Non-exhaustive, Overlapping Clustering via Low-Rank Semidefinite Programming

Yangyang Hou[1]*, Joyce Jiyoung Whang[2]*
David F. Gleich[1] Inderjit S. Dhillon[2]

[1]Purdue University

[2]The University of Texas at Austin

(* first authors)

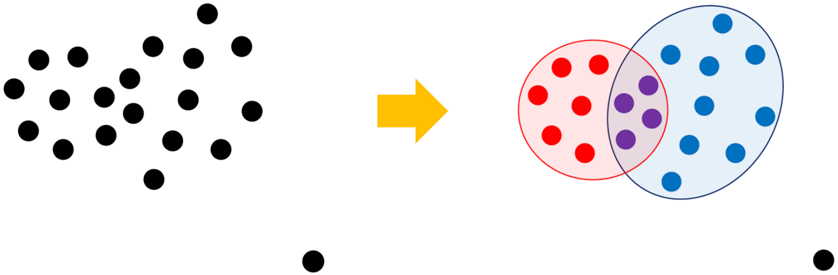
ACM SIGKDD Conference on Knowledge Discovery and Data Mining
Aug. 10 – 13, 2015.

Contents

- Non-exhaustive, Overlapping Clustering
  - NEO-K-Means Objective
  - NEO-K-Means Algorithm
- Semidefinite Programming (SDP) for NEO-K-Means
- Low-Rank SDP for NEO-K-Means
- Experimental Results
- Conclusions

Clustering

- Clustering: finding a set of cohesive data points
- Traditional disjoint, exhaustive clustering (e.g., *k*-means)
  - 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.
  - Clusters are allowed to overlap with each other.



NEO-K-Means (Non-Exhaustive, Overlapping K-Means) [1]

- The NEO-K-Means objective function
  - Overlap and non-exhaustiveness - handled in a unified framework

$$\begin{array}{ll} \min\limits_{U} & \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij} \|\mathbf{x}_i - \mathbf{m}_j\|^2, \text{ where } \mathbf{m}_j = \frac{\sum_{i=1}^{n} u_{ij}\mathbf{x}_i}{\sum_{i=1}^{n} u_{ij}} \\ \text{s.t.} & trace(U^T U) = (1+\alpha)n, \ \sum_{i=1}^{n} \mathbb{I}\{(U\mathbf{1})_i = 0\} \leq \beta n. \end{array}$$

  - α: overlap, β: non-exhaustiveness
  - $\alpha = 0, \beta = 0$: equivalent to the standard k-means objective

$$U = \begin{array}{cc} \begin{matrix} C_1 & C_2 & C_3 \end{matrix} & \\ \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} & \begin{matrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{matrix} \end{array} \quad U^T U = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 1 \end{bmatrix} \quad U\mathbf{1} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \boxed{\begin{bmatrix} 1 \\ 2 \\ 0 \\ 2 \end{bmatrix}}$$

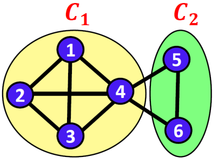
cluster sizes

no. of clusters a data point belongs to

[1] J. J. Whang, I. S. Dhillon, and D. F. Gleich. Non-exhaustive, overlapping k-means. SDM, 2015.

NEO-K-Means (Non-Exhaustive, Overlapping K-Means) [1]

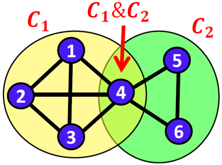
- Normalized Cut for Overlapping Community Detection



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

(a) Disjoint communities:

$$\texttt{ncut}(G) = \frac{2}{14} + \frac{2}{4}$$



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

(b) Overlapping communities:

$$\texttt{ncut}(G) = \frac{2}{14} + \frac{3}{9}$$

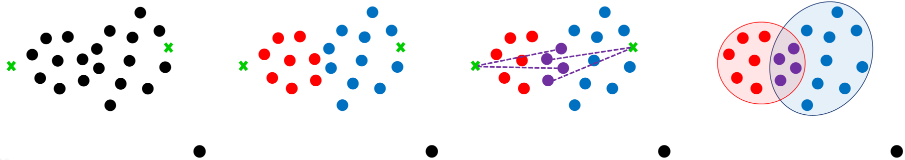
- Weighted Kernel NEO-K-Means objective is equivalent to the extended normalized cut objective.

