Proceedings of Machine Learning Research vol 336:1–63, 2026 39th Annual Conference on Learning Theory

Phase Transition for Stochastic Block Model with more than $\sqrt{n}$ Communities

Alexandra Carpentier ALEXANDRA.CARPENTIER@UNI-POTSDAM.DE
Institut für Mathematik – Universität Potsdam, Potsdam, Germany

Christophe Giraud CHRISTOPHE.GIRAUD@UNIVERSITE-PARIS-SACLAY.FR
Laboratoire de Mathématiques d'Orsay, Université Paris-Saclay, CNRS, France

Nicolas Verzelen NICOLAS.VERZELEN@INRAE.FR
INRAE, Institut Agro, MISTEA, Univ. Montpellier, France

Editors: Steve Hanneke and Tor Lattimore

Abstract

Predictions from statistical physics postulate that recovery of the communities in the Stochastic Block Model (SBM) with a fixed number K of communities is possible in polynomial time above, and only above, the Kesten-Stigum (KS) threshold. This conjecture has given rise to a rich literature, proving that non-trivial community recovery is indeed possible in SBM above the KS threshold. Failure of low-degree polynomials (LDP) below the KS threshold was also proven, as long as $K \ll \sqrt{n}$, where n is the number of nodes in the observed graph.

When $K \geq \sqrt{n}$, Chin et al. (2025) recently proved that, in a *sparse regime*, community recovery in polynomial time is possible below the KS threshold by counting non-backtracking paths. This breakthrough led them to postulate a new threshold for the many-communities regime $K \geq \sqrt{n}$. In this work, we provide evidence supporting their conjecture:

1- We prove that, for *any graph density*, LDP fail to recover communities below the threshold postulated by Chin et al. (2025) ;

2- We prove that community recovery is possible in polynomial time above the postulated threshold, not only in the *sparse regime* considered in Chin et al. (2025), but also in *moderately sparse regimes*, by counting occurrences of some specific motifs inspired by the LDP analysis.

In particular, counting self-avoiding paths of length $\log(n)$—which is closely related to spectral algorithms based on the Non-Backtracking operator—is optimal only in the sparse regime. More complex motifs based on the blow-up of a cycle must be considered in denser regimes.

Keywords: Stochastic Block Model, Low Degree Polynomials, Computational Barrier

1. Introduction

Network analysis investigates random interactions among individuals or objects. A typical network over n individuals can be represented as an undirected graph with n nodes, where edges encode the observed binary interactions between individuals. Examples of such networks include social networks (interactions being e.g., friendships, links, or followers), biological networks (e.g., gene—gene or gene—protein interactions), and information networks (e.g., email communication networks, internet, or citation networks). The *Stochastic Block Model* (SBM), introduced by Holland, Laskey, and Leinhardt Holland et al. (1983) is a popular statistical model for network analysis. In the SBM, the set of nodes is partitioned into $K < n$ disjoint groups, or *communities*, according to a random assignment. Then, conditional on the (unobserved) community assignments, edges are generated independently, with probabilities determined solely by the community memberships of the nodes they connect. In its simplest version, the probability of connection is p if both nodes belong to the same community, and q otherwise. When $p > q$, nodes within the same community tend to exhibit stronger connectivity among themselves than with nodes outside their community. Because of its simple analytic structure, its ability to capture community formation, and its interesting theoretical properties, the SBM has become an iconic model for network data, and it has received significant attention from theoreticians in statistics, computer science and probability Abbe (2018).

The central task in SBM analysis is to recover the latent community memberships of the nodes from a single observed network. For a fixed number of communities K, a community assignment uniformly at random, a diverging population size n, and probabilities of connection p and q scaling as $1/n$, the seminal paper of Decelle, Krzakala, Moore and Zdeborová Decelle et al. (2011) conjectured with the replica heuristic from statistical physics, that, while non-trivial community recovery is possible above the information threshold

$$\frac{n\lambda^2}{K\lambda + K^2 q} \gtrsim \frac{\log(K)}{K}, \quad \text{where } \lambda = p - q \; ; \tag{1}$$

community recovery is possible in polynomial time only above the Kesten–Stigum (KS) threshold

$$\lambda > \lambda_{KS}, \quad \text{with } \lambda_{KS} \text{ defined as the solution to } \quad \frac{n\lambda_{KS}^2}{K\lambda_{KS} + K^2 q} = 1 \, . \tag{2}$$

This conjecture then suggests that, for $K \geq 5$, there is an information-computation gap for community recovery in SBM in this sparse regime, which means that the minimal separation λ required for community recovery in polynomial-time is larger than the minimal separation required without algorithmic constraints. This conjecture has raised a high interest among theoreticians, and a lot of efforts have been devoted to confirm this conjecture.

On the one hand, non-trivial recovery in polynomial time above the KS threshold was established in this setting in Massoulié (2014); Mossel et al. (2018); Bordenave et al. (2015); Abbe and Sandon (2015); Chin et al. (2025), see Abbe (2018) for a more exhaustive review. On the other hand, a strong effort was devoted to prove the impossibility to recover the communities in poly-time below the KS threshold. Due to the random nature of the observed graph in the SBM, the classical notions of worst-case hardness, such as P, NP, etc are not suitable for proving computational hardness in this framework. In such random settings, other notions are considered, such as reduction to other hard statistical problems, like planted clique Brennan and Bresler (2020); Berthet and Rigollet (2013); Brennan et al. (2018), or computational hardness in some specific models of computations, such as SoS Hopkins et al. (2017); Barak et al. (2019), overlap gap property Gamarnik (2021), statistical query Kearns (1998); Brennan et al. (2021), and low-degree polynomials (LDP) Hopkins (2018); Kunisky et al. (2019); Schramm and Wein (2022); Sohn and Wein (2025). Among these, LDP bounds have gained attention for establishing state-of-the-art lower bounds in tasks like community detection Hopkins and Steurer (2017), spiked tensor models Hopkins and Steurer (2017); Kunisky et al. (2019), sparse PCA Ding et al. (2024), Gaussian clustering Even et al. (2025), and many other models Wein (2025). In this model of computation, only estimators that are multivariate polynomials of degree at most D are considered. It is conjectured that, for many problems, failure of degree $D = O(\log n)$ polynomials enforces failure of any polynomial-time algorithm Kunisky et al. (2019) (LDP conjecture). The first LDP hardness results for SBM were for the detection problem Hopkins and Steurer (2017); Bandeira et al. (2021); Kunisky (2024), which is the problem of testing $\lambda = 0$ against $\lambda > 0$. LDP hardness below the KS threshold (2) for community recovery was proven in Luo and Gao (2024); Sohn and Wein (2025); Ding et al. (2025).

The setting where the number K of communities grows with n has attracted much less attention until recently. A polynomial time algorithm performing non-trivial recovery close to the KS threshold was already proposed in Chen and Xu (2016). Yet, for growing K, non-trivial community recovery in poly-time just above the KS threshold was only obtained recently in Stephan and Zhu (2024); Chin et al. (2025). Lower bounds for growing K have been derived in Luo and Gao (2024); Ding et al. (2025); Chin et al. (2025), proving failure of LDP below the KS threshold (2) when $K = o(\sqrt{n})$, both in the sparse case discussed here (q scaling like $1/n$), and in the denser case where probabilities of connection grow faster than $1/n$. In particular, Chin et al. (2025) proves that, conditional on the LDP conjecture, the prediction from Decelle et al. (2011) holds in the sparse case as long as $K = o(\sqrt{n})$. Indeed, in addition to proving failure of LDP below the KS threshold in any regime, they also prove the existence of a poly-time algorithm based on non-backtracking statistics which succeeds to partially recover the communities above the KS threshold in the sparse regime (where q scales like $1/n$).

The regime with a higher number of communities, $K \geq \sqrt{n}$ remains mostly not understood. Chin, Mossel, Sohn and Wein Chin et al. (2025) made an important breakthrough by proving that when $K \gg \sqrt{n}$, a poly-time algorithm based on non-backtracking statistics succeeds to partially recover the communities below the KS threshold in the *sparse regime* (where q scales like $1/n$). A similar phenomenon had already been observed in Even et al. (2024) for the related problem of clustering in Gaussian mixture: When $K \geq \sqrt{n}$, a simple hierarchical clustering algorithm succeeds to recover the clusters below the KS threshold (of Gaussian mixture models), and it has been proved to be optimal up to possible log factors Even et al. (2024, 2025). For the SBM in the sparse regime (q scaling like $1/n$), Chin et al. (2025) prove that, when $K \gg \sqrt{n}$, partial recovery of the communities is possible in polynomial time when

$$\lambda \gtrsim_{\log} \left(q + \frac{\lambda}{K}\right)^{1-\log_n(K)}, \tag{3}$$

where $\log_n$ is the logarithm in base n (i.e. $K = n^{\log_n(K)}$) and where $\gtrsim_{\log}$ hides a poly-log factor, which is proportional to the square of the natural logarithm of the average density. The Condition (3) can be simplified as

$$\lambda \gtrsim_{\log} q^{1-\log_n(K)}, \tag{4}$$

for $q \geq 1/n$. This upper-bound for the sparse regime led them to pose the following open questions (Question 1.9 in Chin et al. (2025), that we transpose with our notation)

Let $\log_n(K) = \frac{1}{2} + \Omega(1)$. Does the phase transition for efficient weak recovery occur at $\lambda \gtrsim_{\log} \left(q + \frac{\lambda}{K}\right)^{1-\log_n(K)}$, where $\gtrsim_{\log}$ hides poly-log factors? Moreover, is there a sharp or coarse transition in the recovery accuracy as λ varies as a function of the average density?

Let us (partially) rephrase these two questions for convenience of the discussion. When $K \gg \sqrt{n}$:

Q1: Does the phase transition occur at (4) in the sparse regime (q scaling like $1/n$)?

Q2: Where does the phase transition occur in denser regimes?

For the first question Q1, one may guess that the answer should be positive. Indeed, non-backtracking statistics are tightly related to Belief-Propagation and the Bethe free-energy, the latter being known to be a valid approximation of the log-likelihood in the sparse regime, while the former is expected to be optimal again in the sparse regime Krzakala et al. (2013); Moore (2017). Hence, as far as the analysis of Chin et al. (2025) is tight, we can expect that (4) is the threshold where the phase transition occurs in the sparse regime.

The answer to the second question Q2 seems much more open. First, the upper-bound of Chin et al. (2025) is for the sparse regime, and we are not aware of any result proving non-trivial recovery below the KS threshold in a denser regime. Second, even if the upper-bound of Chin et al. (2025) were to hold in a denser regime, there is no strong indication that non-backtracking statistics should be optimal in this denser regime. In particular, in the light of the results for the Gaussian clustering problem Even et al. (2024), we may guess that some other algorithms could be more powerful in denser regimes.

1.1. Main contributions

In this paper, we provide an answer to these two questions. We prove that, when $K \geq \sqrt{n}$

1. LDP fail for partial recovery when $\lambda \lesssim_{\log} q^{1-\log_n(K)}$, for all density regimes;
2. Community recovery in poly-time is possible in moderately sparse regimes when (4) holds.

These two results provide evidence that, when $K \geq \sqrt{n}$, the phase transition for community recovery in poly-time holds around the Threshold (4) conjectured by Chin et al. (2025). Let us present our two main contributions with more details.

Contribution 1: Low degree polynomial lower bound. Our first main result is a LDP lower bound for community recovery for any $K \leq n$. Low-degree polynomials are not well-suited for directly outputting a partition of the nodes, which is combinatorial by nature. Instead, as standard for clustering or community recovery Luo and Gao (2024); Even et al. (2024); Sohn and Wein (2025); Chin et al. (2025); Even et al. (2025), we focus on the problem of estimating $x_{ij} = \mathbf{1}\{i, j \text{ are in the same community}\} - 1/K$, which can be directly related to the problem of clustering Even et al. (2024, 2025). Theorem 6, Appendix A, states a precise version of the following.

Theorem 1 (Informal) *When*

$$\lambda \lesssim_{\log} \lambda_c, \quad \text{with } \lambda_c \text{ solution to} \quad \sup_{r\geq 1} \frac{n\lambda_c^{2r}}{K\lambda_c^r + K^2 q^r} = 1\,, \tag{5}$$

any $O(\log(n))$-degree polynomial f fails to estimate x_{ij} significantly better than $f = \mathbb{E}[x_{ij}] = 0$.

At first sight, the Condition (5) looks unrelated to the Threshold (4) identified by Chin et al. (2025). Next lemma, proved in Appendix J, highlights that the Condition (5) is closely related both to the Kesten-Stigum threshold (2) when $K \leq \sqrt{n}$ and the Threshold (4) when $K \geq \sqrt{n}$.

Lemma 2 *Assume that $q = n^{-\alpha_n}$ with $\alpha_n \in (0, +\infty)$, and $K = n^{\frac{1+\delta_n}{2}}$ —i.e. $\log_n(K) = (1+\delta_n)/2$— with $\delta_n \in (-1, 1)$. Define $\bar{r}_n$ as the value $r \geq 1$ achieving the maximum in (5). Then, for any small $\epsilon > 0$, and n large*

1. *If* $K \leq (1-\epsilon)\sqrt{n}$, *then* $\bar{r}_n = 1$ *and* $\lambda_c = \lambda_{KS} := \frac{K}{2n}\left(1+\sqrt{1+2nq}\right)$.

2. *If* $(1+\epsilon)\sqrt{n} \leq K \leq n^{1-\epsilon}$, *then*

   - *If* $q \leq \frac{1-\delta_n^2}{4\delta_n^2} \times \frac{1-o(1)}{n}$, *then* $\bar{r}_n = 1$ *and* $\lambda_c = \lambda_{KS}$;

   - $q \geq \frac{1-\delta_n^2}{4\delta_n^2} \times \frac{1-o(1)}{n}$, *then* $\bar{r}_n \sim \alpha_n^{-1}\left(1 - \frac{\log\left(\frac{1-\delta_n^2}{4\delta_n^2}\right)}{\log(n)} - O\left(\frac{1}{(\log(n))^2}\right)\right)$ *and*

$$\lambda_c \sim q^{(1-\delta_n)/2}\left(\frac{(1+\delta_n)^{1+\delta_n}}{(1-\delta_n)^{1-\delta_n}(2\delta_n)^{2\delta_n}}\right)^{\alpha_n/2}. \tag{6}$$

When $K \leq \sqrt{n}$, we have $\lambda_c = \lambda_{KS}$, and we recover that low-degree polynomials fail to estimate x_{ij} better than 0 below the Kesten-Stigum threshold, as already proved in Chin et al. (2025). When $K \geq \sqrt{n}$ and $q \gtrsim n^{-1}$, then the threshold value (6) matches the Threshold (4) since $(1-\delta_n)/2 = 1 - \log_n(K)$.

Theorem 1 establishes that the Threshold (4) identified by Chin et al. (2025) is correct, up to a poly-logarithmic factor, in the sparse regime $q \asymp 1/n$ with $K \gg \sqrt{n}$. Moreover, it serves as a lower bound on the minimal separation required for polynomial-time recovery in denser regimes. Our second main contribution is to leverage our LDP analysis for deriving novel optimal motif counts. These counts enable polynomial-time community recovery in the intermediate dense regime $q \asymp n^{-\alpha}$ (for $\alpha \in (0,1) \cap \mathbb{Q}$), provided that $K \geq \sqrt{n}$ and (4) are satisfied.

Contribution 2: Successful recovery above threshold λ_c with motif counts. Chin et al. (2025) prove that, in the sparse regime (q scaling like $1/n$), partial recovery of the communities is possible above the threshold (4) with non-backtracking statistics, when $K \gg \sqrt{n}$. Our second contribution — Corollaries 10 ; 13 and 17 Appendix B — is to prove that partial recovery of the communities is possible above the Threshold (6), with an algorithm based on (weighted) motif counting.

Theorem 3 (Informal) *Assume that* $q \asymp n^{-\alpha}$ *with* $\alpha \in (0,1) \cap \mathbb{Q}$. *When* $\lambda \gtrsim_{\log} \lambda_c$, *with* λ_c *given in* (6), *community recovery is possible in polynomial time with an algorithm based on counting a motif, whose shape depends on* α.

Combining Theorem 1 with Theorem 3 —and the result of Chin et al. (2025) for the case $\alpha = 1$—, we get that for $q \asymp n^{-\alpha}$ with $\alpha \in (0,1] \cap \mathbb{Q}$, the phase transition for weak recovery in poly-time occurs at level λ_c, up to possible log factors. When α is irrational, it follows from our results that community recovery is possible either above the threshold λ_c with a polynomial of degree $\log(n)$ —but with $O(n^{\log(n)})$ algorithmic complexity—, or in polynomial time if $\lambda \geq n^{\epsilon}\lambda_c$, where $\epsilon > 0$ can be arbitrarily small. See the discussion after Corollary 17 (page 27) for details.

As mentioned in Theorem 3, the motifs involved in the log-optimal algorithm depends on the power α of the between group probability of connection $q = n^{-\alpha}$. For some very specific values of α, some log-optimal motifs are very simple. Indeed, we prove in Corollaries 10 and 13, Appendix B, the following.

Corollary 4 (Informal) *Assume that* $K \geq \sqrt{n}$ *and* $\lambda \gtrsim_{\log} \lambda_c$.

1. When $q \asymp n^{-\frac{2}{m+1}}$, *with* $m \in \{3,4,5,\dots\}$, *community recovery is possible in polynomial time with an algorithm based on counting cliques of size* m.

2. *When $q \asymp n^{-\frac{m-2}{m-1}}$, with $m \in \{3, 4, 5, \dots\}$, community recovery is possible in polynomial time with an algorithm based on counting self-avoiding paths of length $m - 1$.*

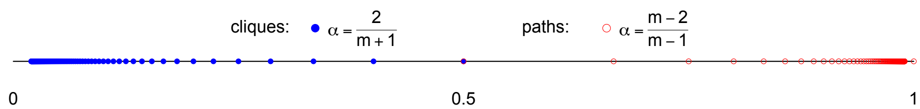

Figure 1: Values of $\alpha \in \left\{\frac{2}{m+1} : m = 3, 4, \dots\right\} \cup \left\{\frac{m-2}{m-1} : m = 3, 4, \dots\right\}$ corresponding to densities $q \asymp n^{-\alpha}$ for which m-cliques counting (blue solid markers) and $(m-1)$-self-avoiding paths counting (red circles) succeed at the (log-)optimal threshold $\lambda \gtrsim_{\log} \lambda_c$.

We notice that taking m slowly diverging[1] with n, we recover the success of self-avoiding paths counting above the threshold (4) when $q \asymp \frac{1}{n}$, as already proved in Chin et al. (2025). We underline yet, that counting length-$\log(n)$ self-avoiding paths —which is closely related to spectral algorithms based on the Non-Backtracking operator— is optimal only in the sparse regime $q \asymp \frac{1}{n}$. Other motifs counts should be considered in Theorem 3 in denser regimes.

