

## Graph Clustering

**Graph clustering** involves the identification of clusters within a graph, where nodes exhibit denser connections among themselves than with the rest of the network. This task is crucial for conducting comprehensive macro- and mesoscopic analyses of large complex systems, such as social networks. As exact graph clustering is recognized as an NP-hard problem, numerous clustering methods have been developed [2]. Notably, model-based approaches fit a mathematical model to a graph, providing an explanatory framework for observed connectivity patterns. In [1], J.-J. Daudin *et al.* investigate the renowned **Stochastic Block Model (SBM)**, proposing a **variational Expectation-Maximization (EM)** algorithm to optimize model fit for a given graph.

### The Stochastic Block Model

**Stochastic Block Model:** A mixture model for graphs where each node is assigned to a *class* and the edge probability between nodes are conditioned on their class memberships.

**Formally**, let  $\mathcal{G}$  be an undirected graph with  $n$  nodes and no self-loop, and  $X$  its adjacency matrix, *i.e.*  $X_{ij} = 1$  if an edge exists between  $i$  and  $j$ , and  $X_{ij} = 0$  otherwise.

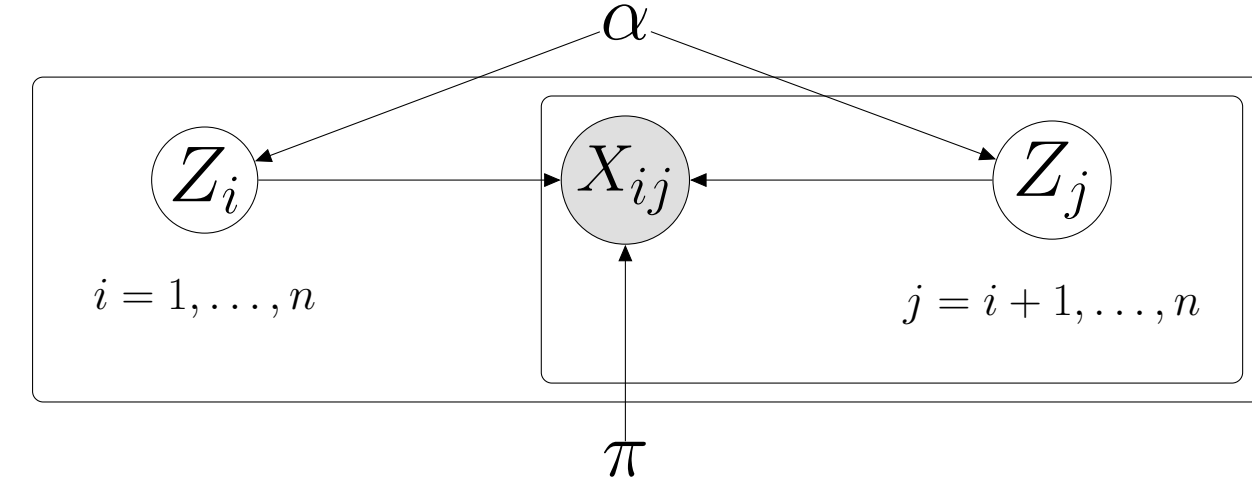


Figure 1. Graphical model of the SBM

The model is parametrized by  $Q$  the number of classes,  $\alpha \in [0, 1]^Q$  the prior distribution on the classes, and  $\pi \in [0, 1]^{Q \times Q}$  the probability of an edge between two nodes of different classes. We also introduce the random variables  $Z_i \in \{0, 1\}^Q$  for  $i \in \llbracket 1, n \rrbracket$ , that represent the membership of node  $i$  to each class. The prior distributions on  $Z$  and  $X$  are given by :

$$\begin{cases} \forall i \in \llbracket 1, n \rrbracket, & \sum_{q=1}^Q Z_{iq} = 1 \quad (\text{unique class}) \\ \forall q \in \llbracket 1, Q \rrbracket, & \mathbb{P}(Z_{iq} = 1) = \alpha_q \quad (\text{class distribution}) \\ \forall q, l \in \llbracket 1, Q \rrbracket, & \forall i \neq j \in \llbracket 1, n \rrbracket, \quad \mathbb{P}(X_{ij} = 1 | Z_{iq} = 1, Z_{jl} = 1) = \pi_{ql} \quad (\text{edge probability}) \end{cases}$$