[1] **J. J. Whang, I. S. Dhillon, and D. F. Gleich. Non-exhaustive, overlapping k-means. SDM, 2015.**

NEO-K-Means (Non-Exhaustive, Overlapping K-Means) [1]

- The NEO-K-Means Algorithm is a simple iterative algorithm that monotonically decreases the NEO-K-Means objective.
  - $\alpha = 0, \beta = 0$: identical to the standard k-means algorithm

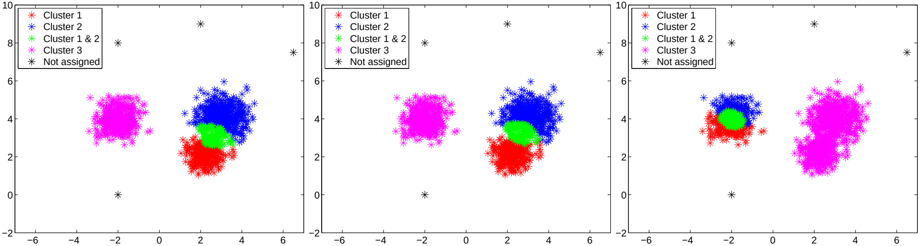
- Example ($n = 20, \alpha = 0.15, \beta = 0.05$)
  - Assign $n - \beta n$ (=19) data points to their closest clusters.
  - Make $\beta n + \alpha n$ (=4) assignments by taking minimum distances.



[1] **J. J. Whang, I. S. Dhillon, and D. F. Gleich. Non-exhaustive, overlapping k-means. SDM, 2015.**

Motivation

- NEO-K-Means Algorithm
  - Fast iterative algorithm
  - Susceptible to initialization
  - Can be trapped in local optima



(a) Ground-truth clusters

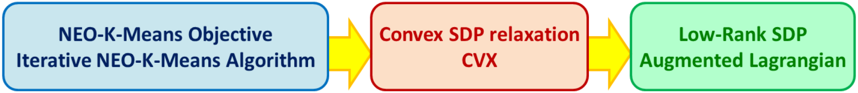
(b) Success of *k*-means initialization

(c) Failure of *k*-means initialization

- LRSDP initialization allows the NEO-K-Means algorithm to consistently produce a reasonable clustering structure.

Overview

- Goal: more accurate and more reliable solutions than the iterative NEO-K-Means algorithm by paying additional computational cost



NEO-K-Means Objective
Iterative NEO-K-Means Algorithm

Convex SDP relaxation
CVX

Low-Rank SDP
Augmented Lagrangian

Fast and scalable
Trapped in local optima

Slow and not scalable
Globally optimized

Faster than CVX
Locally optimized

Background: Semidefinite Programs (SDPs)

- Semidefinite Programming (SDP)
  - Convex problem (→ globally optimized via a variety of solvers)
  - The number of variables is quadratic in the number of data points.
  - 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

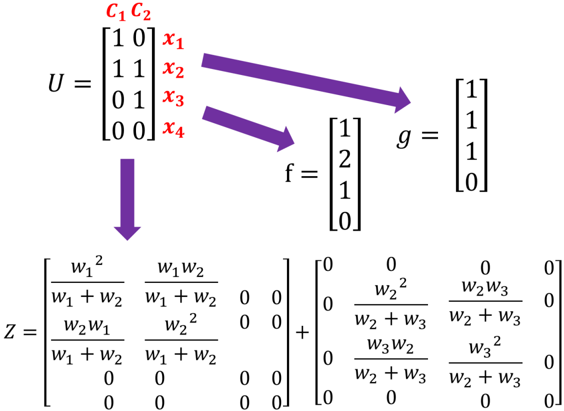
$$\begin{array}{ll} \text{maximize} & \operatorname{trace}(\mathbf{CX}) \\ \text{subject to} & \mathbf{X} \succeq 0, \mathbf{X} = \mathbf{X}^T, \\ & \operatorname{trace}(\mathbf{A}_i\mathbf{X}) = b_i \\ & \quad i = 1, \ldots, m \end{array}$$

Low-rank SDP

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

NEO-K-Means as an SDP

- Three key variables to model the assignment structure U
  - Co-occurrence matrix $\mathbf{Z} = \sum_{c=1}^{k} \frac{\mathbf{W}\mathbf{u}_c(\mathbf{W}\mathbf{u}_c)^T}{\mathbf{u}_c^T\mathbf{W}\mathbf{u}_c}$
  - **f**: overlap, **g**: non-exhaustiveness



$$U = \begin{array}{c c} c_1 & c_2 & \\ \end{array}$$

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

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

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

SDP-like Formulation for NEO-K-Means

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

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

Z must arise from an assignment matrix

Overlap & non-exhaustiveness constraints

Combinatorial problem

SDP for NEO-K-Means

- Convex relaxation of NEO-K-Means
  - Any local optimal solution must be a global solution.

