Spectral Clustering for Directed Graphs via Likelihood Estimation on Stochastic Block Models

Ning Zhang
Department of Statistics
University of Oxford

Xiaowen Dong
Department of Engineering Science
University of Oxford

Mihai Cucuringu
Department of Mathematics
UCLA

Abstract

Graph clustering is a fundamental task in unsupervised learning with broad real-world applications. While spectral clustering methods for undirected graphs are well-established and guided by a minimum cut optimization consensus, their extension to directed graphs remains relatively underexplored due to the additional complexity introduced by edge directions. In this paper, we leverage statistical inference on stochastic block models to guide the development of a spectral clustering algorithm for directed graphs. Specifically, we study the maximum likelihood estimation under a widely used directed stochastic block model, and derive a global objective function that aligns with the underlying community structure. Building on its spectral relaxation, we propose two novel spectral clustering algorithms for directed graphs and establish theoretical guarantees for their misclustering error. Extensive experiments on synthetic and real-world datasets demonstrate significant performance gains over existing baselines.

1 INTRODUCTION

Graph clustering is a fundamental problem in unsupervised learning, providing tools for uncovering hidden structure in complex relational data from social networks (Oliveira and Gama, 2012), economics (Bennett et al., 2022), and neuroscience (Priebe et al., 2017). In graph clustering, the goal is to group vertices into "structurally equivalent" communities, where vertices from the same community interact with the rest of the graph in similar ways. Among various approaches, spectral clustering has become one of the most popular methods due to its computational efficiency. Most spectral clustering methods focus on undirected graphs, building on well-established optimization objectives such as cut minimization (Shi and Malik, 2000; Von Luxburg, 2007) or Girvan-Newman modularity maximization (Newman, 2013, 2016).

In many real-world networks, however, interactions are inherently directional as seen in causal relationships (Pearl and Verma, 1987), interbank debt (Acemoglu et al., 2015), migration pattern, and neuron synapses (Priebe et al., 2017). In such settings, edge directionality carries important structural information and is often tightly correlated with community membership. For instance, in neuron-neuron interaction networks, synaptic direction typically depends on the types of pre- and post-synaptic cells (Eichler et al., 2017). Such strong coupling between edge orientation and community membership is naturally modeled by a directed stochastic block model (DSBM) (Holland and Leinhardt, 1981; Nowicki and Snijders, 2001; Cucuringu et al., 2020), motivating our development of a model-based approach that leverages statistical inference on DSBMs to guide the design of clustering algorithms for directed graphs.

Existing spectral methods for clustering directed graphs are largely heuristic, relying on singular value decomposition (SVD) (Rohe et al., 2016; Wang et al., 2020), symmetrization techniques (Satuluri and Parthasarathy, 2011), or manually prespecified optimization objectives (Leicht and Newman, 2008; Fanuel et al., 2017; Laenen and Sun, 2020; Meilă and Pentney, 2007; Hayashi et al., 2022). Despite the empirical success of these approaches, it remains unclear whether the underlying optimization criteria are well-justified, or whether the detected communities reflect meaningful structure or mere artifacts. In contrast, clustering criteria for undirected graphs benefit from rich theoretical frameworks grounded in studying stochastic block models (SBMs) and their variants (Newman, 2016; Abbe et al., 2015; Bickel and Chen, 2009; Abbe et al., 2015; Hajek et al., 2016). For directed graphs, comparable theoretically

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

grounded studies remain underdeveloped. Motivated by this gap, we develop a principled model-based spectral clustering framework through likelihood estimation under a well-established DSBM. We summarize our main contributions as follows:

A principled optimization criterion for directed community detection (Section 3). We study the maximum likelihood estimator (MLE) under the DSBM for recovering planted communities. This model-based formulation provides a statistically grounded optimization criterion for clustering directed graphs (Theorem 1) in contrast to existing heuristics. The resulting objective admits a flow-based interpretation, which jointly considers edge density and edge orientation. Moreover, we show that this objective takes a Hermitian quadratic form, which naturally enables spectral algorithms.

Spectral algorithms with guarantees (Section 4). Building on the MLE objective, we develop two spectral algorithms that approximate it in different ways. `DirHSC` is a simple one-shot method that assigns equal importance to edge density and directionality, and is particularly effective when directional signal is strong. `LEHSC` is a data-driven, self-adaptive algorithm that iteratively estimates model parameters via pseudo-likelihood and updates the spectral embedding. For both algorithms, we provide rigorous analysis on their misclustering error (Corollary 2 and Theorem 3). Extensive experiments on synthetic and real-world data sets demonstrate significant performance gains over baseline methods. For example, on U.S. migration data (Figure 4), our methods identify an economically significant cluster of counties that are not captured by existing approaches.[1]

Related work. A number of classical spectral clustering methods for directed graphs rely on the singular vectors (Rohe et al., 2016, 2012; Wang et al., 2020) or eigenvectors of symmetrized matrix representations (Satuluri and Parthasarathy, 2011), with underlying optimization on graph connectivity patterns. For example, in spectral clustering on the bibliographic coupling matrix AA^T, the symmetrized matrix representation involves the number of shared offspring vertices. Another more recent line of spectral methods adopts complex-valued Hermitian matrices to represent directed graphs. In (Cucuringu et al., 2020), the authors cluster directed graphs using eigenvectors of a Hermitian matrix, employing $\pm i$ to represent directed edges, with a flow optimization heuristic (Hayashi et al., 2022; Laenen, 2019). Laenen and Sun (2020) proposes a different Hermitian matrix, where the k-th roots of unity are used to represent directed edges, leading to a higher-order flow optimization scheme. However, most existing studies lack a solid theoretical underpinning of why and when a heuristic approach might work, and this is exactly the gap we address in this study through a rigorous model-based analysis.

[1]Code is available at: https://github.com/NingZhang-Git/DirGraph_Cluster

2 PRELIMINARIES

Let $G(\mathcal{V}, \mathcal{E})$ be a directed graph on vertex set $\mathcal{V}$ and edge set $\mathcal{E}$. For a pair of vertices $u, v \in \mathcal{V}$, we denote $u \rightsquigarrow v$ if there is an edge pointing from u to v and we denote $u \not\sim v$ if there is no edge between u and v. A directed graph can be represented by its adjacency matrix $A \in \{0,1\}^{N\times N}$, where $A_{uv} = 1$ if and only if $u \rightsquigarrow v$. For a Hermitian matrix H, it has n real eigenvalues, and throughout this paper, the eigenvalues are consistently organized in descending order of magnitude, i.e., $|\lambda_1(H)| \geq |\lambda_2(H)| \geq \cdots |\lambda_n(H)|$. We use H^T to denote its transpose and H^* to denote the conjugate transpose. We use I to denote the identity matrix and J to denote the square all-one matrix. We use $\|H\|$ to denote the spectral norm, and $\|H\|_F$ to denote the Frobenius norm. For ease of reference, we refer to the summary in Table 2 of notions used in this paper.

2.1 Directed Stochastic Block Model

The canonical stochastic block model for directed graphs was introduced by Holland and Leinhardt (1981), known as the p_1 model. In this paper, we study an instance of the p_1 model, the DSBM (Cucuringu et al., 2020), which considers only directed (nonreciprocal) edges. Our motivation is that, in the context of clustering directed graphs, the main challenge lies in handling edge directionality, while clustering with undirected (reciprocal) edges is relatively well understood (Abbe, 2018). We note that this focus on directed edges does not preclude integration with undirected clustering methods; rather, it provides a complementary perspective that could be synthesized with existing undirected approaches in future hybrid models.

DSBMs with multiple communities involve intricate higher-order interactions between communities (see Figure 2a for an example), and explicitly modeling such interactions may introduce bias by enforcing specific structural assumptions. For generality and interpretability, we begin with the most basic two-community setup, with a source community $\mathcal{C}_1$ of size n_1 and a sink community $\mathcal{C}_2$ of size n_2. While our theoretical analysis focuses on this basic setting, the derived clustering algorithm naturally extends to multi-community problems via iterative bipartitioning (see Section 5 for details).

In a two-community DSBM, the community memberships are treated as fixed but unknown parameters (rather than as latent variables). We introduce a vector σ to indicate the community assignments,

where $\sigma_u = \sigma_v$ if and only if vertices u, v belong to the same community. Given a directed graph adjacency matrix sampled from the two-community DSBM (n_1, n_2, p, q, η), its entries are generated independently, conditioning on the community labeling σ as follows: For $u, v \in \mathcal{C}_1$ or $u, v \in \mathcal{C}_2$:

$$\begin{cases} A_{uv} = 1, A_{vu} = 0 & \text{w.p. } p/2, \\ A_{uv} = 0, A_{vu} = 1 & \text{w.p. } p/2, \\ A_{uv} = A_{vu} = 0 & \text{w.p. } 1-p. \end{cases}$$

For $u \in \mathcal{C}_1, v \in \mathcal{C}_2$:

$$\begin{cases} A_{uv} = 1, A_{vu} = 0 & \text{w.p. } (1-\eta)q\ , \\ A_{uv} = 0, A_{vu} = 1 & \text{w.p. } \eta q, \\ A_{uv} = A_{vu} = 0 & \text{w.p. } 1-q. \end{cases}$$

Here, the parameter p denotes the probability of forming an edge within communities, while q represents the edge probability between communities. Within-community edges are assigned directions uniformly at random, making their orientation symmetric in expectation. In contrast, the directionality of inter-community edges is controlled by the parameter $\eta \in (0, 0.5]$, where an edge from $\mathcal{C}_1$ to $\mathcal{C}_2$ occurs with probability $1-\eta$. This model captures community structure through both heterogeneous edge densities and asymmetric, community-dependent edge directions.

3 METHODOLOGY

Given a directed graph generated from a DSBM, our goal is to recover the underlying community structure. A natural approach is to formulate this task as a maximum likelihood estimation (MLE) problem, seeking the community assignment that best explains the observed edge patterns. Formally, the MLE is given by $\hat{\sigma}_{\text{MLE}} = \arg\max_\sigma \mathcal{L}(A; \sigma)$ with log-likelihood function $\mathcal{L}(A; \sigma) = \sum_{u<v} \log\left(\mathbb{P}(A_{u,v} \mid \sigma_u, \sigma_v)\right)$. As we show below in Theorem 1, this objective admits a compact reformulation that leads naturally to a spectral algorithm. The full proof is deferred to Appendix B.

Theorem 1 (MLE on DSBMs)**.** *Let A be the adjacency matrix of a directed graph sampled from the DSBM(n_1, n_2, p, q, η). Define the indicator vector $\mathbf{x} \in \{1, i\}^N$ such that $\mathbf{x}_u = i$ if $u \in \mathcal{C}_1$ and $\mathbf{x}_u = 1$ if $u \in \mathcal{C}_2$. Then, the MLE problem is equivalent to the following optimization problem*

$$\max_{\mathbf{x} \in \{1,i\}^N} \quad \mathbf{x}^* H_\theta \mathbf{x} \qquad \text{(Herm-MLE)}$$

where H_θ is a Hermitian matrix given by

$$H_\theta = w_r \left(A + A^T\right) + i w_i \left(A - A^T\right) + w_c (J - I), \quad (1)$$

with real-valued weights

$$w_r = \log\left(\frac{p^2(1-q)^2}{4\eta(1-\eta)q^2(1-p)^2}\right), \quad w_i = \log\left(\frac{1-\eta}{\eta}\right)$$
$$w_c = 2\log\left(\frac{1-p}{1-q}\right). \quad (2)$$

Theorem 1 shows that the MLE objective can be expressed as a quadratic form induced by a Hermitian matrix. This formulation separates the contributions of edge density, captured by the symmetric component $A + A^T$, and edge directionality, captured by the skew-symmetric component $A - A^T$. The corresponding weights w_r and w_i reflect the relative signal strength of these two sources of information. We make this connection more explicit through the following flow-based optimization interpretation.

A flow optimization view for MLE.

Definition 1. Given two clusters $\mathcal{C}_1, \mathcal{C}_2$ in a directed graph, we use $\mathbf{TF}(\mathcal{C}_1, \mathcal{C}_2)$ to denote the *total flow* between $\mathcal{C}_1$ and $\mathcal{C}_2$ and $\mathbf{NF}(\mathcal{C}_1, \mathcal{C}_2)$ to denote the *net flow* from $\mathcal{C}_1$ to $\mathcal{C}_2$, where

$$\mathbf{TF}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{u \in \mathcal{C}_1 v \in \mathcal{C}_2} (A_{uv} + A_{vu})$$
$$\mathbf{NF}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{u \in \mathcal{C}_1 v \in \mathcal{C}_2} (A_{uv} - A_{vu}).$$

Recall from Theorem 1 that the MLE reduces to maximizing

$$w_r \mathbf{x}^*(A + A^T)\mathbf{x} + i w_i \mathbf{x}^*(A - A^T)\mathbf{x} + w_c \mathbf{x}^*(J - I)\mathbf{x},$$

where $\mathbf{x} \in \{1, i\}^N$ encodes the cluster assignment. Each term admits a natural interpretation:

- **Edge density:** The first term, $\mathbf{x}^*(A + A^T)\mathbf{x}$, computes twice the number of edges within clusters. Equivalently, $\mathbf{x}^*(A + A^T)\mathbf{x} = 2|\mathcal{E}| - 2\mathbf{TF}(\mathcal{C}_1, \mathcal{C}_2)$, where we denote $|\mathcal{E}|$ the total number of edges.
- **Directional flow:** The second term $i\mathbf{x}^*(A - A^T)\mathbf{x}$ captures the net flow from $\mathcal{C}_1$ to $\mathcal{C}_2$, and simplifies to $i\mathbf{x}^*(A - A^T)\mathbf{x} = 2\mathbf{NF}(\mathcal{C}_1, \mathcal{C}_2)$.
- **Cluster size:** The last term, $\mathbf{x}^*(J - I)\mathbf{x}$, computes $|\mathcal{C}_1|^2 + |\mathcal{C}_2|^2 - N$, which is regularization term that penalizes imbalanced partitions.

Combining these terms, the MLE objective is equivalent to
$$-w_r \mathbf{TF}(\mathcal{C}_1, \mathcal{C}_2) + w_i \mathbf{NF}(\mathcal{C}_1, \mathcal{C}_2) + \frac{w_c}{2}(|\mathcal{C}_1|^2 + |\mathcal{C}_2|^2),$$

which balances total flow, directional asymmetry, and cluster size. The weights w_r and w_i reflect the relative importance of edge density and directionality, while the last term has limited impact since $w_c = O(|p - q|)$ is typically small.

Weights as signal strength. The directionality weight $w_i = \log\left(\frac{1-\eta}{\eta}\right)$ is non-negative for $\eta \in (0, 0.5]$, and decreases monotonically as direction noise η increases. This reflects a reduced emphasis on directionality when edge orientations become less reliable. The edge density weight w_r captures both the uncertainty in edge directions and the disparity between intra- and inter-cluster connectivity. It takes the form $w_r = \log\left(\frac{1}{4\eta(1-\eta)}\right) + 2\log\left(\frac{p(1-q)}{q(1-p)}\right)$. This term increases when η decreases (i.e., when edge directions are more informative), and when the difference between p and q becomes more pronounced.

Relation to other Hermitian formulations. The statistical inference formulation provides a model-based justification for the clustering criterion and the corresponding matrix representation. Our newly derived Hermitian matrix arises naturally by optimizing sufficient statistics (total flow and net flow) weighted by their relative informativeness. Such a framework enables a principled approach to designing new Hermitian data matrices for clustering, even without assuming a full generative model, requiring only coarse prior knowledge. about community structure. For example, when edge density is uninformative ($p \approx q$) and directionality is highly informative ($\eta \approx 0$), the Hermitian matrix $H_0 = A + A^T + i(A - A^T)$ serves as a practical approximation to the MLE-derived formulation in (1). Prior work by Cucuringu et al. (2020) proposed the matrix $H = i(A - A^T)$, which corresponds to a special case of our formulation. Their approach focuses solely on net flow and can be interpreted as MLE restricted to the cross-community block of the DSBM.

4 ALGORITHMS

The MLE formulation in Theorem 1 provides a principled optimization criterion for community detection in directed graphs. However, its direct application is hindered by two key challenges: computational intractability and unknown model parameters $\theta = (p, q, \eta)$. We address these challenges in a progressive manner. In Section 4.1, we develop a spectral relaxation of the MLE objective, which resolves the computational intractability with provable error bounds. To handle unknown model parameters, we propose in Section 4.2 a parameter-free approximation algorithm, and in Section 4.3 a self-adaptive algorithm that estimates the parameters from data.

4.1 Hermitian spectral clustering

Spectral relaxation. The optimization problem (Herm-MLE) is combinatorial due to the discrete constraint $\mathbf{x} \in \{1, i\}^N$, and is therefore NP-hard to solve exactly. To obtain a tractable formulation, we relax $\mathbf{x}$ to take values in the continuous complex domain. Specifically, we consider

$$\begin{aligned} \max \quad & \mathbf{x}^* H_\theta \mathbf{x} \\ \text{s.t.} \quad & \mathbf{x} \in \mathbb{C}^N, \|\mathbf{x}\|_2^2 = N. \end{aligned} \tag{SC-MLE}$$

This continuous problem (SC-MLE) is analytically solvable, and solution is given by the leading eigenvector of H_θ (scaled to have norm $\sqrt{N}$).

Projection step. Note that the leading eigenvector of H_θ is not unique, as multiplication by $e^{i\alpha}$ yields another solution for any $\alpha \in [0, 2\pi]$. To obtain discrete labels that are invariant to this global phase ambiguity, we embed each vertex into $\mathbb{R}^2$ using the real and imaginary parts of its corresponding entry. We then apply k-means clustering (with $k = 2$) to these points to recover the final partition.

Error analysis. The success of spectral relaxation clustering relies on the fact that the population matrix $\mathbb{E}[H_\theta]$ exhibits a clear community structure, where its entries depend only on the community membership. As a result, the leading eigenvector $\mathbf{v}_1(\mathbb{E}[H_\theta])$ encodes the true labels through two distinct values, which can be viewed as cluster centroids. In practice, we observe the data sample H_θ, which can be treated as a perturbed version of the unknown $\mathbb{E}[H_\theta]$. Classical matrix perturbation theory ensures that if H_θ is sufficiently close to $\mathbb{E}[H_\theta]$, the leading eigenvector $\hat{\mathbf{v}} = \mathbf{v}_1(H_\theta)$ remains well aligned with the top eigenvector of $\mathbb{E}[H_\theta]$, thus enabling accurate recovery of the community structure. Building on this intuition, we present in Theorem 2 an upper bound on the misclustering error. The proofs are provided in Appendix C.

