\mathcal{C}_1, \mathcal{C}_2, \cdots \mathcal{C}_k} \sum_{j=1}^{k} \frac{\texttt{links}(\mathcal{C}_j, \mathcal{V} \backslash \mathcal{C}_j)}{\texttt{links}(\mathcal{C}_j, \mathcal{V})} = \max_{\mathbf{y}_1, \mathbf{y}_2, \cdots \mathbf{y}_k} \sum_{j=1}^{k} \frac{\mathbf{y}_j^T A \mathbf{y}_j}{\mathbf{y}_j^T D \mathbf{y}_j}$$



$$A = \begin{bmatrix} 0&1&1&1&0&0 \\ 1&0&1&1&0&0 \\ 1&1&0&1&0&0 \\ 1&1&1&0&1&1 \\ 0&0&0&1&0&1 \\ 0&0&0&1&1&0 \end{bmatrix}$$

$$\texttt{ncut}(G) = \frac{\texttt{links}(\mathcal{C}_1, \mathcal{V} \backslash \mathcal{C}_1)}{\texttt{links}(\mathcal{C}_1, \mathcal{V})} + \frac{\texttt{links}(\mathcal{C}_2, \mathcal{V} \backslash \mathcal{C}_2)}{\texttt{links}(\mathcal{C}_2, \mathcal{V})} = \frac{2}{14} + \frac{2}{4}$$

Community Detection Using NEO-K-Means

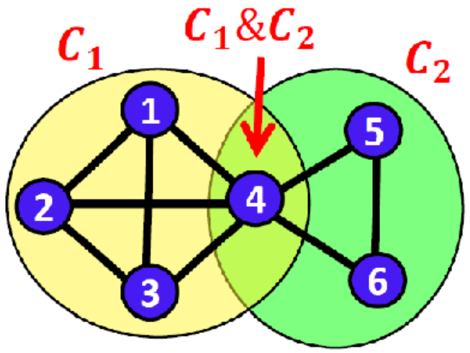
- Extending Normalized Cut to **Non-exhaustive, Overlapping Clustering**

$$\max_{Y} \sum_{j=1}^{k} \frac{\mathbf{y}_j^T A \mathbf{y}_j}{\mathbf{y}_j^T D \mathbf{y}_j}$$

$\boldsymbol{\alpha = 0, \beta = 0}$**: equivalent to the traditional normalized cut**

$$\text{s.t.} \quad trace(Y^T Y) = (1+\alpha)n, \ \sum_{i=1}^{n} \mathbb{I}\{(Y\mathbf{1})_i = 0\} \le \beta n.$$

Constraints in NEO-K-Means



$$A = \begin{bmatrix} 0&1&1&1&0&0\\ 1&0&1&1&0&0\\ 1&1&0&1&0&0\\ 1&1&1&0&1&1\\ 0&0&0&1&0&1\\ 0&0&0&1&1&0 \end{bmatrix}$$

$$\texttt{ncut}(G) = \frac{\texttt{links}(\mathcal{C}_1, \mathcal{V}\backslash\mathcal{C}_1)}{\texttt{links}(\mathcal{C}_1, \mathcal{V})} + \frac{\texttt{links}(\mathcal{C}_2, \mathcal{V}\backslash\mathcal{C}_2)}{\texttt{links}(\mathcal{C}_2, \mathcal{V})} = \frac{2}{14} + \frac{3}{9}$$

Community Detection Using NEO-K-Means

- Weighted Kernel **NEO-K-Means**
  - **Weight (W)** : non-negative weight for each data point
  - **Kernel (K)** : inner products of two data points in a higher-dimensional space

$$\min_U \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} w_i \|\phi(\mathbf{x}_i) - \mathbf{m}_j\|^2 = \min_U \sum_{j=1}^{k} \left(\sum_{i=1}^{n} u_{ij} w_i K_{ii} - \frac{\mathbf{u}_j{}^T W K W \mathbf{u}_j}{\mathbf{u}_j{}^T W \mathbf{u}_j} \right),$$

$$\text{where } \mathbf{m}_j = \frac{\sum_{i=1}^{n} u_{ij} w_i \phi(\mathbf{x}_i)}{\sum_{i=1}^{n} u_{ij} w_i}, \text{ s.t.} trace(U^T U) = (1+\alpha)n, \ \sum_{i=1}^{n} \mathbb{I}\{(U\mathbf{1})_i = 0\} \leq \beta n.$$