For general values of α, our LDP analysis dictates a more complex family of (log-)optimal motifs. Writing $\alpha^{-1} = \gamma + a$, where $\gamma \geq 1$ is an integer and a is a rational in $(0, 1)$, the motif is constructed by considering a γ-blow-up of a cycle of length κ, which is then linked evenly to the reference nodes i and j with $a\kappa\gamma$ edges —referred to as fasteners in the following. The length κ of the cycle is chosen in such a way that $a\kappa\gamma$ is a sufficiently large even integer. We refer to the discussion Section 1.2.2 for some intuition on the choice of this motif, and to Section 1.2.3 page 12 for the formal description.

It is worth mentioning that counting self-avoiding paths of length 2 or cliques of size $m = 3$ is closely related to the hierarchical clustering algorithm in Even et al. (2024), which is log-optimal for Gaussian mixture when $K \gg \sqrt{n}$. Yet, counting 3-cliques is only log-optimal for SBM in a narrow regime of density where $q \asymp n^{-1/2}$. Indeed, contrary to the Gaussian case, the minimal separation λ remains bounded, and in denser regimes, better results can be obtained by averaging the observations over a wider m-clique, to reduce the variance.

Open questions. While our results, combined with those of Chin et al. (2025), provide a better understanding of the phase transition for community recovery when $K \geq \sqrt{n}$, some questions remain open. A first question is relative to the exact location of the phase transition when $K \geq \sqrt{n}$. Indeed, while Chin et al. (2025) prove that the phase transition occurs exactly at $\lambda_c = \lambda_{KS}$ when $K \leq \sqrt{n}$ (and $q \asymp 1/n$ for the upper bound), the upper and lower bounds for $K \geq \sqrt{n}$ involve (un-matching) poly-log factors.

Question 1 *Is the threshold λ_c defined in (5) —equivalently (6)— the* exact *threshold above which weak recovery is possible in polynomial time?*

A second question is relative to optimal motifs for successful recovery. Indeed, while we show that, when $\lambda \gtrsim_{\log} \lambda_c$, we can recover the communities with the family of blow-up motifs introduced in Section 1.2.3, the question of sharp optimality (in terms of log factors) remains open.

1. When m diverges with n, counting length-$(m-1)$ self-avoiding paths has a complexity super-polynomial in n. This issue can be overcome by counting non-backtracking paths instead Chin et al. (2025), which can be done efficiently.

Question 2 *What are the optimal poly-time algorithms for weakly recovering the communities above the computational barrier (with sharp constants)?*

1.2. Proof technique, insights, and limitation

Let us denote by $Y^* \in \mathbb{R}^{n\times n}$ the adjacency matrix of the observed graph on n vertices. The graph being undirected and with no self-loops, the matrix Y^* is symmetric with zero on the diagonal. We assume that Y^* has been generated according to the stochastic block model: Each node i is assigned uniformly at random to one among K communities, and then nodes within a same community are connected with probability p, and those between different communities with probability q.

More formally, let K be an integer in $[2, n]$, and $p, q \in (0, 1)$, with $p > q$. We assume that Y^* is sampled according to the following distribution:

1. $z_1, \ldots, z_n \overset{i.i.d.}{\sim}$ Uniform $\{1, \ldots, K\}$;
2. Conditionally on $z_1, \ldots, z_n$, the entries $(Y^*_{ij})_{i<j}$ are sampled independently, with Y^*_{ij} distributed as a Bernoulli random variable with parameter $q + \lambda \mathbf{1}\{z_i = z_j\}$, where $\lambda = p - q$.

To get simpler formulas, in the remaining of the paper, we will work with the (almost)-recentered adjacency matrix

$$Y_{ij} = Y^*_{ij} - q \text{ , for any } 1 \le i < j \le n. \tag{7}$$

We will also recurrently use the notation

$$\bar{q} = q(1-q), \quad \text{and} \quad \bar{p} = \bar{q} + \lambda(1-2q). \tag{8}$$

1.2.1. LOW-DEGREE LOWER BOUND

Let $\mathbb{P}_\lambda$ denote the distribution of the SBM with a prescribed separation λ. Let us consider the problem of estimating $x_{12} = \mathbf{1}\{z_1 = z_2\} - 1/K$. We say that degree-$D$ polynomials fail to estimate x_{12} in $L^2(\mathbb{P}_\lambda)$-norm, when no degree-D polynomial can estimate x_{12} significantly better than the constant estimate $\mathbb{E}_\lambda[x_{12}] = 0$. In mathematical language, proving failure of degree-D polynomials for estimating x_{12} amounts to prove that

$$MMSE_D := \inf_{f:\deg(f)\le D} \mathbb{E}_\lambda\left[(f - x_{12})^2\right] = \mathbb{E}_\lambda[x_{12}^2] - \sup_{f:\deg(f)\le D} \frac{\mathbb{E}_\lambda[f x_{12}]^2}{\mathbb{E}_\lambda[f^2]} = (1+o(1))MMSE_0, \tag{9}$$

where the inf and sup are over all the polynomials of the observation (the adjacency matrix in our case) with degree at most D. Since $MMSE_0 = \mathbb{E}_\lambda[x_{12}^2] - \mathbb{E}_\lambda[x_{12}]^2 = \mathbb{E}_\lambda[x_{12}^2]$, the problem reduces to proving that

$$\mathrm{Corr}^2_{\le D} := \sup_{f:\deg(f)\le D} \frac{\mathbb{E}_\lambda\left[x_{12} f\right]^2}{\mathbb{E}_\lambda\left[f^2\right]} \le o\left(\mathbb{E}_\lambda\left[x_{12}^2\right]\right) = o\left(\frac{1}{K}\right). \tag{10}$$

Low-degree polynomials theory has emerged in a sequence of works Hopkins et al. (2017); Hopkins and Steurer (2017); Hopkins (2018); Kunisky et al. (2019) focusing on detection problems, where the goal is to perform a test with a simple null distribution (typically with independent entries). For SBM, the detection problem amounts to test $\lambda = 0$ against $\lambda > 0$. In practice, it reduces to proving $\mathrm{Corr}^2_{\le D} = o(1)$, with $\lambda = 0$ and x_{12} replaced by the likelihood ratio between $\mathbb{P}_\lambda$ and $\mathbb{P}_0$.

The core strategy for detection problems is to expand the polynomials f on a $L^2(\mathbb{P}_0)$-orthonormal basis of low-degree polynomials, and then to solve the resulting optimization problem explicitly. This approach has been successfully implemented for SBM in Hopkins and Steurer (2017). A similar approach cannot be implemented for estimation (i.e. recovery) problems, as no simple explicit $L^2(\mathbb{P}_\lambda)$-orthonormal basis of low-degree polynomials is known for $\lambda > 0$. Three strategies have been proposed to circumvent this issue:

1. Schramm and Wein Schramm and Wein (2022) schematically apply a partial Jensen inequality with respect to the latent variables (the community assignment in SBM), and expand the polynomials f on a $L^2(\mathbb{P}_0)$-orthonormal basis. Then, the optimization problem (10) can be upper-bounded recursively. This approach has been implemented by Luo and Gao Luo and Gao (2024) for SBM.

2. To get tighter bounds, Sohn and Wein Sohn and Wein (2025) expand the polynomials f over a basis which is orthonormal in an enlarged probability space including the latent variables. While the analysis is much more delicate, it leads to sharp results. It has been implemented for SBM in Sohn and Wein (2025); Chin et al. (2025) to prove failure of low-degree polynomials below the KS threshold (2) when $K = o(\sqrt{n})$.

3. To provide a simple and more direct analysis, Carpentier, Giancola, Giraud and Verzelen Carpentier et al. (2025a) build a basis for permutation invariant polynomials which is almost $L^2(\mathbb{P}_\lambda)$-orthonormal when the signal is low, i.e. typically when recovery is impossible. Compared to Sohn and Wein (2025), a counterpart of the simplicity of the analysis, is the derivation of less tight lower bounds. Indeed, similarly as for the Schramm and Wein Schramm and Wein (2022) approach, bounds derived from the approach of Carpentier et al. (2025a) are typically tight only up to polylogarithmic factors.

To prove Theorem 1, we leverage the ideas introduced in Carpentier et al. (2025a), that we adapt to our setting. We adjust the basis proposed in Carpentier et al. (2025a) by using a better renormalisation. On top of the proof strategy of Carpentier et al. (2025a), the main technical result is Proposition 22, which controls the correlation between two basis elements. We highlight below the main steps and ideas. All the details are provided in the appendices. We recall that we consider the problem of estimating $x_{12} = \mathbf{1}\{z_1 = z_2\} - \frac{1}{K}$.

Almost orthonormal basis for permutation-invariant polynomials. A first key idea is to notice that, since the distribution of x_{12} and Y is invariant by permutation of the indices — except 1,2— the maximizer of $\mathrm{Corr}^2_{\leq D}$ —defined by (10)— must be a polynomial invariant by permutation of the indices, except 1,2. Hence, we only need to focus on such polynomials, we call *invariant polynomials*.

We can build a simple family spanning the set of degree-D invariant polynomials. Let $G = (V, E)$ be a graph on a node set $V = \{v_1, \ldots, v_D\}$, and Π_V be the set of injections $\pi : V \to \{1, \ldots, n\}$ fulfilling $\pi(v_1) = 1$ and $\pi(v_2) = 2$. Then, any degree-D invariant polynomials can be decomposed on the family $(P_G)_G$ where

$$P_G = \sum_{\pi \in \Pi_V} P_{G,\pi}, \quad \text{with} \quad P_{G,\pi}(Y) = \prod_{(i,j)\in E} Y_{\pi(i),\pi(j)}.$$

We observe that, when applied to the vanilla adjacency matrix Y^*, the polynomial $P_G(Y^*)$ merely counts the number of occurrence of the motif G "attached" at 1,2. The family $(P_G)_G$ itself is not an almost-orthonormal basis of degree-D invariant polynomials. First, it is overcomplete, so we need to extract a maximal free subfamily $(P_G)_{G\in\mathcal{G}_{\le D}}$. Second, $(P_G)_{G\in\mathcal{G}_{\le D}}$ is neither almost-orthogonal, nor almost-normalized. To get almost-orthogonality, we apply some correction as in Carpentier et al. (2025a) (see Appendix C), and to get normality, we rescale the resulting basis by some normalizing factor, providing an almost-orthonormal basis $(\Psi_G)_{G\in\mathcal{G}_{\le D}}$. Improving the normalization process compared to Carpentier et al. (2025a) is decisive for deriving our results, as explained below.

Normalization of the polynomials. For simplicity, let us drop the correction applied to P_G to achieve approximate orthogonality. Our goal is then to normalize the polynomial P_G by its standard deviation $\sqrt{\mathrm{var}(P_G)}$. The second moment of P_G can be expanded as

$$\mathbb{E}\left[P_G^2\right] = \sum_{\pi^{(1)},\pi^{(2)}\in\Pi_V} \mathbb{E}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right]. \tag{11}$$

Assume that the dominating terms in this sum are those for which $\pi^{(1)}(G) = \pi^{(2)}(G)$, where $\pi(G) = (\pi(V), \{(\pi(u),\pi(v)) : (u,v)\in E\})$ represents the graph G with nodes labeled by π. This yields the simplification

$$\mathbb{E}\left[P_G^2\right] \asymp \sum_{\pi\in\Pi_V} |\mathrm{Aut}(G)|\,\mathbb{E}\left[P_{G,\pi}^2\right] \asymp n^{|V|}|\mathrm{Aut}(G)|\,\mathbb{E}\left[P_{G,\pi}^2\right], \tag{12}$$

where $|\mathrm{Aut}(G)|$ denotes the number of automorphisms of G. While Carpentier et al. (2025a) uses $q^{|E|}$ as a simple proxy for the second moment $\mathbb{E}[P_{G,\pi}^2]$ to greatly simplify their analysis, this choice is not precise enough to derive tight bounds in our context.

To understand why the proxy $q^{|E|}$ is insufficient, let us consider a clique $G = (V,E)$, for which approximate computations can be performed explicitly. Let $|E^{\neq}|$ denote the number of edges in $\pi(G)$ between nodes of distinct communities, and let $\ell(G)$ denote the number of distinct communities in $\pi(G)$. Direct computation yields $\mathbb{P}[\ell(G)=\ell] \asymp K^{\ell-|V|}$ for $1\le \ell\le |V|$ and

$$\begin{aligned}\mathbb{E}\left[P_{G,\pi}^2\right] &\asymp \lambda^{|E|}\mathbb{E}\left[\left(\frac{q}{\lambda}\right)^{|E^{\neq}|}\right] \asymp \sum_{\ell=1}^{|V|}\frac{\lambda^{|E|}}{K^{|V|-\ell}}\,\mathbb{E}\left[\left(\frac{q}{\lambda}\right)^{|E^{\neq}|}\middle|\ell(G)=\ell\right]\\ &\asymp \sum_{\ell=1}^{|V|}\frac{\lambda^{|E|}}{K^{|V|-\ell}}\,\left(\frac{q}{\lambda}\right)^{(\ell-1)(|V|-\ell/2)},\end{aligned}$$

where the last equivalence follows because $q<\lambda$, and, conditioned on $\ell(G)=\ell$, the exponent $|E^{\neq}|$ is minimized when all but $\ell-1$ vertices belong to the same community. Since the overall exponent in the sum is quadratic in ℓ, the sum is typically dominated by its first ($\ell=1$) or last ($\ell=|V|$) terms, leading to

$$\mathbb{E}\left[P_{G,\pi}^2\right] \asymp \frac{\lambda^{|E|}}{K^{|V|-1}} + q^{|E|}. \tag{13}$$

Hence, for a clique G, we observe an additional term $\lambda^{|E|}/K^{|V|-1}$ compared to the proxy $q^{|E|}$, which cannot be neglected in our setting.

Consequently, instead of normalizing the polynomials P_G by the proxy $n^{|V|/2}|\mathrm{Aut}(G)|^{1/2}\,q^{|E|/2}$ as in Carpentier et al. (2025a), we apply an implicit normalization by $n^{|V|/2}|\mathrm{Aut}(G)|^{1/2}\mathbb{E}[P_{G,\pi}^2]^{1/2}$. The price to pay for this nearly exact normalization is a significantly more involved analysis to prove the approximate orthonormality of the basis $(\Psi_G)_{G\in\mathcal{G}_{\le D}}$, particularly when controlling the correlation terms.

Controlling correlation terms. Since the polynomials are now normalized by $\mathbb{E}[P_{G,\pi}^2]^{1/2}$, a central challenge in proving the approximate orthonormality of the basis $(\Psi_G)_{G\in\mathcal{G}_{\leq D}}$ is to tightly upper-bound the absolute value of the correlation

$$\frac{\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]}{\sqrt{\mathbb{E}\left[P^2_{G^{(1)},\pi^{(1)}}\right]\mathbb{E}\left[P^2_{G^{(2)},\pi^{(2)}}\right]}}$$

between $P_{G^{(1)},\pi^{(1)}}(Y)$ and $P_{G^{(2)},\pi^{(2)}}(Y)$ for two given motifs $G^{(1)}, G^{(2)}$ and two injections $\pi^{(1)}, \pi^{(2)} \in \Pi_V$. This correlation is upper-bounded in Proposition 22, which forms the cornerstone of our analysis. With this bound in hand, the proof of the approximate orthonormality of $(\Psi_G)_{G\in\mathcal{G}_{\leq D}}$ follows a path similar to that in Carpentier et al. (2025a).

It is worth noticing that for proving success of counting a motif G, like e.g. a clique of size m, a key feature is to upper-bound a variant of the correlation $\mathbb{E}[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}]/\mathbb{E}[P_{G,\pi^{(1)}}]^2$, see Appendix F. So, understanding the scalar product terms $\mathbb{E}[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}]$ appears to be crucial for understanding the nature of the phase transition in SBM.

Identifying the optimal polynomials. Once we have proven that $(\Psi_G)_{G\in\mathcal{G}_{\leq D}}$ is almost-orthonormal, we have

$$\mathrm{Corr}^2_{\leq D} \approx \sum_{G\in\mathcal{G}_{\leq D}} \mathbb{E}[x_{12}\Psi_G]^2 .$$

Furthermore, setting $\bar{q} = q(1-q)$, we prove in Appendix D that

$$\mathbb{E}[x_{12}\Psi_G]^2 \leq \frac{1}{n}\left[\frac{n\lambda^{2r(G)}}{(K\lambda^{r(G)}) \vee (K^2\bar{q}^{r(G)})}\right]^{|V|-1}, \qquad \text{with} \quad r(G) := \frac{|E|}{|V|-1}, \tag{14}$$

when G is connected, and $\mathbb{E}[x_{12}\Psi_G] = 0$ otherwise. Hence $\mathrm{Corr}^2_{\leq D} = O(1/n)$ when $\lambda \lesssim_D \lambda_c$, with λ_c defined in (5).

1.2.2. OPTIMAL COMMUNITY RECOVERY WITH MOTIF COUNTS.

In addition to provide a simple and direct proof of the lower-bound, the construction of the almost orthonormal basis allows to get insights on the best degree-D polynomials, and hence on the optimal poly-time procedures. We underline that, schematically,

1. the Bound (14) is valid only when the bracket in the right-hand side is small compared to 1;
2. non-trivial recovery becomes possible when this bracket becomes larger than 1 for some G.

The bound (14) then provides some interesting information on the optimal motifs G. Indeed, taking for granted that (14) is quite tight below the Threshold (4) — it should be so—, we observe that the motifs G maximizing $\mathbb{E}[x_{12}\Psi_G]^2$ are those being connected and maximizing

$$\left[\frac{n\lambda^{2r(G)}}{(K\lambda^{r(G)}) \vee (K^2\bar{q}^{r(G)})}\right]^{|V|-1}, \qquad \text{with} \quad r(G) := \frac{|E|}{|V|-1}.$$

If this remains true at the phase transition, where this quantity is close to 1, it provides insights on the choice of the motif G for optimal recovery.