Before presenting the main result, we introduce the following standard assumption

$$N p_{\max} = \Omega(\log N) \tag{A1}$$

with $p_{\max} = \max\{p, q\}$ and assume this holds throughout the rest of the paper. This assumption is to allow good concentration properties of the random graph (see Lemma 6), which is common in the realm of spectral methods equipped with theoretical guarantees.

Let σ be the true community assignment and $\hat{\sigma}_\theta$ be the $(1+\epsilon)$-approximate solution obtained by applying k-means++ on the spectral embedding of H_θ. We measure the misclustering error by

$$l(\sigma, \hat{\sigma}_\theta) = \sum_{u \in \mathcal{V}} \mathbb{1}\{\sigma_u \neq \hat{\sigma}_u\}.$$

Theorem 2 (Error bound of SC-MLE)**.** *For graphs generated from the DSBM (n_1, n_2, p, q, η), there exists $C = \Theta\left(\sqrt{w_r^2 + w_i^2}\right)$ (see (23)) and an absolute constant c_0, such that with probability at least $1 - N^{-c_0}$, the misclustering rate for spectral relaxation of MLE is*

$$\frac{l(\sigma, \hat{\sigma}_\theta)}{N} \leq \frac{64(2+\epsilon)C^2 p_{\max} \log N}{d^2 \Delta^2}. \tag{3}$$

Here Δ and d depend only on the population matrix $\mathbb{E}[H_\theta]$. Specifically, Δ denotes the eigengap $\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])$ (see (21)), and d denotes the distance between the two cluster centroids (see (30)).

Theorem 2 shows that recovery improves with increasing average degree and stronger separation in the population matrix, as captured by the eigengap and centroid distance. We next characterize the recovery regimes this implies, following standard notions in community detection (Abbe, 2018).

Definition 2. Exact recovery (strong consistency) requires all vertices to be correctly clustered, i.e.,

$$\sigma = \hat{\sigma} \quad \text{with probability } 1 - o(1).$$

Almost exact recovery (weak consistency) requires the fraction of misclassified vertices to vanish, i.e.,

$$l(\sigma, \hat{\sigma}) = o(N) \quad \text{with probability } 1 - o(1).$$

Partial recovery requires that for some constant $\beta \in (0, 1)$

$$l(\sigma, \hat{\sigma}) \leq \beta N \quad \text{with probability } 1 - o(1).$$

Remark 1. Theorem 2 characterizes the conditions under which different recovery regimes can be achieved.

Exact recovery: $\dfrac{Np_{\max} \log N}{d^2 \Delta^2} = o(1)$..

Almost exact recovery: $\dfrac{p_{\max} \log N}{d^2 \Delta^2} = o(1)$..

Partial recovery: $\dfrac{p_{\max} \log N}{d^2 \Delta^2} \leq \dfrac{\beta}{64(2+\epsilon)C^2}$.

Next, to better interpret the connection between the error bound in Theorem 2 and the noise level in DSBM, we consider a simplified case that admits an explicit analytical form for the error bound. Specifically, in Corollary 1, we specialize the error bound for a symmetric DSBM with homogeneous edge probabilities, a setting that lies below the detection threshold for undirected graphs. The proof of Corollary 1 can be found in Appendix C.8.

Corollary 1. *Consider directed graphs generated from the DSBM $(N/2, N/2, p, p, \eta)$ under the assumption $Np = \Omega(\log N)$. As $N \to \infty$, the misclustering error of spectral clustering is such that*

$$\frac{l(\sigma, \hat{\sigma}_\theta)}{N} = \Theta\left(\frac{\log N}{NpL^2}\right), \tag{4}$$

where $L = L(\eta)$ is a continuous, monotonically decreasing function of the edge directionality parameter η, with $L = 0$ when $\eta = 0.5$. The explicit form of $L(\eta)$ is provided in (48), and its behavior is illustrated in Figure 5b in Appendix C.

From (4), we observe that the upper bound on the misclustering error is inversely proportional to the average degree Np, which aligns with the intuition that a larger average degree provides more edge observations, rendering the generated graph more informative. Furthermore, as the edge direction noise η increases, the function $L(\eta)$ decreases, resulting in a higher misclustering error. In particular, as $\eta \to 0.5$, we have $L \to 0$, indicating a sharp increase in errors as directional information gradually disappears. This behavior is consistent with the intuition that a higher noise level in the edge orientations weakens the structural information to identify communities, rendering the recovery task more difficult. This homogeneous edge density setting highlights the importance of exploiting edge orientation in directed graph clustering: if edge directions are ignored, the model becomes statistically indistinguishable from an Erdős–Rényi graph, making community detection impossible.

4.2 Parameter-free approximation algorithm: `DirHSC`

As noted, the spectral relaxation of (Herm-MLE) described above requires knowledge of the model parameters $\theta = (p, q, \eta)$. To circumvent this limitation, we introduce `DirHSC` (Directional Hermitian Spectral Clustering), a parameter-free approximation of the MLE-based spectral method. Instead of using the weighted Hermitian matrix H_θ given by (1), we construct a simplified matrix by setting $w_r = w_i = 1$ and dropping the constant term in (1), yielding

$$H_0 = A + A^T + i(A - A^T). \tag{5}$$

This construction preserves the two key signals for clustering: edge density through $A + A^T$, and edge direction through $A - A^T$, while discarding the parameter-dependent weights. The resulting one-shot spectral algorithm, based on the leading eigenvector of H_0, is summarized in Algorithm 2.

Error analysis for `DirHSC`. We apply the same perturbation analysis as in Theorem 2 and present the clustering error bound for `DirHSC` in Corollary 2. For simplicity, we assume balanced communities $n_1 = n_2$; the more general result with imbalanced community size can be found in the proof in Appendix D.

Corollary 2. *For graphs generated from the DSBM$(N/2, N/2, p, q, \eta)$, the community assignment $\hat{\sigma}_0$ obtained by `DirHSC` satisfies*

$$\frac{l(\sigma, \hat{\sigma}_0)}{N} \leq \frac{Cp_{\max} \log N}{Nq^2(1 - 2\eta + 2\eta^2)d_0^2}. \tag{6}$$

where $d_0 = \left|1 - \frac{1-(1-2\eta)i}{|1+(1-2\eta)i|}\right|$.

We observe that the error bound (6) depends strongly on the edge direction parameter η, while being insensi-

tive to the strength of community signal in edge density, i.e., the disparity $|p-q|$. In particular, as $\eta \to 1/2$, the rescaled centroid distance $d_0 \to 0$, indicating that the spectral embedding induced by H_0 becomes poorly separated. Consequently, the error bound diverges even when edge density provides strong community signal, highlighting that `DirHSC` relies heavily on directional signal. This is consistent with our empirical observations (see Figure 1).

This behavior can be understood through the implicit inductive bias of `DirHSC`. From the MLE formulation in (2), we have

$$\lim_{\eta\to 0} w_r = \infty, \quad \lim_{\eta\to 0} w_i = \infty, \quad \lim_{\eta\to 0} \frac{w_r}{w_i} = 1.$$

This shows that when directional signal is strong (small η), the MLE objective assigns equal importance to edge density and direction, consistent with the equal weighting enforced by H_0. In contrast, as directional signal disappears (i.e., $\eta \to 0.5$), we have $w_i \to 0, w_r \to \log \frac{p^2(1-q)^2}{q^2(1-p)^2}$, and the MLE naturally downweights the uninformative directional component, relying increasingly on edge density. However, any fixed data matrix, including H_0, cannot adapt to this change in signal strength, as its weights are determined at construction time rather than inferred from data. This limitation motivates the self-adaptive algorithm `LEHSC` introduced in the next section.

4.3 Self-adaptive algorithm: `LEHSC`

The limitation identified above calls for an algorithm that can adaptively estimate the model parameters $\theta = (p, q, \eta)$ from data, and adjust the spectral embedding accordingly. We propose the `LEHSC` (Likelihood Estimation Hermitian Spectral Clustering), a self-adaptive algorithm that alternates between two steps inspired by Gong and Samaniego (1981); Newman (2016): given the current community assignment, it estimates parameters (p, q, η) using simple moment-based estimates, and then updates the community labels via spectral clustering on H_θ constructed from the estimated parameters. The overall procedure is summarized in Algorithm 1.

Error analysis for `LEHSC`. We denote the parameter estimated at step t by $\hat{\theta}^{(t)} = (\hat{p}^{(t)}, \hat{q}^{(t)}, \hat{\eta}^{(t)})$, and the corresponding clustering by $\hat{\sigma}^{(t)}$ with misclustering rate $\varepsilon^{(t)} \triangleq l\left(\sigma, \hat{\sigma}^{(t)}\right)/N$. To analyze how the error evolves across iterations of `LEHSC`, we impose the following assumptions:

$$\min\{n_1, n_2\} \;\geq\; \rho N \tag{A2}$$

$$\varepsilon^{(t)} \;\leq\; c_\rho \, q/p_{\max} \tag{A3}$$

$$q = \Theta(p_{\max}) \tag{A4}$$

where $\rho \in (0, 1/2]$ is a fixed constant and c_ρ is a small universal constant depending on ρ.

Algorithm 1: Likelihood Estimation Hermitian Spectral Clustering (`LEHSC`)

Input : Directed graph, maximum iteration T, initial parameters p_0, q_0, η_0
Output : Community labels $\hat{\sigma}$

for $t = 1$ **to** T **do**
  Compute H_θ using (1);
  Compute the top eigenvector $\hat{\mathbf{v}}$ of H_θ;
  Apply k-means to the embedding $[\Re(\hat{\mathbf{v}}), \Im(\hat{\mathbf{v}})]$ to partition into two clusters: $\mathcal{C}_1$ and $\mathcal{C}_2$;
  Update DSBM parameters based on current clustering;
    $p \leftarrow \frac{|\mathcal{E}| - \mathbf{TF}(\mathcal{C}_1,\mathcal{C}_2)}{\binom{|\mathcal{C}_1|}{2}+\binom{|\mathcal{C}_2|}{2}}$;
    $q \leftarrow \frac{\mathbf{TF}(\mathcal{C}_1,\mathcal{C}_2)}{|\mathcal{C}_1|\cdot|\mathcal{C}_2|}$;
    $\eta \leftarrow \min\left\{\frac{|\mathcal{C}_1\to\mathcal{C}_2|}{\mathbf{TF}(\mathcal{C}_1,\mathcal{C}_2)}, \; \frac{|\mathcal{C}_2\to\mathcal{C}_1|}{\mathbf{TF}(\mathcal{C}_1,\mathcal{C}_2)}\right\}$;
end
return *Final community labels* $\hat{\sigma}$.

The balance-community condition (A2) ensures that the two communities have sizes of the same order, i.e., $\Theta(n_1) = \Theta(n_2)$. The warm-start condition (A3) guarantees that the current partition has a non-trivial overlap with both communities, thereby avoiding degenerate or highly imbalanced assignments. Finally, condition (A4) ensures that cross-community edges are not too sparse, which is necessary for reliable parameter estimation. This assumption is mild, as when $q = o(p)$, clustering can be easily recovered using density-based methods (Abbe, 2018).

Theorem 3 (Recursive error update of `LEHSC`)**.** *Assume* (A1)–(A4) *hold. Let* $\hat{\sigma}^{(t)}$ *denote the community labels at iteration* t*, with misclustering rate* $\varepsilon^{(t)}$*. Then, with probability* $1 - o(1)$ *as* $N \to \infty$*, the parameter estimation errors satisfy*

$$\left|\hat{p}^{(t+1)} - p\right| \leq C_1 \, \varepsilon^{(t)} \, |p-q| \;+\; C_2 \, \frac{\sqrt{p_{\max} \log N}}{N}, \tag{7}$$

$$\left|\hat{q}^{(t+1)} - q\right| \leq C_1 \, \varepsilon^{(t)} \, |p-q| \;+\; C_2 \, \frac{\sqrt{p_{\max} \log N}}{N}, \tag{8}$$

$$\left|\hat{\eta}^{(t+1)} - \eta\right| \leq C_1 \; \varepsilon^{(t)} \frac{p_{\max}}{q} \;+\; C_2 \, \frac{\sqrt{\log N}}{N\sqrt{q}}. \tag{9}$$

Here C_1, C_2 *are constants depending on the balance and warm-start conditions. Consequently, applying the spectral step with* $H_{\hat{\theta}^{(t+1)}}$ *yields*

$$\frac{l(\sigma, \hat{\sigma}^{(t+1)})}{N} \leq \frac{64(2+\epsilon)C^2 \, p_{\max} \log N}{d^2_{\hat{\theta}^{(t+1)}} \Delta^2_{\hat{\theta}^{(t+1)}}}, \tag{10}$$

where $d^2_{\hat{\theta}^{(t+1)}}$ *and* $\Delta^2_{\hat{\theta}^{(t+1)}}$ *denote the centroid distance and eigengap of the population matrix* $\mathbb{E}[H_{\hat{\theta}^{(t+1)}}]$.

The proof details of Theorem 3 can be found in Appendix E. This theorem characterizes the recursive interplay between parameter estimation and clustering accuracy in LEHSC. In particular, the estimation error of $\hat{\theta}^{(t)}$ is controlled by the current misclustering rate $\varepsilon^{(t)}$, up to a vanishing statistical error term. As a result, improved community assignments lead to more accurate parameter estimates, which in turn yield a more informative spectral embedding. We note that a formal convergence guarantee does not follow directly from this analysis, as maintaining the warm-start condition across iterations is not guaranteed in general, particularly in noisy or sparse regimes. We consider this as an interesting direction for future work.

To provide a reliable starting point for LEHSC and place the algorithm in a regime where iterative refinement improves both parameter estimation and clustering accuracy, initialize the labels using an existing spectral method such as DirHSC, DI-SIM Rohe et al. (2012) and Herm Cucuringu et al. (2020). Our empirical studies suggest that the LEHSC is relatively robust to random initializations (see Appendix F.2).

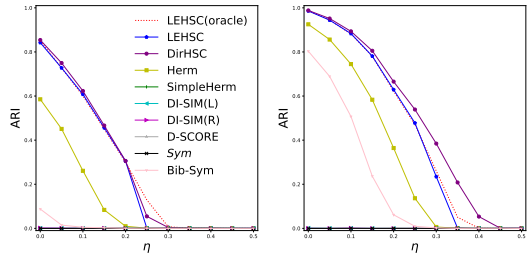
Complexity analysis for DirHSC and LEHSC. For both DirHSC and LEHSC, the main computational cost arises from computing the leading eigenvector of a Hermitian matrix, which requires $\mathcal{O}(|\mathcal{E}|)$ operations (see Appendix F.3 for details). In the setting of multiple clusters, we can further extend both algorithms to recursively bi-partition the largest remaining cluster at each step (see Algorithm 3 in Appendix F.4). The overall computational complexity is $\mathcal{O}(|\mathcal{E}|k)$ for DirHSC and $\mathcal{O}(|\mathcal{E}|kT)$ for LEHSC, where the maximum number of iterations T is typically set to 10 in practice.

5 EMPIRICAL STUDY

5.1 Synthetic DSBM graphs

Data generation overview. We conduct experiments on directed graphs sampled from the DSBM ensemble, with each community having a fixed size of 1000, and varying the number of communities as well as model parameters p, q, and η. Since spectral clustering typically performs well for dense graphs, we specifically focus on the more challenging sparse regime, where the edge probabilities p and q are slightly above $\log N/N$, the connectivity threshold of random graphs.

Baselines. We compare several baseline spectral clustering algorithms: (a) Hermitian-based methods: Herm (Cucuringu et al., 2020) and SimpHerm (Laenen and Sun, 2020); (b) SVD based methods: DI-SIM(L) and DI-SIM(R) (Rohe et al., 2016), and D-SCORE (Wang et al., 2020); and (c) symmetrization-based methods: naive symmetrization Sym (using $A + A^T$), bibliometric symmetrization Bib-Sym (using $AA^T + A^TA$) (Satuluri and Parthasarathy, 2011).



(a) $p = q = 0.5\%$ (b) $p = q = 1\%$

(c) $p = 1\%$, $q = 0.5\%$ (d) $p = 2\%$, $q = 1\%$

Figure 1: Experiments on two-community DSBM with varying model parameters.

Evaluation metric. We assess the clustering performance using the Adjusted Rand Index (ARI) (Gates and Ahn, 2017), which quantifies the similarity between the clustering outcomes and ground-truth labels. The ARI ranges from -1 to 1, with higher values indicating better clustering performance: ARI value of 1 indicates perfect recovery and 0 implies that the recovery is almost like a random guess. In each synthetic experiment, we independently sample 10 directed graphs with a fixed parameter set, and report the averaged ARIs over these graph samples.

Computation. All experiments were conducted on a MacBook Pro with an Apple M2 chip and 24 GB of RAM. For graphs with thousands of vertices, both DirHSC and LEHSC run in a few seconds.

5.1.1 Two-community DSBMs.

We conduct experiments on two-community DSBMs with varying model parameters and summarize the results in Figure 1. Overall, our algorithm LEHSC, consistently outperforms existing methods by a significant margin. In particular, we compare LEHSC with its oracle version, where the true parameters p, q, and η are provided, bypassing the iterative pseudo-maximum likelihood estimation. The nearly identical performance between LEHSC and LEHSC(oracle) suggests that the iterative procedure effectively recovers the underlying model parameters and approaches the oracle solution.

In the homogeneous cases ($p = q$, Figure 1a and Figure 1b), community recovery can only rely on the infor-

(a) 3-community DSBM (b) 4-community DSBM

Figure 2: Meta-graph for multiple-community DSBM

mation attached to the edge directions. In this regime, both `DirHSC` and `LEHSC` outperform all baseline methods, with `DirHSC` achieving slightly better performance due to its strong directional inductive bias.

In the inhomogeneous setting ($p \neq q$, Figure 1c and Figure 1d), clusters are informed by both the direction and the density of directed edges. In this setting, `LEHSC` consistently outperforms all other methods, as it adaptively balances these two sources of information. While `DirHSC` performs well when directional information is strong (small η), its performance deteriorates rapidly as η increases. As directional information diminishes (i.e., as η approaches 0.5), `DirHSC` fails to recover any meaningful clusters, which is consistent with our theoretical analysis in Corollary 2. In contrast, `LEHSC` gradually approaches the performance of symmetrization-based baselines, demonstrating its ability to adapt across different regimes.