- **Weighted Kernel NEO-K-Means objective $\equiv$ the extended normalized cut objective**
  - $W \coloneqq D, K \coloneqq \sigma D^{-1} + D^{-1} A D^{-1}$

Community Detection Using NEO-K-Means

	Data Clustering	Graph Clustering
Traditional Idea	K-Means	Normalized Cut
New Objectives	NEO-K-Means Weighted Kernel NEO-K-Means	Extended Normalized Cut
New Algorithms	Fast Iterative NEO-K-Means LRSDP NEO-K-Means	

These two objectives are mathematically Equivalent.

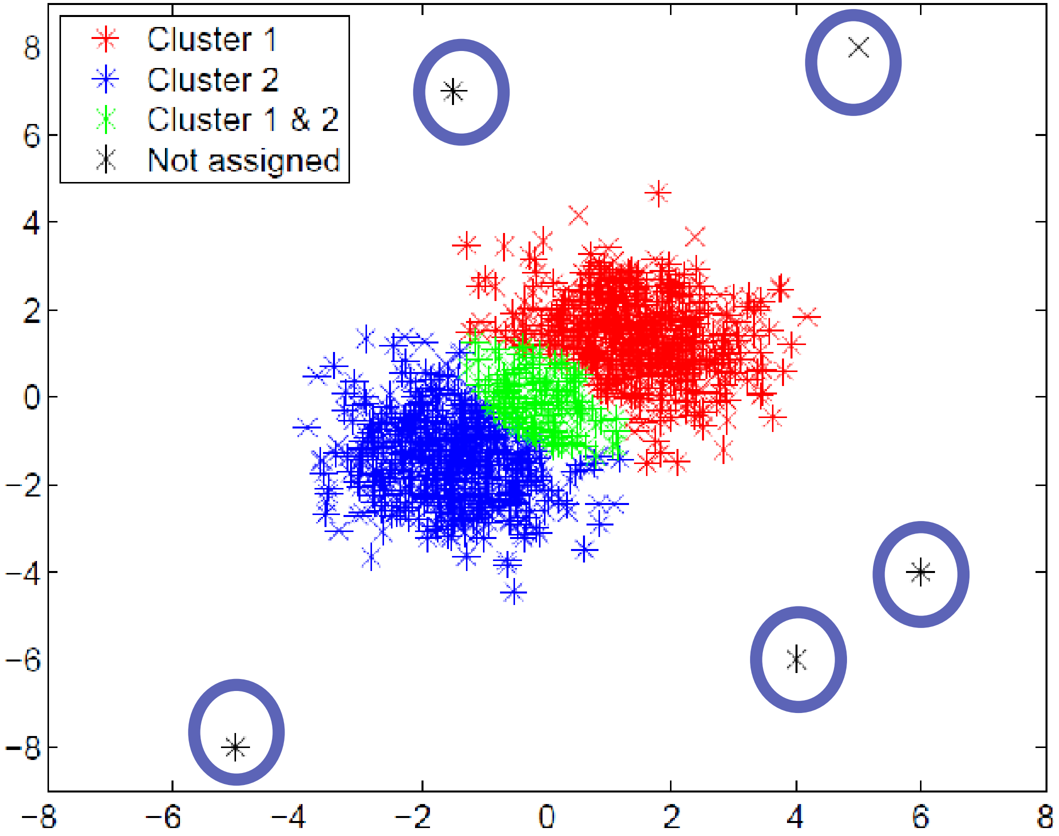
Community Detection Using NEO-K-Means

	Data Clustering	Graph Clustering
Traditional Idea	K-Means	Normalized Cut
New Objectives	NEO-K-Means Weighted Kernel NEO-K-Means	Extended Normalized Cut
New Algorithms	Fast Iterative NEO-K-Means LRSDP NEO-K-Means	Weighted Kernel NEO-K-Means LRSDP WK-NEO-K-Means

Since the objectives are equivalent,
NEO-K-Means can be applied to overlapping community detection.