When $K \geq \sqrt{n}$ and $q \asymp n^{-\alpha}$, Lemma 2 ensures that the value $\bar{r}$ achieving the maximum in (5) fulfills $\bar{r} \sim (\alpha^{-1} \vee 1)$. Accordingly, let us focus on the density regime $q \asymp n^{-1/r}$ with $r > 1$, and with signal condition $\lambda \geq_{\log} \lambda_c \asymp q^{1-\log_n(K)}$. In particular,

$$\lambda^r \geq_{\log} q^r K \asymp \frac{K}{n}. \tag{15}$$

Equation (14) suggests to count the occurrence of a connected motif G fulfilling $r = |E|/(|V|-1)$, where r is such that $q \asymp n^{-1/r}$. In fact, since our objective is to test whether 1 and 2 belong to the same community rather than to estimate x_{12}, this ratio must be slightly adjusted to (see Equation (18) below)

$$r = \frac{|E|}{|V|-2}\,. \tag{16}$$

Introducing the notation $\mathbb{P}_{12} := \mathbb{P}\left[\cdot|z_1 = z_2\right]$ and $\mathbb{P}_{\not{1}2} := \mathbb{P}\left[\cdot|z_1 \neq z_2\right]$, a direct computation gives $\mathbb{E}_{\not{1}2}\left[P_G\right] = 0$. So, for testing $z_1 = z_2$ against $z_1 \neq z_2$ with a small error, we seek for a connected motif G fulfilling

$$\mathbb{E}_{12}\left[P_G\right]^2 \gg \mathrm{var}_{12}(P_G) \vee \mathrm{var}_{\not{1}2}(P_G). \tag{17}$$

The expectation can be simply evaluated (see Equations (72) and (65))

$$\mathbb{E}_{12}\left[P_G\right] \asymp \left(\frac{n\lambda^{|E|/(|V|-2)}}{K}\right)^{|V|-2} = \left(\frac{n\lambda^r}{K}\right)^{|V|-2}, \tag{18}$$

when G is connected and (16) holds. So, according to (15), we have $\mathbb{E}_{12}\left[P_G\right] \gtrsim 1$. For analyzing the variance terms, we proceed as in (12). We recall that $|E^{\neq}|$ denotes the number of edges in $\pi(G)$ between nodes of distinct communities, $\bar{q} = q(1-q)$ and $\bar{p} = \bar{q} + \lambda(1-2q)$, with $\bar{p} \sim \lambda$ since $\lambda \geq_{\log} K^{1/r} q$ according to (15). A direct computation gives (see Equation (71))

$$\mathbb{E}_{12}\left[P_{G,\pi}^2\right] = \mathbb{E}_{12}\left[\prod_{\substack{(i,j)\in\pi(E)\\ z_i = z_j}} \bar{p} \prod_{\substack{(i,j)\in\pi(E)\\ z_i \neq z_j}} \bar{q}\right] \asymp \lambda^{|E|}\, \mathbb{E}_{12}\left[\left(\frac{q}{\lambda}\right)^{|E^{\neq}|}\right]. \tag{19}$$

We expect that the leading terms in the variance decomposition $\mathrm{var}_{12}(P_G) = \sum_{\pi,\pi'} \mathrm{cov}_{12}(P_{\pi,G}, P_{\pi',G})$, are those for which $\pi(G) = \pi'(G)$. Writing $|\mathrm{Aut}(G)| \leq |V|^{|V|}$ for the number of automorphisms of G, and recalling (19) and $|E| = r(|V|-2)$, we then expect to have

$$\mathrm{var}_{12}(P_G) \vee \mathrm{var}_{\not{1}2}(P_G) \overset{?}{\lesssim} \sum_{\pi\in\Pi_V} |\mathrm{Aut}(G)|\, \mathrm{var}_{12}(P_{G,\pi}) \;\lesssim\; (n\lambda^r)^{|V|-2}\; \mathbb{E}_{12}\left[(q/\lambda)^{|E^{\neq}|}\right], \tag{20}$$

where $\lesssim$ hides some factors depending on $|E|$ and $|V|$ only.

We recall that $\ell(G)$ denotes the number of distinct communities in $\pi(G)$. Combining (20) with (15) and $\mathbb{P}_{12}[\ell(G) = \ell] \asymp K^{-(|V|-1-\ell)}$, we then get

$$\mathrm{var}_{12}(P_G) \vee \mathrm{var}_{\not{1}2}(P_G) \lesssim \left(\frac{n\lambda^r}{K}\right)^{|V|-2} \sum_{\ell=1}^{|V|-1} K^{\ell-1}\mathbb{E}_{12}\left[K^{-|E^{\neq}|/r}\middle|\ell(G) = \ell\right]. \tag{21}$$

Taking for granted that the upper-bound in (21) is valid up to constants or log factors, comparing (21) and (18), a test based on P_G will fulfill (17), as soon as the sum in (21) remains small. This will occur in particular if the motif G fulfills

$$|E^{\neq}| \geq r(\ell - 1) , \tag{22}$$

for any partition of the nodes of G into ℓ communities, with v_1, v_2 in the same community.

Cliques (after removing the edge (v_1, v_2)) and self-avoiding paths (SAP) between v_1 and v_2, are two families of motifs fulfilling the Condition (22). To prove Corollary 4, we provide a rigorous proof of the heuristic sketched above, and we combine cliques/SAP counting with a Median-of-Means (MoM) post-processing step. The benefit of the MoM post-processing step is that we only need to control the expectation and the variance of the clique/SAP counting — see Lemma 8 and 14 — in order to get good enough concentration inequalities. The main drawback is that we lose a constant factor in the variance, so we cannot expect our analysis to provide upper bounds with sharp constants.

Cliques and SAP counting only cover a small subset of the exponents $\alpha \in (0,1) \cap \mathbb{Q}$. To prove Theorem 3 for any rational $r > 1$:
- we construct in Section 1.2.3 a connected motif G —based on the blow-up of a cycle— fulfilling the Conditions (22) and (16);
- we prove that when $q \asymp n^{-1/r}$ and $\lambda \gtrsim_{\log} \lambda_c$ hold, we can recover the communities in the SBM with an algorithm based on P_G.

1.2.3. CONSTRUCTION OF THE CYCLE BLOW-UP WITH FASTENERS

Consider any rational number $a \in (0,1)$, any positive integer γ. Fix $\kappa \geq 3 \vee (2a^{-1})$ an integer such that $a\kappa\gamma$ is an even integer. In the following, we construct the cycle blow-up with fasteners $G_{\kappa,\gamma,a} = (V, E)$.

Let C_κ be a simple cycle on vertices $[\kappa]$. For each vertex $\omega \in [\kappa]$, define a *layer*

$$L_\omega := \{v_{\omega,t} : t = 1, \dots, \gamma\},$$

consisting of γ vertices that represent the "blow-up" of the cycle node ω. The collection of all such vertices forms the *cycle nodes* of the graph: $V_{\text{cyc}} := \bigcup_{\omega=1}^{\kappa} L_\omega$, with $|V_{\text{cyc}}| = \kappa\gamma$. Between every two consecutive layers L_ω and $L_{\omega+1}$ (with the convention $L_{\kappa+1} = L_1$), we insert a complete bipartite graph: $E_{\text{cyc}} := \bigcup_{\omega=1}^{\kappa} \{(v_{\omega,t}, v_{\omega+1,t'}) : 1 \leq t, t' \leq \gamma\}$, so that every vertex in a layer L_ω is connected to all γ vertices in $L_{\omega-1}$ and all γ vertices in $L_{\omega+1}$. This produces a total of $|E_{\text{cyc}}| = \kappa\gamma^2$ edges and the graph induced by $G_{\text{cyc}} = (V_{\text{cyc}}, E_{\text{cyc}})$ is 2γ-regular. We refer to G_{cyc} as the cycle part of the graph G.

We then introduce the two additional *distinguished* vertices v_1 and v_2, and define the set of nodes $V := V_{\text{cyc}} \cup \{v_1, v_2\}$. We connect v_1, v_2 to V_{cyc} in such a way that $|E| = r(|V| - 2) = (\gamma + a)(|V| - 2)$. Accordingly, we connect v_1 or v_2 to each node in V_{cyc} with frequency a. We denote by $V_{\text{fst}} \subseteq V_{\text{cyc}}$ the cyclic vertices to be connected to v_1 or v_2, which we call the *fastener nodes*. Informally, the set V_{fst} whose cardinality is equal to $a\kappa\gamma$ is defined such that the nodes V_{fst} are as evenly spread as possible. More formally, for each layer $\omega = 1, \dots, \kappa$, we define the integers s_ω by

$$s_0 = 0, \qquad s_\omega = \lfloor a\omega\gamma \rfloor - (s_0 + \dots + s_{\omega-1}) . \tag{23}$$

Note, that $s_\omega \in \{\lfloor a\gamma \rfloor, \lceil a\gamma \rceil\}$. In layer L_ω, exactly s_ω vertices are selected into V_{fst}. By enumerating V_{fst} in the lexicographic order and alternatively assigning these nodes to two sets $V_{\text{fst},1}$ or to $V_{\text{fst},2}$, we partition V_{fst} into two subsets of size $a\kappa\gamma/2$ –which is an integer, as we have assumed $a\kappa\gamma$ to be an even integer.

Finally, we introduce the collection E_{fst} of *fastener edges* by

$$E_{\text{fst}} := \{ (v_1, v) : v \in V_{\text{fst},1} \} \cup \{ (v_2, v) : v \in V_{\text{fst},2} \} \ .$$

The complete edge set of the graph is then $E := E_{\text{cyc}} \cup E_{\text{fst}}$. This construction is illustrated in Figure 2.

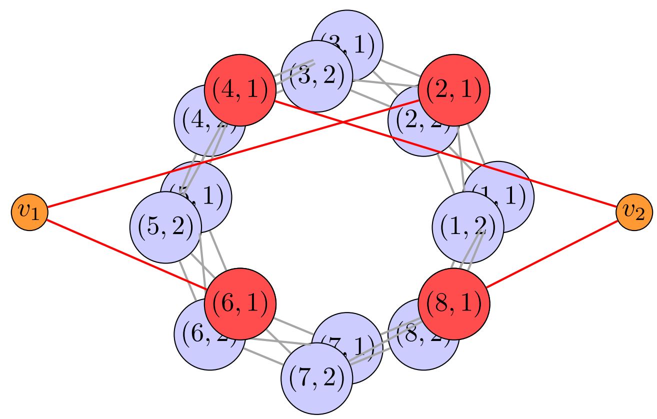

Figure 2: Blow-up graph with fasteners $G_{\kappa,\gamma,a}$, in the case where $\gamma = 2, \kappa = 8, a = 0.25$. The distinguished nodes v_1, v_2 are in orange, the *fastener nodes* (in V_{fst}) are in red, the other cycle nodes are in blue. The *fastener edges* (in E_{fst}) are in red, while the other edges are in gray.

Interestingly, we observe that $|V| = \kappa\gamma + 2$ and $|E| = \kappa\gamma^2 + a\kappa\gamma$, so that the graph $G_{\kappa,\gamma,a}$ satisfies $\frac{|E|}{|V|-2} = \gamma + a$. The following proposition, proved in Appendix H, ensures that the motif $G_{\kappa,\gamma,a}$ fulfills the Condition (22) we are looking for. We recall that $\kappa \geq (3 \vee 2/a)$ is such $a\kappa\gamma$ is an even integer.

Proposition 5 *Fix any positive integer $I \leq |V_{\text{cyc}}|$ and consider any partition of V into $I + 1$ communities such that both v_1 and v_2 are in the same community. Define $E^{\neq} \subset E$ as the set of edges between nodes of distinct communities. Then, we have $|E^{\neq}| \geq I \frac{|E|}{|V|-2}$.*

1.3. Organisation and notation.

In Appendix A, we describe more precisely the statistical setting, and the low-degree lower bound is precisely stated in Appendix A.2. Appendix B gathers the results on community recovery above the threshold λ_c: cliques counting is analyzed in Appendix B.1, self-avoiding paths counting in Appendix B.2, and cycle blow-up counting in Appendix B.3. Proofs are deferred to the last Appendices.

We denote by $\log(\cdot)$ the natural logarithm, and by $\log_n(\cdot)$ the logarithm in base n, i.e. $K = n^{\log_n(K)}$. We write $\mathbf{1}\{A\}$ for the indicator function $\mathbf{1}_A$ of the set A and $[D]$ for the set $\{1, \ldots, D\}$. In the discussion of the results, we use the symbol $a \lesssim_D b$ (respectively $a \lesssim_{\log} b$) to state that a is smaller than b up to a possible polynomial factor in D (resp. up to a poly-log factor), and we use $a \asymp b$ to state that $a \lesssim b \lesssim a$.

Acknowledgments

The work of the three authors has been partially supported by ANR-21-CE23-0035 (ASCAI, ANR). The work of A. Carpentier is partially supported by the Deutsche Forschungsgemeinschaft (DFG)-Project-ID 318763901 - SFB1294 "Data Assimilation", Project A03, by the DFG on the Forschungsgruppe FOR5381 "Mathematical Statistics in the Information Age - Statistical Efficiency and Computational Tractability", Project TP 02 (Project-ID 460867398), and by the DFG on the French-German PRCI ANR-DFG ASCAI CA1488/4-1 "Aktive und Batch-Segmentierung, Clustering und Seriation: Grundlagen der KI" (Project-ID 490860858), by the European Research council (ERC) through the ERC Consolidator Grant SOCE (Nr: 101229569). The work of A. Carpentier and N. Verzelen is also supported by the Université franco-allemande (UFA) through the college doctoral franco-allemand CDFA-02-25 "Statistisches Lernen für komplexe stochastische Prozesse". The work of C. Giraud has been partially supported by ANR-19-CHIA-0021-01 (BiSCottE, ANR).

References

Emmanuel Abbe. Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research*, 18(177):1–86, 2018. URL http://jmlr.org/papers/v18/16-480.html.

Emmanuel Abbe and Colin Sandon. Community Detection in General Stochastic Block Models: Fundamental Limits and Efficient Algorithms for Recovery. In *Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (FOCS)*, FOCS '15, pages 670–688, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8191-8. URL http://dx.doi.org/10.1109/FOCS.2015.47.

Afonso S Bandeira, Jess Banks, Dmitriy Kunisky, Christopher Moore, and Alex Wein. Spectral planting and the hardness of refuting cuts, colorability, and communities in random graphs. In *Conference on Learning Theory*, pages 410–473. PMLR, 2021.

Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K. Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019. URL https://doi.org/10.1137/17M1138236.

Quentin Berthet and Philippe Rigollet. Complexity Theoretic Lower Bounds for Sparse Principal Component Detection. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 1046–1066, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL http://proceedings.mlr.press/v30/Berthet13.html.

Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 1347–1357. IEEE, 2015.

Matthew Brennan and Guy Bresler. Reducibility and statistical-computational gaps from secret leakage. In *Conference on Learning Theory*, pages 648–847. PMLR, 2020.

Matthew Brennan, Guy Bresler, and Wasim Huleihel. Reducibility and computational lower bounds for problems with planted sparse structure. In *Conference On Learning Theory*, pages 48–166. PMLR, 2018.

Matthew Brennan, Guy Bresler, Samuel B Hopkins, Jerry Li, and Tselil Schramm. Statistical query algorithms and low-degree tests are almost equivalent. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 774–774. PMLR, 15–19 Aug 2021.

Alexandra Carpentier, Simone Maria Giancola, Christophe Giraud, and Nicolas Verzelen. Low-degree lower bounds via almost orthonormal bases. *arXiv preprint arXiv:2509.09353*, 2025a.

Alexandra Carpentier, Christophe Giraud, and Nicolas Verzelen. Phase transition for stochastic block model with more than $\sqrt{n}$ communities. *arXiv preprint arXiv:2509.15822*, 2025b.

Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *Journal of Machine Learning Research*, 17(27):1–57, 2016.

Byron Chin, Elchanan Mossel, Youngtak Sohn, and Alexander S. Wein. Stochastic block models with many communities and the Kesten–Stigum bound - extended abstract. In Nika Haghtalab and Ankur Moitra, editors, *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pages 1253–1258. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/chin25a.html.

Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011. URL https://link.aps.org/doi/10.1103/PhysRevE.84.066106.

Jingqiu Ding, Yiding Hua, Lucas Slot, and David Steurer. Low degree conjecture implies sharp computational thresholds in stochastic block model. *arXiv preprint arXiv:2502.15024*, 2025.

Yunzi Ding, Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Subexponential-time algorithms for sparse pca. *Foundations of Computational Mathematics*, 24(3):865–914, 2024.

Bertrand Even, Christophe Giraud, and Nicolas Verzelen. Computation-information gap in high-dimensional clustering. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 1646–1712. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/even24a.html.

Bertrand Even, Christophe Giraud, and Nicolas Verzelen. Computational lower bounds in latent models: clustering, sparse-clustering, biclustering. *arXiv preprint arXiv:2506.13647*, 2025. URL https://arxiv.org/abs/2506.13647.

David Gamarnik. The overlap gap property: A topological barrier to optimizing over random structures. *Proceedings of the National Academy of Sciences*, 118(41):e2108492118, 2021.

Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

Samuel Hopkins. *Statistical inference and the sum of squares method*. PhD thesis, Cornell University, 2018.

Samuel B. Hopkins and David Steurer. Efficient bayesian estimation from few samples: Community detection and related problems. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 379–390, 2017. URL https://ieeexplore.ieee.org/abstract/document/8104074.

Samuel B. Hopkins, Pravesh K. Kothari, Aaron A. Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer. The power of sum-of-squares for detecting hidden structures. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 720–731, Los Alamitos, CA, USA, oct 2017. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/FOCS.2017.72.

Michael Kearns. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)*, 45(6):983–1006, 1998.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

Dmitriy Kunisky. Low coordinate degree algorithms ii: Categorical signals and generalized stochastic block models. *arXiv preprint arXiv:2412.21155*, 2024.

Dmitriy Kunisky, Alexander S Wein, and Afonso S Bandeira. Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio. In *ISAAC Congress (International Society for Analysis, its Applications and Computation)*, pages 1–50. Springer, 2019. URL https://link.springer.com/chapter/10.1007/978-3-030-97127-4_1.

Yuetian Luo and Chao Gao. Computational lower bounds for graphon estimation via low-degree polynomials. *The Annals of Statistics*, 52(5):2318–2348, 2024.

Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 694–703, 2014.

Cristopher Moore. The Computer Science and Physics of Community Detection: Landscapes, Phase Transitions, and Hardness. *CoRR*, abs/1702.00467, 2017. URL http://arxiv.org/abs/1702.00467.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

Tselil Schramm and Alexander S Wein. Computational barriers to estimation from low-degree polynomials. *The Annals of Statistics*, 50(3):1833–1858, 2022. URL https://projecteuclid.

org/journals/annals-of-statistics/volume-50/issue-3/Computational-barriers-to-estimation-from-low-degree-polynomials/10.1214/22-AOS2179.short.

Youngtak Sohn and Alexander S. Wein. Sharp phase transitions in estimation with low-degree polynomials. *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 891–902, 2025. URL https://arxiv.org/abs/2502.14407.

Ludovic Stephan and Yizhe Zhu. Community detection with the bethe-hessian. *arXiv preprint arXiv:2411.02835*, 2024.

Alexander S. Wein. Computational complexity of statistics: New insights from low-degree polynomials. *arXiv preprint arXiv:2506.10748*, 2025. URL https://arxiv.org/abs/2506.10748.

