Spectral methods for the clustering of cyclic and acyclic graphs

Jacopo Palumbo, Dimosthenis Pasadakis, Albert-Jan Yzelman*, Olaf Schenk
*Computing Systems Lab, Huawei Zürich Research Center.

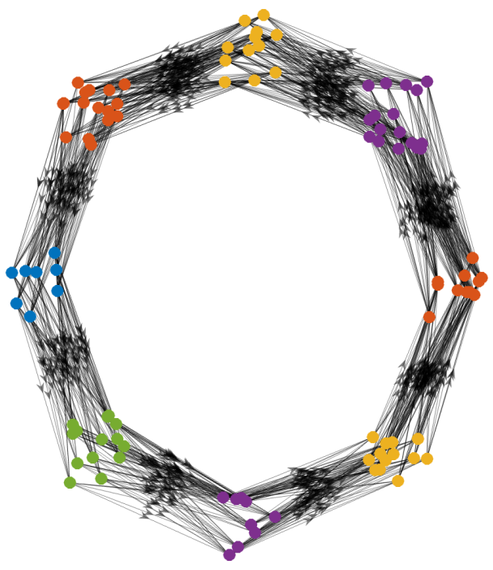
1. H. Van Lierde, T. W. S. Chow, and J. C. Delvenne, "Spectral clustering algorithms for the detection of clusters in block-cyclic and block-acyclic graphs," Journal of Complex Networks, May 2018.
2. Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe, "A consistent adjacency spectral embedding for stochastic blockmodel graphs.", Journal of the American Statistical Association, 2012.
3. K. Hayashi, S. G. Aksoy and H. Park, "Skew-Symmetric Adjacency Matrices for Clustering Directed Graphs," IEEE International Conference on Big Data (Big Data), 2022, pp. 555-564.
4. M. Cucuringu, H. Li, H. Sun, and L. Zanetti, "Hermitian matrices for clustering directed graphs: insights and applications," in AISTATS, 2020, pp. 983–992.

What are cyclic graphs?

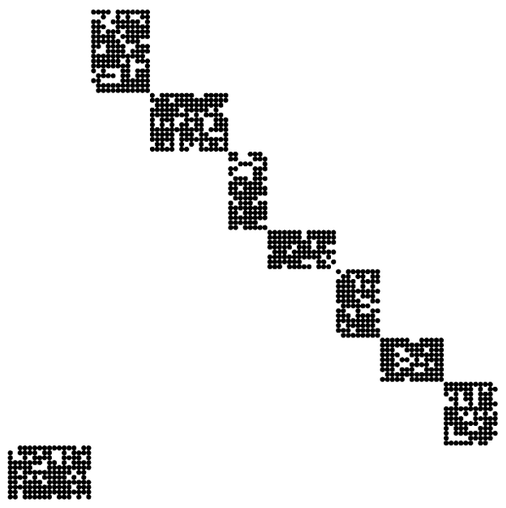
Definition: Given V a set of nodes, $E \in V \times V$ the set of directed edges, and $W \in \mathbb{R}^{n\times n}$ a matrix of positive edge weights (i.e. the **adjacency matrix**).
A directed graph $G = (V, E, W)$ is a block-cycle of k blocks if it contains at least one directed cycle of length k and if there exists a function $\tau : V \to \{1, \dots, k\}$ partitioning the nodes of V into k non-empty subsets, such that

$$E \subseteq \{(u, v) : (\tau(u), \tau(v)) \in \mathcal{C}\} \tag{1}$$

where $\mathcal{C} = \{(1, 2), (2, 3), \dots, (k-1, k), (k, 1)\}$.



a) Cyclic graph with 8 blocks and 100 nodes.

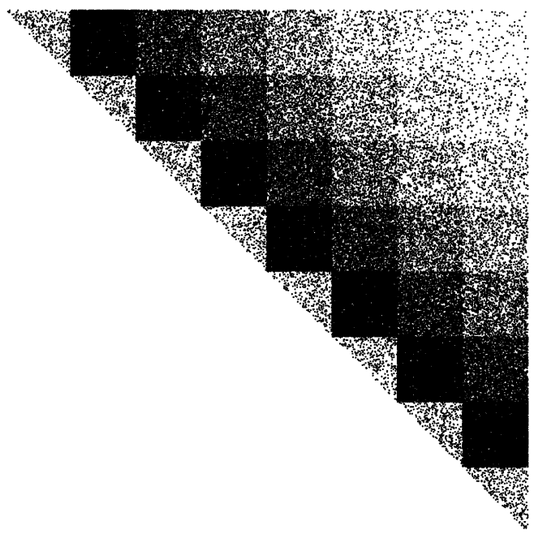


b) Sparsity pattern of the adjacency matrix.