5.1.2 Multi-community DSBMs

For DSBMs with multiple communities, interactions between different community pairs are independent of one another and may vary from pair to pair. For clarity, we use a meta-graph to represent these community-level relationships. In the meta-graph, each vertex corresponds to a community, and for each pair of vertices, a weighted and directed edge encode the orientation parameter $1-\eta$, i.e., the probability of an edge is pointing from the source community to the sink community. The absence of a meta-edge between a community pair implies $\eta = 0.5$, corresponding to uniformly random edge directions. Figures 2 and Figure 7 show example meta-graphs describing such community–community interaction patterns.

While exhaustive testing on all possible multi-community DSBMs is infeasible, we present experimental results on DSBMs with meta-graph structures illustrated in Figure 2. Overall, `LEHSC` achieves consistently competitive performance across all settings, highlighting its robustness compared to non-adaptive methods. `DirHSC` exhibits greater variability, yet it remains competitive across different regimes. We provide more empirical studies and discussions in Appendix F.5 to better understand the strengths and limitations of our method in multi-community settings.

Method	(a-1)	(a-2)	(b-1)	(b-2)
`Sym`	0.00	**0.66**	0.00	0.57
`Bib-Sym`	0.14	0.66	0.20	**0.73**
`DI-SIM`	0.00	0.28	0.01	0.32
`D-SCORE`	0.00	0.02	0.00	0.01
`Herm`	0.29	0.20	**0.44**	0.32
`SimpHerm`	0.00	0.01	0.04	0.05
`LEHSC`	**0.40**	**0.85**	**0.28**	**0.67**
`DirHSC`	**0.41**	0.44	0.31	0.59

Table 1: ARIs on clustering multi-community DSBMs. The prefix in the column header denotes corresponding meta-graphs from Figure 2, while suffix "1" denotes $p = q = 1\%$ and suffix "2" denotes $p = 2\%, q = 1\%$.

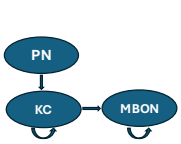
5.2 Larval Drosophila mushroom body connectome

In neuron-neuron interaction networks, vertices represent neurons and directed edges correspond to synapses, where the edge orientation often reflects the functional roles of neuron types. We study the Larval Drosophila mushroom body connectome, a real-world neuron connectivity graph that contains three types of neurons: Kenyon Cells (KC), Output Neurons (MBON), and Projection Neurons (PN). According to neuroscience research (Eichler et al., 2017; Priebe et al., 2017), these neuron types exhibit characteristic patterns of directed connectivity, and can be modeled using a meta-graph shown in Figure 3a. As illustrated in Figure 3b, the underlying community structure is strongly aligned with edge directionality: cross-community synapses follow consistent orientations, corresponding to a regime with minimal directional noise ($\eta = 0$).

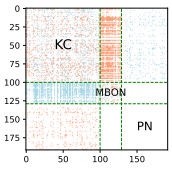
We apply spectral algorithms to recover the neuron types based solely on the observed directed graph. The adjacency matrices, reordered by the inferred cluster membership, along with the corresponding ARI scores, are shown in Figure 3c. Among the nine spectral algorithms evaluated, our methods `DirHSC` and `LEHSC` achieve the highest recovery accuracy. The reordered adjacency matrices reveal clear directional structures between the (1,3) and (2,3) blocks, corresponding to strong directed connections from KC to MBON and PN to KC. The strong performance of our methods, especially `DirHSC`, is consistent with their inductive bias toward settings in which community structure is primarily encoded through edge directionality.

5.3 US migration network

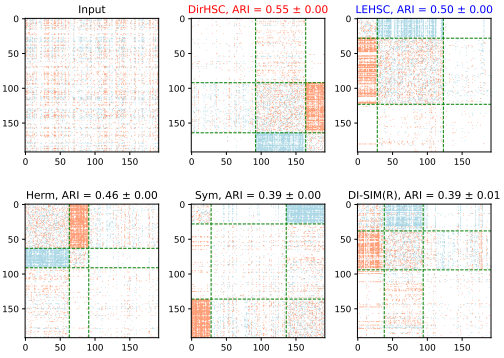
We study migration patterns in the United States using data from the 2000 U.S. Census, which documents county-to-county migration between 1995 and 2000 (U.S. Census Bureau, 2002; Cucuringu et al., 2020). We include 3074 mainland U.S. counties, and represent



(a) Relation between neuron types



(b) Ground truth

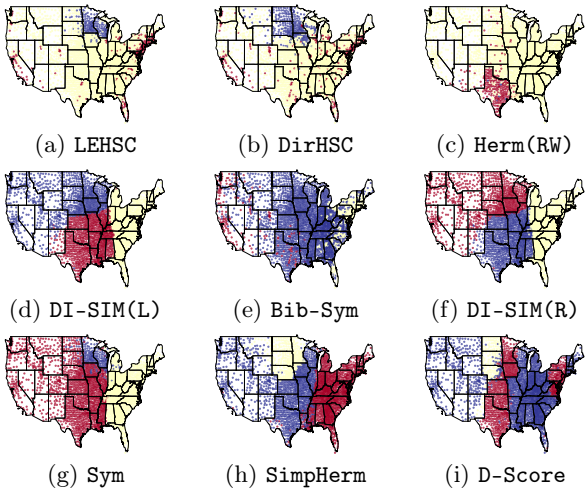


(c) Graph adjacency matrices reordered by clusters (only the best five algorithms). A red pixel indicates an outgoing edge from the vertex indexed by the row to the vertex indexed by the column, and blue pixel is inverse direction.

Figure 3: Cluster mushroom body connectome

migration flows between counties using a directed and weighted graph. In this graph, each edge weight corresponds to the number of individuals migrating from one county to another. To avoid a biased result dominated by extremely high degree vertices, we normalize the directed graph and use $D^{-1/2}AD^{-1/2}$, with the degree matrix D accounting for both incoming and outgoing edges. We then apply spectral clustering methods to partition the graph into $k = 3$ clusters and visualize the outcomes in Figure 4. Additional details and results for $k = 2, 5$ and 10 are provided in Appendix F.6.

Notably, `LEHSC` and `DirHSC` are the only two algorithms, that identify a distinct cluster of economically significant metropolitan areas (the red cluster in Figure 4a and Figure 4b). This includes regions surrounding New York State, major cities in California, and urban hubs like Seattle, Dallas, Atlanta, Chicago, and Denver. The presence of these distant yet economically vibrant regions within the same cluster suggests that those migrations were likely to be driven more by economic opportunity than by geographic proximity, an insight not captured by existing baseline methods. The other two clusters identified by `LEHSC` and `DirHSC` (shown in blue and yellow) exhibit more geographically cohesive patterns, consistent with regional migration trends. We also observe slight differences between the clusters produced by `LEHSC` and `DirHSC`, including surrounding counties of metropolitan hubs, disagreements on smaller hubs, and the spread of the blue cluster. These discrepancies can be attributed to the differing



(a) `LEHSC` (b) `DirHSC` (c) `Herm(RW)`
(d) `DI-SIM(L)` (e) `Bib-Sym` (f) `DI-SIM(R)`
(g) `Sym` (h) `SimpHerm` (i) `D-Score`

Figure 4: US counties clustered by migration data.

inductive biases of the two algorithms: `DirHSC` emphasizes directional patterns, whereas `LEHSC` adaptively balances both direction and density. Together, these results underscore the strength of `LEHSC` and `DirHSC` in detecting latent patterns in human mobility, revealing insights into migration dynamics beyond what existing baselines capture.

6 DISCUSSION

This paper develops a model-based framework for clustering directed graphs, providing a principled alternative to existing spectral methods that are largely heuristic. By deriving the clustering objective from MLE under the DSBM, we show that effective community recovery arises from jointly leveraging edge density and directional asymmetry induced by community structure. This perspective naturally leads to two complementary algorithms. `DirHSC` is a one-shot method that emphasizes directional information and performs well when edge orientation provides a strong signal. `LEHSC`, in contrast, is a self-adaptive algorithm that iteratively estimates model parameters from data and refines the spectral embedding, enabling it to balance density and directionality across different regimes. For both algorithms, we establish theoretical error guarantees and demonstrate strong empirical performance on synthetic and real-world datasets.

Despite these contributions, several limitations remain. Our theoretical analysis focuses on the two-community setting, extending the framework to the multi-community case with formal guarantees is an important direction for future work. In addition, our model does not fully capture real-world complexities such as degree heterogeneity, which may affect performance in practice. Extending the framework to more realistic models, such as degree-corrected SBMs, is a promising direction for future work.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. NZ acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) and IBM.

References

E. Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018.

E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *IEEE Transactions on information theory*, 62(1):471–487, 2015.

D. Acemoglu, A. Ozdaglar, and A. Tahbaz-Salehi. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608, 2015.

S. Bennett, M. Cucuringu, and G. Reinert. Lead–lag detection and network clustering for multivariate time series with an application to the us equity market. *Machine Learning*, 111(12):4497–4538, 2022.

P. J. Bickel and A. Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

Y. Chen, Y. Chi, J. Fan, C. Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14 (5):566–806, 2021.

M. Cucuringu, H. Li, H. Sun, and L. Zanetti. Hermitian matrices for clustering directed graphs: insights and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2020.

C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

K. Eichler, F. Li, A. Litwin-Kumar, Y. Park, I. Andrade, C. M. Schneider-Mizell, T. Saumweber, A. Huser, C. Eschbach, B. Gerber, et al. The complete connectome of a learning and memory centre in an insect brain. *Nature*, 548(7666):175–182, 2017.

M. Fanuel, C. M. Alaiz, and J. A. Suykens. Magnetic eigenmaps for community detection in directed networks. *Physical Review E*, 95(2):022302, 2017.

A. J. Gates and Y.-Y. Ahn. The impact of random models on clustering similarity. *arXiv preprint arXiv:1701.06508*, 2017.

G. Gong and F. J. Samaniego. Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics*, pages 861–869, 1981.

B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Transactions on Information Theory*, 62(5): 2788–2797, 2016.

K. Hayashi, S. G. Aksoy, and H. Park. Skew-symmetric adjacency matrices for clustering directed graphs. *arXiv preprint arXiv:2203.01388*, 2022.

P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76 (373):33–50, 1981.

W. R. Inc. Mathematica, Version 14.0.

S. Laenen. Directed graph clustering using hermitian laplacians. *Master's thesis*, 2019.

S. Laenen and H. Sun. Higher-order spectral clustering of directed graphs. *Advances in neural information processing systems*, 33:941–951, 2020.

J. Lei and A. Rinaldo. Consistency of spectral clustering in stochastic block models. 2015.

E. A. Leicht and M. E. Newman. Community structure in directed networks. *Physical review letters*, 100 (11):118703, 2008.

M. Meilă and W. Pentney. Clustering by weighted cuts in directed graphs. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 135–144. SIAM, 2007.

M. E. Newman. Spectral methods for community detection and graph partitioning. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 88(4):042822, 2013.

M. E. Newman. Equivalence between modularity optimization and maximum likelihood methods for community detection. *Physical Review E*, 94(5):052315, 2016.

K. Nowicki and T. A. B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American statistical association*, 96(455):1077–1087, 2001.

M. Oliveira and J. Gama. An overview of social network analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2):99–115, 2012.

J. Pearl and T. Verma. The logic of representing dependencies by directed graphs. In *Proceedings of the sixth National conference on Artificial intelligence-Volume 1*, pages 374–379, 1987.

C. E. Priebe, Y. Park, M. Tang, A. Athreya, V. Lyzinski, J. T. Vogelstein, Y. Qin, B. Cocanougher, K. Eich-

ler, M. Zlatic, et al. Semiparametric spectral modeling of the drosophila connectome. *arXiv preprint arXiv:1705.03297*, 2017.

K. Rohe, T. Qin, and B. Yu. Co-clustering for directed graphs: the stochastic co-blockmodel and spectral algorithm di-sim. *arXiv preprint arXiv:1204.2296*, 2012.

K. Rohe, T. Qin, and B. Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45):12679–12684, 2016.

V. Satuluri and S. Parthasarathy. Symmetrizations for clustering directed graphs. In *Proceedings of the 14th international conference on extending database technology*, pages 343–354, 2011.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

J. A. Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

U.S. Census Bureau. County-to-county migration flow files. https://www.census.gov/data/tables/2000, 2002. Accessed: 2025-05-15.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17:395–416, 2007.

Z. Wang, Y. Liang, and P. Ji. Spectral algorithms for community detection in directed networks. *Journal of Machine Learning Research*, 21(153):1–45, 2020.

H. Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Summary on notations

Notation	Definition
$G(\mathcal{V}, \mathcal{E})$	Graph with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$
$u \rightsquigarrow v$	There is an edge pointing from vertex u to vertex v
$u \not\sim v$	There is no edge between vertex u and vertex v
A	Graph adjacency matrix, $A \in \{0,1\}^{n\times n}$ and $A_{uv} = 1$ iff $u \rightsquigarrow v$
$\mathcal{C}_1$	Source community
$\mathcal{C}_2$	Target community
$\mathcal{V}$	Set of all vertices, $\mathcal{V} = \mathcal{C}_1 \cup \mathcal{C}_2$
H_θ	Hermitian matrix derived from MLE on DSBM (see (1))
σ	A general community indicator vector, $\sigma_u = \sigma_v$ iff $c(u) = c(v)$
$\mathcal{L}(A;\sigma)$	Log-likelihood function
$\mathbf{TF}(\mathcal{C}_1, \mathcal{C}_2)$	Total flow between $\mathcal{C}_1$ and $\mathcal{C}_2$, $\mathbf{TF}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{u\in\mathcal{C}_1 v\in\mathcal{C}_2} (A_{uv} + A_{vu})$
$\mathbf{NF}(\mathcal{C}_1, \mathcal{C}_2)$	Net flow from $\mathcal{C}_1$ to $\mathcal{C}_2$, $\mathbf{NF}(\mathcal{C}_1, \mathcal{C}_2) = \sum_{u\in\mathcal{C}_1 v\in\mathcal{C}_2} (A_{uv} - A_{vu})$
$\lvert\mathcal{C}_1 \to \mathcal{C}_2\rvert$	Number of edges from $\mathcal{C}_1$ to $\mathcal{C}_2$
H^T	Transpose of H
H^*	Conjugate transpose of H
H_{j*}	The j-th row vector of H
$[H_1, H_2]$	Concatenating columns of H_1 and H_2
$\lVert H\rVert$	Spectral norm of H, $\lVert H\rVert = \lvert\lambda_1(H)\rvert$
$\lVert H\rVert_F$	Frobenius norm of H, $\lVert H\rVert_F = \sqrt{\sum_j \lambda_j^2(H)}$
$\langle H_1, H_2\rangle$	Frobenius inner product, $\langle H_1, H_2\rangle = \mathrm{Tr}(H_1^* H_2)$
$\mathrm{diag}(H)$	Create a diagonal matrix by taking the main diagonal elements of H
$\Re(H)$	Take the real part of the matrix H;
$\Im(H)$	Take the imaginary part of the matrix H
$\mathbf{v_j}(H)$	The j-th eigenvector of H
$\lambda_j(H)$	The j-th eigenvalue of H
ARI	Adjusted Rand Index
$M \in \{0,1\}^{N\times 2}$	Membership matrix
$\mathbb{1}(\cdot)$	Indicator function, $\mathbb{1}(p) = 1$ if claim p is true, otherwise $\mathbb{1}(p) = 0$
$\mathbb{1}_{\mathcal{C}_1}$	Binary indicator vector for community $\mathcal{C}_1$, $\mathbb{1}_u = 1$ if $u \in \mathcal{C}_1$ otherwise $\mathbb{1}_u = 0$
$\mathbb{1}_{\mathcal{C}_2}$	Binary indicator vector for community $\mathcal{C}_2$, $\mathbb{1}_u = 1$ if $u \in \mathcal{C}_2$ otherwise $\mathbb{1}_u = 0$
$g_n = o(f_n)$	g_n is asymptotically dominated by f_n, i.e., $\lim_{n\to\infty} \frac{g_n}{f_n} = 0$
$g_n = O(f_n)$	g_n is asymptotically bounded above by f_n, i.e., $\limsup_{n\to\infty} \frac{g_n}{f_n} < \infty$
$g_n = \Theta(f_n)$	$\limsup_{n\to\infty} \frac{g_n}{f_n} < \infty$ and $\liminf_{n\to\infty} \frac{g_n}{f_n} > 0$
$g_n = \Omega(f_n)$	g_n bounded below by f_n asymptotically, i.e., $\limsup_{n\to\infty} \frac{g_n}{f_n} > 0$
$g_n = \omega(f_n)$	g_n dominate f_n asymptotically, i.e., $\limsup_{n\to\infty} \frac{g_n}{f_n} = \infty$

Table 2: Summary on notations

B Proof of MLE

We detail the derivation of the optimization problem (MLE) from the maximum likelihood estimator. To start with, we explicitly express the likelihood function as a matrix, which simply relies on subdividing the likelihood function according to which community an edge belongs to.

Consider a directed graph with adjacency matrix A sampled from the model DSBM(n_1, n_2, p, q, η). Then, applying the maximum likelihood estimation is equivalent to solving the following combinatorial optimization

problem

$$\begin{aligned} \max \quad & \frac{1}{2}\left\langle M_{intra}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T \right\rangle + \left\langle M_{inter}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T \right\rangle \\ s.t. \quad & \mathbb{1}_{\mathcal{C}_1} \in \{0,1\}^N \\ & \mathbb{1}_{\mathcal{C}_1} + \mathbb{1}_{\mathcal{C}_2} = \mathbb{1} \end{aligned} \tag{MLE}$$

where $\mathbb{1}_{\mathcal{C}_1}, \mathbb{1}_{\mathcal{C}_2} \in \{0,1\}^N$ are the indicator vectors for cluster $\mathcal{C}_1$ and $\mathcal{C}_2$ separately, and

$$\begin{aligned} M_{intra} &= \log(1/2p)(A + A^T) + \log(1-p)(J - I - A - A^T), \\ M_{inter} &= \log(q(1-\eta))A + \log(\eta q)A^T + \log(1-q)(J - I - A - A^T), \end{aligned}$$

are derived from the log-likelihood functions for intra-community and inter-community edges.

