Parametrized Power Iteration Clustering for Directed Graphs

Gwendal Debaussart-Joniec, Harry Sevi, Matthieu Jonckheere & Argyris Kalogeratos.
Centre Borelli · École Normale Supérieure Paris-Saclay · CNRS · Université Paris-Saclay | LAAS-CNRS

Abstract

Clustering directed graphs remains difficult because edge directionality breaks the assumptions behind classical spectral clustering. Existing methods often rely on expensive eigen-decomposition, graph symmetrization, or teleportation mechanisms that distort the original dynamics. As the directed nature breaks the reversibility of random walks, adaptation of diffusion-based methods to digraphs is non-trivial.

We propose Parametrized Power-Iteration Clustering **ParPIC**, an *eigen-free framework* that constructs a parametrized reversible random walk operator for any weakly connected digraph, automatically selects the diffusion time via an entropic criterion, and computes low-dimensional embeddings through random projections. On both synthetic and real-world directed graphs, ParPIC matches or outperforms 10 competing methods while scaling significantly better than spectral-based approaches.

Motivation

Classical diffusion geometry and spectral clustering assume:
- Reversible random walks
- Real-valued spectra
- Strong connectivity

These assumptions fail for many directed graphs:
- Random walks are often non-reversible
- Spectra may become complex-valued
- Weak connectivity breaks convergence guarantees

Existing workarounds introduce limitations:

Workarounds	**Limitations**
Symmetrization	Loose directional information
Teleportation	Distortion of original dynamics, loss of sparsity
Hermitianization	No diffusion interpretation

Goal. Enabling diffusion-based clustering for directed graphs without eigen-decomposition, while preserving edge directionality and scalability.

Notations

$G = (V, E)$: digraph
$\mathbf{A}$: adjacency matrix
$\mathbf{d}^+, \mathbf{d}^-$: out-degree and in-degree
$\mathbf{D}_{ii} = \mathbf{d}^+(i)$: out-degree matrix
$\mathbf{P} = \mathbf{D}^{-1}\mathbf{A}$: natural random walk
$\mathbf{A}_S = (\mathbf{A} + \mathbf{A}^T)/2$: symmetrized $\mathbf{A}$
k: number of clusters
N: number of vertices
d : embedding dimension
$\nu : V \to \mathbb{R}, \|\nu_i\|_1 = 1$: vertex measure
$\mathbf{D}_\nu$: diagonal matrix of ν
$\mathbf{P}_S = (\mathbf{D}_S)^{-1}\mathbf{A}_S$: random walk of $\mathbf{A}_S$

Method

Parametrized random walk operator for directed graphs:

$$\mathbf{P}_\nu = (\mathbf{D}_\xi + \mathbf{D}_\nu)^{-1}(\mathbf{D}_\nu \mathbf{P} + \mathbf{P}^\top \mathbf{D}_\nu), \quad \xi = \mathbf{P}^\top \nu.$$

Key properties. under mild assumptions:
- Reversible with respect to ν
- Converges to a unique stationary distribution π_ν
- Real-valued spectrum
- Continuous over ν

Vertex measure design

$$\nu(i) = \gamma \mathbf{d}^+(i) + (1-\alpha)\mathbf{d}^-(i), \quad \gamma \in [0,1].$$

It is assumed that $\mathbf{d}^+$ and $\mathbf{d}^-$ are normalized to sum to 1. Experiments show that $\gamma \approx 0.5$ is a good default choice and that extreme values of γ can lead to poor clustering performance.

Diffusion time selection

We propose an entropy-based criterion to automatically select the diffusion time t that balances local and global structure in the embedding space, avoiding the need for manual tuning.

$$\mathcal{H}(P) = \sum_i \mathcal{H}_i(P), \quad \mathcal{H}_i(\mathbf{P}) = -\sum_j \mathbf{P}(i,j)\log \mathbf{P}(i,j),$$

$$t^* = \underset{t \in \mathbb{N}}{\text{Elbow}} \;\; \mathcal{H}(\mathbf{P}_\nu^t)$$

Randomized embeddings

To scale to large graphs, we use randomized methods to compute low-dimensional embeddings based on the diffusion operator *without explicit eigen-decomposition*.