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

Z must arise from an assignment matrix

Overlap & non-exhaustiveness constraints

Relaxation

Low-Rank SDP for NEO-K-Means

- Low-Rank SDP
  - Low-rank factorization of $\mathbf{Z}$: $\mathbf{Y}\mathbf{Y}^T$ ($\mathbf{Y}$: $n \times k$, non-negative)
  - $\mathbf{s}, r$: slack variables
  - Lose convexity but only requires linear memory

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

Solving the NEO-K-Means Low-Rank SDP

- LRSDP: optimize the NEO-K-Means Low-Rank SDP
- Augmented Lagrangian method:
  minimizing an augmented Lagrangian of the problem that includes
  - Current estimate of the Lagrange multipliers
  - Penalty term that derives the solution towards the feasible set

All the details of the augmented Lagrangian method are in the paper.

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

Algorithmic Validation

- Comparison of SDP and LRSDP
  - LRSDP is roughly an order of magnitude faster than CVX.
  - The objective value are different in light of the solution tolerances.
  - dolphins [1]: 62 nodes, 159 edges, les miserables [2]: 77 nodes, 254 edges

		Objective value		Run time	
		SDP	LRSDP	SDP	LRSDP
dolphins	k=2, α=0.2, β=0	-1.968893	-1.968329	107.03 secs	2.55 secs
	k=2, α=0.2, β=0.05	-1.969080	-1.968128	56.99 secs	2.96 secs
	k=3, α=0.3, β=0	-2.913601	-2.915384	160.57 secs	5.39 secs
	k=3, α=0.3, β=0.05	-2.921634	-2.922252	71.83 secs	8.39 secs
les miserables	k=2, α=0.2, β=0	-1.937268	-1.935365	453.96 secs	7.10 secs
	k=2, α=0.3, β=0	-1.949212	-1.945632	447.20 secs	10.24 secs
	k=3, α=0.2, β=0.05	-2.845720	-2.845070	261.64 secs	13.53 secs
	k=3, α=0.3, β=0.05	-2.859959	-2.859565	267.07 secs	19.31 secs

[1] D. Lusseau et al., *Behavioral Ecology and Sociobiology*, 2003.

[2] D. E. Knuth. The Stanford GraphBase: A Platform for Combinatorial Computing. Addison-Wesley, 1993.

Rounding Procedure & Practical Improvements

- Problem $\rightarrow$ Relaxation $\rightarrow$ Rounding $\rightarrow$ Refinement

- Rounding procedure
  - $\mathbf{Y}$: normalized assignment matrix
  - $\mathbf{f}$: the number of clusters each data point is assigned to
  - $\mathbf{g}$: which data points are not assigned to any cluster

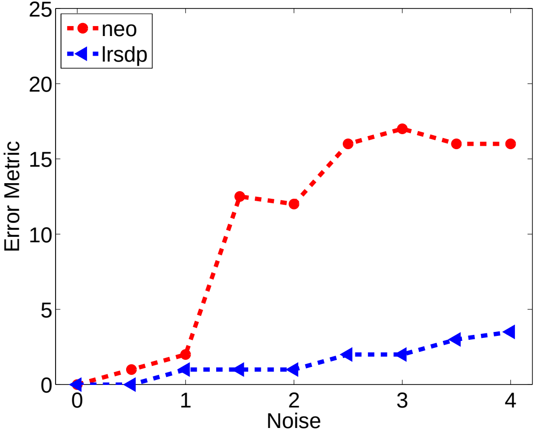
- Refinement
  - Use LRSDP solution as the initial cluster assignment for the iterative NEO-K-Means algorithm

- Sampling
  - Run LRSDP on a 10% sample of the data points

- Two-level hierarchical clustering
  - First level: $k' = \sqrt{k}$, $\alpha' = \sqrt{1+\alpha} - 1$ and unchanged β
  - Second level: k', α' and $\beta' = 0$ for each cluster at level 1

Experimental Results on Synthetic Problems

- Overlapping community detection on a Watts-Strogatz cycle graph
  - LRSDP initialization lowers the errors.