Proof. Let A be the adjacency matrix of a directed graph generated from DSBM(n_1, n_2, p, q, η). For a particular clusterization of the graph, we use c to denote its community labeling function $c : \mathcal{V} \to \{\mathcal{C}_1, \mathcal{C}_2\}$, and we use the vectors $\mathbb{1}_{\mathcal{C}_1}, \mathbb{1}_{\mathcal{C}_2} \in \{0,1\}^N$ to indicate community $\mathcal{C}_1$ and $\mathcal{C}_2$ separately, where $\mathbb{1}_{\mathcal{C}_1} + \mathbb{1}_{\mathcal{C}_2} = \mathbb{1}$. The log likelihood function of A given $\mathbb{1}_{\mathcal{C}_1}$ and $\mathbb{1}_{\mathcal{C}_2}$ can be decomposed as follows

$$\begin{aligned} \mathcal{L}(A;\sigma) &= \log \mathbb{P}(A|\mathbb{1}_{\mathcal{C}_1}, \mathbb{1}_{\mathcal{C}_2}) = \sum_{u<v} \log \mathbb{P}(A_{uv}|c(u), c(v)) \\ &= \sum_{\substack{u<v \\ c(u)=c(v)}} \log \mathbb{P}(A_{uv}|c(u), c(v)) + \sum_{\substack{u<v \\ c(u)=\mathcal{C}_1, c(v)=\mathcal{C}_2}} \log \mathbb{P}(A_{uv}|c(u), c(v)), \end{aligned} \tag{11}$$

where the first term in (11) is only summing over intra-community pair, and the second term handles the inter-community pair.

For an intra-community vertex pair u, v, the log-likelihood function is

$$\log \mathbb{P}(A_{uv}|c(u) = c(v)) = \begin{cases} \log(1/2p) & \text{if } u \rightsquigarrow v, \\ \log(1/2p) & \text{if } v \rightsquigarrow u, \\ \log(1-p) & \text{if } u \not\sim v. \end{cases}$$

This intra-community log-likelihood function coincides with the matrix

$$M_{intra} \triangleq \log(1/2p)(A + A^T) + \log(1-p)(J - I - A - A^T), \tag{12}$$

on the entries that represent intra-community pairs, thus allowing us to convert the intra-community summation in (11) into the following matrix multiplication form

$$\sum_{\substack{u<v \\ c(u)=c(v)}} \log \mathbb{P}(A_{uv}|c(u), c(v)) = \frac{1}{2}\left\langle M_{intra}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T \right\rangle. \tag{13}$$

For an inter-community vertex $u \in \mathcal{C}_1, v \in \mathcal{C}_2$, the log-likelihood function is

$$\log \mathbb{P}(A_{uv}|c(u) = \mathcal{C}_1, c(v) = \mathcal{C}_2) = \begin{cases} \log((1-\eta)q) & \text{if } u \rightsquigarrow v, \\ \log(\eta q) & \text{if } v \rightsquigarrow u, \\ \log(1-q) & \text{if } u \not\sim v. \end{cases}$$

Similar to the approach followed for the intra-community case, we convert the inter-community summation in (11) into the matrix multiplication form

$$\sum_{\substack{u<v \\ c(u)=\mathcal{C}_1, c(v)=\mathcal{C}_2}} \log \mathbb{P}(A_{uv}|c(u), c(v)) = \left\langle M_{inter}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T \right\rangle. \tag{14}$$

where

$$M_{inter} \triangleq \log\left((1-\eta)q\right)A + \log\left(\eta q\right)A^T + \log\left(1-q\right)(J - I - A - A^T). \tag{15}$$

Combining (13), (14) and (11), we have

$$\log \mathbb{P}(A|\mathbb{1}_{\mathcal{C}_1}, \mathbb{1}_{\mathcal{C}_2}) = \frac{1}{2}\left\langle M_{intra}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle + \left\langle M_{inter}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle. \tag{16}$$

□

To arrive at a more compact expression for the optimization formulation, we introduce an equivalent Hermitian matrix optimization framework. The transformation from the real-valued matrix optimization to the Hermitian optimization builds on the following observation.

Lemma 1. *Consider an arbitrary Hermitian matrix $H = \Re(H) + i\Im(H)$, where $\Re(H) \in \mathbb{R}^{n\times n}$ with all 0 diagonal entries is symmetric, and $\Im(H) \in \mathbb{R}^{n\times n}$ is skew-symmetric. Let $\mathbf{x} \in \{i, 1\}^n$ be the complex community indicator vector, where $\mathbf{x}_u = i$ for $u \in \mathcal{C}_1$. Then, the quadratic form $\mathbf{x}^* H\mathbf{x}$ is the sum of entries in $\Re(H)$ that are in the same community, plus the sum of entries in $\Im(H)$ that belong to different communities, i.e.,*

$$\mathbf{x}^* H\mathbf{x} = \sum_{\substack{u,v\in\mathcal{C}_1 \\ or\ u,v\in\mathcal{C}_2}} \Re(H)_{uv} + \sum_{\substack{u\in\mathcal{C}_1 v\in\mathcal{C}_2 \\ or\ u\in\mathcal{C}_2, v\in\mathcal{C}_1}} \Im(H)_{uv} = 2\sum_{\substack{u<v \\ u,v\in\mathcal{C}_1 \\ or\ u,v\in\mathcal{C}_2}} \Re(H)_{uv} + 2\sum_{\substack{u<v, \\ u\in\mathcal{C}_1 v\in\mathcal{C}_2 \\ or\ u\in\mathcal{C}_2, v\in\mathcal{C}_1}} \Im(H)_{uv}.$$

In other words,

$$\mathbf{x}^* H\mathbf{x} = \langle \Re(H), \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\rangle + 2\langle \Im(H), \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\rangle. \tag{17}$$

With the real-valued MLE optimization formula derived in Lemma B and the observation made in Lemma 1, we are ready to prove the Hermitian optimization formulation of the MLE on DSBM in Theorem 1.

Theorem 1 (MLE on DSBMs)**.** *Let A be the adjacency matrix of a directed graph sampled from the DSBM(n_1, n_2, p, q, η). Define the indicator vector $\mathbf{x} \in \{1, i\}^N$ such that $\mathbf{x}_u = i$ if $u \in \mathcal{C}_1$ and $\mathbf{x}_u = 1$ if $u \in \mathcal{C}_2$. Then, the MLE problem is equivalent to the following optimization problem*

$$\max_{\mathbf{x}\in\{1,i\}^N} \quad \mathbf{x}^* H_\theta \mathbf{x} \tag{Herm-MLE}$$

where H_θ is a Hermitian matrix given by

$$H_\theta = w_r\left(A + A^T\right) + iw_i\left(A - A^T\right) + w_c(J - I), \tag{1}$$

with real-valued weights

$$\begin{aligned} w_r &= \log\left(\frac{p^2(1-q)^2}{4\eta(1-\eta)q^2(1-p)^2}\right), \quad w_i = \log\left(\frac{1-\eta}{\eta}\right) \\ w_c &= 2\log\left(\frac{1-p}{1-q}\right). \end{aligned} \tag{2}$$

Proof. From Lemma B, we derive the log-likelihood

$$\log \mathbb{P}(A|\mathbb{1}_{\mathcal{C}_1}, \mathbb{1}_{\mathcal{C}_2}) = \frac{1}{2}\left\langle M_{intra}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle + \left\langle M_{inter}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle. \tag{18}$$

Compared this equation with the observation (17) in Lemma 1, we need the matrix corresponding to the intra-cluster log-likelihood matrix to be a symmetric matrix and need the inter-cluster log-likelihood matrix to be a skew-symmetric matrix, so that we can directly apply (17) to convert the real-valued objective into a more compact complex-valued expression.

From the definition in (12) and (15), we have that $M_{intra} = M_{intra}^T$, however the inter-cluster log-likelihood matrix M_{inter} is not skew-symmetric. Note that we can be decomposed the inter-cluster log-likelihood matrix M_{inter} into a symmetric matrix plus a skew-symmetric matrix as follows

$$M_{inter} = \frac{1}{2}(M_{inter} + M_{inter}^T) + \frac{1}{2}(M_{inter} - M_{inter}^T),$$

where $M_{inter} + M_{inter}^T$ is symmetric and $M_{inter} - M_{inter}^T$ is skew-symmetric. Correspondingly, we have

$$\left\langle M_{inter}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T \right\rangle = \frac{1}{2}\langle M_{inter} + M_{inter}^T, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\rangle + \frac{1}{2}\langle M_{inter} - M_{inter}^T, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\rangle, \tag{19}$$

where the first term sums over the inter-cluster entries of a symmetric matrix and the second term sums over the inter-cluster entries of a skew-symmetric matrix. Due to the symmetry in the first term of (19), we can further write it as

$$\frac{1}{2}\langle M_{inter} + M_{inter}^T, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\rangle = \frac{1}{4}\langle M_{inter} + M_{inter}^T, J - \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T - \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\rangle \tag{20}$$

Combining (19), (20) and (18), we can rewrite the log-likelihood objective as

$$\begin{aligned}
&\log \mathbb{P}(A|\mathbb{1}_{\mathcal{C}_1}, \mathbb{1}_{\mathcal{C}_2}) = \frac{1}{2}\left\langle M_{intra}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle + \left\langle M_{inter}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle \\
&= \frac{1}{2}\left\langle M_{intra}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle + \frac{1}{2}\langle M_{inter} + M_{inter}^T, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\rangle + \frac{1}{2}\langle M_{inter} - M_{inter}^T, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\rangle \\
&= \frac{1}{2}\left\langle M_{intra}, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle + \frac{1}{4}\langle M_{inter} + M_{inter}^T, J - \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T - \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\rangle \\
&\quad + \frac{1}{2}\langle M_{inter} - M_{inter}^T, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\rangle
\end{aligned}$$

Because the term $\langle M_{inter} + M_{inter}^T, J\rangle$ is always a constant and resealing the objective function by a constant factor 4 does not affect the optimal solution, therefore solving the (MLE) in Lemma B is equivalent to solve the following

$$\begin{aligned}
\max \quad & \left\langle 2M_{intra} - (M_{inter} + M_{inter}^T), \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_1}^T + \mathbb{1}_{\mathcal{C}_2}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle + 2\left\langle M_{inter} - M_{inter}^T, \mathbb{1}_{\mathcal{C}_1}\mathbb{1}_{\mathcal{C}_2}^T\right\rangle \\
s.t. \quad & \mathbb{1}_{\mathcal{C}_1} \in \{0,1\}^N \\
& \mathbb{1}_{\mathcal{C}_1} + \mathbb{1}_{\mathcal{C}_2} = \mathbb{1}
\end{aligned}$$

Using (17) from Lemma 1, we convert the above real-valued optimization problem to the following complex-valued equivalence

$$\begin{aligned}
\max \quad & \mathbf{x}^* H_\theta \mathbf{x} \\
s.t. \quad & \mathbf{x} \in \{i,1\}^N
\end{aligned}$$

where the Hermitian matrix H_θ has

$$\begin{aligned}
\Re(H_\theta) &= 2M_{intra} - (M_{inter} + M_{inter}^T) \\
&= \log\frac{p^2}{4\eta(1-\eta)q^2}(A + A^T) + 2\log\frac{1-p}{1-q}(J - I - A - A^T) \\
&= \log\left(\frac{p^2(1-q)^2}{4\eta(1-\eta)q^2(1-p)^2}\right)(A + A^T) + 2\log\left(\frac{1-p}{1-q}\right)(J - I), \\
\Im(H_\theta) &= M_{inter} - M_{inter}^T \\
&= \log\left(\frac{1-\eta}{\eta}\right)(A - A^T).
\end{aligned}$$

□

C Proofs for Theorem 2

C.1 Proof overview

The intuition behind the success of spectral relaxation clustering is that the expected Hermitian matrix $\mathbb{E}[H_\theta]$ has community-dependent structure, and its leading eigenvector $\mathbf{v}^* = \mathbf{v}_1(\mathbb{E}[H_\theta])$ exactly encodes the true community labels through two distinct values. In practice, we observe the empirical matrix H_θ, which serves as a perturbed version of $\mathbb{E}[H_\theta]$. Classical matrix perturbation theory ensures that if H_θ is sufficiently close to $\mathbb{E}[H_\theta]$, the leading eigenvector $\hat{\mathbf{v}} = \mathbf{v}_1(H_\theta)$ remains informative and enables accurate recovery of the community structure. Our proof is structured as follows:

(i) Using matrix perturbation theory, we first bound the eigenvector perturbation $\|\mathbf{v}\mathbf{v}^* - \hat{\mathbf{v}}\hat{\mathbf{v}}^*\|_F$;

(ii) Using the Matrix-Bernstein inequality from random matrix theory, we provide a high-probability upper bound on the random perturbation $\|H_\theta - \mathbb{E}[H_\theta]\|$;

(iii) Combining the results from the above two steps, we perform an error analysis on the k-means clustering step, and derive the final spectral clustering error bound.

We first characterize the eigenspace properties of $\mathbb{E}[H_\theta]$ in the following Lemma 2.

Lemma 2. *For the DSBM(n_1, n_2, p, q, η), the population matrix $\mathbb{E}[H_\theta]$ has a unique largest eigenvalue. The top eigenvector $\mathbf{v}$ has exactly two distinct values that indicate the community labels, where the distance d between them can be easily computed using* (30)*. Moreover, the eigengap is*

$$\Delta = |\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])| = \min\{2\Delta_0, |1/2N(w_r p + w_c)| + \Delta_0\} \geq \Delta_0, \tag{21}$$

where

$$\Delta_0 = 1/2\sqrt{N^2(w_r p + w_c)^2 - 4n_1 n_2\left((w_r p + w_c)^2 - |w_r + w_c + i w_i(1-2\eta)q|^2\right)}. \tag{22}$$

We consider H_θ as a perturbed version of $\mathbb{E}[H_\theta]$, and denote $R = H_\theta - \mathbb{E}[H_\theta]$. The perturbation on the eigenspace and eigenvalues is characterized by the following two well-known results, namely the Davis-Kahan perturbation bound (Theorem 4) and Weyl's inequality (Theorem 5), from which we derive an upper bound on the eigenspace misalignment distance $\|\mathbf{v}\mathbf{v}^* - \hat{\mathbf{v}}\hat{\mathbf{v}}^*\|_F$.

Lemma 3. *Given a directed graph from DSBM(n_1, n_2, p, q, η) and its Hermitian matrix representation H_θ, the projection matrix of the top eigenvector is such that*

$$\|\mathbf{v}\mathbf{v}^* - \hat{\mathbf{v}}\hat{\mathbf{v}}^*\|_F \leq 2\sqrt{2}\frac{\|R\|}{\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])}.$$

We apply the Matrix Bernstein inequality for Hermitian matrices, and obtain a upper bound on $\|R\|$.

Lemma 4. *Consider a directed graph from DSBM (n_1, n_2, p, q, η) and its Hermitian matrix representation H_θ. Assume that $Np_{\max} = \Omega(\log N)$. Then there exists an absolute constant ϵ and*

$$C = (2+\epsilon)\sqrt{w_r^2 + w_i^2}\left(\frac{\log N}{Np_{\max}} + 1\right) = \Theta\left(\sqrt{w_r^2 + w_i^2}\right), \tag{23}$$

such that the random perturbation $R = H_\theta - \mathbb{E}[H_\theta]$ has

$$\mathbb{P}(\|R\| \geq C\sqrt{Np_{\max}\log N}) \leq N^{-\epsilon}. \tag{24}$$

C.2 Error analysis Theorem 2

Theorem 2 (Error bound of SC-MLE)**.** *For graphs generated from the DSBM (n_1, n_2, p, q, η), there exists $C = \Theta\left(\sqrt{w_r^2 + w_i^2}\right)$ (see* (23)*) and an absolute constant c_0, such that with probability at least $1 - N^{-c_0}$, the misclustering rate for spectral relaxation of MLE is*

$$\frac{l(\sigma, \hat{\sigma}_\theta)}{N} \leq \frac{64(2+\epsilon)C^2 p_{\max}\log N}{d^2\Delta^2}. \tag{3}$$

Here Δ and d depend only on the population matrix $\mathbb{E}[H_\theta]$. Specifically, Δ denotes the eigengap $\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])$ (see (21)*), and d denotes the distance between the two cluster centroids (see* (30)*).*

Proof. Recall that the key steps of our spectral clustering algorithm involve: first compute the top eigenvector of H_θ, and then cluster the vertices using k-means on the embedding space given by the real and imaginary part of the top eigenvector. We use $\hat{U} = [\Re(\hat{\mathbf{v}}), \Im(\hat{\mathbf{v}})]$ to denote the embedding space given by the concatenation of the real and imaginary part of the top eigenvector of H_θ and $U = [\Re(\mathbf{v}), \Im(\mathbf{v})]$ for that of $\mathbb{E}[H_\theta]$, where both $U, \hat{U} \in \mathbb{R}^{N\times 2}$. For the clustering outcomes, we denote by $\hat{\sigma}_{\text{SC-MLE}}$ the clustering result using H_θ, and use σ to represent the clustering given by $\mathbb{E}[H_\theta]$. First, note that from Lemma 2, we conclude that the leading eigenvector of $\mathbb{E}[H_\theta]$ perfectly recovers the true community membership, therefore σ is the true community membership vector. Next, given that the k-means clustering step achieves a $(1+\epsilon)$ approximation, using the error bound (41) from Lemma 7, we have that

$$l(\sigma, \hat{\sigma}_{\text{SC-MLE}})d^2 \leq 4(4+2\epsilon)\|\hat{U} - U\|_F^2,$$

where d is the distance between the two cluster centroids of the population version $\mathbb{E}[H_\theta]$, with its expression provided in (30). Given that a rotation of $\hat{U}$ does not change the k-means clustering result, the tightest upper bound we can obtain is

$$l(\sigma, \hat{\sigma}_{\text{SC-MLE}})d^2 \leq 4(4+2\epsilon) \min_{O\in\mathcal{O}_2} \|\hat{U} - OU\|_F^2 \tag{25}$$

$$= 4(4+2\epsilon) \min_{r\in\mathbb{C}_1} \|\mathbf{v} - r\hat{\mathbf{v}}\|_F^2 \leq 4(4+2\epsilon)\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\|_F^2, \tag{26}$$

where (25) follows from the fact that $\|U - \hat{U}\|_F = \|\mathbf{v} - \hat{\mathbf{v}}\|_F$ and the inequality in (26) follows from Lemma 5.