Experimental Results

- Output of NEO-K-Means on a synthetic dataset



Green data points: overlap
→ A natural overlapping clustering structure

Black data points: outliers
→ Perfectly finds the five outliers

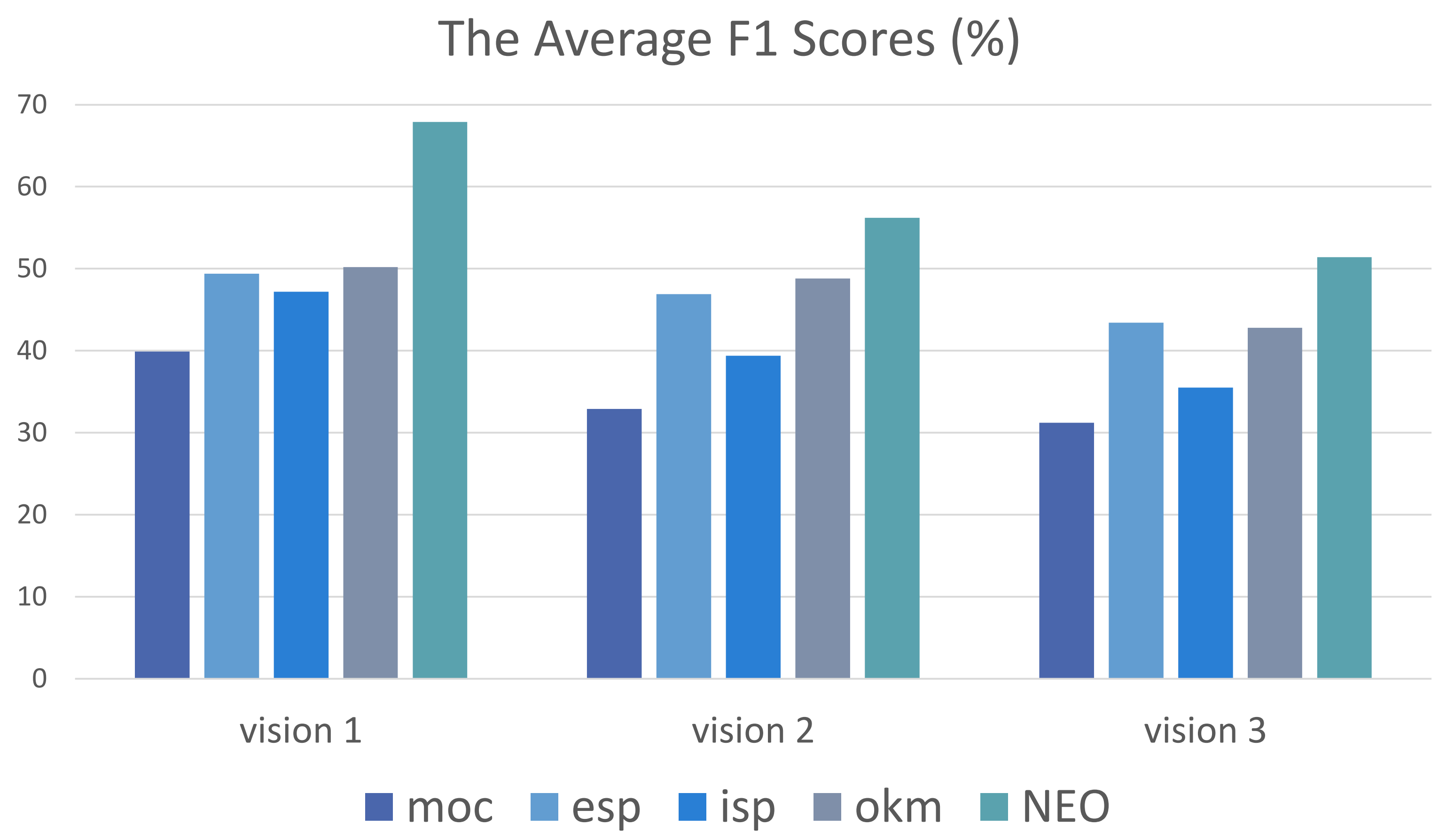
Experimental Results

- Applications in **Computer Vision**
  - An image is annotated by various attributes such as parts (e.g., “arm”, “wing”) and materials (e.g., “glass”, “plastic”). Each attribute corresponds to a cluster.