What are acyclic graphs?

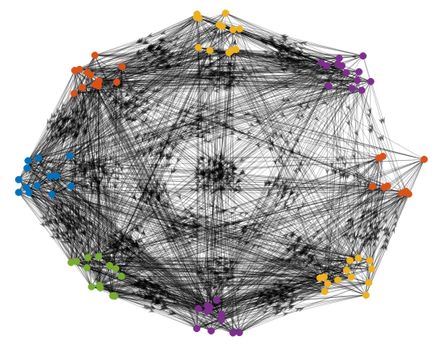
Definition: Given V a set of nodes, $E \in V \times V$ the set of directed edges, and $W \in \mathbb{R}^{n\times n}$ a matrix of positive edge weights (i.e. the **adjacency matrix**).
A directed graph $G = (V, E, W)$ is block acyclic with k blocks if there exists a function $\tau : V \to \{1, \dots, k\}$ partitioning the nodes of V into k non-empty subsets, such that

$$E \subseteq \{(u, v) : (\tau(u) < \tau(v))\}. \tag{2}$$



Sparsity pattern of the adjacency matrix of an acyclic graph with 8 blocks and 1000 nodes.

On this type of graphs we can perform **flow-based clustering**: the edges between clusters tend to be oriented in one direction. This type of graph is representative for many real world scenarios: migration data, food webs, and trade data.



Acyclic graph with 8 blocks and 100 nodes.

Spectral clustering for block-cyclic graphs algorithm

Algorithm 1 Block-Cyclic Spectral (BCS) Clustering Algorithm [1]

Input: Adjacency matrix $W \in \{0, 1\}^{n\times n}$ such that all nodes have nonzero out-degree.
Parameters: Number of clusters $k \in \{2, 3, \dots, n\}$.
Step 1: Compute Transition Matrix
Compute the transition matrix P:

$$P = D_{\text{out}}^{-1} W,$$

where $D_{\text{out}} = \text{diag}(d_{\text{out},1}, \dots, d_{\text{out},n})$, with $d_{\text{out},i} = \sum_{j=1}^{n} w_{ij}$.
Step 2: Extract Cycle Eigenvalues and Eigenvectors
Find the k eigenvalues of P with the largest modulus (cycle eigenvalues). Store the corresponding eigenvectors (cycle eigenvectors) as columns of a matrix $\Gamma \in \mathbb{C}^{n\times k}$.
Step 3: Cluster Points in Spectral Space
Consider each row of Γ as a point in $\mathbb{C}^k$. Apply the k-means algorithm to cluster these points. Let $\phi : \{1, \dots, n\} \to \{1, \dots, k\}$ be the function assigning each row of Γ to a cluster.
Step 4: Compute Block Membership Function
Compute the block membership function $\tilde{\tau} = \phi(u), \quad \forall u \in \{1, \dots, n\}$.
Output: Estimation of block membership function $\tilde{\tau} : \{1, \dots, n\} \to \{1, \dots, k\}$.

Efficient eigenvalue problem

Eigenvalue Problem: The power method computes the k eigenvalues of P with largest modulus and the associated eigenvectors. For a sparse transition matrix with nz non-zero elements, the per-iteration complexity is $O(kn^2 + nk^2)$ This approach ensures computational efficiency, especially for large and sparse graphs.

Spectral methods for acyclic graphs (flow-based clustering)

• **Adjacency Spectral Embedding (ASE or SVD)** [2]
Embeds nodes by performing a truncated singular value decomposition (SVD) of the adjacency matrix W, with $W \approx U\Sigma V^\top$. The embedding uses either the *scaled* form $[U\Sigma^{1/2} \mid V\Sigma^{1/2}]$ or the *unscaled* form $[U \mid V]$. The scaled version preserves inner products and is consistent under the random dot product graph (RDPG) model. ASE is widely used but does not explicitly model edge directionality.

• **Skew-symmetric ASE (SKEW)** [3]
Constructs $S = W - W^\top$ to isolate antisymmetric components. Since S has purely imaginary eigenvalues, embeddings are derived from the imaginary parts of its eigenvectors, capturing directed cycles and asymmetries in a real-valued space.

• **Hermitian spectral clustering (HERM)** [4]
Defines the Hermitian matrix $H = i(W - W^\top)$, with real eigenvalues and orthonormal eigenvectors. Non-zero eigenvectors serve as features for clustering, retaining directional information in a form compatible with standard spectral methods.