Combining matrix perturbation analysis from Lemma 2, Lemma 3 and Lemma 4, we derive that, for an absolute constant ϵ_0, with probability at least $1 - N^{-\epsilon_0}$, the misalignment distance is upper bounded by

$$\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\|_F \leq \frac{2\sqrt{2}C\sqrt{Np_{\max}\log N}}{\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])} = \frac{2\sqrt{2}C\sqrt{Np_{\max}\log N}}{\Delta}. \tag{27}$$

Combining (26) and (27), we eventually obtain that with probability at least $1 - N^{-\epsilon_0}$

$$\frac{l(\sigma, \hat{\sigma}_{\text{SC-MLE}})}{N} \leq \frac{64(2+\epsilon)C^2 p_{\max}\log N}{d^2\Delta^2}.$$

□

C.3 Eigenspace of the population matrix $\mathbb{E}[H_\theta]$)

Lemma 2. *For the DSBM(n_1, n_2, p, q, η), the population matrix $\mathbb{E}[H_\theta]$ has a unique largest eigenvalue. The top eigenvector $\mathbf{v}$ has exactly two distinct values that indicate the community labels, where the distance d between them can be easily computed using* (30)*. Moreover, the eigengap is*

$$\Delta = |\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])| = \min\{2\Delta_0, |1/2N(w_r p + w_c)| + \Delta_0\} \geq \Delta_0, \tag{21}$$

where

$$\Delta_0 = 1/2\sqrt{N^2(w_r p + w_c)^2 - 4n_1 n_2\left((w_r p + w_c)^2 - |w_r + w_c + i w_i(1-2\eta)q|^2\right)}. \tag{22}$$

Proof. Recall that the population version of H_θ has a block structure and can be written as

$$\mathbb{E}[H_\theta] = MQM^T - (pw_r + w_c)I$$

$$Q = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} + w_r \begin{bmatrix} p & q \\ q & p \end{bmatrix} + w_c \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix},$$

where $M \in \{0,1\}^{N\times 2}$ is the community membership matrix, and $M_{uc} = 1$ denotes that vertex u belongs to community c. We further normalize the columns of M as follows

$$MQM^T = MD^{-1}DQD(MD^{-1})^T$$

$$D = \begin{bmatrix} \sqrt{n_1} & 0 \\ 0 & \sqrt{n_2}. \end{bmatrix}$$

Here the normalized matrix MD^{-1} has orthonormal column vectors.

Let $DQD = U\Lambda U^*$ be the eigendecomposition on the 2×2 matrix, Then, the $N \times N$ matrix MQM^T can be diagonalized as

$$MQM^T = (MD^{-1}U)\Lambda(MD^{-1}U)^*,$$

where $\mathrm{diag}(\Lambda)$ contains the eigenvalues of MQM^T and the columns of $MD^{-1}U \in \mathbb{R}^{N\times 2}$ are the orthonormal eigenvectors. Therefore, the problem of computing the eigenpairs of $\mathbb{E}[H_\theta]$ reduces to compute the eigenpairs of the 2×2 matrix DQD where

$$DQD = \begin{bmatrix} n_1(w_r p + w_c) & \sqrt{n_1 n_2}(w_r q + w_c + i(1-2\eta)q) \\ \sqrt{n_1 n_2}(w_r q + w_c - i(1-2\eta)q) & n_2(w_r p + w_c) \end{bmatrix}$$

For the eigenvalues, via a simple calculation we arrive at

$$\lambda_1(DQD) = \frac{1}{2}\left(N(w_r p + w_c) + 2\Delta_0\right),$$
$$\lambda_2(DQD) = \frac{1}{2}\left(N(w_r p + w_c) - 2\Delta_0\right),$$

where

$$2\Delta_0 = \sqrt{N^2(w_r p + w_c)^2 - 4n_1 n_2((w_r p + w_c)^2 - |w_r q + w_c + i w_i q(1-2\eta)|^2)}.$$

Therefore, we obtain the eigenvalues of $\mathbb{E}[H] = MQM^T$

$$\lambda_1(\mathbb{E}[H_\theta]) = \frac{1}{2}\left(N(w_r p + w_c) + 2\Delta_0\right) - (w_r p + w_c)$$
$$\lambda_2(\mathbb{E}[H_\theta]) = \frac{1}{2}\left(N(w_r p + w_c) - 2\Delta_0\right) - (w_r p + w_c)$$
$$\lambda_3(\mathbb{E}[H_\theta]) = \ldots = \lambda_N(\mathbb{E}[H_\theta]) = -(w_r p + w_c).$$

The eigenvalue that obtains the largest magnitude is unique and it is $\lambda_1(\mathbb{E}[H_\theta])$ when $N(w_r p + w_c) \geq 0$, or $\lambda_2(\mathbb{E}[H_\theta])$ when $N(w_r p + w_c) < 0$. The gap between the largest and the second largest eigenvalue is

$$\min\{2\Delta_0, |1/2N(w_r p + w_c)| + \Delta_0\}. \tag{28}$$

One can easily verify that the eigengap Δ lies in $[\Delta_0, 2\Delta_0]$. Therefore, the lower bound Δ_0 is a good approximation to the spectral gap in the sense that they are of the same order.

Next, we move on to compute the top eigenvector of $\mathbb{E}[H_\theta]$. We use $\mathbf{x} = (x_1, x_2) \in \mathbb{C}^2$ to denote the top eigenvector of DQD, and we have that $x_1 \neq x_2$. Then, the top eigenvector of $\mathbb{E}[H_\theta]$ can be easily computed through $\mathbf{v} = MD^{-1}\mathbf{x}$, and it has two distinct values

$$\mathbf{v}(u) = \begin{cases} x_1/\sqrt{n_1} & \text{if } u \in \mathcal{C}_1, \\ x_2/\sqrt{n_2} & \text{if } u \in \mathcal{C}_2. \end{cases} \tag{29}$$

The distance between the two cluster centroids d is simply

$$d = \left| \frac{x_1}{\sqrt{n_1}} - \frac{x_2}{\sqrt{n_2}} \right|. \tag{30}$$

□

C.4 Useful theorems from matrix perturbation analysis

Theorem 4 (Davis-Kahan's perturbation bound Davis and Kahan (1970)). *Let $H, R \in \mathcal{H}$ be two Hermitian matrices. Then, for any $a \leq \beta$ and $\delta > 0$ it holds that*

$$\left\| P_{[\alpha,\beta]}(H) - P_{(\alpha-\delta,\beta+\delta)}(H+R) \right\| \leq \frac{\|R\|}{\delta}.$$

Here $P_{[\alpha,\beta]}(H)$ denotes the projection matrix on the subspace spanned by eigenvectors of H with corresponding eigenvalues lie between $[\alpha, \beta]$, and $P_{(\alpha-\delta,\beta+\delta)}(H+R)$ is the projection matrix on the subspace spanned by eigenvectors of $H+R$ with eigenvalues lie between $(\alpha-\delta, \beta+\delta)$.

Theorem 5 (Weyl's inequality Weyl (1912)). *Let $H, R \in \mathcal{H}$ be two Hermitian matrices. Then for every $1 \leq j \leq n$, the j-th largest eigenvalues of H and $H + R$ obey*

$$|\lambda_j(H) - \lambda_j(H + R)| \leq \|R\|.$$

In addition to the eigenspace perturbation bound in Theorem 4, we summarize in Lemma 5 comparisons over two different representations of the eigenspace distance, which will useful in the error analysis of k-means.

Lemma 5. *[Adapted From Lemma 2.1 in Chen et al. (2021)] For any $U, \tilde{U} \in \mathbb{C}^{\mathbb{N}\times k}$, we have*

$$\min_{O\in\mathcal{O}_{k\times k}} \|U - O\tilde{U}\|_F \leq \|UU^* - \tilde{U}\tilde{U}^*\|_F$$

C.5 Eigenspace perturbation

Lemma 3. *Given a directed graph from DSBM(n_1, n_2, p, q, η) and its Hermitian matrix representation H_θ, the projection matrix of the top eigenvector is such that*

$$\|\mathbf{v}\mathbf{v}^* - \hat{\mathbf{v}}\hat{\mathbf{v}}^*\|_F \leq 2\sqrt{2}\frac{\|R\|}{\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])}.$$

Proof on Lemma 3. From the Davis-Kahan's pertubation bound, we have

$$\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\| \leq \frac{\|R\|}{|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(H_\theta)|}. \tag{31}$$

Using Weyl's inequality, we have that

$$|\lambda_2(\mathbb{E}[H_\theta]) - \lambda_2(H_\theta)| \leq \|R\|$$

Therefore, the denominator in (31) can be further lower bounded by $|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])| - \|R\|$, and we obtain

$$\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\| \leq \frac{\|R\|}{|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])| - \|R\|}. \tag{32}$$

The denominator in (32) involves comparing the spectral gap $|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])|$ and $\|R\|$, which further requires an extra condition on the denominator being positive, to allow the inequality to hold. To circumvent this limitation, we divide the comparison into two cases

- if $\|R\| \geq \frac{1}{2}|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])|$, then we have

$$\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\| \leq 1 \leq \frac{2\|R\|}{|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])|}.$$

- if $\|R\| \leq \frac{1}{2}|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])|$, then we use the perturbation bound (32) and get

$$\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\| \leq \frac{\|R\|}{|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])| - \|R\|} \leq \frac{2\|R\|}{|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])|}.$$

Combining the two cases, we obtain that for any $\|R\|$, the following upper bound always holds

$$\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\| \leq \frac{2\|R\|}{|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])|}.$$

Since for any rank r Hermitian matrix H, $\|H\|_F \leq \sqrt{r}\|H\|$. Therefore, we have that

$$\|\mathbf{v}\mathbf{v}^* - \hat{\mathbf{v}}\hat{\mathbf{v}}^*\|_F \leq \sqrt{2}\|\mathbf{v}\mathbf{v}^* - \hat{\mathbf{v}}\hat{\mathbf{v}}^*\| \leq \frac{2\sqrt{2}\|R\|}{|\lambda_1(\mathbb{E}[H_\theta]) - \lambda_2(\mathbb{E}[H_\theta])|}.$$

□

C.6 Bound the random perturbation $\|R\|$

Lemma 6 (Matrix Bernstein Tropp et al. (2015)). *Consider a finite sequence $\{S_k\}$ of independent, random matrices with dimension d. Assume that*

$$\mathbb{E}S_k = 0 \text{ and } \|S_k\| \leq L \text{ for each index } k.$$

For the random matrix $Z = \sum_k S_k$, let $v(Z)$ be the matrix variance statistic of the sum:

$$v(Z) = \max\left\{\left\|\sum_k \mathbb{E}\left(S_k S_k^*\right)\right\|, \left\|\sum_k \mathbb{E}\left(S_k^* S_k\right)\right\|\right\}.$$

Then, for all $t \geq 0$,

$$\mathbb{P}\{\|Z\| \geq t\} \leq 2d \exp\left(\frac{-t^2/2}{v(Z) + Lt/3}\right).$$

Lemma 4. *Consider a directed graph from DSBM (n_1, n_2, p, q, η) and its Hermitian matrix representation H_θ. Assume that $Np_{\max} = \Omega(\log N)$. Then there exists an absolute constant ϵ and*

$$C = (2+\epsilon)\sqrt{w_r^2 + w_i^2}\left(\frac{\log N}{Np_{\max}} + 1\right) = \Theta\left(\sqrt{w_r^2 + w_i^2}\right), \tag{23}$$

such that the random perturbation $R = H_\theta - \mathbb{E}[H_\theta]$ has

$$\mathbb{P}(\|R\| \geq C\sqrt{Np_{\max}\log N}) \leq N^{-\epsilon}. \tag{24}$$

Proof. Recall that by definition random perturbation $R = H_\theta - \mathbb{E}[H_\theta]$ is Hermitian. We first decompose it into summation of perturbations on different entries $R = \sum_{j<l} R^{jl}$ where R^{jl} is also a random Hermitian and only has non-zero entries at (j,l) and (l,j). If j, l belongs to the same community $\sigma(j) = \sigma(l)$

$$R_{jl}^{jl} = \begin{cases} w_r(1-p) + iw_i & w.p.\ \ p/2 \\ w_r(1-p) - iw_i & w.p.\ \ p/2 \\ -w_r p & w.p.\ \ 1-p. \end{cases} \tag{33}$$

If j, l belongs to different communities $\sigma(j) \neq \sigma(l)$, and without loss of generality we assume $j \in \mathcal{C}_1, l \in \mathcal{C}_2$

$$R_{jl}^{jl} = \begin{cases} w_r(1-q) + iw_i(1-(1-2\eta)q) & w.p.\ \ q(1-\eta) \\ w_r(1-q) - iw_i(1+(1-2\eta)q) & w.p.\ \ q\eta \\ -w_r q - iw_i(1-2\eta)q & w.p.\ \ 1-q \end{cases} \tag{34}$$

From the Matrix Bernstein's inequality in Lemma 6, we have for any $t \geq 0$

$$\mathbb{P}(\|R\| \geq t) \leq 2N \exp\left(\frac{-t^2/2}{\mathrm{Var}(R) + Lt/3}\right),$$

where L is an upper upper of $\|R^{jl}\|$ and $\mathrm{Var}(R)$ is the variance.

For computing L, recall that by definition

$$\|R^{jl}\| \leq L, \ \ \forall j \neq l.$$

Here the matrix spectral norm can be simplified to be upper bounded by $|R_{jl}|$ because the spectral norm is always upper bounded by the maximum absolute values of each entry. Therefore, it suffices to take L as an upper bound on $\max_{j\neq l}|R_{jl}|$. From (33), we have that, if $\sigma(j) = \sigma(l)$, then $|R_{jl}| \leq \sqrt{w_r^2(1-p)^2 + w_i^2}$; if $\sigma(j) \neq \sigma(l)$, then $|R_{jl}| \leq \sqrt{w_r^2(1-q)^2 + w_i^2(1+(1-2\eta)q)^2}$. Combining the two, it suffices for us to take

$$L = \sqrt{w_r^2(1-p_{min})^2 + w_i^2(1+(1-2\eta)q)^2} \leq 2\sqrt{w_r^2 + w_i^2}. \tag{35}$$

To compute the variance term $\mathrm{Var}(R)$, first recall by definition

$$\mathrm{Var}(R) = \max\left\{\left\|\sum_{j<l}\mathbb{E}[R^{jl}(R^{jl})^*]\right\|, \left\|\sum_{j<l}\mathbb{E}[(R^{jl})^* R^{jl}]\right\|\right\}.$$

For each $j<l$, we have $R^{jl}(R^{jl})^* = (R^{jl})^* R^{jl}$ and we use M^{jl} to denote the product matrix. In M^{jl}, the only two non zero entries are M^{jl}_{jj} and M^{jl}_{ll} and $M^{jl}_{jj} = M^{jl}_{ll} = R_{jl}\overline{R_{jl}}$. Therefore, $\mathbb{E}[R^{jl}(R^{jl})^*]$ also only has two non-zero entries at (j,j) and (l,l) for every $j<l$ and thus, the spectral norm of the matrix summation is simply the largest diagonal element, i.e.,

$$\mathrm{Var}(R) = \max_{j\in[N]}\sum_{l\neq j}\mathbb{E}[M^{jl}_{jj}]. \tag{36}$$

In (36), M^{jl}_{jj} is a real random variable whose distribution can be derived from (33) and (34), and we have that, if $\sigma(j)=\sigma(l)$, then

$$M^{jl}_{jj} = \begin{cases} w_r^2(1-p)^2 + w_i^2 & w.p.\ \ p \\ w_r^2 p^2 & w.p.\ \ 1-p, \end{cases}$$

and $\mathbb{E}[M^{jl}_{jj}] = w_r^2 p(1-p) + p w_i^2$.

If $\sigma(j)\neq\sigma(l)$, then

$$M^{jl}_{jj} = \begin{cases} w_r^2(1-q)^2 + w_i^2(1-(1-2\eta)q)^2 & w.p.\ \ q(1-\eta) \\ w_r^2(1-q)^2 + w_i^2(1+(1-2\eta)q)^2 & w.p.\ \ q\eta \\ w_r^2 q^2 + w_i^2(1-2\eta)^2 q^2 & w.p. 1-q, \end{cases}$$

and $\mathbb{E}[M^{jl}_{jj}] = w_r^2 q(1-q) + w_i^2 q(1-(1-2\eta)^2 q)$. Since, without loss of generality, we assume $p, q \leq 0.5$, thus for all $j\neq l$ we have $\mathbb{E}[M^{jl}_{jj}] \leq w_r^2 p_{\max}(1-p_{\max}) + w_i^2 p_{\max}$. Therefore, from (36), we arrive at

$$\mathrm{Var}(R) \leq N(w_r^2 p_{\max}(1-p_{\max}) + w_i^2 p_{\max}) \leq N p_{\max}(w_r^2 + w_i^2). \tag{37}$$

Using the Matrix Bernstein's inequality, we have for $t = C\sqrt{N p_{\max}\log N}$,

$$\begin{aligned}
\mathbb{P}(\|R\|\geq t) &\leq 2\exp\left(-\frac{C^2 N p_{\max}\log N}{2\mathrm{Var}(R) + 2LC\sqrt{N p_{\max}\log N}/3} + \log N\right) \\
&\leq 2\exp\left(-\frac{C^2 N p_{\max}\log N}{2N p_{\max}(w_r^2+w_i^2) + 2L\sqrt{N p_{\max}\log N}/3} + \log N\right) \qquad (38) \\
&= 2\exp\left(-\frac{C^2}{2(w_r^2+w_i^2) + 2LC/3\sqrt{\log N/N p_{\max}}}\log N + \log N\right).
\end{aligned}$$

Here (38) follows from the analysis on $\mathrm{Var}(R)$ in (37). From (38), if there exist an absolute ϵ, such that

$$\frac{C^2}{(w_r^2+w_i^2) + LC/3\sqrt{\log N/N p_{\max}}} \geq 2+\epsilon, \tag{39}$$

then we have $\mathbb{P}(\|R\|\geq t) \leq N^{-\epsilon}$, which conclude the proof. It turns out that we can always find an absolute constant C such that (39) holds. To see this, first note that (39) is equivalent to

$$C \geq (1+\epsilon/2)L\sqrt{\frac{\log N}{N p_{\max}}} + \sqrt{(2+\epsilon)(w_r^2+w_i^2) + (1+\epsilon/2)^2 L^2\frac{\log N}{N p_{\max}}}.$$

Since $a^2+b^2 \leq (a+b)^2$ for $a,b>0$, it suffices to let

$$C = (2+\epsilon)L\frac{\log N}{N p_{\max}} + (2+\epsilon)\sqrt{w_r^2+w_i^2}.$$

Since $L \leq 2\sqrt{w_r^2 + w_i^2}$, we have

$$C \leq (2+\epsilon)\sqrt{w_r^2 + w_i^2}\left(\frac{\log N}{Np_{\max}} + 1\right) = \Theta\left(\sqrt{w_r^2 + w_i^2}\right), \tag{40}$$

where the last equality is due to the connectivity assumption $Np_{\max} = \Omega(\log N)$.