### The variational Expectation-Maximization algorithm

In [1], Daudi et. al. propose a variational EM algorithm to fit the SBM.

$$\log \mathcal{L}(X, Z) = \sum_i \sum_q Z_{iq} \log \alpha_q + \frac{1}{2} \sum_{i \neq j} \sum_{q, l} Z_{iq} Z_{jl} \log \left( \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \right)$$

⇒ No closed-form solution for the parameters of the model from the likelihood.

**Problem :** Using an EM algorithm, no closed-form solution for the E-step either.

**Solution:** Daudin et. al. search for an approximated distribution  $R_X$  of  $Z$  given  $X$ , among the family of product of multinomial distributions, thus introducing the random variables  $\tau_i \in [0, 1]^Q$  for  $i \in \llbracket 1, n \rrbracket$ . This approximation leads to maximizing the following lower bound of  $\log \mathcal{L}(X)$  :

$$\mathcal{J}(R_X) = \log \mathcal{L}(X) - \text{KL}[R_X(\cdot), P(\cdot|X)]$$

The authors derive a closed form solution for the M-step, and a **fixed-point relation** for the E-step :

$$\begin{cases} (\text{E-step}) & \hat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_l \left[ \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \right]^{\hat{\tau}_{jl}} \\ (\text{M-step}) & \hat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq} \quad \text{and} \quad \hat{\pi}_{ql} = \frac{\sum_{i \neq j} \tau_{iq} \tau_{jl} X_{ij}}{\sum_{i \neq j} \tau_{iq} \tau_{jl}} \end{cases}$$

#### Algorithm 1 Variational Expectation-Maximization Algorithm

- 1: **Input:** Adjacency matrix  $X$ , number of communities  $Q$
- 2: **Initialize:** Initialize  $\tau_{iq}$  for  $i \in \llbracket 1, n \rrbracket$ ,  $q \in \llbracket 1, Q \rrbracket$
- 3: **while** not converged **do**
- 4:   **E-step:**
- 5:   **for**  $i \in \llbracket 1, n \rrbracket$  **do**
- 6:    **for**  $q \in \llbracket 1, Q \rrbracket$  **do**
- 7:      **Fixed-point algorithm:**
- 8:        $\hat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_l \left[ \pi_{ql}^{X_{ij}} (1 - \pi_{ql})^{1 - X_{ij}} \right]^{\hat{\tau}_{jl}}$    ▷ Update class memberships
- 9:   **M-step:**
- 10:   **for**  $q \in \llbracket 1, Q \rrbracket$  **do**
- 11:      $\hat{\alpha}_q = \frac{1}{n} \sum_i \tau_{iq}$    ▷ Update class proportions
- 12:      $\hat{\pi}_{ql} = \frac{\sum_{i \neq j} \tau_{iq} \tau_{jl} X_{ij}}{\sum_{i \neq j} \tau_{iq} \tau_{jl}}$    ▷ Update edge probabilities
- 13: **Return:**  $\hat{\alpha}, \hat{\pi}, \hat{\tau}$

### Some alternative algorithms

- **Newman et al. [3]** use a finer definition for clusters. While Daudin *et al.* define a cluster-cluster affinity (nodes belonging to the same cluster behave the same way), the model of Newman introduces a cluster-node affinity. Therefore, edges between different nodes of the same two clusters are not necessarily equivalent. ⇒  $\Pi_{q,i}$  denotes the probability of an edge between a node in cluster  $q$  and node  $i$ . The expected log-likelihood (with respect to  $Z$ ) is

$$\mathcal{L}(X) = \sum_{i=1}^n \sum_{q=1}^Q \tau_{iq} \left[ \log(\alpha_q) + \sum_{j=1}^n X_{ij} \log(\Pi_{q,j}) \right] \quad \text{where} \quad \tau_{iq} = \mathbb{P}(Z_{iq} = 1 | X, \alpha, \Pi).$$