• **Block-Acyclic Spectral (BAS) clustering** [1]
Targets hierarchical structure using the row-stochastic matrix

$$(P_a)_{ij} = \begin{cases} W_{ij}/d_i^{\text{out}}, & d_i^{\text{out}} > 0, \\ 1/n, & d_i^{\text{out}} = 0, \end{cases}$$

with $d_i^{\text{out}} = \sum_j W_{ij}$. Node embeddings are built from the eigenvectors of P_a with largest modulus, revealing top-down organization.

Numerical results

Directed Stochastic Block Model (DSBM)

We consider the Directed Stochastic Block Model (DSBM) dataset presented in [3], which is defined by the tuple (k, p, q, c, F). Here:

- k is the number of clusters,
- p is the intra-cluster edge probability,
- q is the inter-cluster edge probability,
- $c \in \mathbb{N}^k$ specifies the number of nodes in each cluster,
- $F \in [0, 1]^{k\times k}$ encodes cluster-level orientation probabilities.

Each node u is assigned to a cluster via $\tau(u) \in \{1, \dots, k\}$. Then, for any pair of nodes u, v:

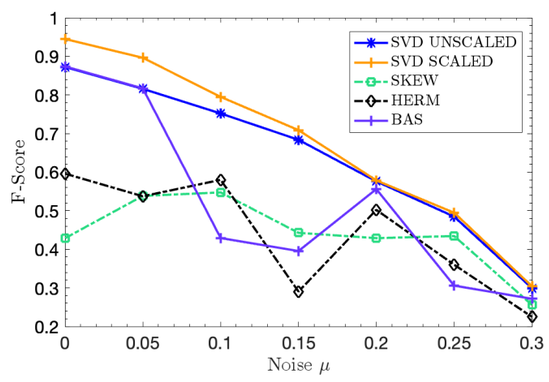
$$\mathbb{P}(A_{uv} = 1) = \begin{cases} p, & \text{if } \tau(u) = \tau(v), \\ q, & \text{if } \tau(u) \neq \tau(v), \end{cases}$$

and direction is assigned with probability $F_{\tau(u),\tau(v)}$. We set $p = q = 0.008\}$, $n = 5000$, $k = 5$, and assign $c_i = 1000$ for all i. Directional structure is controlled by a noise parameter $\mu \in [0, 0.3]$. The matrix $F \in [0, 1]^{k\times k}$ is constructed as follows:

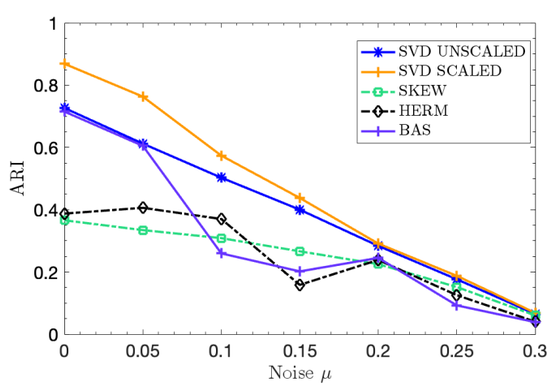
$$F_{uv} = \begin{cases} \mu, & \text{if } v = u + 1 \text{ or } v = u + 2, \\ 1 - \mu, & \text{if } v = u - 1 \text{ or } v = u - 2, \\ \frac{1}{2}, & \text{otherwise.} \end{cases}$$

As $\mu \to 0.5$, edge directions become increasingly random, making cluster recovery more difficult.

Results on the synthetic DSBM



a) Obtained F-Score for the presented methods on the graph generated using the parameters in the DSBM section.



b) Obtained ARI for the presented methods on the graph generated using the parameters in the DSBM section.

Real world results

F-score combines precision and recall into a single measure of clustering quality. Higher F-scores indicate better alignment between predicted and true cluster assignments.

$$\text{F-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Adjusted Rand Index (ARI) measures the similarity between predicted and true clusterings. It ranges from 1 to 1, with 1 indicating perfect alignment and 0 corresponding to random labeling.

$$\text{ARI} = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where n_{ij}, a_i, b_j are values from the contingency table.

email-Eu-core is a directed network representing email exchanges between individuals at European research institutions. An edge (u, v) indicates that person u sent at least one email to person v. The dataset includes 42 institutions, which define the ground-truth communities. We evaluate clustering methods on two-community subgraphs: *email-Eu-core12* (1st and 2nd largest communities) and *email-Eu-core23* (2nd and 3rd largest communities).

	email-Eu-core12		email-Eu-core23	
Method	F-Score	ARI	F-Score	ARI
SVD Unscaled	0.8364	0.4754	0.7174	0.2679
SVD Scaled	0.8307	0.4617	0.7895	0.3308
SKEW	0.6035	0.0992	0.3695	0.0060
HERM	0.6035	0.0992	0.3695	0.0020
BAS	0.3516	-0.0016	0.4485	0.0296