(a) Wood & Glass (b) Wood (c) Glass

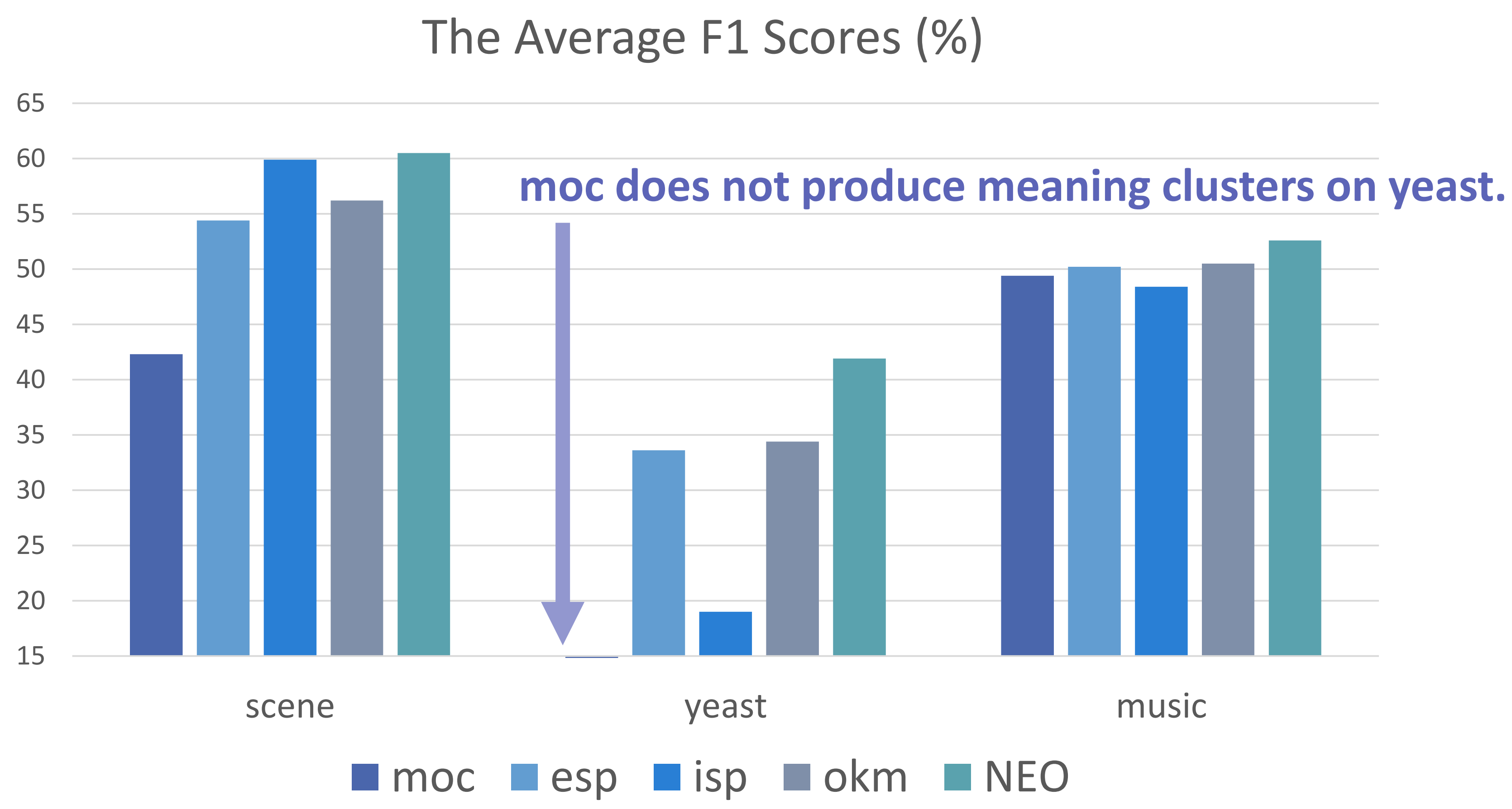
Experimental Results

- Applications in **Computer Vision**