Experimental Results on Data Clustering

- Comparison of NEO-K-Means objective function values
  - Real-world datasets from Mulan[3]
  - By using the LRSDP solution as the initialization of the iterative algorithm, we can achieve better objective function values.

		worst	best	avg.
yeast	kmeans+neo	9611	9495	9549
	lrsdp+neo	**9440**	9280	**9364**
	slrsdp+neo	9471	**9231**	9367
music	kmeans+neo	87779	70158	77015
	lrsdp+neo	**82323**	**70157**	**75923**
	slrsdp+neo	82336	70159	75926
scene	kmeans+neo	18905	**18745**	**18806**
	lrsdp+neo	18904	18759	18811
	slrsdp+neo	**18895**	18760	18810

[3] http://mulan.sourceforge.net/datasets.html

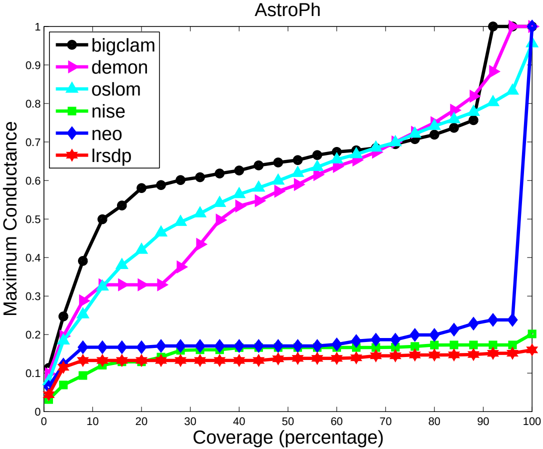
Experimental Results on Data Clustering

- F_1 scores on real-world vector datasets
  - NEO-K-Means-based methods outperform other methods.
  - Low-rank SDP method improves the clustering results.

		moc	*esp*	*isp*	*okm*	kmeans+neo	lrsdp+neo	slrsdp+neo
yeast	worst	-	0.274	0.232	0.311	0.356	**0.390**	0.369
	best	-	0.289	0.256	0.323	0.366	**0.391**	**0.391**
	avg.	-	0.284	0.248	0.317	0.360	**0.391**	0.382
music	worst	0.530	0.514	0.506	0.524	0.526	0.537	**0.541**
	best	0.544	0.539	0.539	0.531	0.551	**0.552**	**0.552**
	avg.	0.538	0.526	0.517	0.527	0.543	0.545	**0.547**
scene	worst	0.466	0.569	0.586	0.571	0.597	**0.610**	0.605
	best	0.470	0.582	0.609	0.576	**0.627**	0.614	0.625
	avg.	0.467	0.575	0.598	0.573	0.610	**0.613**	**0.613**

Experimental Results on Graph Clustering

- Conductance-vs-graph coverage
  - The lower curve indicates better communities.



Experimental Results on Graph Clustering

- AUC of conductance-vs-graph coverage
  - Real-world networks from SNAP[4]
  - LRSDP produces the best quality communities in terms of AUC score.
  - The largest graph: AstroPh (17,903 nodes, 196,972 edges)

	Facebook1	Facebook2	HepPh	AstroPh
bigclam	0.830	0.640	0.625	0.645
demon	0.495	0.318	0.503	0.570
oslom	0.319	0.445	0.465	0.580
nise	0.297	0.293	0.102	0.153
neo	0.285	0.269	0.206	0.190
LRSDP	**0.222**	**0.148**	**0.091**	**0.137**

[4] http://snap.stanford.edu/

Conclusions

- We propose a convex SDP relaxation of a k-means-like objective that handles non-exhaustive, overlapping clustering problems.
- We formulate a low-rank factorization of the SDP problem and implement the scalable LRSDP algorithm.
- We also propose a series of initialization and rounding strategies that accelerate the convergence of our optimization procedures.
- Experiments show that our LRSDP approach gives reliable solutions on both data clustering and overlapping community detection problems.

http://www.cs.utexas.edu/∼joyce/