$$\mathbf{Z}_t = \mathbf{P}_\nu \mathbf{Z}_{t-1}, \quad \mathbf{Z}_0 \sim \frac{1}{\sqrt{N}}\mathcal{N}(0, \mathbf{I}), \quad \mathbf{Z}_t \in \mathbb{R}^{N \times d},$$

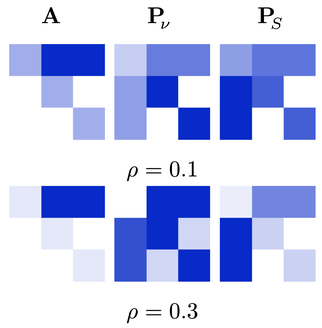
where $d = \sqrt{N}$ is the embedding dimension. For larger graphs, it can be further reduced to $d = c\log(N)$ with minimal loss in clustering performance, with c being a 'small' constant.

Pipeline

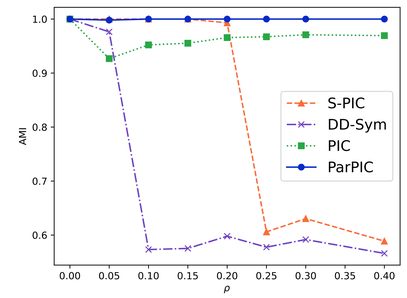
1. Compute $\mathbf{P}$ from the directed graph
2. Compute $\mathbf{P}_\nu$ from $\mathbf{P}$ and ν
3. Select diffusion time t using entropy-based criterion
4. Compute low-dimensional embedding
5. Cluster using k-means++

Experiments

Synthetic graphs



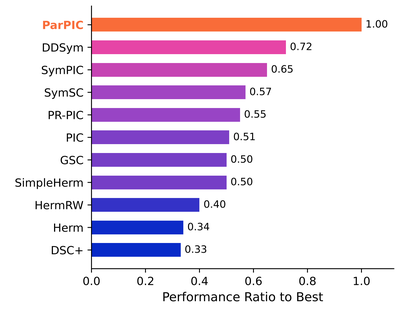
Expected matrices for a DiSBM.



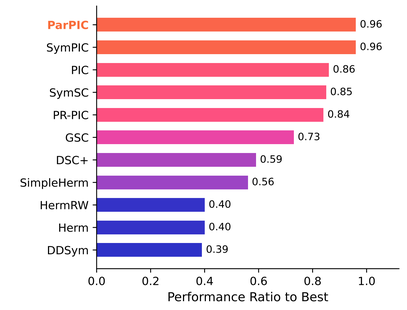
Sensitivity to out-degree heterogeneity.

The probability of an edge between the first cluster and the two other clusters is denoted ρ.

Clustering performance



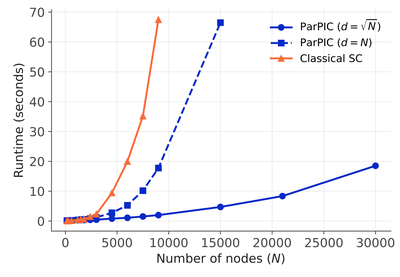
Directed graphs



k-NN directed graphs

Method families — Hermitian spectral: Herm, HermRW, SimpleHerm I RW-spectral: DDsym, DSC+, GSC, SymSC I Power-Iteration: PIC, PR-PIC, S-PIC, ParPIC. Performance Ratio to Best is defined as the ratio of the performance of each method to the best performance among all methods. The mean is shown and computed with respect to AMI.

Runtime compared to spectral clustering



Runtimes on synthetic graphs.

Take-aways. ParPIC provides a scalable and performant, diffusion-based clustering framework by extending PIC to directed graphs.

Interested ?
Check out our paper!

[1] Generalized Dirichlet Energy and Graph Laplacians for Clustering Directed and Undirected Graphs (2026), H Sevi, G Debaussart-Joniec, M Hacini, M Jonckheere, A Kalogeratos. *TMLR*.
[2] Power iteration clustering (2010), Lin, F., & Cohen, W. W. *ICML*.
[3] Hermitian matrices for clustering directed graphs: insights and applications (2020), Cucuringu, M., Li, H., Sun, H., & Zanetti, L. *AISTATS*