Appendix A. Low-degree lower bound

A.1. Setting

Community recovery. Let $\mathcal{C}^* = \{\mathcal{C}_1^*, \ldots, \mathcal{C}_K^*\}$ be the partition of $\{1, \ldots, n\}$ induced by the community assignment $z_1, \ldots, z_n$, i.e.

$$\mathcal{C}_k^* := \{i \in [n] : z_i = k\}, \quad k = 1, \ldots, K.$$

Let $\widehat{\mathcal{C}}$ be a partition estimating $\mathcal{C}^*$. Exact recovery corresponds to the event $\widehat{\mathcal{C}} = \mathcal{C}^*$. Non-trivial recovery corresponds to a recovery of $\mathcal{C}^*$ better than random guessing. More precisely, let us define the clustering error as

$$err(\widehat{\mathcal{C}}, \mathcal{C}^*) := \min_{\pi \in \mathrm{Perm}(\{1,\ldots,K\})} \frac{1}{2n} \sum_{k=1}^{K} |\mathcal{C}_k^* \triangle \widehat{\mathcal{C}}_{\pi(k)}|,$$

where $\triangle$ is the symmetric difference, and where the permutation π accounts for the non-identifiability of the labels k of the partitions elements $\mathcal{C}_k^*$. A zero clustering error $err(\widehat{\mathcal{C}}, \mathcal{C}^*) = 0$ corresponds to perfect recovery, while non trivial recovery holds when $err(\widehat{\mathcal{C}}, \mathcal{C}^*) = o(1)$.

Membership matrix estimation. Low degree polynomials cannot directly output a partition $\hat{\mathcal{C}}$. For this reason, the usual alternative objective to clustering or community recovery Luo and Gao (2024); Even et al. (2024); Sohn and Wein (2025); Chin et al. (2025); Even et al. (2025), is to estimate the (centered) membership matrix

$$[x_{ij}]_{1 \le i < j \le n} = [\mathbf{1}\{i, j \text{ are in the same community}\} - 1/K]_{1 \le i < j \le n}$$

in L^2-norm. As explained in Even et al. (2024, 2025), this estimation problem can be directly related to the problem of clustering / community recovery.

The distribution of Y being invariant by permutation of the indices, we have

$$\inf_{f_{ij}:\deg(f_{ij})\leq D} \mathbb{E}\left[\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}(f_{ij}(Y)-x_{ij})^2\right] = \inf_{f_{12}:\deg(f_{12})\leq D} \mathbb{E}\left[(f_{12}(Y)-x_{12})^2\right]$$
$$= \mathbb{E}[x_{12}^2] - \sup_{f_{12}:\deg(f_{12})\leq D} \frac{\mathbb{E}[f_{12}(Y)x_{12}]^2}{\mathbb{E}[f_{12}(Y)^2]}.$$

Hence, to prove that degree-D polynomials cannot perform significantly better than degree-0 polynomials, all we need is to prove that the supremum is $o(\mathbb{E}[x_{12}^2]) = o(1/K)$. To simplify the notation, in the remaining of the paper we write $x := x_{12}$ and

$$\mathrm{Corr}_{\leq D}^2 := \sup_{f:\deg(f)\leq D} \frac{\mathbb{E}[f(Y)x]^2}{\mathbb{E}[f(Y)^2]}. \tag{24}$$

Proving that $\mathrm{Corr}_{\leq D}^2 = o(1/K)$ for D on the order of $\log(n)$ provides strong evidence of computational hardness, as it is conjectured to imply that no polynomial-time algorithm in n and K can estimate the membership matrix significantly better than the trivial estimator $\hat{x} = 0$.

A.2. Low-degree lower bound

Our first main result is a low-degree lower bound for the problem of estimating the membership matrix x. We recall that $\bar{q} = q(1-q)$.

Theorem 6 *Let $c \geq 14$, $D \geq 2$, $q \leq 1/2$, $q + 2\lambda \leq 1$, and $K \leq n$. Let $\bar{\lambda} = 2D^{16c}\lambda$. Assume that*

$$\lambda \leq 2D^{16c}\lambda_c, \quad \text{with } \lambda_c \text{ solution to} \quad \sup_{r\geq 1} \frac{n\lambda_c^{2r}}{K\lambda_c^r + K^2\bar{q}^r} = 1\ . \tag{25}$$

Then, $\mathrm{Corr}_{\leq D}^2$ defined by (24) fulfills

$$\mathrm{Corr}_{\leq D}^2 \leq \frac{4}{n}D^{-15c} \quad .$$

As mentioned in the proof sketch Section 1.2, the strategy to prove Theorem (6) is to build an almost-orthonormal basis of degree-D polynomials, and then to essentially “solve” the optimization problem (24). We refer to Appendix D for all the details.

Since, we “solve” almost exactly the optimization problem (24), we can provide some interpretation of the Condition (25) in terms of algorithms. Let us consider a connected graph G made of e edges and v nodes. As sketch in Section 1.2, schematically, the term with $r \approx e/v$ in Condition (25) ensures the failure of an algorithm based on counting — within the observed graph Y — the occurrence of motifs G involving the nodes 1 and 2. For example, the term $r \approx 1$ is related to counting self-avoiding paths, while the term $r \approx m/2$ is related to counting m-cliques. We refer to Section 1.2 for a more detailed discussion of this point.

Remarks:

- Condition (25) with $r = 1$ is the KS condition $\lambda^2 \leq \frac{K}{n}\lambda + \frac{K^2}{n}q$, or equivalently $\lambda^2 \lesssim \frac{K^2}{n^2} + \frac{K^2}{n}q$. The "first condition" namely
$$\lambda \lesssim_D \frac{K}{n},$$
corresponds to the Information-Theoretic barrier for diverging K —see Theorem 1.7 in Chin et al. (2025). All the other terms (including the second one for $r = 1$) correspond to an additional computational barrier.

- When $K \leq \sqrt{n}$, according to Lemma 2, (i) the maximum in (25) is achieved for $r = 1$, which suggests that counting self-avoiding walks is optimal, which is already known from Chin et al. (2025), (ii) our Condition (25) is merely the KS threshold (2) — up to a poly-D factor. Our result is yet not as tight as the lower bound from Chin et al. (2025), which proves failure with sharp constants, while we are loosing a poly-D factor.

- When $K \geq \sqrt{n}$, according to Lemma 2, (i) the supremum in Condition (25) can be achieved for $r > 1$, leading to a phase transition below the KS threshold, (ii) for $q \gtrsim 1/n$ the threshold λ_c fulfills $\lambda_c \asymp q^{1-\log_n(K)}$ as conjectured by Chin et al. (2025) for the sparse case $q \asymp 1/n$.

Appendix B. Community recovery above the Threshold λ_c

B.1. Community recovery above the threshold λ_c with clique-counting

The proof of Theorem 6 suggests that counting cliques of size m might be optimal in some regimes, where the supremum in (25) is achieved for $r \approx m/2$. In this subsection, we confirm this insight by proving that, for $q \asymp n^{\frac{-2}{m+1}}$, we can recover the communities above the threshold $\lambda \gtrsim_{\log} \lambda_c$ by counting m-cliques.

Counting m-cliques. We want to determine if i and j are in the same community, i.e. to estimate $x_{ij} = \mathbf{1}\{z_i = z_j\} - 1/K$. Let us modify the observed graph Y^* by adding an edge between the nodes i and j, if there is no such an edge in the initial graph. Schematically, our strategy consists in counting the number of m-cliques involving the nodes i and j in the modified graph.

More precisely, let $V = \{v_1, \ldots, v_m\}$ be a node set, and $G = (V, E)$ be a clique on V, where we remove the edge (v_1, v_2). For $i < j$, we define $\Pi_{i,j}$ as the set of injections $\pi : V \to \{1 \ldots, n\}$ such that $\pi(v_1) = i$ and $\pi(v_2) = j$, and we set

$$S_{ij} = \sum_{\pi \in \Pi_{i,j}} P_{G,\pi}(Y), \quad \text{with} \quad P_{G,\pi}(Y) = \prod_{(v,v') \in E} Y_{\pi(v),\pi(v')}, \tag{26}$$

where Y is the "centered" adjacency matrix (7). Should $P_{G,\pi}$ be applied to the initial adjacency matrix Y^* instead of the "centered" one Y, the sum S_{ij} would be equal to $(m-2)!$ times the number of m-cliques involving the nodes i and j in the modified graph (where we have enforced an edge between i and j). The time complexity to compute S_{ij} is $O(m^2(en/m)^m)$, so it can be computed in polynomial time, as long as m is considered as a constant.

Our strategy to estimate whether i and j are in the same community is, essentially, to compare S_{ij} to a threshold to be determined. To analyse this strategy, we need to compute some concentration bounds on S_{ij} conditionally on $z_i = z_j$ and $z_i \neq z_j$. A first step in this direction is to compute the conditional means and variances of S_{ij} given these two events. It will be convenient to use the notation $\bar{q} = q(1-q)$,

$$\mathbb{P}_{ij} := \mathbb{P}\left[\cdot | z_i = z_j\right], \quad \text{and} \quad \mathbb{P}_{i\not j} := \mathbb{P}\left[\cdot | z_i \neq z_j\right]. \tag{27}$$

Proposition 7 *Assume that $q \leq 1/2$, $q + 2\lambda \leq 1$, and $3 \leq m \leq K$. Let $i < j$ and let S_{ij} be defined by (26). We have*

$$\mathbb{E}_{ij}\left[S_{ij}\right] = \frac{(n-2)!}{(n-m)!}\left(\frac{\lambda^{\frac{m+1}{2}}}{K}\right)^{m-2}, \tag{28}$$

$$\mathbb{E}_{i\not j}\left[S_{ij}\right] = 0. \tag{29}$$

In addition, if for some $\rho > 0$

$$\frac{(n-2)\lambda^{\frac{m+1}{2}}}{K} \geq \rho(m-2)^2\left(1+\frac{\bar{q}}{\lambda}\right)^{\frac{m+1}{2}} \tag{30}$$

$$\text{and} \qquad \frac{n-2}{K^2}\left(\frac{\lambda^2}{\bar{q}}\right)^{\frac{m+1}{2}} \geq \rho(m-2), \tag{31}$$

then, we have

$$\mathrm{var}_{ij}(S_{ij}) \leq \mathbb{E}_{ij}\left[S_{ij}\right]^2\left[\left(\frac{2}{\rho}\right)^{m-2} + \left((m-2)\left(1+\frac{\bar{q}}{\lambda}\right)^{m+1} + \frac{1}{\rho}\right)^{m-2}\left(1+\frac{\bar{q}}{\lambda}\right)^{1/2}\lambda^{1/2}\right] \tag{32}$$

$$\mathrm{var}_{i\not j}(S_{ij}) \leq \mathbb{E}_{ij}\left[S_{ij}\right]^2\left[\left(\frac{2}{\rho}\right)^{m-2} + \left(\frac{1}{\rho}\right)^{m-2}\right]. \tag{33}$$

Proof [Proof of Proposition 7] The core of the proof is to provide some unconditional upper-bounds on the variance, the proof of which is deferred to Appendix F.

Lemma 8 *Assume that $q \leq 1/2$, $q + 2\lambda \leq 1$, $3 \leq m \leq K$, and set $\bar{p} = \bar{q} + \lambda(1-2q)$. Then, for any $1 \leq i < j \leq n$, we have*

$$\mathbb{E}_{ij}\left[S_{ij}\right] = \frac{(n-2)!}{(n-m)!}\left(\frac{\lambda^{\frac{m+1}{2}}}{K}\right)^{m-2}, \tag{34}$$

$$\mathbb{E}_{i\not j}\left[S_{ij}\right] = 0, \tag{35}$$

$$\mathrm{var}_{ij}(S_{ij}) \leq \frac{(n-2)!(m-2)!}{(n-m)!}\bar{p}^{\frac{(m+1)(m-2)}{2}}\left(\frac{m-2}{K} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2} + \frac{(n-2)!(m-2)!}{(n-m)!}\frac{\bar{p}^{(m+1)(m-2)}}{K^{2m-4}}\left(n-m+\frac{K}{\bar{p}^{\frac{m+1}{2}}}\right)^{m-2}\bar{p}^{1/2} \tag{36}$$

$$\mathrm{var}_{i\not j}(S_{ij}) \leq \frac{(n-2)!(m-2)!}{(n-m)!}\bar{p}^{\frac{(m+1)(m-2)}{2}}\left(\left(\frac{m-2}{K} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2}\left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m-1}{2}} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{(m+1)(m-2)}{2}}\right). \tag{37}$$

Let us upper-bound $\mathrm{var}_{ij}(S_{ij})$ and $\mathrm{var}_{i\not j}(S_{ij})$ under the conditions (30) and (31). Since $\bar{p} \leq \lambda + \bar{q}$, we have

$$\begin{aligned}
\mathrm{var}_{ij}(S_{ij}) &\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \left(\frac{m-2}{n-2}\right)^{m-2} \left[\left(\bar{p}^{\frac{m+1}{2}} \frac{m-2}{K} + \bar{q}^{\frac{m+1}{2}}\right)^{m-2} + \left(\bar{p}^{m+1}\frac{n-m}{K^2} + \frac{\bar{p}^{\frac{m+1}{2}}}{K}\right)^{m-2} \bar{p}^{1/2}\right] \\
&\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \left(\frac{\lambda^{\frac{m+1}{2}}}{K}\right)^{2(m-2)} \left[\left(\frac{2}{\rho}\right)^{m-2} + \left((m-2)\left(1+\frac{\bar{q}}{\lambda}\right)^{m+1} + \frac{1}{\rho}\right)^{m-2} \left(1+\frac{\bar{q}}{\lambda}\right)^{1/2} \lambda^{1/2}\right] \\
&\leq \mathbb{E}_{ij}\left[S_{ij}\right]^2 \left[\left(\frac{2}{\rho}\right)^{m-2} + \left((m-2)\left(1+\frac{\bar{q}}{\lambda}\right)^{m+1} + \frac{1}{\rho}\right)^{m-2} \left(1+\frac{\bar{q}}{\lambda}\right)^{1/2} \lambda^{1/2}\right].
\end{aligned}$$

Similarly, since $\bar{p} \geq \bar{q}$,

$$\begin{aligned}
\mathrm{var}_{i\not j}(S_{ij}) &\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \left(\frac{m-2}{n-2}\right)^{m-2} \left[\left(\bar{p}^{\frac{m+1}{2}} \frac{m-2}{K} + \bar{q}^{\frac{m+1}{2}}\right)^{m-2} \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m-1}{2}} + \bar{q}^{\frac{(m+1)(m-2)}{2}}\right] \\
&\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \left(\frac{\lambda^{\frac{m+1}{2}}}{K}\right)^{2(m-2)} \left[\left(\frac{2}{\rho}\right)^{m-2} \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m-1}{2}} + \left(\frac{1}{\rho}\right)^{m-2}\right] \\
&\leq \mathbb{E}_{ij}\left[S_{ij}\right]^2 \left[\left(\frac{2}{\rho}\right)^{m-2} + \left(\frac{1}{\rho}\right)^{m-2}\right],
\end{aligned}$$

which concludes the proof of Proposition 7. ■

Adding a Median-of-Means post-processing step. In Proposition 7, we control the mean and the variance of the clique count S_{ij}. A concentration inequality based on a second moment Markov inequality would not be tight enough to ensure meaningful results. Indeed, with a second moment inequality, we can ensure a correct result only with a probability approximately $1-\sqrt{\lambda}$, which is worse than the probability $1-K^{-1}$ of correctness of the trivial estimator $\hat{x}^0_{ij} = -K^{-1}$. So we need a better concentration inequality. To avoid computing higher moments, we add a "Median-of-Means post-processing step" to get concentration bounds good enough for our purpose.

Assume that $m \leq n/(24\log(n))$ and $1 \leq i < j \leq n$. Let $\Lambda = 24\log(n)$ and assume for simplicity[2] that $(n-2)/\Lambda$ is a positive integer larger than $m-2$. We define $N = (n-2)/\Lambda + 2$, and we partition the set of nodes $\{1,\ldots,n\}\setminus\{i,j\}$ into L disjoint parts $J^{(1)},\ldots,J^{(\Lambda)}$ with the same cardinality $N-2$. For $\ell = 1,\ldots,\Lambda$, we define $\Pi^{(\ell)}_{i,j}$ has the set of injections $\pi : V \to \{i,j\} \cup J^{(\ell)}$, such that $\pi(v_1) = i$ and $\pi(v_2) = j$, and we introduce the partial clique count

$$S^{(\ell)}_{ij} := \sum_{\pi \in \Pi^{(\ell)}_{i,j}} P_{G,\pi}(Y) \ . \tag{38}$$

2. when $(n-2)/\Lambda$ is not an integer, we partition $\{1,\ldots,n\} \setminus \{i,j\}$ into Λ disjoint sets of cardinality $\lfloor (n-2)/\Lambda \rfloor$ and $\lceil (n-2)/\Lambda \rceil$. The only change in Theorem 9 is that $(n-2)/(24\log(n))$ is replaced by $\lfloor (n-2)/(24\log(n)) \rfloor$.

Finally, we define M_{ij} as a median of the set $\left\{S_{ij}^{(1)}, \ldots, S_{ij}^{(\Lambda)}\right\}$ and we estimate $x_{ij} = \mathbf{1}_{z_i=z_j} - \frac{1}{K}$ by

$$\hat{x}_{ij} = \mathbf{1}\left\{M_{ij} > \frac{(N-2)!}{2(N-m)!}\left(\frac{\lambda^{\frac{m+1}{2}}}{K}\right)^{m-2}\right\} - \frac{1}{K}, \quad \text{with} \quad N = \frac{n-2}{24\log(n)} + 2. \tag{39}$$

We can now state our second main result.

Theorem 9 *Assume that $q \le 1/2$, $N = 2+(n-2)/(24\log(n))$ is an integer, and $3 \le m \le K \wedge N$. When*

$$\frac{(n-2)\lambda^{\frac{m+1}{2}}}{K\log(n)} \ge 48(32)^{\frac{1}{m-2}}(m-2)^2\left(1+\frac{\bar{q}}{\lambda}\right)^{\frac{m+1}{2}} \tag{40}$$

$$\frac{n-2}{K^2\log(n)}\left(\frac{\lambda^2}{\bar{q}}\right)^{\frac{m+1}{2}} \ge 48(32)^{\frac{1}{m-2}}(m-2) \tag{41}$$

$$\lambda \le 32^{-2}\left(1+\frac{\bar{q}}{\lambda}\right)^{-2(m+1)(m-2)-1}(2m-4)^{-(2m-4)}, \tag{42}$$

we have for $\hat{x}_{ij}$ defined by (39)

$$\mathbb{P}\left(\hat{x}_{ij} = x_{ij}\right) \ge 1 - n^{-3}.$$

Before proving Theorem 9, let us comment this result. We recall that m is a fixed integer. Since $\bar{q} \le \lambda$, the Condition (42) merely requires λ to be smaller than the constant

$$\eta_m := 2^{-11}\left(2^{(m+2)}(m-2)\right)^{-(2m-4)}, \tag{43}$$

that depends only on m. As for Conditions (40) and (41), they roughly correspond to the opposite of Condition (70) (which is another formulation of Condition (25)) in Theorem 21 for $r = (m+1)/2$. Let us relate these two conditions to the threshold λ_c. First, in light of Theorem 6 and Conditions (40)-(41), we emphasize that m-cliques counting can only be optimal if the maximum in Condition (25) is achieved for $r = (m+1)/2$. According to Lemma 2, this happens when $q \asymp n^{-\frac{2}{m+1}}$. The next (immediate) corollary of Theorem 9 confirms that m-cliques counting is successful above the threshold λ_c in this density regime. We omit the proof, which is a straightforward check that Conditions (40) and (41) hold under (44).