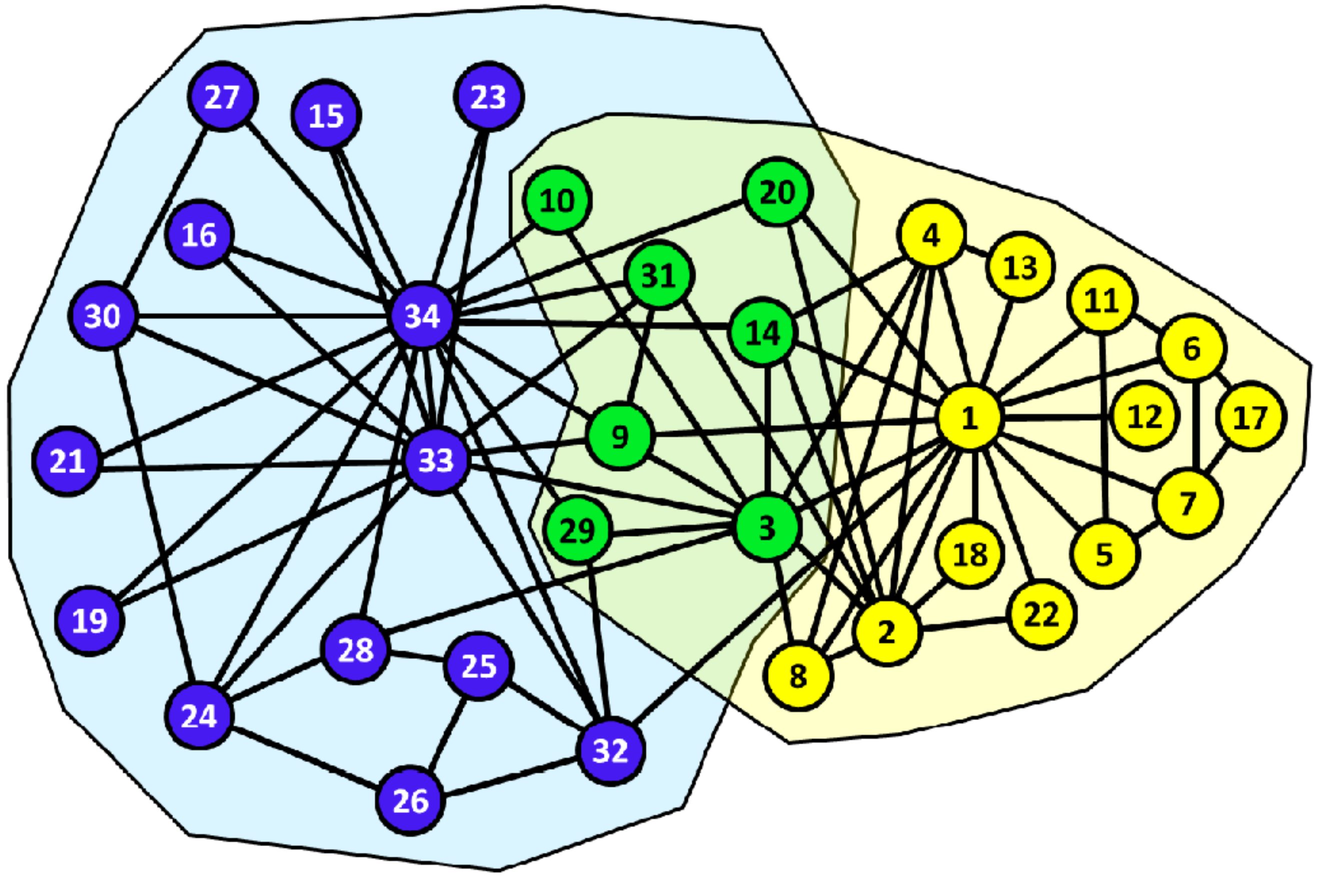
Experimental Results

- Applications in **Computer Vision**, **Bioinformatics**, and **Music Classification**