The EM updates are,

$$(\text{E-step}) \quad \hat{\tau}_{iq} = \frac{\alpha_q \prod_{j=1}^n \Pi_{qj}^{X_{ij}}}{\sum_{s=1}^Q \alpha_s \prod_{j=1}^n \Pi_{sj}^{X_{ij}}}; \quad (\text{M-step}) \quad \hat{\alpha}_q = \frac{1}{n} \sum_{i=1}^n \tau_{iq} \quad \text{and} \quad \hat{\Pi}_{qj} = \frac{\sum_{i=1}^n X_{ij} \tau_{iq}}{\sum_{i=1}^n \tau_{iq}}$$

The update of  $\tau$  (E-step) is now **deterministic**, with no resort to fixed-point methods.

- **Spectral clustering [4]** uses the eigenvectors of the Laplacian matrix of the graph to cluster the nodes. The  $k$  first eigenvectors of  $L = D - X$  are stacked in a matrix  $U \in \mathbb{R}^{n \times k}$  and a  $k$ -means clustering is performed on the rows of  $U$ .

## References

- [1] J.-J. Daudin, F. Picard, and S. Robin. A mixture model for random graphs. *Statistics and Computing*, 18(2):173–183, June 2008.
- [2] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, February 2010.
- [3] M. E. J. Newman and E. A. Leicht. Mixture models and exploratory analysis in networks. *PNAS*, 104, 2007.
- [4] Andrew Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, 2001.
- [5] Prithviraj Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3), 2008.

## Experiments: SBM dataset

**Algorithm 1** is implemented in PyTorch. To test it, a first dataset is created by sampling the SBM for parameters given in **Table 1**. For each experiment, 10 graphs are generated and the algorithm is run for 100 iterations on each graph.

**Figure 2** shows the results.

#	Experiment Name	Hyper-parameters		Parameters	
		$n$	$Q$	$\alpha$	$\pi$
1	Random	500	3	$\alpha \sim \text{Dir}(1.5)$	$\pi_{ij} \sim \mathcal{U}([0, 1])$
2	Homophilic	150	5	$\alpha_i = (\frac{1}{Q})_i$	$\pi_{ii} \sim \mathcal{U}([0.5, 1])$ , $\pi_{ij} = 0.01$
3	Heterophilic	150	3	$\alpha_i = (\frac{1}{Q})_i$	$\pi_{ii} = 0.01$ , $\pi_{ij} = 0.9$
4	Dense minority	150	2	$\alpha^T = (\frac{9}{10}, \frac{1}{10})$	$\pi = \begin{pmatrix} 0.01 & 0.7 \\ 0.7 & 0.8 \end{pmatrix}$

Table 1. Parameters for generating the SBM dataset. Dir is the Dirichlet law. In experiment 1,  $\pi$  is scaled and made symmetric.

- **Metric for model-fitting capacity:** Sum of (cluster-permutation-invariant) Euclidean distances between predicted  $(\alpha, \pi)$  and the ground truth of Table 1. The result is normalized by the number of classes.
- **Metric for clustering capacity:** Normalized Mutual information (NMI) and Rand Index (RI).