Corollary 10 *Assume that $N = 2 + (n-2)/(24\log(n))$ is an integer, $m \in \{3, 4, \ldots, K \wedge N\}$, $\bar{q} = n^{\frac{-2}{m+1}}$, and $\lambda \le \eta_m$, with η_m defined in* (43)*. When*

$$\lambda \ge w_m \log(n)^{\frac{2}{m+1}} q^{1-\log_n(K)}, \quad \textit{with } w_m = 2\left(96(m-2)^2(32^{\frac{1}{m-2}})\right)^{\frac{2}{m+1}}, \tag{44}$$

the estimator $\hat{x}_{ij}$ defined by (39) *fulfills*

$$\mathbb{P}\left(\hat{x}_{ij} = x_{ij}\right) \ge 1 - n^{-3}.$$

We now turn to the proof of Theorem 9.

Proof [Proof of Theorem 9] Without loss of generality, we focus on the case $(i,j) = (1,2)$ to reduce the number of indetermined indices. We recall that $L = 24\log(n)$ and $N = (n-2)/\Lambda + 2$. We observe that the statistic $S_{12}^{(\ell)}$ is simply the statistic (26) applied to the graph Y restricted to the node set $\{1,2\} \cup J^{(\ell)}$, whose cardinality is N. Since the graph Y restricted to this node set follows a SBM with the same parameters K, q, λ, the results proven for S_{12} hold for $S_{12}^{(\ell)}$ with n replaced by N. From Proposition 7, when (40), (41), and(42) hold, we have for all $\ell = 1, \ldots, \Lambda$,

$$\mathbb{E}_{\not{1}2}\left[S_{12}^{(\ell)}\right] = 0, \qquad \mathbb{E}_{12}\left[S_{12}^{(\ell)}\right] = E_{12}[N] := \frac{(N-2)!}{(N-m)!}\left(\frac{\lambda^{\frac{m+1}{2}}}{K}\right)^{m-2}, \tag{45}$$

and

$$\mathrm{var}_{12}(S_{12}^{(\ell)}) \vee \mathrm{var}_{\not{1}2}(S_{12}^{(\ell)}) \leq V[N] := \frac{E_{12}[N]^2}{16}.$$

As a consequence, we have from Markov inequality

$$\mathbb{P}_{\not{1}2}\left[S_{12}^{(\ell)} \geq 2\sqrt{V[N]}\right] \leq \frac{1}{4}, \quad \text{and} \quad \mathbb{P}_{12}\left[S_{12}^{(\ell)} \leq E_{12}[N] - 2\sqrt{V[N]}\right] \leq \frac{1}{4}\,. \tag{46}$$

A key feature of the partial clique counts $S_{12}^{(1)}, \ldots, S_{12}^{(L)}$ is that they are independent both under $\mathbb{P}_{12}$ and $\mathbb{P}_{\not{1}2}$. Indeed, conditionally on $z_1 = z_2$ (resp. $z_1 \neq z_2$), the random variables $W_1, \ldots, W_\Lambda$ defined by $W_\ell = \{Y_{1j}, Y_{2j}\}_{j \in J^{(\ell)}} \cup \{Y_{ij}\}_{i<j \in J^{(\ell)}}$ are independent, and $S_{12}^{(\ell)}$ is $\sigma(W_\ell)$-measurable. Hence, the number of $S_{12}^{(\ell)}$ smaller than $E_{12}[N] - 2\sqrt{V[N]}$ under $\mathbb{P}_{12}$ (resp. exceeding $2\sqrt{V[N]}$ under $\mathbb{P}_{\not{1}2}$) is stochastically dominated by a binomial distribution with parameter $(\Lambda, 1/4)$, and so M_{12} being a median of the $\left\{S_{12}^{(1)}, \ldots, S_{12}^{(\Lambda)}\right\}$, we get

$$\mathbb{P}_{\not{1}2}\left[M_{12} \geq 2\sqrt{V[N]}\right] \leq e^{-\Lambda/8} = \frac{1}{n^3}, \quad \text{and} \quad \mathbb{P}_{12}\left[M_{12} \leq E_{12}[N] - 2\sqrt{V[N]}\right] \leq e^{-\Lambda/8} = \frac{1}{n^3}\,. \tag{47}$$

Since $V[N] = (E_{12}[N]/4)^2$, we conclude that

$$\mathbb{P}_{\not{1}2}\left[M_{12} < \frac{E_{12}[N]}{2}\right] \geq 1 - \frac{1}{n^3}, \quad \text{and} \quad \mathbb{P}_{12}\left[M_{12} > \frac{E_{12}[N]}{2}\right] \geq 1 - \frac{1}{n^3}\,. \tag{48}$$

The estimator

$$\hat{x}_{12} = \mathbf{1}\left\{M_{12} > \frac{E_{12}[N]}{2}\right\} - \frac{1}{K},$$

with $E_{12}[N]$ defined in (45), then fulfills $\hat{x}_{12} = x_{12}$ with probability at least $1 - n^{-3}$. ■

Remark: for the simplicity of the exposition, we have assumed that $(n-2)/(24\log(n))$ is an integer. In general, we can set $N = \left\lfloor \frac{n-2}{24\log(n)} \right\rfloor + 2$ and splits the set $\{1, \ldots, n\} \setminus \{i,j\}$ into $\Lambda = 24\log(n)$ subsets with cardinality N or $N+1$. Then, all the results hold with $\frac{n-2}{24\log(n)}$ replaced by $N-2$.

Successful recovery. Theorem 9 readily gives a non-trivial bound for the estimation of x_{ij} by m-cliques couting. We recall that $MMSE_0 = \mathbb{E}[x_{ij}^2] = K^{-1}(1-K^{-1})$. Corollary 10 ensures that $\mathbb{E}\left[(\hat{x}_{ij}-x_{ij})^2\right] \leq n^{-3} = o(K^{-1})$, when $\bar{q} \asymp n^{\frac{-2}{m+1}}$ and (44) holds.

Once we have estimated x_{ij}, we still need a last step to output communities. Let us define the matrix $\widehat{X} \in \{0,1\}^{n\times n}$ by $\widehat{X}_{ii} = 0$ for $i = 1,\ldots,n$, and

$$\widehat{X}_{ij} = \widehat{X}_{ji} = \hat{x}_{ij} + \frac{1}{K}, \quad \text{for } i < j.$$

Seeing $\widehat{X}$ as the adjacency matrix of a graph $\bar{X}$, we estimate the communities by the connected components of $\bar{X}$. The overall complexity is $O(n^2m^2(en/m)^m)$. When $\bar{q} \asymp n^{\frac{-2}{m+1}}$ and (44) holds, Corollary 10 ensures that we recover the communities with probability at least $1-1/n$.

B.2. Counting self-avoiding paths

Our analysis suggests that counting self-avoiding paths of length $m-1$ should be optimal for the density $q \asymp n^{-\frac{m-2}{m-1}}$. In this subsection, we confirm this prediction. Our analysis follows the same lines as the one for m-cliques counting. We expose it more succinctly.

Let $V = \{v_1,\ldots,v_m\}$ be a node set and $G = (V,E)$ be the self-avoiding path $v_1, v_3, v_4, \ldots, v_m, v_2$ on V. Define

$$T_{ij} = \sum_{\pi\in\Pi_{i,j}} P_{G,\pi}(Y), \quad \text{with} \quad P_{G,\pi}(Y) = \prod_{(v,v')\in E} Y_{\pi(v),\pi(v')} \quad . \tag{49}$$

Should $P_{G,\pi}$ be applied to the initial adjacency matrix Y^* instead of the "centered" one Y, the sum T_{ij} would be equal to the number of self-avoiding path of length $m-1$ with end nodes i and j. The time complexity to compute T_{ij} is $O(m(en/m)^m)$, so it can be computed in polynomial time, as long as m is considered as a constant.

As for the clique counting, our strategy to estimate whether i and j are in the same community is, essentially, to compare T_{ij} to a threshold to be determined. To analyse it, we first compute the conditional means and variances of T_{ij} given the two events $z_i = z_j$ and $z_i \neq z_j$. We remind the Definition (27) of $\mathbb{P}_{ij}$ and $\mathbb{P}_{i\not j}$, as well as the notation $\bar{q} = q(1-q)$.

Proposition 11 *Assume that $q \leq 1/2$, $q+2\lambda \leq 1$, and $3 \leq m \leq (K \wedge n/2)$. Let $i<j$ and let T_{ij} be defined by (49). We have*

$$\mathbb{E}_{ij}\left[T_{ij}\right] = \frac{(n-2)!}{(n-m)!}\cdot\frac{\lambda^{m-1}}{K^{m-2}}\ , \qquad \mathbb{E}_{i\not j}\left[T_{ij}\right] = 0\ .$$

In addition, if for some $\rho > 1$, we have $n \geq 6\rho K m^3$,

$$\begin{aligned} &\lambda \geq 8\left[1+\frac{\bar{q}}{\lambda}\right]\left[2\rho\frac{m^3K}{n} + e^3m^2\rho^{1/(m-1)}\left(\frac{K}{n}\right)^{1-1/(m-1)}\right] \\ &\frac{\lambda^2}{\bar{q}} \geq 4\rho\frac{Km^3}{n} + 4\rho^{1/(m-1)}e^3m^2\left(\frac{K^2}{n}\right)^{1-1/(m-1)}, \end{aligned} \tag{50}$$

then we have $\mathrm{var}_{ij}(T_{ij}) \vee \mathrm{var}_{i\not j}(T_{ij}) \leq 6\rho^{-1}\mathbb{E}_{ij}^2\left[T_{ij}\right]$.

As for the clique counting problem, relying on T_{ij} alone together with a Markov type bound is not sufficient. For this reason, we rely again on a Median-of-Means post-processing step and we use the same notation as in that section. As in the previous subsection, fix $\Lambda = 24\log(n)$ and we assume for simplicity that $(n-2)/\Lambda$ is an integer. Recall the definition of $N = (n-2)/\Lambda + 2$ and let $J^{(1)}, \ldots, J^{(\Lambda)}$ be a partition of $\{1, \ldots, n\} \setminus \{i, j\}$ into L disjoint parts. Then, as in (38) for clique counts, we introduce the partial self-avoiding path count $T_{ij}^{(\ell)} := \sum_{\pi \in \Pi_{i,j}^{(\ell)}} P_{G,\pi}(Y)$, and we define M_{ij} as a median of the set $\left\{T_{ij}^{(1)}, \ldots, T_{ij}^{(\Lambda)}\right\}$. We estimate $x_{ij} = \mathbf{1}_{z_i = z_j} - \frac{1}{K}$ by

$$\hat{x}_{ij} = \mathbf{1}\left\{M_{ij} > \frac{(N-2)!}{2(N-m)!}\frac{\lambda^{m-1}}{K^{m-2}}\right\} - \frac{1}{K}, \quad \text{where} \quad N = \frac{n-2}{24\log(n)} + 2. \tag{51}$$

We can now state our second main result.

Theorem 12 *Assume that $q \leq 1/2$, $q + 2\lambda \leq 1$, $N = 2 + (n-2)/(24\log(n))$ is an integer, and $3 \leq m \leq K \wedge N/2$. When $n \geq c_0 K m^3 \log(n)$,*

$$\frac{n\lambda^{1+1/(m-2)}}{K\log(n)} \geq c_1 m^3 \left(1 + \frac{\bar{q}}{\lambda}\right)^{1+1/(m-2)} ;$$
$$\frac{n}{K^2\log(n)}\left(\frac{\lambda^2}{\bar{q}}\right)^{1+1/(m-2)} \geq c_2 m^2 ,$$

we have for $\hat{x}_{ij}$ defined by (51) that $\mathbb{P}(\hat{x}_{ij} = x_{ij}) \geq 1 - n^{-3}$.

Proof [Proof of Theorem 12] The proof follows exactly the same lines as that of Theorem 9 to the difference that we build upon Proposition 11 instead of Proposition 7. We skip the details. ■

Corollary 13 *Let $\bar{q} = n^{-\frac{m-2}{m-1}}$ for some $m \in \{3, 4, \ldots, K \wedge N/2\}$. When $n \geq c_0 K m^3 \log(n)$, $\lambda \leq 1 - 2q$, and*

$$\lambda \geq w'_m \log(n)^{\frac{m-2}{m-1}} \bar{q}^{1-\log_n(K)}, \tag{52}$$

with w'_m depending only on m, then the estimator $\hat{x}$ based on the number of self-avoiding path recovers the communities with probability at least $1 - 1/n$.

Proof [Proof of Corollary 13] Define $r = (m-1)/(m-2)$. In this corollary, we consider the regime $\bar{q} = n^{-1/r}$. It follows from Theorem 12 that $\hat{x}$ recovers the communities with probability at least $1 - 1/n$ as long as $q + 2\lambda \leq 1$ and

$$\lambda \geq \eta'_m \log(n)^{1/r}\left(\frac{K}{n}\right)^{1/r} = \eta'_m \log(n)^{1/r} \bar{q}^{1-\log_n(K)} ,$$

where η'_m depends only on m. The result follows. ■

Proof [Proof of Proposition 11] Similarly as for the proof of Proposition 7, the proof of Proposition 11 is a direct consequence of the following lemma proved in Appendix G.

Lemma 14 *Assume that $q \leq 1/2$, $q + 2\lambda \leq 1$, $3 \leq m \leq K$, and set $\bar{p} = \bar{q} + \lambda(1-2q)$. Then, for any $1 \leq i < j \leq n$, we have*

$$\mathbb{E}_{ij}\left[T_{ij}\right] = \frac{(n-2)!}{(n-m)!}\left(\frac{\lambda^{m-1}}{K^{m-2}}\right), \tag{53}$$

$$\mathbb{E}_{i\not{j}}\left[T_{ij}\right] = 0, \tag{54}$$

$$\mathrm{var}_{ij}(T_{ij}) \leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \frac{\lambda^{2(m-1)}}{K^{2m-4}} m\left[\frac{8Km^2\bar{p}}{(n-m)\lambda^2} + \frac{2m^2K\bar{q}}{(n-m)\lambda^2} + 3m^2\frac{K}{n-m} + 3\left(\frac{m^2K}{n-m}\right)^{m-2} + \frac{2nm^2}{K}\left(\frac{4Km^2\bar{p}}{(n-m)\lambda^2}\right)^{m-1} + \frac{nm^2}{K^2}\left(\frac{2m^2K^2\bar{q}}{(n-m)\lambda^2}\right)^{m-1}\right]. \tag{55}$$

$$\mathrm{var}_{i\not{j}}(T_{ij}) \leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \frac{\lambda^{2(m-1)}}{K^{2m-4}} m\left[\frac{4m^2\bar{p}}{(n-m)\lambda^2} + \frac{2m^2K\bar{q}}{(n-m)\lambda^2} + 2m^2\frac{K}{n-m} + 2\left(\frac{m^2K}{n-m}\right)^{m-2} + \frac{nm^2}{K}\left(\frac{4Km^2\bar{p}}{(n-m)\lambda^2}\right)^{m-1} + \frac{nm^2}{K^2}\left(\frac{2m^2K^2\bar{q}}{(n-m)\lambda^2}\right)^{m-1}\right]. \tag{56}$$

■

B.3. Counting Blow-up Motifs

Consider a cycle blow-up with fasteners $G = G_{\kappa,\gamma,a}$, as defined in Section 1.2.3. For $i < j$, we remind that $\Pi_{i,j}$ is the set of injections $\pi : V \to \{1\ldots, n\}$ such that $\pi(v_1) = i$ and $\pi(v_2) = j$, and we set

$$R_{ij} = \sum_{\pi \in \Pi_{i,j}} P_{G,\pi}(Y), \quad \text{with} \quad P_{G,\pi}(Y) = \prod_{(v,v')\in E} Y_{\pi(v),\pi(v')}, \tag{57}$$

where Y is the "centered" adjacency matrix (7).

In the following proposition, we control the mean and the variance of R_{ij} both when $z_i = z_j$ and when $z_i \neq z_j$. We recall that the conditional probabilities $\mathbb{P}_{ij}$ and $\mathbb{P}_{i\not{j}}$ are defined in (27).

Proposition 15 *Assume that $q \leq 1/4$, $q + 2\lambda \leq 1$, and $n \geq 2\kappa\gamma + 4$. We also assume that $a\kappa\gamma$ is an even integer and that $\kappa \geq 3 \vee 2/a$. Let $i < j$ and let R_{ij} be defined by (57). We have*

$$\mathbb{E}_{ij}\left[R_{ij}\right] = \frac{(n-2)!}{(n-\kappa\gamma-2)!}\left(\frac{\lambda^{\gamma+a}}{K}\right)^{\kappa\gamma}, \qquad \mathbb{E}_{i\not{j}}\left[R_{ij}\right] = 0\ . \tag{58}$$

In addition, if for some $\rho > 1$,

$$\left(\frac{\lambda^2}{2\bar{q}}\right)^{\gamma+a} \geq \frac{2K^2(\kappa\gamma)^5\rho}{n}\ ; \qquad \lambda^{\gamma+a} \geq \frac{2K(\kappa\gamma)^5\rho}{n} \tag{59}$$

then, we have

$$\mathrm{var}_{ij}(R_{ij}) \bigvee \mathrm{var}_{i\not{j}}(R_{ij}) \leq \frac{4}{\rho}\cdot \mathbb{E}^2_{ij}\left[R_{ij}\right]\ . \tag{60}$$

As for the clique counting or self-avoiding path counting, relying on R_{ij} alone together with a Markov type bound is not sufficient. For this reason, we add again a Median-of-Means post-processing and we use the same notation as before. In particular, we fix $\Lambda = 24\log(n)$ and we assume for simplicity that $(n-2)/\Lambda$ is an integer. Recall the definition of $N = (n-2)/\Lambda + 2$ and let $J^{(1)}, \ldots, J^{(\Lambda)}$ be a partition of $[n] \setminus \{i, j\}$ into Λ disjoint parts. For $\ell = 1, \ldots, \Lambda$, we define $\Pi_{i,j}^{(\ell)}$ has the set of injections $\pi : V \to \{i, j\} \cup J^{(\ell)}$, such that $\pi(v_1) = i$ and $\pi(v_2) = j$. Then, we introduce the blow-up count $R_{ij}^{(\ell)} := \sum_{\pi \in \Pi_{i,j}^{(\ell)}} P_{G,\pi}(Y)$, and we define M_{ij} as a median of the set $\left\{R_{ij}^{(1)}, \ldots, R_{ij}^{(\Lambda)}\right\}$. We estimate $x_{ij} = \mathbf{1}_{z_i = z_j} - \frac{1}{K}$ by

$$\hat{x}_{ij} = \mathbf{1}\left\{M_{ij} > \frac{(N-2)!}{2(N-\kappa\gamma-2)!}\left(\frac{\lambda^{\gamma+a}}{K}\right)^{\kappa\gamma}\right\} - \frac{1}{K}, \quad \text{where} \quad N = \frac{n-2}{24\log(n)} + 2. \tag{61}$$

We can now state our main result.