Experimental Results

- Output of NEO-K-Means on Karate Club Network



Experimental Results

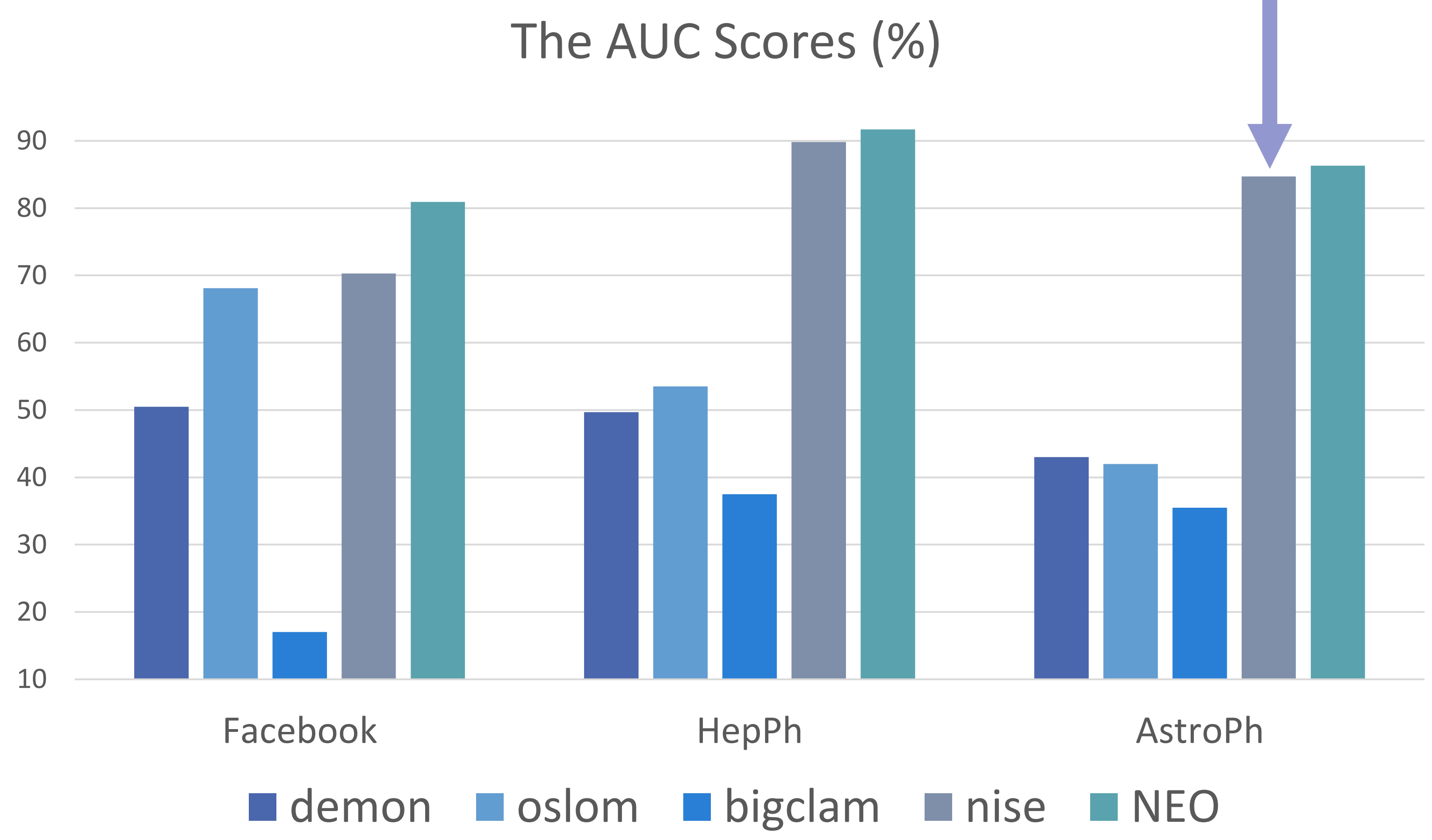
- Real-world Graph Datasets

Category	Graph	No. of Nodes	No. of Edges
Social Network	Facebook	348	2,866
	Orkut	731,332	21,992,171
Collaboration Network	HepPh	11,204	117,619
	AstroPh	17,903	196,972
	CondMat	21,363	91,286
Product Network	Amazon	334,863	925,872