**NB:** The fixed-point algorithm does not always converge within 1000 iterations per graph.

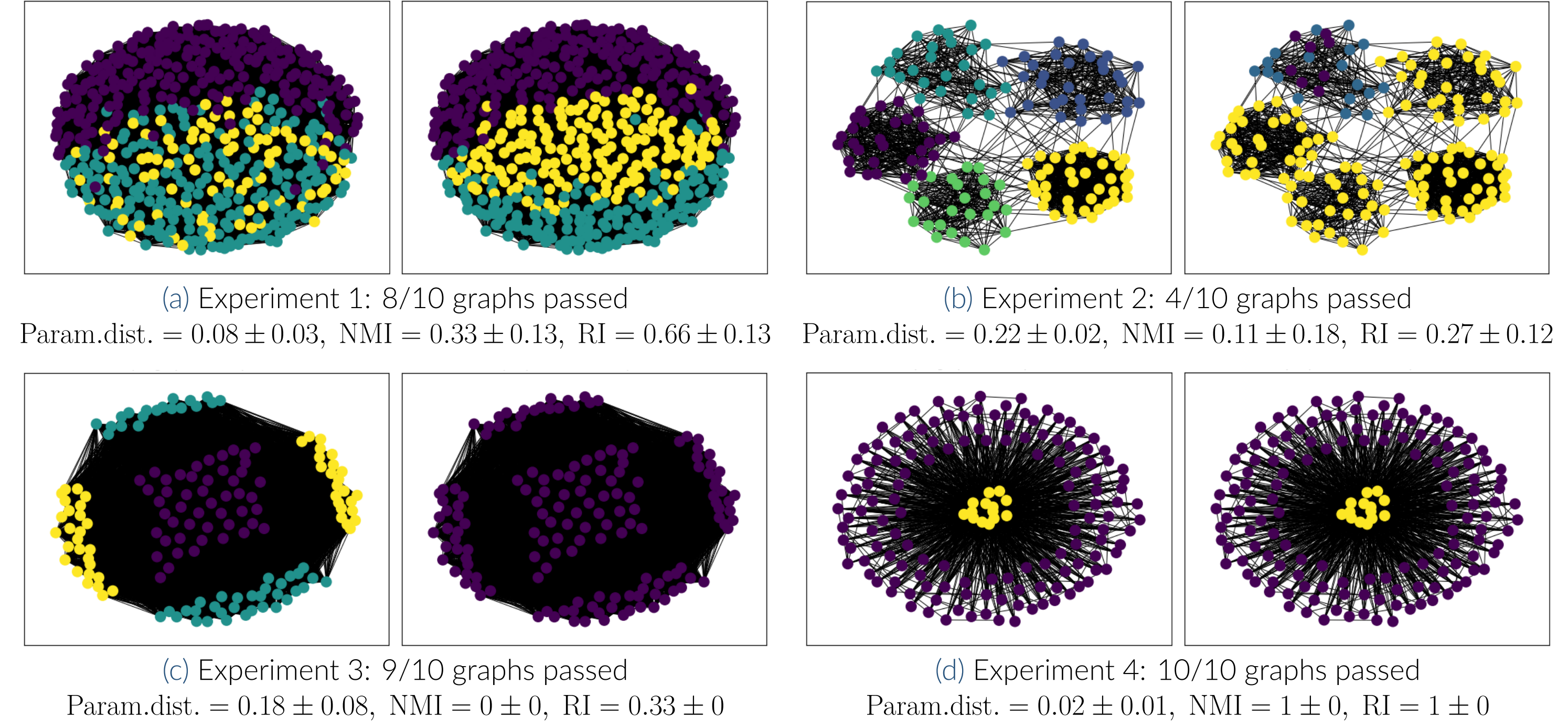


Figure 2. Results for the SBM dataset experiments. Left: sample graph; right: associated prediction.

## Experiments: Cora dataset

**The Cora dataset [5]** (citation network of scientific publications): Undirected graph of 2708 nodes (=papers) labeled with 7 different classes and 5429 edges(=citations).