Theorem 16 *There exists a numerical constant c such that the following holds. Assume that $q \leq 1/4$, $q + 2\lambda \leq 1$, that $N := 2 + (n-2)/(24\log(n))$ is an integer, and $N \geq 2\kappa\gamma + 4$, and that $\kappa\gamma a$ is an an even integer. Provided that*

$$\left(\frac{\lambda^2}{2\bar{q}}\right)^{\gamma+a} \geq c\frac{K^2(\kappa\gamma)^5\log(n)}{n} \; ; \qquad \lambda^{\gamma+a} \geq c\frac{K(\kappa\gamma)^5\log(n)}{n} \;\; ,$$

we have for $\hat{x}_{ij}$ defined by (61) that $\mathbb{P}(\hat{x}_{ij} = x_{ij}) \geq 1 - n^{-3}$.

Proof [Proof of Theorem 16] The proof follows exactly the same lines as that of Theorem 9 to the difference that we build upon Proposition 15 instead of Proposition 7. We skip the details. ■

Corollary 17 *Let $\bar{q} = n^{-1/r}$ for some $r = \gamma + \theta/\beta$ where γ, θ, and β are positive integers with $\theta < \beta$. Consider the blow-up graph with fasteners $G_{\kappa,\gamma,a}$ with $\kappa = 2\beta\gamma$ and $a = \theta/\beta$. When $n \geq c_0\beta\gamma^2\log(n)$, $\lambda \leq 1 - 2q$, and*

$$\lambda \geq w_r' \log^{1/r}(n)\, \bar{q}^{1-\log_n(K)}, \tag{62}$$

with w_r' depending only on r, then the estimator $\hat{x}$ based on the number of blow-up motifs recovers the communities with probability at least $1 - 1/n$.

If $q = n^{-1/r}$, with r a rational number, recovering the communities above the threshold (4) is feasible with a polynomial of fixed degree, only depending on r. When $r > 1$ is not a rational number, we can still count blow-up motifs for some rational number $\bar{r}$ close to r. Indeed, consider any $\epsilon < 1$. There exists an integer $\overline{\beta} \leq 2/[\epsilon r^2] \vee 1$ and a rational number $\overline{r} = \overline{\gamma} + \overline{\alpha}/\overline{\beta}$ such that $|\overline{r} - r| \leq \epsilon r^2/2$. Here, we have $\bar{\gamma} = \lfloor r \rfloor$. Choosing $\overline{\kappa} = 2\overline{\beta}\overline{\gamma}$, we deduce from Theorem 16, that the estimator $\hat{x}$ based on counting occurrence of blow-up motifs $G_{\bar{\kappa},\bar{\gamma},\bar{\alpha}/\bar{\beta}}$ recovers the communities with probability at least $1 - 1/n$ as long as

$$\lambda \geq w_r'' \epsilon^{-5} \log^{1/r}(n)\, \bar{q}^{1-\log_n(K)} n^{\epsilon} \; .$$

The corresponding polynomial has a degree of the order $\epsilon^{-1}r$. In particular, if we take $\epsilon = \log(\log(n))/\log(n)$, we establish that, for any irrational r, it is possible to recover the communities above the threshold (4) with a polynomial of degree $O(\log(n)/\log(\log(n)))$. Whether this polynomial can be computed (or well-approximated) in polynomial-time remains an open question.

Appendix C. Construction of the almost orthonormal basis and preliminary results

C.1. Construction of the permutation-invariant basis

In this section, we closely follow the construction of the polynomial basis in Section 3 of Carpentier et al. (2025a) up to a few (but important) changes.

Definition of invariant polynomials. Given a permutation $\sigma : [n] \mapsto [n]$, we define the matrix Y_σ by $(Y_\sigma)_{ij} = Y_{\sigma(i)\sigma(j)}$. A function f is said to invariant by permutation of $[n]$ up to 1 and 2, if for any permutation $\sigma : [n] \mapsto [n]$ such that $\sigma(1) = 1$ and $\sigma(2) = 2$, we have $f(Y) = f(Y_\sigma)$.

Lemma 18 *Fix any any degree $D > 0$. Then, the minimum low-degree risk $\mathrm{Corr}_{\leq D}$ is achieved by a function f that is invariant by permutation up to individuals 1 and 2.*

Proof This result is established is the proof of Lemma 3.5 in Carpentier et al. (2025a). It is a consequence of the permutation invariance of the distribution $\mathbb{P}$. ■

As a consequence, we need to build a suitable basis of invariant polynomials. We follow the same approach as in Section 3.2 of Carpentier et al. (2025a). In what follows, we consider simple undirected graphs $G = (V, E)$ where $V = \{v_1, \ldots v_r\}$ is the set of nodes and where E is the set of edges.

Let $G^{(1)} = (V^{(1)}, E^{(1)})$ and $G^{(2)} = (V^{(2)}, E^{(2)})$ be two graphs. We say $G^{(1)}$ and $G^{(2)}$ are equivalent if there exist a bijection $\sigma : V^{(1)} \mapsto V^{(2)}$ such that that $\sigma(v_1^{(1)}) = v_1^{(2)}$, $\sigma(v_2^{(1)}) = v_2^{(2)}$, and σ preserves the edges. In other words, the graphs $G^{(1)}$ and $G^{(2)}$ are isomorphic with the additional constraint that the corresponding bijection maps the two first nodes.

Definition 19 (Collection $\mathcal{G}_{\leq D}$) *Let $\mathcal{G}_{\leq D}$ be any maximum collection of graphs $G = (V, E)$ such that (i) $|V| \geq 2$, (ii) G does not contain any isolated node to the possible exceptions of v_1, v_2, (iii) $1 \leq |E| \leq D$, and (iv) no graphs in $\mathcal{G}_{\leq D}$ are equivalent.*

The collection $\mathcal{G}_{\leq D}$ corresponds to the collections of equivalence classes of all graphs with at most D edges and at least 2 nodes, and without isolated nodes (except maybe the first two nodes), if we keep the first two nodes fixed. Henceforth, we refer to $\mathcal{G}_{\leq D}$ as the collection of *templates*. In fact, $\mathcal{G}_{\leq D}$ corresponds to $\mathcal{G}^{(1,2)}_{\leq D}$ in Section 3.2 Carpentier et al. (2025a)– here we drop the exponent $(1, 2)$ in the notation of the templates and of the polynomials because we only consider a basis for estimation.

Consider a template $G = (V, E) \in \mathcal{G}_{\leq D}$. We define Π_V the set of injective mappings from $V \to [n]$ that satisfy $\pi(v_1) = 1, \pi(v_2) = 2$. An element $\pi \in \Pi_V$ corresponds to a labeling of the generic nodes in V by elements in $[n]$. For $\pi \in \Pi_V$, we define the polynomials

$$P_{G,\pi}(Y) = \prod_{(i,j)\in E} Y_{\pi(i),\pi(j)} \quad \text{and} \qquad P_G = \sum_{\pi\in\Pi_V} P_{G,\pi}. \tag{63}$$

For short, we sometimes write P_G for $P_G(Y)$ when there is no ambiguity. For the invariant polynomials P_G, we say that G is the **template** (graph) that indexes the polynomial.

Consider a template $G \in \mathcal{G}_{\leq D}$ with c connected components $(G_1, G_2, \ldots, G_c)$ that contain at least one edge. To improve the orthogonality of the family $(P_G)_{\mathcal{G}_{\leq D}}$, we apply a correction as in

Carpentier et al. (2025a)

$$\overline{P}_G := \sum_{\pi\in\Pi_V} \overline{P}_{G,\pi}\ ; \quad \overline{P}_{G,\pi} := \prod_{l=1}^{c} [P_{G_l,\pi} - \mathbb{E}[P_{G_l,\pi}]]\ .$$

Define $\mathrm{Aut}(G)$ is group of automorphisms of the graph G that let v_1 and v_2 fixed. Then, given $\pi \in \Pi_V$, we will normalize $\bar{P}_G$ with the variance proxy

$$\mathbb{V}(G) = \frac{(n-2)!}{(n-|V|)!}|\mathrm{Aut}(G)|\mathbb{E}[P^2_{G,\pi}]\ , \tag{64}$$

where, by permutation invariance, $\mathbb{E}[P^2_{G,\pi}]$ does not depend on the specific choice of π. Importantly, the definition of $\mathbb{V}(G)$ in (64) is the only difference with the original construction in Sect.3.5 of Carpentier et al. (2025a). In the latter work, we used a smaller variance proxy which turns out to be a lose lower bound of $\mathbb{E}[\overline{P}^2_G]$. The rationale with this new choice (64) of $\mathbb{V}(G)$ is that $\mathbb{E}[\overline{P}^2_G]$ turns out to be of the same order as $\mathbb{E}[P^2_G] = \sum_{\pi_1\pi_2}\mathbb{E}[P_{G,\pi_1}P_{G,\pi_2}]$. The largest $\mathbb{E}[P_{G,\pi_1}P_{G,\pi_2}]$ are achieved for labelings (π_1,π_2) such that $P_{G,\pi_1} = P_{G,\pi_2}$. As there are

$$|\Pi_V||\mathrm{Aut}(G)| = \frac{(n-2)!}{(n-|V|)!}|\mathrm{Aut}(G)| \tag{65}$$

such labelings, we arrive at (64).

Finally, we introduce the polynomial Ψ_G by $\Psi_G := \frac{\overline{P}_G}{\sqrt{\mathbb{V}(G)}}$. Since the $(1, (\Psi_G)_{G\in\mathcal{G}_{\leq D}})$ is a basis of permutation-invariant (up to 1 and 2) polynomials –see Carpentier et al. (2025a) for details. We readily deduce from Lemma 18 the following result.

Lemma 20 *We have*

$$\mathrm{Corr}^2_{\leq D} = \sup_{\alpha_\emptyset,(\alpha_G)_{G\in\mathcal{G}_{\leq D}}} \frac{\mathbb{E}\left[x\left(\alpha_\emptyset + \sum_{G\in\mathcal{G}_{\leq D}}\alpha_G\Psi_G\right)\right]^2}{\mathbb{E}\left[\left[\alpha_\emptyset + \sum_{G\in\mathcal{G}_{\leq D}}\alpha_G\Psi_G\right]^2\right]}\ .$$

C.2. Some central notation and definition

The crux of the proof is to establish that, in relevant regimes, the basis $(1,\Psi_G)$ is almost orthogonal. For that purpose, we need to introduce some notation. Those are similar to those in Section 5 in Carpentier et al. (2025a).

Labeled graph. For a template $G = (V, E)$ and a labeling $\pi \in \Pi_V$, we define the labeled graph $\pi(G)$ as the graph with node set $\{\pi(v) : v \in V\}$ and edge set $\{(\pi(v), \pi(v')) : (v, v') \in E\}$.

Matching of nodes. Consider two templates $G^{(1)} = (V^{(1)}, E^{(1)})$ and $G^{(2)} = (V^{(2)}, E^{(2)})$. Given labelings $\pi^{(1)}$ and $\pi^{(2)}$, we say that two nodes $v^{(1)}$ and $v^{(2)}$ are matched if $\pi^{(1)}(v^{(1)}) = \pi^{(2)}(v^{(2)})$. More generally, a matching $\mathbf{M}$ stands for a set of pairs of nodes $(v^{(1)}, v^{(2)}) \in V^{(1)} \times V^{(2)}$ where no node in $V^{(1)}$ or $V^{(2)}$ appears twice. We denote $\mathcal{M}$ for the collection of all possible node matchings. For $\mathbf{M} \in \mathcal{M}$, we define the collection of labelings that are compatible with $\mathbf{M}$ by

$$\Pi(\mathbf{M}) = \Big\{\pi^{(1)} \in \Pi_{V^{(1)}}, \pi^{(2)} \in \Pi_{V^{(2)}} : \forall (v^{(1)}, v^{(2)}) \in V^{(1)} \times V^{(2)},$$
$$\{\pi^{(1)}(v^{(1)}) = \pi^{(2)}(v^{(2)})\} \iff \{(v^{(1)}, v^{(2)}) \in \mathbf{M}\}\Big\}\ .$$

Importantly, as $\mathbb{P}$ is permutation invariant, $\mathbb{E}[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}]$ is the same for all $(\pi^{(1)},\pi^{(2)})$ in $\Pi(\mathbf{M})$. Given a matching $\mathbf{M}$, we write that two edges $e\in E^{(1)}$ and $e'\in E^{(2)}$ are matched if the corresponding incident nodes are matched.

Merged graph $G_\cup$, intersection graph $G_\cap$, and symmetric difference graph G_Δ. Consider two templates $G^{(1)}$ and $G^{(2)}\in\mathcal{G}_{\leq D}$ and two labelings $\pi^{(1)}$ and $\pi^{(2)}$. Then, the merged graph $G_\cup=(V_\cup,E_\cup)$ is defined as the union of $\pi^{(1)}(G^{(1)})$ and $\pi^{(2)}(G^{(2)})$, with the convention that two same edges are merged into a single edge. Similarly, we define the intersection graph $G_\cap=(V_\cap,E_\cap)$ and the symmetric difference graph $G_\Delta=(V_\Delta,E_\Delta)$ so that $E_\Delta=E_\cup\setminus E_\cap$. Here, $V_\cap$ (resp. V_Δ) is the set of nodes induced by the edges $E_\cap$ (resp. E_Δ) so that $G_\cap$ (resp. G_Δ) does not contain any isolated node. We also have $|E_\cup|=|E^{(1)}|+|E^{(2)}|-|E_\cap|$ and $|V_\cup|=|V^{(1)}|+|V^{(2)}|-|\mathbf{M}|$ for $(\pi^{(1)},\pi^{(2)})\in\Pi(\mathbf{M})$. Note that, for a fixed matching $\mathbf{M}$, all graphs $G_\cup$ (resp. $G_\cap$, G_Δ) are isomorphic for $(\pi^{(1)},\pi^{(2)})\in\Pi(\mathbf{M})$ and, we shall refer to quantities such as $|E_\Delta|$, $|V_\Delta|$,... associated to a matching $\mathbf{M}$. Finally, we write $\#\mathrm{CC}_\Delta$ for the number of connected components in G_Δ.

Sets of unmatched nodes and of semi-matched nodes. Write $U^{(1)}$, resp. $U^{(2)}$ for the set of nodes in $\pi^{(1)}(G^{(1)})$, resp. $\pi^{(2)}(G^{(2)})$ that are not matched, namely the **unmatched nodes**, that is

$$U^{(1)}=\pi^{(1)}(V^{(1)})\setminus\pi^{(2)}(V^{(2)})\,;\qquad U^{(2)}=\pi^{(2)}(V^{(2)})\setminus\pi^{(1)}(V^{(1)})\,.$$

Again, $|U^{(1)}|$ and $|U^{(2)}|$ only depend on $(\pi^{(1)},\pi^{(2)})$ through the matching $\mathbf{M}$. We have, for $i\in\{1,2\}$,

$$|V^{(i)}|=|\mathbf{M}|+|U^{(i)}|. \tag{66}$$

Write also $\mathbf{M}_{\mathrm{SM}}=\mathbf{M}_{\mathrm{SM}}(\mathbf{M})\subset\mathbf{M}$, for the set of node matches of $(G^{(1)},G^{(2)})$ that are matched, and yet that are not pruned when creating the symmetric difference graph G_Δ. This is the set of **semi-matched nodes** - i.e. at least one of their incident edges is not matched. The remaining pairs of nodes $\mathbf{M}\setminus\mathbf{M}_{\mathrm{SM}}$ are said to be **perfectly matched** as all the edges incident to them are matched. We write $\mathbf{M}_{\mathrm{PM}}=\mathbf{M}_{\mathrm{PM}}(\mathbf{M})$ for the set of perfect matches in $\mathbf{M}$. Note that

$$|V^{(1)}|+|V^{(2)}|=|V_\Delta|+|\mathbf{M}_{\mathrm{SM}}|+2|\mathbf{M}_{\mathrm{PM}}|\ . \tag{67}$$

Definition of some relevant sets of nodes matchings. Given $(G^{(1)},G^{(2)})$ and a matching $\mathbf{M}$, we define the pruned matching $\mathbf{M}^-$ as the matching $\mathbf{M}$ to which we remove $(v_1^{(1)},v_1^{(2)})$ (resp. $(v_2^{(1)},v_2^{(2)})$) if either $v_1^{(1)}$ (resp. $v_2^{(1)}$) is isolated in $G^{(1)}$ or $v_1^{(2)}$ (resp. $v_2^{(2)}$) is isolated in $G^{(2)}$. This definition accounts for the fact that, when isolated, the nodes $v_i^{(j)}$, for $i=1,2$ or $j=1,2$ do not play a role in the corresponding polynomials. Then, we define $\mathcal{M}^\star\subset\mathcal{M}$ for the collection of matchings $\mathbf{M}$ such that all connected components of $G^{(1)}$ and of $G^{(2)}$ intersect with $\mathbf{M}^-$. Finally, we introduce $\mathcal{M}_{\mathrm{PM}}\subset\mathcal{M}$ for the collection of perfect matchings, that is matchings $\mathcal{M}$ such that all the nodes in $V^{(1)}$ and $V^{(2)}$ are **perfectly matched**. Note that, if $\mathbf{M}\in\mathcal{M}_{\mathrm{PM}}$, then G_Δ is the empty graph (with $E_\Delta=\emptyset$). Besides, $\mathcal{M}_{\mathrm{PM}}\neq\emptyset$ if and only $G^{(1)}$ and $G^{(2)}$ are isomorphic, which is equivalent to $G^{(1)}=G^{(2)}$ when $G^{(1)},G^{(2)}\in\mathcal{G}_{\leq D}$.