Experimental Results

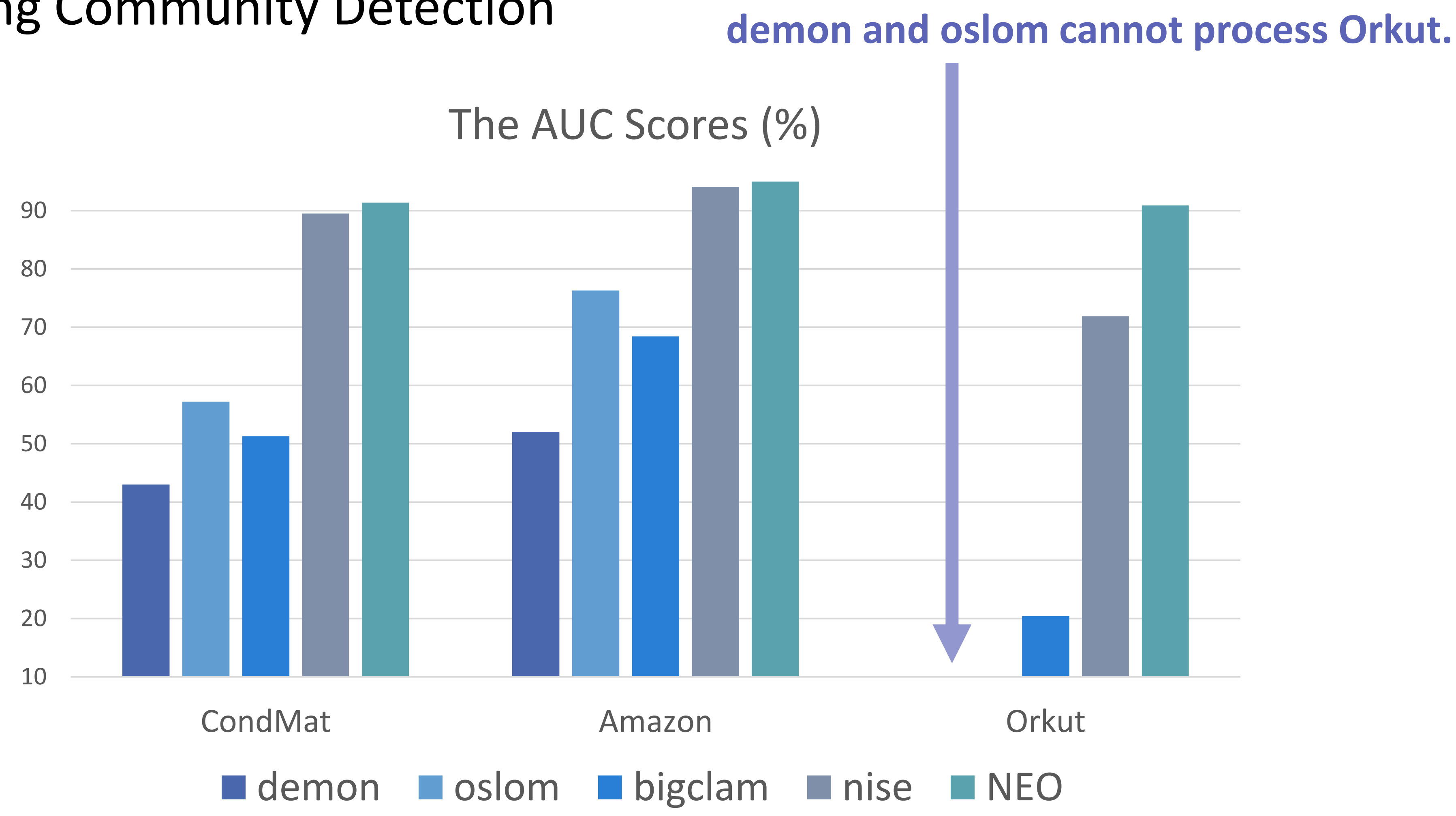
- Overlapping Community Detection

nise is also my algorithm (*TKDE* 2016 & *CIKM* 2013).



Experimental Results

- Overlapping Community Detection



Summary

- **NEO-K-Means**
  - **Overlap** and **non-exhaustiveness**: handled in a unified framework
  - Simple and intuitive objective function
  - **LRSDP** boosts up the performance of the iterative NEO-K-Means
- Weighted Kernel NEO-K-Means
  - **Overlapping Community Detection**
- Effective in **identifying the ground-truth clusters** in both data clustering and overlapping community detection

More Information: **http://bigdata.cs.skku.edu/**