□

C.7 Useful theorem in k-means error analysis

Lemma 7 (k-means error adapted from Lemma 5.3 in Lei and Rinaldo (2015))**.** *For $\epsilon > 0$ and any two matrices $\hat{U}, U$, such that $U = MX$ with $M \in \{0,1\}^{N\times 2}$ be the indicator matrix and $X \in \mathbb{R}^{2\times 2}$ have its row vectors representing the centroids of two clusters, let $(\hat{M}, \hat{X})$ be a $(1+\epsilon)$ solution to the k-means problem and $\bar{U} = \hat{M}\hat{X}$. For $\delta = \|X_{1*} - X_{2*}\|$, define $S = \{j \in [N] : \|\bar{U}_{j*} - U_{j*}\|\} \geq \delta/2$ then*

$$|S|\delta^2 \leq 4(4+2\epsilon)\|\hat{U} - U\|_F^2. \tag{41}$$

C.8 Proof on Corollary 1

Corollary 1. *Consider directed graphs generated from the DSBM $(N/2, N/2, p, p, \eta)$ under the assumption $Np = \Omega(\log N)$. As $N \to \infty$, the misclustering error of spectral clustering is such that*

$$\frac{l(\sigma, \hat{\sigma}_\theta)}{N} = \Theta\left(\frac{\log N}{NpL^2}\right), \tag{4}$$

where $L = L(\eta)$ is a continuous, monotonically decreasing function of the edge directionality parameter η, with $L = 0$ when $\eta = 0.5$. The explicit form of $L(\eta)$ is provided in (48), and its behavior is illustrated in Figure 5b in Appendix C.

Proof. From Theorem 2, we have the general upper bound on the error bound

$$\frac{l(\sigma, \hat{\sigma})}{N} \leq \frac{64(2+\epsilon)C^2 p_{\max}\log N}{d^2\Delta^2} = \Theta\left(\frac{C^2 p_{\max}\log N}{d^2\Delta^2}\right), \tag{42}$$

where d and Δ depends on $\mathbb{E}[H_\theta]$. When $p = q$, we have that

$$w_r = \log\left(\frac{1}{4\eta(1-\eta)}\right), \quad w_i = \log\left(\frac{1-\eta}{\eta}\right), \quad w_c = 0.$$

Moreover, notice that normalizing H_θ does not affect the clustering error. For the rest of the discussion, we consider $1/w_i H_\theta$ as the input Hermitian matrix for spectral relaxation of MLE, and correspondingly we denote the updated coefficient as

$$\tilde{w_r} = \log\left(\frac{1}{4\eta(1-\eta)}\right) / \log\left(\frac{1-\eta}{\eta}\right), \quad \tilde{w_i} = 1, \quad \tilde{w_c} = 0.$$

Because $\tilde{w_r} \leq 1$ and $\tilde{w_i} = 1$, the term C^2 in (42) has $C^2 = \Theta(\tilde{w_r}^2 + \tilde{w_i}^2) = \Theta(1)$. Therefore, we can further simplify (42) as follows

$$\frac{l(\sigma, \hat{\sigma})}{N} \leq \Theta\left(\frac{p\log N}{d^2\Delta^2}\right).$$

For analyzing the asymptotic behaviour of the error bound, we are only left with computing the centroid distance d and eigengap bound Δ.

Following from the definition of Δ_0 in (22), we have

$$\Delta \geq \Delta_0 = \frac{Np}{2}\sqrt{\tilde{w_r}^2 + \tilde{w_i}^2(1-2\eta)^2}. \tag{43}$$

For computing the centroid distance d, recall that the population matrix can be written as

$$\mathbb{E}[H_\theta] = MQM^T - \tilde{w_r} pI,$$

where M is the community indicator matrix and the 2×2 matrix Q has

$$Q = \begin{bmatrix} \tilde{w_r} & \tilde{w_r} + (1-2\eta)i \\ \tilde{w_r} - (1-2\eta)i & \tilde{w_r} \end{bmatrix} p.$$

Since $n_1 = n_2 = N/2$, we have that the two distinct values in $v_1(\mathbb{E}[H_\theta])$ (see (29)) to be the values of $v_1(Q)$ divided by $\sqrt{N/2}$. We can easily compute that the top eigenvector has

$$v_1(\mathbb{E}[H_\theta]) = \begin{cases} \frac{w_r + w_i(1-2\eta)i}{\sqrt{N}|-w_r + w_i(1-2\eta)i|} & \text{for } u \in \mathcal{C}_1 \\ 1/\sqrt{N} & \text{for } u \in \mathcal{C}_2. \end{cases} \tag{44}$$

We denote $\bar{c_1}, \bar{c_2}$ the cluster centroids of $\mathcal{C}_1, \mathcal{C}_2$ in the embedding given by the top eigenvector of $\mathbb{E}[H_\theta]$. The locations of two cluster centroids in the complex plane are exactly the two distinct values in (44). We visualize the two cluster centroids in Figure 5a. Let $\theta = \arccos\left(\frac{w_r}{|w_r + w_i(1-2\eta)i|}\right)$ be the angle between the two values in the complex plane. Therefore, we have

$$d^2 = \frac{1}{N}(1 - \cos\theta) = \frac{4\sin^2\theta/2}{N}. \tag{45}$$

Combining (42), (43), and (45) and letting

$$L(\eta) = |\tilde{w_r} + i(1-2\eta)| \sin(\theta/2). \tag{46}$$

We have

$$\frac{l(\sigma, \hat{\sigma})}{N} \leq \Theta\left(\frac{\log N}{NpL^2}\right). \tag{47}$$

From the above inequality (47), the upper bound of misclustering error is determined by two independent variables Np and L^2. The term Np is the average degree of the graph. The term L^2, by definition (46), is a function on η with expression

$$L(\eta) = \left((1-2\eta)^2 + \frac{\left(\log\left(\frac{1}{4\eta - 4\eta^2}\right)\right)^2}{\left(\log\left(\frac{\eta}{1-\eta}\right)\right)^2}\right)\left(1 - \frac{\log\left(\frac{1}{4\eta-4\eta^2}\right)}{\log\left(\frac{\eta}{1-\eta}\right)\sqrt{(1-2\eta)^2 + \frac{\left(\log\left(\frac{1}{4\eta-4\eta^2}\right)\right)^2}{\left(\log\left(\frac{\eta}{1-\eta}\right)\right)^2}}}\right) \tag{48}$$

To see how the value L changes as η varies from 0 to 0.5, we plot $L(\eta)$ in Figure 5b using Mathematica Inc.. We observe that $L(\eta) = \Theta(1)$ when η is bounded away from 0.5. Therefore, if $\eta \leq 0.5 - \epsilon$ with an absolute constant $\epsilon > 0$, then the misclustering error on the spectral relaxation is such that

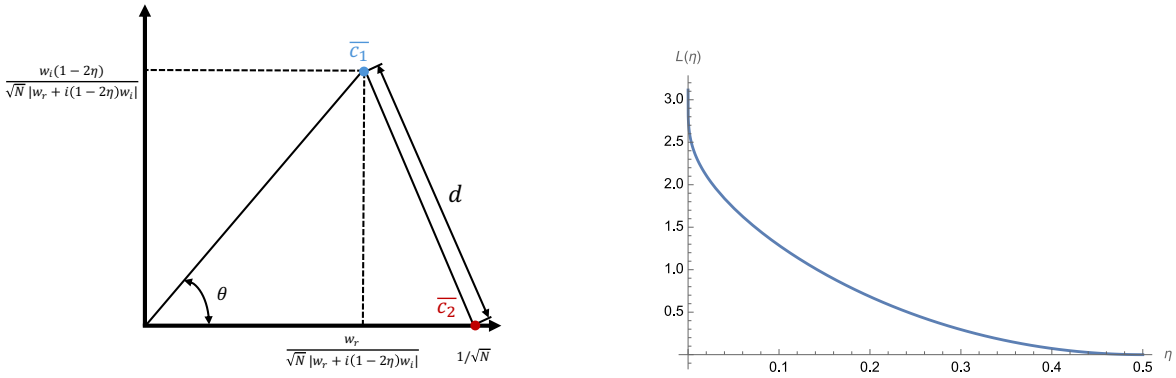
$$\frac{l(\sigma, \hat{\sigma})}{N} = O\left(\frac{\log N}{Np}\right)$$

When η converges to 0.5 (when the imbalance structure disappears), $L(\eta)$ converges to 0. Therefore, if $\eta = 0.5 - o(1)$, we have that

$$\frac{l(\sigma, \hat{\sigma})}{N} = \omega\left(\frac{\log N}{Np}\right).$$

The above results on misclustering error bounds agree with the intuition that lower values of η denote a less noisy problem instance, and thus lead to a lower clustering error.

□



(a) Visualization on l and θ.

(b) Plot of l as function of η.

Figure 5: Visualization of important parameters for representing the error bound.

D Proof for Corollary 2

Corollary 2. *For graphs generated from the DSBM($N/2, N/2, p, q, \eta$), the community assignment $\hat{\sigma}_0$ obtained by* `DirHSC` *satisfies*

$$\frac{l(\sigma, \hat{\sigma}_0)}{N} \leq \frac{C p_{\max} \log N}{N q^2 (1 - 2\eta + 2\eta^2) d_0^2}. \tag{6}$$

where $d_0 = \left|1 - \frac{1-(1-2\eta)i}{|1+(1-2\eta)i|}\right|$.

Proof. The analysis follows the same steps as in Theorem 2, specialized to the Hermitian matrix

$$H_0 = A + A^T + i(A - A^T),$$

for which $w_r = 1$, $w_i = 1$, and $w_c = 0$.

By Lemma 7, the clustering error satisfies

$$\frac{l(\sigma, \sigma_0)}{N} \leq \frac{8(2+\epsilon)\,\|\hat{\mathbf{v}}\hat{\mathbf{v}}^* - \mathbf{v}\mathbf{v}^*\|}{d^2 N},$$

where $\mathbf{v}$ is the leading eigenvector of $\mathbb{E}[H_0]$, and d is the distance between the two population centroids.

Applying the perturbation bound from Lemma 3, we obtain

$$\frac{l(\sigma, \sigma_0)}{N} \leq \frac{64(2+\epsilon)\,\|R\|^2}{d^2 N\,(\lambda_1(\mathbb{E}[H_0]) - \lambda_2(\mathbb{E}[H_0]))^2}.$$

Using the Bernstein concentration inequality, we have with probability at least $1 - N^{-c}$ that

$$\|R\|^2 \leq C' N p_{\max} \log N,$$

for some absolute constant C'. Substituting into the above bound yields

$$\frac{l(\sigma, \sigma_0)}{N} \leq \frac{64(2+\epsilon)C' p_{\max} \log N}{d^2\,(\lambda_1(\mathbb{E}[H_0]) - \lambda_2(\mathbb{E}[H_0]))^2}. \tag{49}$$

It remains to characterize the eigengap and centroid distance of $\mathbb{E}[H_0]$. From Lemma 2, the eigengap is

$$\Delta_0 = \frac{1}{2}\sqrt{N^2 p^2 - 4 n_1 n_2 \left(p^2 - |1 + i(1-2\eta)q|^2\right)}.$$

In the balanced case $n_1 = n_2$, this simplifies to

$$\Delta_0 = \frac{1}{2} N q \sqrt{1 + (1-2\eta)^2}. \tag{50}$$

For the centroid distance, we derived in Lemma 2 the 2×2 community indicator matrix matrix,

$$DQD = N/2 \begin{bmatrix} p & q + i(1-2\eta)q) \\ q - i(1-2\eta)q) & p \end{bmatrix}$$

For the leading eigenvector $\mathbf{x} DQD = \lambda_1 \mathbf{x}$, we have

$$\lambda_1(DQD) = \frac{1}{2}\left(Np + 2\Delta_0\right) = 1/2N(p + q\sqrt{1 + (1-2\eta)^2})$$

Solving the eigenvalue equation yields

$$\frac{N}{2}((q - i(1-2\eta)q)x_1 + px_2) = \frac{N}{2}\left(p + q\sqrt{1 + (1-2\eta)^2}\right) x_2$$
$$\frac{x_1}{x_2} = \frac{\sqrt{1 + (1-2\eta)^2}}{1 - i(1-2\eta)}$$

Setting

$$x_1 = 1;\ x_2 = \frac{1 - i(1-2\eta)}{\sqrt{1 + (1-2\eta)^2}},$$

and normalizing $\|\mathbf{v}\|_2 = 1$, the centroid distance in complex plane is then

$$d^2 = \left| \frac{x_1}{\sqrt{N/2}} - \frac{x_2}{\sqrt{N/2}} \right|^2 = \frac{2}{N}|x_1 - x_2|^2 = \frac{2}{N}\left| 1 - \frac{1 - i(1-2\eta)}{\sqrt{1 + (1-2\eta)^2}} \right|^2 \triangleq \frac{2}{N} d_0^2. \tag{51}$$

Here we define the normalize centroid distance $d_0 = \left| 1 - \frac{1 - i(1-2\eta)}{\sqrt{1 + (1-2\eta)^2}} \right|$. Combining (49), (50), and (51), we obtain

$$\frac{l(\sigma, \sigma_0)}{N} \leq \frac{C \log N}{d_0^2 N q^2 (1 - 2\eta + 2\eta^2)},$$

which completes the proof. □

E Proofs for Theorem 3

Theorem 3 (Recursive error update of `LEHSC`)**.** *Assume* (A1)–(A4) *hold. Let* $\hat{\sigma}^{(t)}$ *denote the community labels at iteration* t*, with misclustering rate* $\varepsilon^{(t)}$*. Then, with probability* $1 - o(1)$ *as* $N \to \infty$*, the parameter estimation errors satisfy*

$$\left|\hat{p}^{(t+1)} - p\right| \leq C_1\, \varepsilon^{(t)}\, |p - q| \ +\ C_2\, \frac{\sqrt{p_{\max} \log N}}{N}, \tag{7}$$

$$\left|\hat{q}^{(t+1)} - q\right| \leq C_1\, \varepsilon^{(t)}\, |p - q| \ +\ C_2\, \frac{\sqrt{p_{\max} \log N}}{N}, \tag{8}$$

$$\left|\hat{\eta}^{(t+1)} - \eta\right| \leq C_1\ \varepsilon^{(t)} \frac{p_{\max}}{q} \ +\ C_2\, \frac{\sqrt{\log N}}{N\sqrt{q}}. \tag{9}$$

Here C_1, C_2 *are constants depending on the balance and warm-start conditions. Consequently, applying the spectral step with* $H_{\hat{\theta}^{(t+1)}}$ *yields*

$$\frac{l(\sigma, \hat{\sigma}^{(t+1)})}{N} \leq \frac{64(2+\epsilon)C^2\, p_{\max} \log N}{d^2_{\hat{\theta}^{(t+1)}}\, \Delta^2_{\hat{\theta}^{(t+1)}}}, \tag{10}$$

where $d^2_{\hat{\theta}^{(t+1)}}$ *and* $\Delta^2_{\hat{\theta}^{(t+1)}}$ *denote the centroid distance and eigengap of the population matrix* $\mathbb{E}[H_{\hat{\theta}^{(t+1)}}]$.

Proof. We start with proof for the parameter estimation error $\hat{\theta}^{(t)} - \theta$.

Step 1. Bound graph counts. Under community assignment $\sigma^{(t)}$, we denote the following graph statistics. Recall that $A \in \{0,1\}^{N\times N}$ is the adjacency with zero diagonal, the true community assignment is $\mathcal{C}_1, \mathcal{C}_2$ with $n_1 = |\mathcal{C}_1|$, $n_2 = |\mathcal{C}_2|$, and the current clusters is $\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2$.