Shadow matchings. Given two sets $W^{(1)} \subset V^{(1)}, W^{(2)} \subset V^{(2)}$ and a set of node matches $\underline{\mathbf{M}} \subset \mathcal{M}$, we define $\mathcal{M}_{\text{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$ as the collection of matchings $\mathbf{M}$ satisfying

$$\mathbf{M}_{\text{SM}}(\mathbf{M}) = \underline{\mathbf{M}}, \quad (\pi^{(1)})^{-1}(U^{(1)}) = W^{(1)} \quad \text{and} \quad (\pi^{(2)})^{-1}(U^{(2)}) = W^{(2)} \tag{68}$$

for any $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$. Note that, as long as (68) is satisfied for one labeling $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$, it it satisfied for all such $(\pi^{(1)}, \pi^{(2)})$. In fact, $\mathcal{M}_{\text{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$ is the collection of all matchings that lead to the set $\underline{\mathbf{M}}$ of semi-matched nodes and such that $W^{(1)}, W^{(2)}$ correspond to unmatched nodes in resp. $G^{(1)}, G^{(2)}$. We say that these matchings satisfy a given **shadow** $(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$. The only thing that can vary between two elements of $\mathcal{M}_{\text{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$ is the matching of the nodes that are not in $W^{(1)}, W^{(2)}$, or part of a pair of nodes in $\underline{\mathbf{M}}$. This matching must however ensure that all of these nodes are perfectly matched.

Edit Distance between graphs. For any two templates $G^{(1)}$ and $G^{(2)}$, we define the so-called edit-distance

$$d(G^{(1)}, G^{(2)}) := \min_{\mathbf{M} \in \mathcal{M}} |E_\Delta| \ . \tag{69}$$

Note that $d(G^{(1)}, G^{(2)}) = 0$ if and only if $G^{(1)}$ and $G^{(2)}$ are isomorphic. As a consequence, if $G^{(1)}$ and $G^{(2)}$ are in $\mathcal{G}_{\leq D}$, the edit distance is equal to 0 if and only if $G^{(1)} = G^{(2)}$.

Appendix D. Proof of Theorem 6

We actually prove a slightly stronger version (in terms of log factors) of Theorem 6.

Theorem 21 *Let $c_s \geq 14$, $D \geq 2$, $q \leq 1/2$, $q + 2\lambda \leq 1$, and $K \leq n$. Assume that*

$$\sup_{1 \leq r \leq D} \left\{ D^{8c_s r} \times \left[\left(\frac{\sqrt{n}}{K} \left(\frac{\lambda}{\sqrt{\bar{q}}} \right)^r \right) \wedge \left(\sqrt{\frac{n}{K}} \lambda^{r/2} \right] \right) \right\} \leq 1. \tag{70}$$

Then, $\text{Corr}^2_{\leq D}$ *defined by* (24) *fulfills*

$$\text{Corr}^2_{\leq D} \leq \frac{4}{n} D^{-15c_s} \quad .$$

Theorem 6 readily follows from Theorem 21 since Condition (25) implies Condition (70).

Let us prove Theorem 21. We first prove that the family $(1, (\Psi_G)_{G \in \mathcal{G}_{\leq D}})$ is almost-orthonormal - see Proposition 25. In turn, this allows us bound $\text{Corr}_{\leq D}$ by simply controlling $\left\| (\mathbb{E}[\Psi_G x])_{G \in \mathcal{G}_{\leq D}} \right\|_2$ - see Equation (76).

Step 1: Proving that $(\Psi_G)_{G \in \mathcal{D}}$ is almost-orthonormal. By definion of Ψ_G and $\overline{P}_G$, we have $\mathbb{E}[1 \cdot \Psi_G] = 0$ for any $G \in \mathcal{G}_{\leq D}$. As a consequence, we only have to prove that $(\Psi_G)_{G \in \mathcal{D}}$ is almost-orthonormal.

It is convenient to use the short notation $X_{ij} = \mathbf{1}\{z_i = z_j\}$. We observe that $\mathbb{E}[Y_{ij}|X] = \lambda X_{ij}$ and $\mathbb{E}\left[Y_{ij}^2|X\right] = \bar{q} + \lambda X_{ij}(1 - 2q)$. Consider two templates $G^{(1)}, G^{(2)}$, some node matching $\mathbf{M} \in \mathcal{M}$ and two injections $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$. Since the Y_{ij} are conditionally independent given X, we have

$$\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right] = \mathbb{E}\left[\prod_{(i,j)\in E_\Delta} Y_{ij} \prod_{(i,j)\in E_\cap} Y_{ij}^2\right]$$
$$= \mathbb{E}\left[\prod_{(i,j)\in E_\Delta} (\lambda X_{ij}) \prod_{(i,j)\in E_\cap} [\bar{q} + \lambda X_{ij}(1-2q)]\right]. \tag{71}$$

In particular

$$\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}\right] = \lambda^{|E^{(1)}|}\frac{1}{K^{|V^{(1)}|-\#\mathrm{CC}_{G^{(1)}}}}\ , \tag{72}$$

where $\#\mathrm{CC}_{G^{(1)}}$ is the number of connected components of $G^{(1)}$.

Given $T \subset [n]$, write $Y_T = (Y_{ij})_{i\in T, j\in T}$. By independence of the z_i as well as the conditional independence of the Y_{ij} given X, it follows that for any set $T_1, T_2 \subset [n]$ such that $T_1 \cap T_2 = \emptyset$, for any functions $f^{(1)}$, $f^{(2)}$

$$\mathbb{E}\left[f^{(1)}(Y_{T_1})f^{(2)}(Y_{T_2})\right] = \mathbb{E}\left[f^{(1)}(Y_{T_1})\right]\mathbb{E}\left[f^{(2)}(Y_{T_2})\right]. \tag{73}$$

As a consequence, when the vertex sets $\pi^{(1)}(V^{(1)})$ and $\pi^{(2)}(V^{(2)})$ are disjoint, the expectation factorizes in (71). When the vertex sets overlap, we have the following upper-bound.

Proposition 22 *For any templates $(G^{(1)}, G^{(2)})$, any node matching $\mathbf{M} \in \mathcal{M}^\star$, and any labeling $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$, we have*

$$\left|\frac{\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]}{\sqrt{\mathbb{E}[P^2_{G^{(1)},\pi^{(1)}}]\mathbb{E}[P^2_{G^{(2)},\pi^{(2)}}]}}\right| \leq \left(\frac{1}{K^{|U^{(1)}|+|U^{(2)}|}}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^{|E_\Delta|}\right) \wedge \left(\frac{\lambda^{|E_\Delta|/2}}{K^{(|U^{(1)}|+|U^{(2)}|)/2}}\right)\ . \tag{74}$$

Besides, for any connected template G, and any labeling π, we have

$$\left|\frac{\mathbb{E}\left[P_{G,\pi}\right]}{\sqrt{\mathbb{E}[P^2_{G,\pi}]}}\right| \leq \left(\frac{1}{K^{|V|-1}}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^{|E|}\right) \wedge \left(\frac{\lambda^{|E|/2}}{K^{(|V|-1)/2}}\right)\ . \tag{75}$$

A first key observation is that the Gram matrix $\left(\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right]\right)_{G^{(1)},G^{(2)}\in\mathcal{G}_{\leq D},\pi^{(1)}\in\Pi_{V^{(1)}},\pi^{(1)}\in\Pi_{V^{(2)}}}$ associated to $(\overline{P}_{G,\pi})_{G\in\mathcal{G}_{\leq D},\pi\in\Pi_V}$ is quite sparse - unlike the one associated to $(P_{G,\pi})_{G\in\mathcal{G}_{\leq D},\pi\in\Pi_V}$ - and that the non-zero entries are quite close of those associated to $(P_{G,\pi})_{G\in\mathcal{G}_{\leq D},\pi\in\Pi_V}$.

Proposition 23 *We have*

1. *if $\mathbf{M} \notin \mathcal{M}^\star$ we have $\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right] = 0$;*
2. *if $\mathbf{M} \in \mathcal{M}^\star$ we have*

$$\left|\frac{\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right] - \mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]}{\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]}\right| \leq 2^{|\mathbf{M}_{\mathrm{SM}}|}D^{-4c_s} + \mathbf{1}_{|\mathbf{M}_{\mathrm{SM}}|>0}2^{|\mathbf{M}_{\mathrm{SM}}|}\ .$$

We then consider the family of invariant polynomials $(\Psi_G)_{G\in\mathcal{G}_{\leq D}}$. We prove that the associated Gram matrix $\left(\Gamma_{G^{(1)},G^{(2)}}\right)_{G^{(1)},G^{(2)}\in\mathcal{G}_{\leq D}} = \left(\mathbb{E}\left[\Psi_{G^{(1)}}\Psi_{G^{(2)}}\right]\right)_{G^{(1)},G^{(2)}\in\mathcal{G}_{\leq D}}$, is close to the identity.

Proposition 24 *Let $D \geq 2$. Consider two templates $G^{(1)}, G^{(2)} \in \mathcal{G}_{\leq D}$. We have*

$$|\Gamma_{G^{(1)},G^{(2)}} - 1| = \left|\mathbb{E}\left[\Psi_{G^{(1)}}\Psi_{G^{(2)}}\right] - \mathbf{1}\{G^{(1)} = G^{(2)}\}\right| \leq 3D^{-c_s d(G^{(1)},G^{(2)})\vee 1} \ .$$

In turn, we deduce from the above bound that Γ is close to the identity matrix in operator norm $\|.\|_{op}$.

Proposition 25 *Let $D \geq 2$. We have*

$$\|\Gamma - \mathrm{Id}\|_{op} \leq 6D^{-c_s/2} \ .$$

From there, we can easily conclude on the bound on $\mathrm{Corr}_{\leq D}$. Note first that the previous proposition together with Lemma 20 and $\mathbb{E}[x] = 0$ implies

$$\mathrm{Corr}^2_{\leq D} = \sup_{\alpha_\emptyset,(\alpha_G)_{G\in\mathcal{G}_{\leq D}}} \frac{\mathbb{E}\left[x\left(\alpha_\emptyset + \sum_{G\in\mathcal{G}_{\leq D}}\alpha_G\Psi_G\right)\right]^2}{\mathbb{E}\left[\left[\alpha_\emptyset + \sum_{G\in\mathcal{G}_{\leq D}}\alpha_G\Psi_G\right]^2\right]} \leq \frac{\sum_{G\in\mathcal{G}_{\leq D}}\mathbb{E}[x\Psi_G]^2}{1 - 6D^{-c_s/2}} \ . \tag{76}$$

So, we just need to bound $\mathbb{E}[x\Psi_G]$ for each $G \in \mathcal{G}_{\leq D}$.

Step 2: Bounding $\mathbb{E}[x\Psi_G]$. For any $G \in \mathcal{G}_{\leq D}$, write c its number of connected components with a least one edge and write $(G_1, \dots, G_c)$ its decomposition in such connected components. It follows from (73) that, for any $\pi \in \Pi_V$, we have

$$\mathbb{E}[x\Psi_G] = \frac{(n-2)!}{(n-|V|)!}\frac{1}{\sqrt{\mathbb{V}(G)}}\mathbb{E}\left[\left(\mathbf{1}_{z_1=z_2} - \frac{1}{K}\right)\prod_{i=1}^{c}\overline{P}_{G_i,\pi}\right] = \frac{(n-2)!}{(n-|V|)!}\frac{1}{\sqrt{\mathbb{V}(G)}}\mathbb{E}\left[\mathbf{1}_{z_1=z_2}\prod_{i=1}^{c}\overline{P}_{G_i,\pi}\right] \ . \tag{77}$$

Case 1: G contains at least one connected component that does not contain v_1 or v_2. By Equation (73), we have

$$\mathbb{E}[x\Psi_G] = 0.$$

So that we restrict to $G \in \mathcal{G}_{\leq D}$ such that each connected component contains v_1 or v_2.
Case 2: v_1 and v_2 are not in the same connected component. Then we know that the graph contains exactly two connected components $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$, one of them being possibly an isolated node. Fix $\pi \in \Pi_V$. If G_1 (resp. G_2 is an isolated node), we use the convention $P_{G_1,\pi} = 1$ (resp. $P_{G_2,\pi} = 1$). Developing the Ψ_G and using Equation (72) we obtain:

$$\begin{aligned}
&\frac{(n-|V|)!}{(n-2)!}\sqrt{\mathbb{V}(G)}\mathbb{E}[x\Psi_G] \\
&\quad= \mathbb{E}[\mathbf{1}_{z_1=z_2}P_{G_1,\pi}P_{G_2,\pi}] + \mathbb{E}[\mathbf{1}_{z_1=z_2}]\mathbb{E}[P_{G_1,\pi}]\mathbb{E}[P_{G_2,\pi}] \\
&\qquad - \mathbb{E}[P_{G_1,\pi}]\mathbb{E}[\mathbf{1}_{z_1=z_2}P_{G_2,\pi}] - \mathbb{E}[\mathbf{1}_{z_1=z_2}P_{G_1,\pi}]\mathbb{E}[P_{G_2,\pi}] \\
&\quad= \lambda^{|E|}\left[\frac{1}{K^{|V_1|+|V_2|-1}} + \frac{1}{K}\cdot\frac{1}{K^{|V_1|-1}}\cdot\frac{1}{K^{|V_2|-1}} - \frac{1}{K^{|V_1|}}\cdot\frac{1}{K^{|V_2|-1}} - \frac{1}{K^{|V_2|}}\cdot\frac{1}{K^{|V_1|-1}}\right] \\
&\quad= 0 \ .
\end{aligned}$$

Case 3: v_1, v_2 are in the same connected component. Then, this implies that $\mathbb{E}[x\Psi_G] = 0$ unless G is connected. It then follows that, for any $\pi \in \Pi_V$,

$$\frac{(n-|V|)!}{(n-2)!}\sqrt{\mathbb{V}(G)}\mathbb{E}[x\Psi_G] = \mathbb{E}[\mathbf{1}_{z_1=z_2}P_{G,\pi}] - \mathbb{E}[\mathbf{1}_{z_1=z_2}]\mathbb{E}[P_{G,\pi}] = [1-K^{-1}]\mathbb{E}[P_{G,\pi}],$$

where the last equality stems from

$$\mathbb{E}[\mathbf{1}_{z_1=z_2}P_{G,\pi}] = \lambda^{[E]}\mathbb{P}\left[z_i \text{ all equal on } V\right] = \mathbb{E}[P_{G,\pi}].$$

Combining the definition of $\mathbb{V}(G)$ and (75) in Proposition 22, we get

$$\begin{aligned}
|\mathbb{E}[x\Psi_G]|| &\leq \sqrt{\frac{(n-2)!}{\mathbb{E}[P_{G,\pi}^2](n-|V|)!|\mathrm{Aut}(G)|}}|\mathbb{E}[P_{G,\pi}]| \\
&\leq \sqrt{\frac{(n-2)!}{(n-|V|)!|\mathrm{Aut}(G)|}}\left[\left(\frac{1}{K^{|V|-1}}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^{|E|}\right) \wedge \left(\frac{\lambda^{|E|/2}}{K^{(|V|-1)/2}}\right)\right] \quad .
\end{aligned}$$

As a consequence, we get

$$\begin{aligned}
|\mathbb{E}[x\Psi_G]|| &\leq n^{(|V|-2)/2}\left[\left(\frac{1}{K^{|V|-1}}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^{|E|}\right) \wedge \left(\frac{\lambda^{|E|/2}}{K^{(|V|-1)/2}}\right)\right] \\
&\leq \frac{1}{\sqrt{n}}\left[\left(\left(\frac{\sqrt{n}}{K}\right)^{|V|-1}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^{|E|}\right) \wedge \left(\left(\sqrt{\frac{n}{K}}\right)^{|V|-1}\lambda^{|E|/2}\right)\right] \\
&\leq \frac{1}{\sqrt{n}}\left[\left(\frac{\sqrt{n}}{K}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^{r}\right) \wedge \left(\sqrt{\frac{n}{K}}\lambda^{r/2}\right)\right]^{|V|-1} \quad ,
\end{aligned}$$

where $r = r(G) = |E|/(|V|-1)$. Recall, that G is a connected graph. Hence, $|E| \geq |V|-1$ and we have $1 \leq r(G) \leq D$, since $|E| \leq D$. Under our signal condition in Equation (70), we conclude that

$$|\mathbb{E}[x\Psi_G]|| \leq \frac{1}{\sqrt{n}}D^{-8c_s|E|}.$$

Conclusion. Combining the above, we arrive at

$$\begin{aligned}
\sum_{G\in\mathcal{G}_{\leq D}}\mathbb{E}^2[x\Psi_G] &\leq \frac{1}{n}\sum_{G\in\mathcal{G}_{\leq D}:|E|\geq 1}D^{-16c_s|E|} = \frac{1}{n}\sum_{v\in[2D],e\in[D]}\quad\sum_{G\in\mathcal{G}_{\leq D}:|V|=v,|E|=e}D^{-16c_s|E|} \\
&\leq \frac{1}{n}\sum_{v\in[2D],e\in[D]}v^{2e}D^{-16c_s e} \leq \frac{2}{n}D^{-15c_s}
\end{aligned}$$

since $c_s \geq 8$ and $D \geq 2$. Together with (76), we conclude that

$$\mathrm{Corr}_{\leq D}^2 \leq \frac{4}{n}D^{-15c_s} \quad .$$

Proof [Proof of Proposition 23]

Proof of 1): Consider any matching $\mathbf{M} \notin \mathcal{M}^\star$. As a consequence, there exists a connected component G' of $G^{(1)}$ or in $G^{(2)}$ such that no nodes in G' belongs to $\mathbf{M}^-$ —recall the definition of $\mathbf{M}^-$ in Section C.2. Without loss of generality, we assume that G' is a connected component of $G^{(1)}$. By definition, we have $\overline{P}_{G^{(1)},\pi^{(1)}} = \overline{P}_{G',\pi^{(1)}}\overline{P}_{\underline{G}^{(1)},\pi^{(1)}}$ where $\underline{G}^{(1)}$ is the complement of G' in $G^{(1)}$. So that by Equation (73), we get

$$\mathbb{E}[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}] = \mathbb{E}[\overline{P}_{\underline{G}^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}]\mathbb{E}[\overline{P}_{G',\pi^{(1)}}] = 0 \ ,$$

as $\mathbb{E}[\overline{P}_{G',\pi^{(1)}}] = 0$.

Proof of 2): Write c_1 (resp. c_2) for the number of connected component of $G^{(1)}$ (resp. $G^{(2)}$) that contain at least one edge. Besides, we write $(G^{(1)}_1, \ldots, G^{(1)}_{c_1})$ and $(G^{(2)}_1, \ldots, G^{(2)}_{c_2})$ for the decomposition of $G^{(1)}$, $G^{(2)}$ into these connected components. Then, we have

$$\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right] = \sum_{S_1\subset[c_1],\ S_2\subset[c_2]} (-1)^{|S_1|+|S_2|}\mathbb{E}\left[\prod_{i\in[c_1]\setminus S_1} P_{G^{(1)}_i,\pi^{(1)}} \prod_{j\in[c_2]\setminus S_2} P_{G^{(2)}_j,\pi^{(2)}}\right] \\ \times \mathbb{E}\left[\prod_{i\in S_1} P_{G^{(1)}_i,\pi^{(1)}}\right]\mathbb{E}\left[\prod_{i\in S_2} P_{G^{(2)}_i,\pi^{(2)}}\right] \ .$$

The following lemma holds.