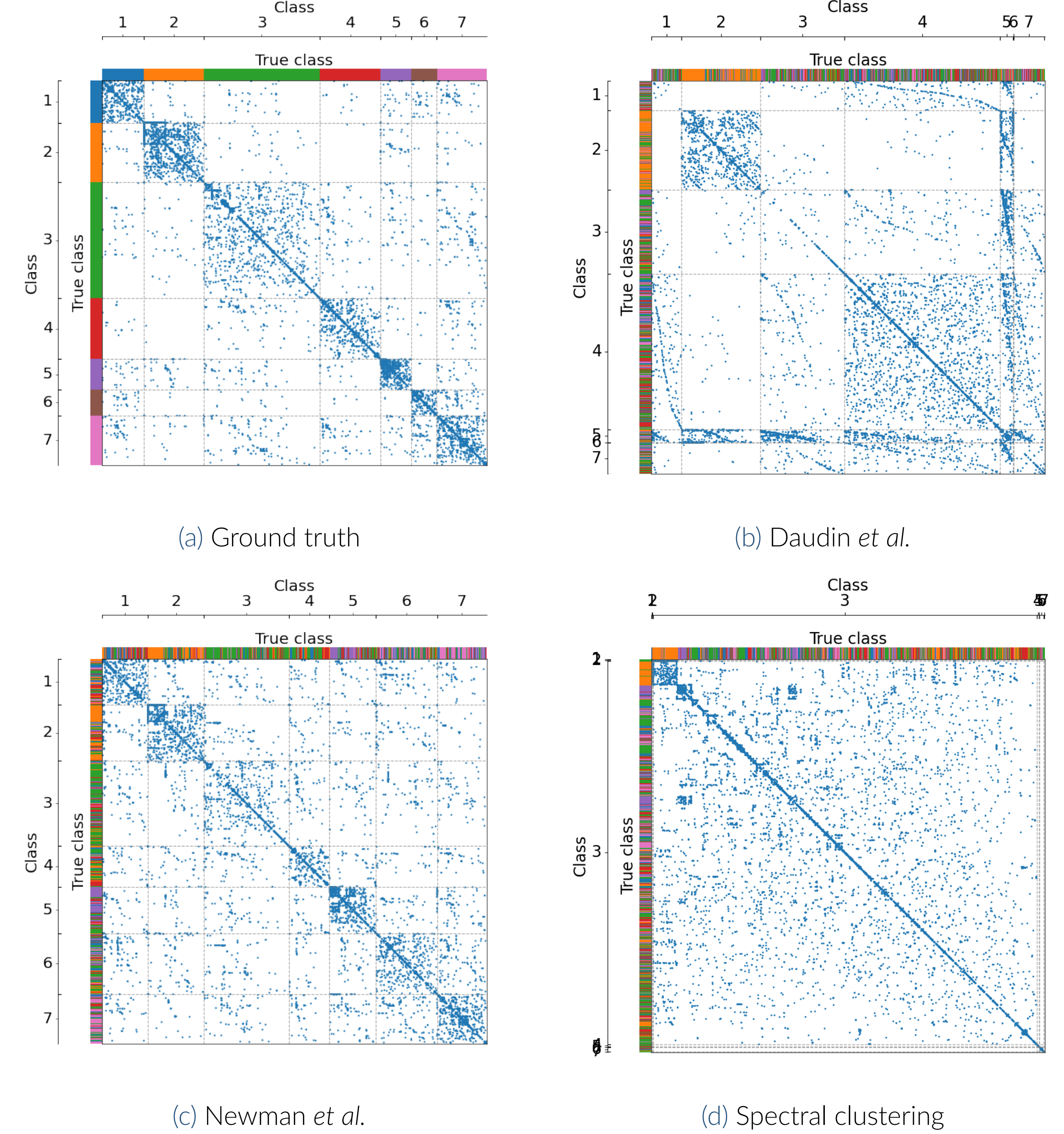
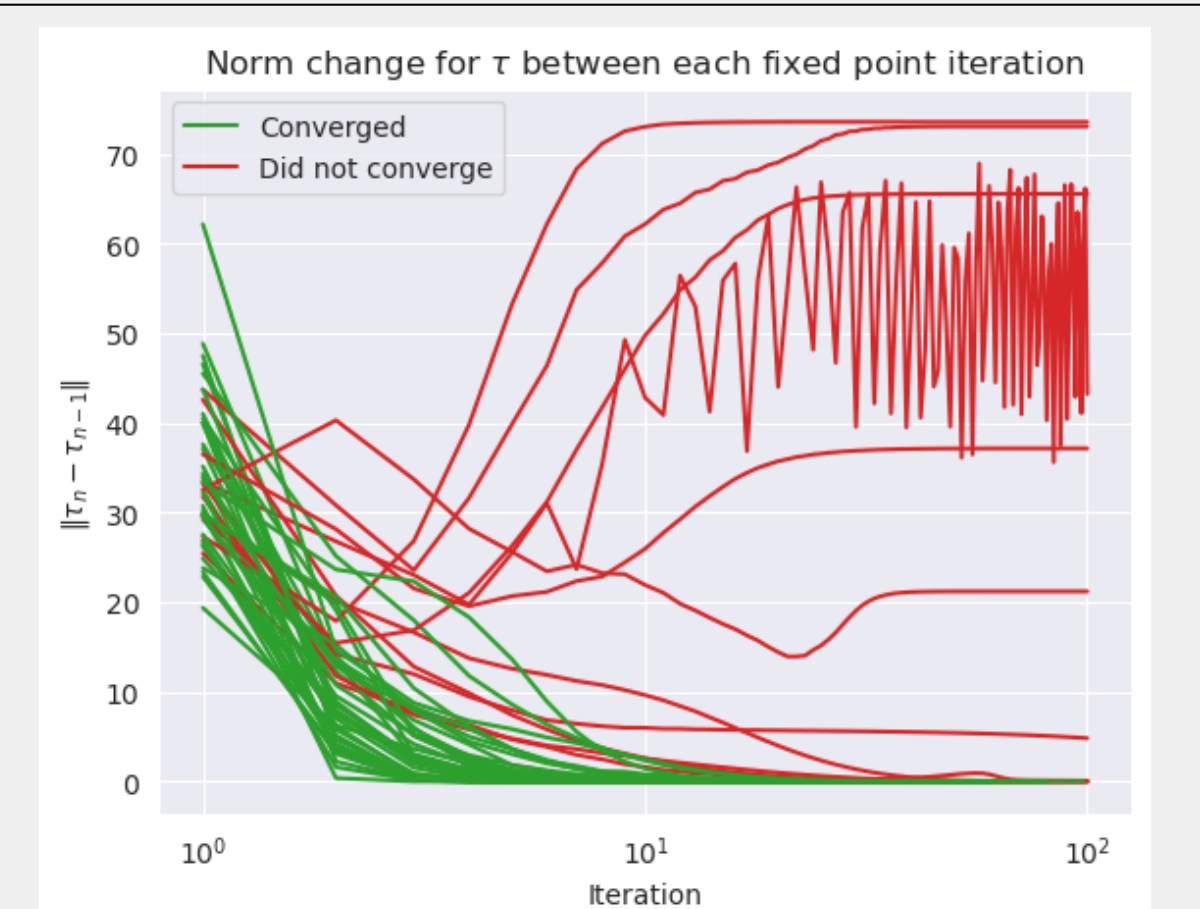


Figure 3. Dot-plot of the results for Cora.

Method	Time (s)	NMI	RI	M	CC	CC (per cluster)						
						1	2	3	4	5	6	7
Ground truth	-	-	-	0.64	0.09	0.19	0.06	0.12	0.23	0.10	0.22	0.16
Daudin-EM	3802	0.15	0.70	0.22	-	0	0.16	0	0.29	0.25	0	0.86
Newman-EM	27	0.18	0.76	0.53	-	0.25	0.06	0.16	0.28	0.12	0.29	0.18
Spectral	1	0.01	0.21	0.02	-	0	0	0.09	0	0	0.66	0

Table 2. Results on Cora.

## Discussion



- While [1] can reveal both homophilic and heterophilic properties of a graph, it is not as