We define the mislabel counts as follows

$$a \triangleq |\mathcal{C}_1 \cap \hat{\mathcal{C}}_2|, \qquad b \triangleq |\mathcal{C}_2 \cap \hat{\mathcal{C}}_1|, \qquad \varepsilon^{(t)} = \frac{a+b}{N}.$$

We denote the collection of within community and between community vertex pairs

$$\mathcal{W} \triangleq \big\{\{i,j\} : i<j,\ i,j \in \mathcal{C}_1 \text{ or } i,j \in \mathcal{C}_2\big\}, \qquad \mathcal{B} \triangleq \big\{\{i,j\} : i<j,\ i \in \mathcal{C}_1,\ j \in \mathcal{C}_2\big\}.$$

$$\hat{\mathcal{W}} \triangleq \big\{\{i,j\} : i<j,\ i,j \in \hat{\mathcal{C}}_1 \text{ or } i,j \in \hat{\mathcal{C}}_2\big\}, \qquad \hat{\mathcal{B}} \triangleq \big\{\{i,j\} : i<j,\ i \in \hat{\mathcal{C}}_1,\ j \in \hat{\mathcal{C}}_2\big\}.$$

Note that an unordered pair $\{i,j\}$ changes status iff exactly one endpoint lies in the set of mislabeled vertices $(\mathcal{C}_1 \cap \hat{\mathcal{C}}_2) \cup (\mathcal{C}_2 \cap \hat{\mathcal{C}}_1)$. The number of such pairs is

$$\big|\mathcal{B}\triangle\hat{\mathcal{B}}\big| = \big|\mathcal{W}\triangle\hat{\mathcal{W}}\big| \ = \ (a+b)\big(N-a-b\big) \ \le \ \varepsilon N^2.$$

Moreover, we have

$$\big|\,|\hat{\mathcal{B}}| - |\mathcal{B}|\,\big| \le \big|\mathcal{B}\triangle\hat{\mathcal{B}}\big| \le \varepsilon N^2,$$
$$\big|\,|\hat{\mathcal{W}}| - |\mathcal{W}|\,\big| \le \big|\mathcal{W}\triangle\hat{\mathcal{W}}\big| \le \varepsilon N^2$$

Step 2: Expected bias (contamination) of the moment updates. Define the directed cross-flows and their total:

$$X(\mathcal{C}) = |\mathcal{C}_1 \to \mathcal{C}_2| = \sum_{i\in\mathcal{C}_1}\sum_{j\in\mathcal{C}_2} A_{ij}, \qquad Y(\mathcal{C}) = |\mathcal{C}_2 \to \mathcal{C}_1| = \sum_{i\in\mathcal{C}_2}\sum_{j\in\mathcal{C}_1} A_{ij}, \qquad W(\mathcal{C}) = |\mathcal{C}_1 \to \mathcal{C}_1| + |\mathcal{C}_2 \to \mathcal{C}_2|.$$

$$X(\hat{\mathcal{C}}) = |\hat{\mathcal{C}}_1 \to \hat{\mathcal{C}}_2| = \sum_{i\in\hat{\mathcal{C}}_1}\sum_{j\in\hat{\mathcal{C}}_2} A_{ij}, \qquad Y(\hat{\mathcal{C}}) = |\hat{\mathcal{C}}_2 \to \hat{\mathcal{C}}_1| = \sum_{i\in\hat{\mathcal{C}}_2}\sum_{j\in\hat{\mathcal{C}}_1} A_{ij}, \qquad W(\hat{\mathcal{C}}) = |\hat{\mathcal{C}}_1 \to \hat{\mathcal{C}}_1| + |\hat{\mathcal{C}}_2 \to \hat{\mathcal{C}}_2|.$$

Consequence, we can write

$$TF(\mathcal{C}) = X(\mathcal{C}) + Y(\mathcal{C}), \qquad TF(\hat{\mathcal{C}}) = X(\hat{\mathcal{C}}) + Y(\hat{\mathcal{C}})$$

Under the current partition $(\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2)$, the estimators are

$$\hat{q} \ = \ \frac{TF(\hat{\mathcal{C}})}{|\hat{\mathcal{B}}|}, \qquad \hat{\eta} \ = \ \min\left\{\frac{X(\hat{\mathcal{C}})}{TF(\hat{\mathcal{C}})}, \frac{Y(\hat{\mathcal{C}})}{TF(\hat{\mathcal{C}})}\right\}, \qquad \hat{p} \ = \ \frac{W(\hat{\mathcal{C}})}{|\hat{\mathcal{W}}|}.$$

Recall that $|\hat{\mathcal{W}}\triangle\mathcal{W}| = |\hat{\mathcal{B}}\triangle\mathcal{B}| = m(N-m) \le \varepsilon N^2$, and by balance assumption (A2) and warm start assumption (A3), we have$|\hat{\mathcal{W}}| \asymp N^2$, $|\hat{\mathcal{B}}| = |\hat{\mathcal{C}}_1||\hat{\mathcal{C}}_2| \asymp N^2$.

Bias of $\hat{p}$. Decompose

$$\mathbb{E}[\hat{p}] = \frac{|\hat{\mathcal{W}} \cap \mathcal{W}|\, p + |\hat{\mathcal{W}} \cap \mathcal{B}|\, q}{|\hat{\mathcal{W}}|} = p + \frac{|\hat{\mathcal{W}} \cap \mathcal{B}|}{|\hat{\mathcal{W}}|}\,(q-p).$$

Note that $|\hat{\mathcal{W}} \cap \mathcal{B}| \le |\hat{\mathcal{W}}\triangle\mathcal{W}| \le \varepsilon N^2$ and by assumption (A2), we have $|\hat{\mathcal{W}}| \ge |\mathcal{W}| - \varepsilon N^2 \ge (1/4 - \varepsilon)N^2$. Let $c_\rho = 1/4 - \varepsilon$.

$$\big|\mathbb{E}[\hat{p}] - p\big| \ \le \ \frac{\varepsilon N^2}{c_\rho N^2}\,|p-q| \ = \ \frac{|p-q|}{c_\rho}\varepsilon,$$

Bias of $\hat{q}$. Similarly,

$$\mathbb{E}[\hat{q}] = \frac{|\hat{\mathcal{B}} \cap \mathcal{B}|\, q + |\hat{\mathcal{B}} \cap \mathcal{W}|\, p}{|\hat{\mathcal{B}}|} = q + \frac{|\hat{\mathcal{B}} \cap \mathcal{W}|}{|\hat{\mathcal{B}}|}\,(p-q),$$

and $|\hat{\mathcal{B}} \cap \mathcal{W}| \le |\hat{\mathcal{B}} \triangle \mathcal{B}| \le \varepsilon N^2$, $|\hat{\mathcal{B}}| \ge |\mathcal{B}| - \epsilon N^2 \ge (\rho(1-\rho) - \epsilon)N^2$. We denote $c'_\rho = \rho(1-\rho) - \varepsilon$ and have

$$\left|\mathbb{E}[\hat{q}] - q\right| \;\le\; \frac{\varepsilon N^2}{c'_\rho N^2}\,|p-q| \;=\; \frac{|p-q|}{c'_\rho}\,\varepsilon.$$

Bias of $\hat{\eta}$. Recall that $X(\hat{\mathcal{C}}) = |\hat{\mathcal{C}}_1 \to \hat{\mathcal{C}}_2|$, $Y(\hat{\mathcal{C}}) = |\hat{\mathcal{C}}_2 \to \hat{\mathcal{C}}_1|$, and $TF(\hat{\mathcal{C}}) = X(\hat{\mathcal{C}}) + Y(\hat{\mathcal{C}})$. We denote

$$y \triangleq \mathbb{E}[Y(\hat{\mathcal{C}})], \quad t \triangleq \mathbb{E}[TF(\hat{\mathcal{C}})], \qquad y_0 \triangleq q\eta\, n_1 n_2, \quad t_0 \triangleq q\, n_1 n_2.$$

Only ordered pairs touching a mislabeled vertex are contaminated, which are of size at most εN^2, so

$$|y - y_0| \le\; p_{\max}\,\varepsilon N^2, \qquad |t - t_0| \le\; p_{\max}\,\varepsilon N^2.$$

For balance community, there exist constant c_0 depending on ρ such that the expectation-level denominator stability $t \ge \frac{1}{2}t_0$ holds (here $1/2$ is arbitrary and choosing different absolute constant does not change the asymptotic analysis). To see this, recall that $t_0 = qn_1n_2 \ge \rho(1-\rho)qN^2$ by the balance assumption (A2). Since $t \ge t_0 - p_{\max}\varepsilon N^2$, it suffices to ensure $p_{\max}\varepsilon N^2 \le \frac{1}{2}t_0$, or equivalently, $\varepsilon \le \frac{\rho(1-\rho)q}{2p_{\max}}$. This holds under the warm-start assumption (A3) with $c_\rho = \frac{\rho(1-\rho)}{2}$, from which denominator stability $t \ge \frac{1}{2}t_0$ follows. Therefore, we have

$$\left|\frac{y}{t} - \frac{y_0}{t_0}\right| = \left|\frac{(y-y_0)t_0 - y_0(t-t_0)}{t_0(t_0 + (t-t_0))}\right| \le \frac{2|y-y_0|}{t_0} + \frac{2y_0}{t_0^2}\,|t-t_0| \le C\,\frac{p_{\max}}{q}\,\varepsilon \le C\varepsilon,$$

where the last inequality holds since $t_0 = qn_1n_2 \asymp qN^2$. The same bound holds for $X(\hat{\mathcal{C}})/TF(\hat{\mathcal{C}})$. Since $\min(\cdot,\cdot)$ is 1-Lipschitz in ℓ_∞, $\left|y/t - \eta\right| \;\le\; C\,\frac{p_{\max}}{q}\,\varepsilon$.

In summary, there exist constants $C_\rho, C'_\rho, C > 0$ (depending only on balance and warm-start) such that

$$\left|\mathbb{E}[\hat{p}] - p\right| \le C_\rho\,\varepsilon\,|p-q|, \qquad \left|\mathbb{E}[\hat{q}] - q\right| \le C'_\rho\,\varepsilon\,|p-q|, \qquad \left|y/t - \eta\right| \le C\,\frac{p_{\max}}{q}\,\varepsilon.$$

The bounds above capture the bias in estimating DSBM parameters θ due to mislabels. The additional randomness of the bias is of smaller order and is handled in the next concentration analysis.

Step 3: Concentration analysis. Condition on the current partition $(\hat{\mathcal{C}}_1, \hat{\mathcal{C}}_2)$. All graph statistics we use are sums of independent Bernoulli variables over index sets of size $\Theta(N^2)$ (with success probabilities bounded by $p_{\max}$ for within/between counts and by q for cross edges). By Bernstein's inequality (24), there exist absolute constants $C, c > 0$ such that, with probability at least $1 - N^{-c}$,

$$\left|W(\hat{\mathcal{C}}) - \mathbb{E}[W(\hat{\mathcal{C}})]\right| \le C\sqrt{|\hat{\mathcal{W}}|\, p_{\max}\, \log N}, \tag{52}$$

$$\left|TF(\hat{\mathcal{C}}) - \mathbb{E}[TF(\hat{\mathcal{C}})]\right| \le C\sqrt{|\hat{\mathcal{B}}|\, p_{\max}\, \log N}, \tag{53}$$

$$\left|X(\hat{\mathcal{C}}) - \mathbb{E}[X(\hat{\mathcal{C}})]\right| \vee \left|Y(\hat{\mathcal{C}}) - \mathbb{E}[Y(\hat{\mathcal{C}})]\right| \le C\sqrt{|\hat{\mathcal{B}}|\, p_{\max}\, \log N}. \tag{54}$$

By the balanced and warm-start assumptions we have $|\hat{\mathcal{W}}| \asymp N^2$ and $|\hat{\mathcal{B}}| = |\hat{\mathcal{C}}_1||\hat{\mathcal{C}}_2| \asymp N^2$. Dividing (52) by the corresponding denominators gives the following sampling errors (write v_W, v_B, v_η for short), with probability at least $1 - N^{-c}$:

$$\left|\hat{p} - \mathbb{E}[\hat{p}]\right| \le C\sqrt{\frac{p_{\max}\log N}{|\hat{\mathcal{W}}|}} \;\le\; C\,\frac{\sqrt{p_{\max}\log N}}{N} = O\left(\frac{\sqrt{\log N}}{N}\right) \to 0, \tag{55}$$

$$\left|\hat{q} - \mathbb{E}[\hat{q}]\right| \le C\sqrt{\frac{p_{\max}\log N}{|\hat{\mathcal{B}}|}} \;\le\; C\,\frac{\sqrt{p_{\max}\log N}}{N} = O\left(\frac{\sqrt{\log N}}{N}\right) \to 0. \tag{56}$$

For $\hat{\eta}$ we again use a stable ratio bound. From (53), we have $TF(\hat{\mathcal{C}}) \geq t - C\sqrt{|\hat{\mathcal{B}}|\, p_{\max} \log N}$. Here the term $\sqrt{|\hat{\mathcal{B}}|\, p_{\max} \log N} = \Theta(N\sqrt{p_{\max} \log N})$. Recall $y = \mathbb{E}[Y(\hat{\mathcal{C}})]$, $t = \mathbb{E}[TF(\hat{\mathcal{C}})] \geq 1/2t_0 = \Theta(N^2 q)$, by balance and warm-start assumption. Note that $q = \Theta(p_{\max}) = \Omega(\log N/N)$, and here the stochastic deviation $\sqrt{|\hat{\mathcal{B}}|\, p_{\max} \log N} = \Theta(N\sqrt{p_{\max} \log N}) \ll t = \Theta(N^2 q)$. Similar to the bias analysis, we can show the denominator is stable $TF(\hat{\mathcal{C}}) \geq (1 - o(1))t = \Theta(N^2 q)$.

$$\left| \frac{Y(\hat{\mathcal{C}})}{TF(\hat{\mathcal{C}})} - \frac{y}{t} \right| = \left| \frac{Y(\hat{\mathcal{C}}) - y}{TF(\hat{\mathcal{C}})} - \frac{y}{t} \frac{TF(\hat{\mathcal{C}}) - t}{TF(\hat{\mathcal{C}})} \right| \leq \left(1 + \frac{y}{t}\right) C \frac{\sqrt{|\hat{\mathcal{B}}|\, p_{\max} \log N}}{TF(\hat{\mathcal{C}})} \lesssim \frac{\sqrt{\log N}}{N\sqrt{q}} \to 0. \tag{57}$$

The same bound holds with X in place of Y, hence by 1-Lipschitzness of $\min(\cdot, \cdot)$ in ℓ_∞,

$$\left| \hat{\eta} - \frac{y}{t} \right| \;\leq\; C\, \frac{\sqrt{\log N}}{N\sqrt{q}}.$$

Putting bias + variance together. Combining Step 2 (expected contamination) with (55)–(57), we obtain, with probability at least $1 - N^{-c}$,

$$\begin{aligned}
\left|\hat{p} - p\right| &\leq C_1\, \varepsilon\, |p - q| \;+\; C_2\, \frac{\sqrt{p_{\max} \log N}}{N}, \\
\left|\hat{q} - q\right| &\leq C_1\, \varepsilon\, |p - q| \;+\; C_2\, \frac{\sqrt{p_{\max} \log N}}{N}, \\
\left|\hat{\eta} - \eta\right| &\leq C_1\, \frac{p_{\max}}{q}\, \varepsilon \;+\; C_2\, \frac{\sqrt{\log N}}{N\sqrt{q}}.
\end{aligned}$$

Here C_1, C_2 depend only on the balance and warm-start constants. Therefore, we derive the parameter updating error bounds used in the one-step misclustering recursion, with $\sigma_N \;:=\; C_\rho \max\left\{ \frac{\sqrt{p_{\max} \log N}}{N},\; \frac{\sqrt{\log N}}{N\sqrt{q}} \right\} \to 0$ as $N \to \infty$.

Clustering error. Now with parameter obtained from moment estimator, we compute the weights $\hat{w}_r^{(t+1)}$, $\hat{w}_i^{(t+1)}$ and $\hat{w}_c^{(t+1)}$, we directly apply the error bound from Theorem 2, and obtain the cluster error for next step

$$\varepsilon^{(t+1)} \leq \frac{64(2 + \epsilon) p_{\max} \log N}{\Delta^2_{\hat{\theta}^{(t+1)}} d^2_{\hat{\theta}^{(t+1)}}}$$

where $d^2_{\hat{\theta}^{(t+1)}}$ and $\Delta^2_{\hat{\theta}^{(t+1)}}$ denote the centroid distance and eigengap of the population matrix $\mathbb{E}[H_{\hat{\theta}^{(t+1)}}]$. The error bound of spectral clustering with estimated $\hat{\theta}^{(t+1)}$ vary as the clustering centroid and eigengap vary with respect to different weights.

□

F More Experimental details

F.1 Parameter-free approximation: `DirHSC`

Here we provide implementation details for the parameter-free approximation `DirHSC`.

F.2 Convergence of iterative learning on model parameters

We conduct empirical tests on this iterative approach on directed graphs generated from the DSBM ensemble. In Figure 6, we show how the learned model parameters vary as one repeats the updating process in `LEHSC`. Through our study on the synthetic data sets from the DSBM, we observe that in most cases, this iterative algorithm converges near the truth model parameters very fast (within 10 iterations).

We also conduct experiments to examine how different initialization strategies influence the clustering outcomes. Below, we summarize the recommended strategies.

Algorithm 2: Directional Hermitian Spectral Clustering (`DirHSC`)

Input : Directed graph
Output : Community labels $\hat{\sigma}$

Compute the top eigenvector $\hat{\mathbf{v}}$ of $H_0 = A + A^T + i(A - A^T)$;
Apply k-means to the embedding $[\Re(\hat{\mathbf{v}}), \Im(\hat{\mathbf{v}})]$ to partition into two clusters: $\mathcal{C}_1$ and $\mathcal{C}_2$;
return *community labels* $\hat{\sigma}$ *given by* $\mathcal{C}_1$ *and* $\mathcal{C}_2$.

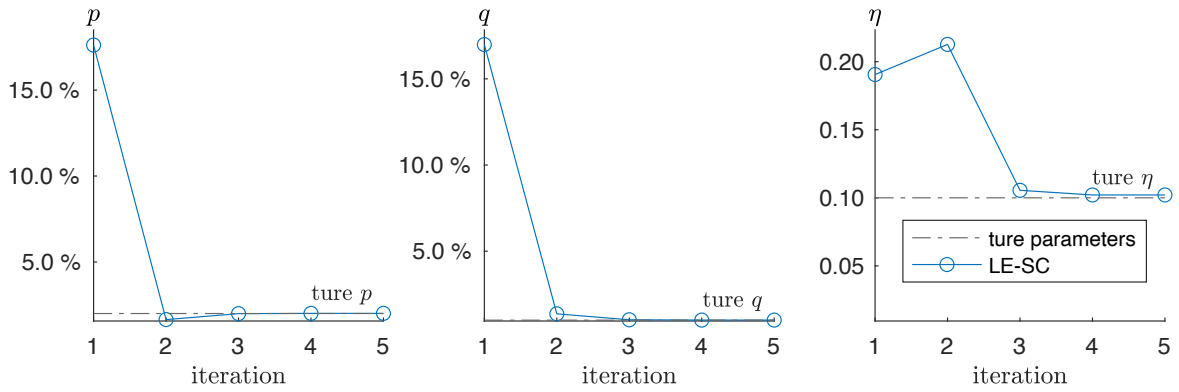


Figure 6: Illustration on the convergence of the iterative algorithm. We first sample a directed graph from DSBM with $n_1 = n_2 = 1000, p = 2\%, q = 1\%, \eta = 0.1$. Then, starting from a random initialization on the model parameters, we apply `LEHSC` to learn the DSBM parameters. The lines with circles represent model parameter learning using the spectral clustering algorithm `LEHSC`.

a. When edge direction is the primary consideration for clustering, we recommend initializing the Hermitian matrix as $H_0 = A + A^T + i(A - A^T)$ (`DirHSC`), which approximates the MLE in the low direction noise regime ($\eta \approx 0$). Alternatively, one may use $H_0 = i(A - A^T)$, which, as discussed in Section 3, corresponds to the net flow optimization approach.

b. When edge density is the main focus, we suggest initializing with $H_0 = A + A^T$, as it aligns with the total flow optimization scheme.

c. We also recommend using alternative clustering algorithms to generate initial clusters, such as `DI-SIM` Rohe et al. (2016), `BibSym` (Satuluri and Parthasarathy, 2011), and `D-SCORE` Wang et al. (2020).

F.3 Complexity analysis for eigenvector computation.

The power method is a fast and scalable algorithm for computing the eigenvectors of sparse matrices through iterative updates. Both algorithms `DirHSC` and `LEHSC` involves computing the eigenvector of a Hermitian matrix of the form

$$w_r(A + A^T) + iw_i(A - A^T) + w_c(J - I),$$

where for `DirHSC` we have $w_r = w_i = 1$ and $w_c = 0$. We decompose matrix of this form into a sparse component $w_r(A+A^T)+iw_i(A-A^T)$ and a dense all-one component w_cJ. This structure enables an efficient implementation of the power method:

1. Randomly initialize b_0
2. For $k \geq 1$, $b_k = (w_r(A + A^T) + iw_i(A - A^T))b_{k-1} + w_c(J - I)b_{k-1}$
3. Repeat until $\|b_k - b_{k-1}\|_2 \leq \epsilon$

The convergence of the above iterative steps depended linearly on the $\lambda_2(H_\theta)/\lambda_1(H_\theta)$. Each iteration involves a matrix-vector multiplication with time complexity $\mathcal{O}(|\mathcal{E}|)$ for the sparse part and $O(N)$ for the all-one matrix. Therefore, the overall computational complexity of computing the top eigenvector of H_θ is $O(|\mathcal{E}|)$, enabling the method to efficiently scale to large and sparse graphs.

F.4 Multi-community clustering

The algorithms introduced in Section 4, `DirHSC` and `LEHSC`, partition a directed graph into two clusters. To further multi-community settings, we consider two approaches: an iterative bi-partitioning scheme (Algorithm 3) and a k-way partitioning method (Algorithm 4).

F.4.1 Iterative bipartition for multiple clusters

Algorithm 3: Iterative algorithm for k clusters

Input : Directed graph $G = (\mathcal{V}, \mathcal{E})$; target number of clusters k; maximum iterations per bipartition T
Output : Community labels $\hat{\sigma} : \mathcal{V} \to \{1, \dots, k\}$

Initialize clustering $\mathcal{C} \leftarrow \{\mathcal{V}\}$ and label counter $\ell \leftarrow 1$;
while $|\mathcal{C}| < k$ **do**
  Select the largest cluster $\mathcal{S} \in \mathcal{C}$;
  Apply `LEHSC` (Algorithm 1) or `DirHSC` Algorithm 2) to subgraph $G[\mathcal{S}]$ with T iterations to bipartition $\mathcal{S}$ into $(\mathcal{S}_1, \mathcal{S}_2)$;
  Replace $\mathcal{S}$ in $\mathcal{C}$ with $\mathcal{S}_1$ and $\mathcal{S}_2$;
end
Assign unique labels $1, \dots, k$ to the final clusters in $\mathcal{C}$ to obtain $\hat{\sigma}$;
return $\hat{\sigma}$

F.4.2 k-way method for multiple clusters

Other than iterative bipartition, another approach to extend to the multi-community setting is to use a k-dimensional spectral embedding based on the top k eigenvectors, commonly referred to as k-way partitioning. This approach is motivated by the Rayleigh quotient, where the leading k eigenvectors provide orthogonal directions that approximately optimize the relaxed objective (SC-MLE). In practice, since θ is unknown, one can conduct iterative parameter estimation as in `LEHSC` (Algorithm 1). We refer to this k-way variant as `LEHSC(`k`-way)`. This yields a simple k-way extension, summarized in Algorithm 4.

Algorithm 4: k-way partition

Input : Directed graph
Output : Community labels $\hat{\sigma}$

Compute the top $\lceil k/2 \rceil$ eigenvector $\hat{\mathbf{v}}$ of H_θ;
Apply k-means to the embedding $[\Re(\hat{\mathbf{v}}), \Im(\hat{\mathbf{v}})]$ to partition into k clusters;
return *community labels* $\hat{\sigma}$.

F.5 Experiment: DSBM with different higher-order structure

When there are multiple communities in the DSBM, the edge probability between communities can vary and form different higher-order structures. We use a directed meta graph to describe such a higher-order structure of community-communitty interaction. In a meta graph, each vertex represents a community, and edges between communities are directed and weighted. Each directed edge in the meta graph indicates a structured edge generating probability, where the edge weight is the probability that an edge is oriented from a source community to a target community. A non-edge means the edge direction between the two communities is totally random, i.e, 1/2 probability for each direction. To give a concrete example, the meta graph in Figure 7 is a 5-cycle. The directed edge between C_1 and C_2 indicates that edges between them are generated with probability $1 - \eta$ pointing from C_1 to C_2 and with probability η backwards.

Figure 7 report clustering performance on multi-community DSBMs with different higher-order community structures, represented by different meta-graphs, and under different levels of directional noise η. Under hierarchical configurations (left columns), both `LEHSC` and `DirHSC` achieve strong recovery performance. In contrast, their performance degrades in the presence of cycle-structured meta-graphs (right column). This limitation is mainly due to bipartition procedure in `LEHSC` and `DirHSC`: every bipartition cutting the cycle in a statistically symmetric way, and there is no meaningful way to enforce a consistent bipartition exploit community structure. This issue can be mitigated by using a k-way variant (e.g., `LEHSC(k-way)` Algorithm 4). A more

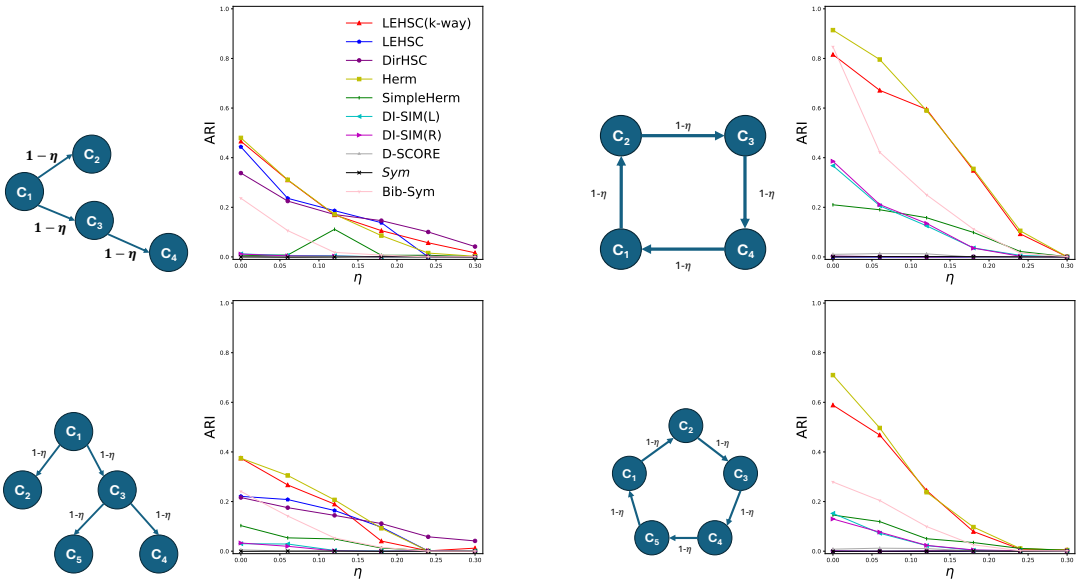
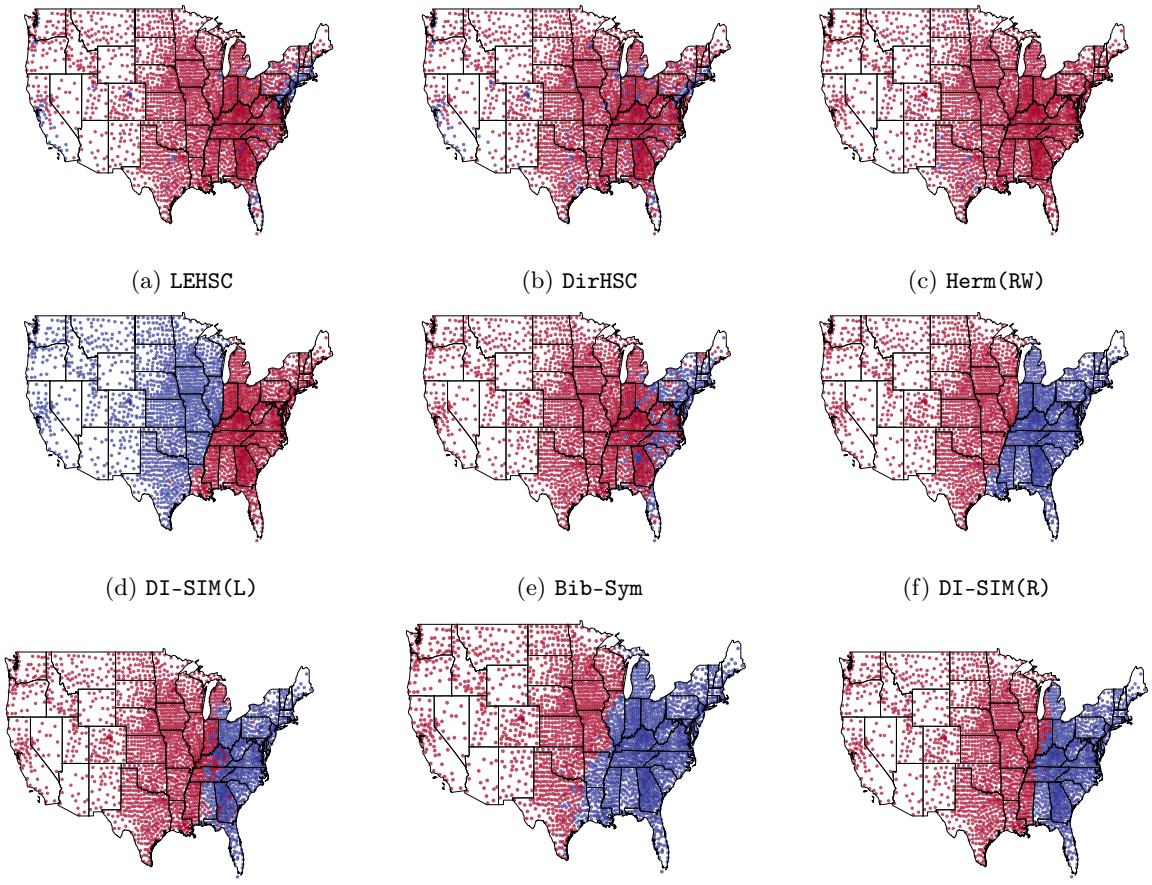


Figure 7: Recovery rate on DSBM ($n_i = 500$, $p = q = 1\%$), where communities form different higher order structure.

rigorous theoretical understanding of multi-community DSBMs with complex higher-order structures remains an important direction for future work.

F.6 Experiment: US migration data

We conduct experiments on migration data comprising 3074 counties in the mainland United States. Migration flows are represented as a directed, weighted graph, where each edge weight represents the number of individuals migrating from one county to another. To avoid a biased result dominated by extremely high degree vertices, we normalize the directed graph and use $D^{-1/2}AD^{-1/2}$, with the degree matrix D accounting for both incoming and outgoing edges. We report the clustering outcomes from different spectral methods. In our experiments, we replace the baseline method `Herm` with its variant `Herm(RW)`, which yields significantly better performance on this dataset (as `Herm` fails to produce meaningful results). We also observe that `SimpHerm` yields trivial clusters when $k \geq 4$.

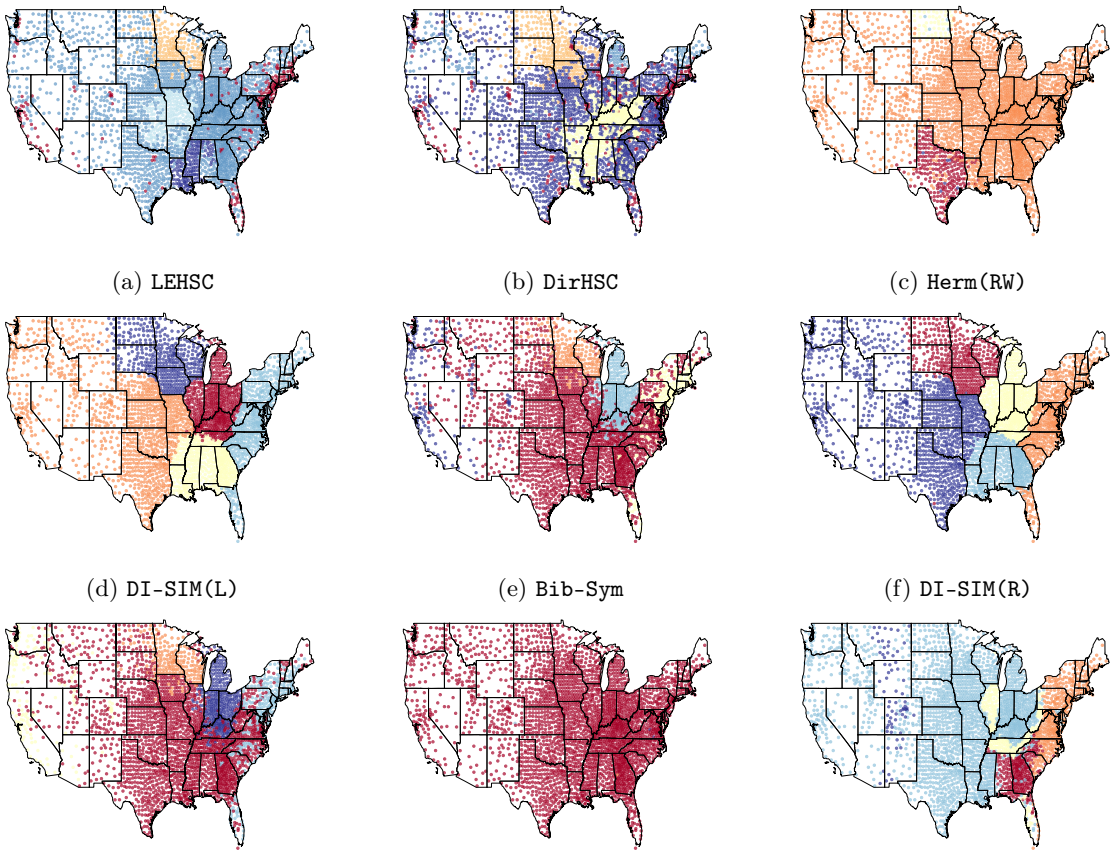


(a) `LEHSC` (b) `DirHSC` (c) `Herm(RW)`

(d) `DI-SIM(L)` (e) `Bib-Sym` (f) `DI-SIM(R)`

(g) `Sym` (h) `SimpHerm` (i) `D-Score`

Figure 8: Counties clustered by migration data ($k = 2$).

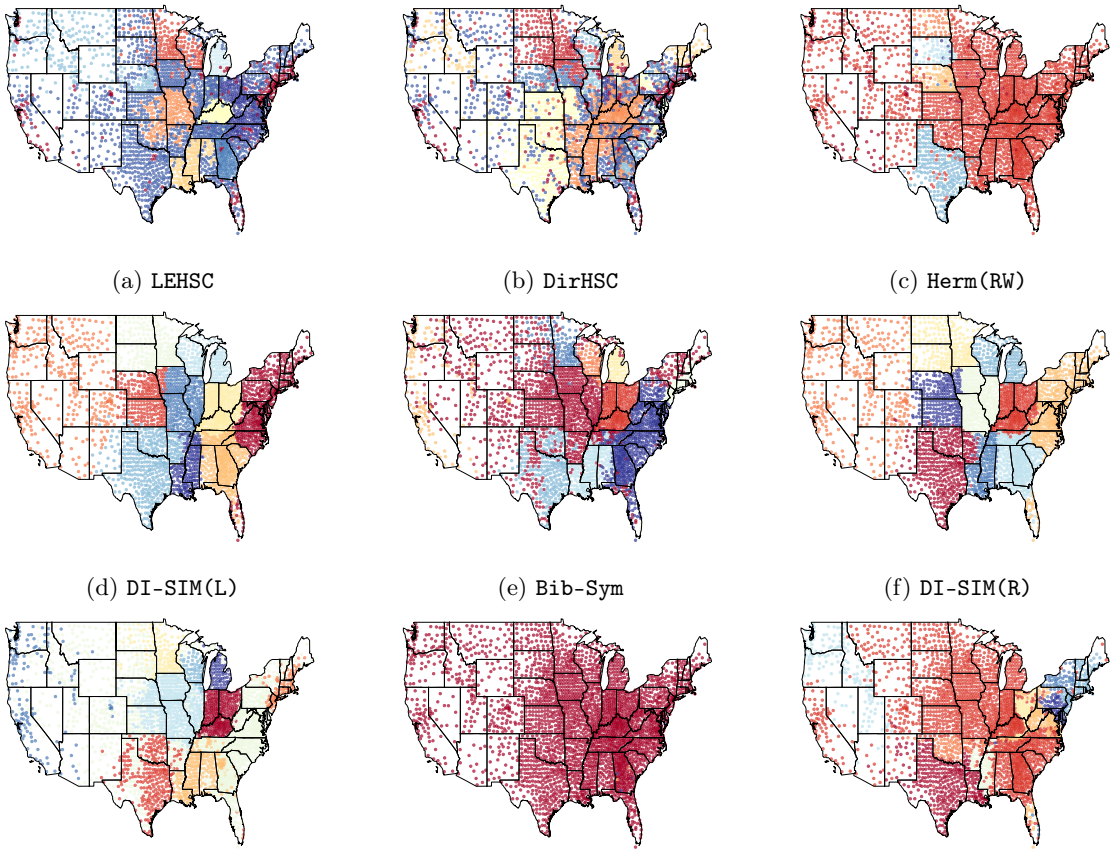


(a) `LEHSC` (b) `DirHSC` (c) `Herm(RW)`

(d) `DI-SIM(L)` (e) `Bib-Sym` (f) `DI-SIM(R)`

(g) `Sym` (h) `SimpHerm` (i) `D-Score`

Figure 9: Counties clustered by migration data ($k = 5$).



(a) `LEHSC` (b) `DirHSC` (c) `Herm(RW)`

(d) `DI-SIM(L)` (e) `Bib-Sym` (f) `DI-SIM(R)`

(g) `Sym` (h) `SimpHerm` (i) `D-Score`

Figure 10: Counties clustered by migration data ($k = 10$).