Lemma 26 *For any $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$ and any $S_1 \subset [c_1]$ and any $S_2 \subset [c_2]$, we have*

$$0 \leq \mathbb{E}\left[\prod_{i\in[c_1]\setminus S_1} P_{G^{(1)}_i,\pi^{(1)}} \prod_{j\in[c_2]\setminus S_2} P_{G^{(2)}_j,\pi^{(2)}}\right]\mathbb{E}\left[\prod_{i\in S_1} P_{G^{(1)}_i,\pi^{(1)}}\right]\mathbb{E}\left[\prod_{i\in S_2} P_{G^{(2)}_i,\pi^{(2)}}\right] \\ \leq \lambda^{|\mathbf{M}^{(S_1,S_2)}_{\mathrm{PM}}|/2}\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right] \ .$$

where $\mathbf{M}^{(S_1,S_2)}_{\mathrm{PM}} \subset \mathbf{M}_{\mathrm{PM}}$ is the set of pairs of (v, v') perfectly matched nodes (to the possible exception of v_1 and v_2 in the case where they are isolated) such that either v belongs to any $(G^{(1)}_i)$ with $i \in S_1$ or v' belongs to any $(G^{(1)}_i)$ with $i \in S_2$.

Note that choosing any set of connected components S_1, S_2 amounts to choosing some specific subsets of node matches in $\mathbf{M} = \mathbf{M}_{\mathrm{PM}} \cup \mathbf{M}_{\mathrm{SM}}$. Since $\mathbf{M} \in \mathcal{M}^\star$, this subset of $\mathbf{M}$ is therefore non empty. This leads us to:

$$\begin{aligned}\frac{\left|\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right]-\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]\right|}{\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]} &\leq \sum_{\substack{S_1\subseteq[c_1],\,S_2\subseteq[c_2]\\ S_1\cup S_2\neq\emptyset}}\lambda^{|\mathbf{M}_{\mathrm{PM}}^{(S_1,S_2)}|/2}\\
&\leq \sum_{\substack{\widetilde{\mathbf{M}}_{\mathrm{PM}}\subseteq\mathbf{M}_{PM},\,\widetilde{\mathbf{M}}_{\mathrm{SM}}\subseteq\mathbf{M}_{SM}\\ \widetilde{\mathbf{M}}_{\mathrm{PM}}\cup\widetilde{\mathbf{M}}_{\mathrm{SM}}\neq\emptyset}}\lambda^{|\widetilde{\mathbf{M}}_{\mathrm{PM}}|/2}\\
&\leq \mathbf{1}_{|\mathbf{M}_{\mathrm{SM}}|>0}2^{|\mathbf{M}_{\mathrm{SM}}|}\sum_{\widetilde{\mathbf{M}}_{\mathrm{PM}}\subseteq\mathbf{M}_{PM}}\lambda^{|\widetilde{\mathbf{M}}_{\mathrm{PM}}|/2}+\mathbf{1}_{|\mathbf{M}_{\mathrm{SM}}|=0}\sum_{\emptyset\subsetneq\widetilde{\mathbf{M}}_{\mathrm{PM}}\subseteq\mathbf{M}_{PM}}\lambda^{|\widetilde{\mathbf{M}}_{\mathrm{PM}}|/2}\\
&\leq \mathbf{1}_{|\mathbf{M}_{\mathrm{SM}}|>0}2^{|\mathbf{M}_{\mathrm{SM}}|}\sum_{r\geq 0}|\mathbf{M}_{\mathrm{PM}}|^r\lambda^{r/2}+\mathbf{1}_{|\mathbf{M}_{\mathrm{SM}}|=0}\sum_{r\geq 1}|\mathbf{M}_{\mathrm{PM}}|^r\lambda^{r/2}\\
&\leq 2^{|\mathbf{M}_{\mathrm{SM}}|}D^{-4c_s}+\mathbf{1}_{|\mathbf{M}_{\mathrm{SM}}|>0}2^{|\mathbf{M}_{\mathrm{SM}}|}\ ,\end{aligned}$$

where we used that $|\mathbf{M}_{\mathrm{PM}}|\leq|V^{(1)}|\leq 2D$, $D\geq 2$, and $\sqrt{\lambda}\leq D^{-8c_s}$ with $c_{\mathrm{s}}>1$ by Condition 70. ■

Proof [Proof of Lemma 26] Define the matching $\bar{\mathbf{M}}\subset\mathbf{M}$ where we have removed all the nodes (v,v') such that either v belongs to any connected component $G_i^{(1)}$ with $i\in S_1$ or v' belongs to any connected component $G_j^{(2)}$ with $j\in S_2$. It is possible to choose labelings $(\bar{\pi}^{(1)},\bar{\pi}^{(2)})\in\Pi(\bar{\mathbf{M}})$ such that $\bar{\pi}^{(1)}=\pi^{(1)}$ and $\bar{\pi}_2(i)=\pi_2(i)$ for $i\in U^{(2)}$, that is unmatched nodes of $G^{(2)}$. We have

$$\mathbb{E}\left[P_{G^{(1)},\bar{\pi}^{(1)}}P_{G^{(2)},\bar{\pi}^{(2)}}\right]=\mathbb{E}\left[\prod_{i\in[c_1]\setminus S_1}P_{G_i^{(1)},\pi^{(1)}}\prod_{j\in[c_2]\setminus S_2}P_{G_j^{(2)},\pi^{(2)}}\right]\mathbb{E}\left[\prod_{i\in S_1}P_{G_i^{(1)},\pi^{(1)}}\right]\mathbb{E}\left[\prod_{i\in S_2}P_{G_i^{(2)},\pi^{(2)}}\right]\ . \tag{78}$$

Write $\bar{G}_\Delta=(\bar{V}_\Delta,\bar{E}_\Delta),\bar{G}_\cap=(\bar{V}_\cap,\bar{E}_\cap),\bar{G}_\cup=(\bar{V}_\cup,\bar{E}_\cup)$ for the resp. symmetric difference, intersection and union graphs corresponding to $(\bar{\pi}^{(1)},\bar{\pi}^{(2)})$. So by Eq. (71)

$$\mathbb{E}\left[P_{G^{(1)},\bar{\pi}^{(1)}}P_{G^{(2)},\bar{\pi}^{(2)}}\right]=\mathbb{E}\left[\prod_{(i,j)\in\bar{E}_\Delta}(\lambda X_{ij})\prod_{(i,j)\in\bar{E}_\cap}[\bar{q}+\lambda X_{ij}(1-2q)]\right]\ . \tag{79}$$

We have chosen $\bar{\pi}^{(1)}=\pi^{(1)}$. Hence, we have $\bar{E}_\cap\subseteq E_\cap$. Since $\bar{q}+(1-2q)\lambda\geq\lambda$, we again deduce from (71) that

$$\begin{aligned}\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]&=\mathbb{E}\left[\prod_{(i,j)\in E_\Delta}(\lambda X_{ij})\prod_{(i,j)\in E_\cap}[\bar{q}+\lambda X_{ij}(1-2q)]\right]\\
&\geq\mathbb{E}\left[\prod_{(i,j)\in E_\cup\setminus\bar{E}_\cap}(\lambda X_{ij})\prod_{(i,j)\in\bar{E}_\cap}[\bar{q}+\lambda X_{ij}(1-2q)]\right]\ ,\end{aligned} \tag{80}$$

Lemma 27 *We have*

$$\mathbb{E}\left[\prod_{(i,j)\in\bar{E}_\Delta} X_{ij}\prod_{(i,j)\in\bar{E}_\cap}[\bar{q}+\lambda X_{ij}(1-2q)]\right]\leq\mathbb{E}\left[\prod_{(i,j)\in E_\cup\setminus\bar{E}_\cap} X_{ij}\prod_{(i,j)\in\bar{E}_\cap}[\bar{q}+\lambda X_{ij}(1-2q)]\right]\ .$$

By Equation (79) and Equation (80), we then deduce that

$$\mathbb{E}\left[P_{G^{(1)},\bar{\pi}^{(1)}}P_{G^{(2)},\bar{\pi}^{(2)}}\right]\leq\lambda^{|\bar{E}_\Delta|-|E_\cup\setminus\bar{E}_\cap|}\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]=\lambda^{|E_\cap\setminus\bar{E}_\cap|}\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]\ , \tag{81}$$

because $|E_\cup\setminus\bar{E}_\cap|=|E_\Delta|+|E_\cap\setminus\bar{E}_\cap|$ and $|\bar{E}_\Delta|=|E_\Delta|+2|E_\cap\setminus\bar{E}_\cap|$. The cardinality $|\mathbf{M}_{\mathrm{PM}}^{(S_1,S_2)}|$ is, by definition, smaller than the number of nodes in S_1 or in S_2 that are not isolated and are perfecty matched. By definition, any such perfectly matched node in S_1 and in S_2 is incident to at least an edge in $E_\cap\setminus\bar{E}_\cap$. We deduce that

$$|\mathbf{M}_{\mathrm{PM}}^{(S_1,S_2)}|\leq 2|E_\cap\setminus\bar{E}_\cap|\ .$$

Coming back to (81), we obtain

$$\mathbb{E}\left[P_{G^{(1)},\bar{\pi}^{(1)}}P_{G^{(2)},\bar{\pi}^{(2)}}\right]\leq\lambda^{|\mathbf{M}_{\mathrm{PM}}^{(S_1,S_2)}|/2}\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]\ ,$$

which, combined with (78), concludes the proof.

■

Proof [Proof of Lemma 27]
Let us start by developing the products of $\bar{q}+\lambda X_{ij}(1-2q)$.

$$\mathbb{E}\left[\prod_{(i,j)\in\bar{E}_\Delta} X_{ij}\prod_{(i,j)\in\bar{E}_\cap}[\bar{q}+\lambda X_{ij}(1-2q)]\right]=\mathbb{E}\left[\sum_{S\subseteq\bar{E}_\cap}\prod_{(i,j)\in\bar{E}_\Delta}X_{ij}\bar{q}^{|\bar{E}_\cap|-|S|}\prod_{(i,j)\in S}[\lambda X_{ij}(1-2q)]\right]\ ;$$

$$\mathbb{E}\left[\prod_{(i,j)\in E_\cup\setminus\bar{E}_\cap} X_{ij}\prod_{(i,j)\in\bar{E}_\cap}[\bar{q}+\lambda X_{ij}(1-2q)]\right]=\mathbb{E}\left[\sum_{S\subseteq\bar{E}_\cap}\prod_{(i,j)\in E_\cup\setminus\bar{E}_\cap}X_{ij}\bar{q}^{|\bar{E}_\cap|-|S|}\prod_{(i,j)\in S}[\lambda X_{ij}(1-2q)]\right]\ .$$

Hence, it order to conclude, if suffices to prove that, for any $S\subseteq\bar{E}_\cap$, we have

$$\mathbb{E}\left[\prod_{(i,j)\in S\cup(E_\cup\setminus\bar{E}_\cap)}X_{ij}\right]\geq\mathbb{E}\left[\prod_{(i,j)\in S\cup\bar{E}_\Delta}X_{ij}\right]\ . \tag{82}$$

To show (82), we fix such a subset $S\subseteq\bar{E}_\cap$. Define the graphs H and $\bar{H}$ respectively induced by the set of edges $S\cup(E_\cup\setminus\bar{E}_\cap)$ and $S\cup\bar{E}_\Delta$. Then denoting $|V(H)|$, $|V(\bar{H})|$ their number of nodes and $\#\mathrm{CC}_H$ and $\#\mathrm{CC}_{\bar{H}}$ their number of connected components; we have

$$\mathbb{E}\left[\prod_{(i,j)\in S\cup(E_\cup\setminus\bar{E}_\cap)}X_{ij}\right]=K^{-|V(H)|+\#\mathrm{CC}_H}\ ,\qquad\mathbb{E}\left[\prod_{(i,j)\in S\cup\bar{E}_\Delta}X_{ij}\right]=K^{-|V(\bar{H})|+\#\mathrm{CC}_{\bar{H}}}\ ,$$

so that (82) is equivalent to

$$|V(H)| - \#\mathrm{CC}_H \le |V(\bar{H})| - \#\mathrm{CC}_{\bar{H}} \ . \tag{83}$$

Thus, we only have to establish this last inequality.

Observe that H is the subgraph of $G_\cup$ where we have removed the edges $\bar{E}_\cap \setminus S$ and have pruned the isolated nodes. Similarly, $\bar{H}$ is the subgraph of $\bar{G}_\cup$ where we have removed the edges $\bar{E}_\cap \setminus S$ and have pruned the isolated nodes. By definition of $\bar{G}_\cup$ at the beginning of the proof of Lemma 26, $G_\cup$ is obtained from $\bar{G}_\cup$ by merging/identifying couples of nodes. If a node has been pruned in the construction of H or $\bar{H}$, then this implies that this node was only incident to edges in $\bar{E}_\cap \setminus S$. As a consequence, pruned nodes in $\bar{H}$ are in correspondence with those in H. Hence, to go from $\bar{H}$ to H, we merge/identify $s \ge 0$ couples of nodes in $\bar{H}$. The operation of merging two nodes decreases the number of nodes of the graph by one and decreases its number of connected components by at most one. Iterating the arguments s times, we conclude that $|V(H)| - \#\mathrm{CC}_H \le |V(\bar{H})| - \#\mathrm{CC}_{\bar{H}}$, which concludes the proof.

■

Proof [Proof of Proposition 24] By definition of Ψ_G, we have

$$\begin{aligned}
\mathbb{E}\left[\Psi_{G^{(1)}}\Psi_{G^{(2)}}\right] &= \sum_{\mathbf{M}\in\mathcal{M}} \sum_{(\pi^{(1)},\pi^{(2)})\in\Pi(\mathbf{M})} \frac{1}{\sqrt{\mathbb{V}(G^{(1)})\mathbb{V}(G^{(2)})}}\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right] \\
&= \sum_{\mathbf{M}\in\mathcal{M}^\star} \sum_{(\pi^{(1)},\pi^{(2)})\in\Pi(\mathbf{M})} \frac{1}{\sqrt{\mathbb{V}(G^{(1)})\mathbb{V}(G^{(2)})}}\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right] ,
\end{aligned}$$

where the second line follows by Proposition 23.

Step 1: Decomposition of the scalar product over $\mathcal{M}^\star \setminus \mathcal{M}_{\mathrm{PM}}$ and $\mathcal{M}_{\mathrm{PM}}$. The collection $\mathcal{M}_{\mathrm{PM}}$ is non-empty only if $G^{(1)} = G^{(2)}$. If $G^{(1)} = G^{(2)}$, we have

$$\begin{aligned}
\sum_{\mathbf{M}\in\mathcal{M}_{\mathrm{PM}}} \sum_{(\pi^{(1)},\pi^{(2)})\in\Pi(\mathbf{M})} \frac{1}{\sqrt{\mathbb{V}(G^{(1)})\mathbb{V}(G^{(2)})}}\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right] &= \frac{(n-2)!|\mathrm{Aut}(G^{(1)})|\mathbb{E}\left[\overline{P}^2_{G^{(1)},\pi}\right]}{\left(n-|V^{(1)}|\right)!\mathbb{V}(G^{(1)})} \\
&= \frac{\mathbb{E}[\overline{P}^2_{G^{(1)},\pi}]}{\mathbb{E}[P^2_{G^{(1)},\pi}]},
\end{aligned}$$

since we have

$$|\mathcal{M}_{\mathrm{PM}}| = |\mathrm{Aut}(G^{(1)})| \quad \text{and} \quad |\Pi(\mathbf{M})| = \frac{(n-2)!}{\left(n-(|V^{(1)}|+|V^{(2)}|-|\mathbf{M}|)\right)!}. \tag{84}$$

Then, it follows from the second part of Proposition 23 that

$$\begin{aligned}
&\left|\mathbb{E}\left[\Psi_{G^{(1)}}\Psi_{G^{(2)}}\right] - \mathbf{1}\{G^{(1)} = G^{(2)}\}\right| \\
&\le \left|\sum_{\mathbf{M}\in\mathcal{M}^\star\setminus\mathcal{M}_{\mathrm{PM}}} \sum_{(\pi^{(1)},\pi^{(2)})\in\Pi(\mathbf{M})} \frac{1}{\sqrt{\mathbb{V}(G^{(1)})\mathbb{V}(G^{(2)})}}\mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right]\right| + D^{-4c_s[d(G^{(1)},G^{(2)})\vee 1]} \\
&:= A + D^{-4c_s[d(G^{(1)},G^{(2)})\vee 1]} \ .
\end{aligned} \tag{85}$$

Step 2: Making A explicit as a sum of $A(\mathbf{M})$. Observe that, for any $\mathbf{M} \in \mathcal{M}$, $\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]$ is identical for any $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$. So that by Proposition 23 and Equation (84)

$$\begin{aligned}
A\sqrt{\mathbb{V}(G^{(1)})\mathbb{V}(G^{(2)})} &= \left|\sum_{\mathbf{M}\in\mathcal{M}^\star\setminus\mathcal{M}_{\mathrm{PM}}}\ \sum_{(\pi^{(1)},\pi^{(2)})\in\Pi(\mathbf{M})} \mathbb{E}\left[\overline{P}_{G^{(1)},\pi^{(1)}}\overline{P}_{G^{(2)},\pi^{(2)}}\right]\right| \\
&\leq \sum_{\mathbf{M}\in\mathcal{M}^\star\setminus\mathcal{M}_{\mathrm{PM}}} \frac{(n-2)!}{\left(n-(|V^{(1)}|+|V^{(2)}|-|\mathbf{M}|)\right)!}\left|\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]\right|\left(2^{|\mathbf{M}_{\mathrm{SM}}|}(1+D^{-4c_s})+1\right) \\
&\leq 2\sum_{\mathbf{M}\in\mathcal{M}^\star\setminus\mathcal{M}_{\mathrm{PM}}} \frac{(n-2)!}{\left(n-(|V^{(1)}|+|V^{(2)}|-|\mathbf{M}|)\right)!}\left|\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]\right|2^{|\mathbf{M}_{\mathrm{SM}}|}\ ,
\end{aligned}$$

since $D \geq 2$.

Observe that $\frac{(n-2)!}{\left(n-(|V^{(1)}|+|V^{(2)}|-|\mathbf{M}|)\right)!}\frac{\sqrt{(n-|V^{(1)}|)!(n-|V^{(2)}|)!}}{(n-2)!} \leq n^{\frac{|U^{(1)}|+|U^{(2)}|}{2}}$. Then, by Proposition 22, we get

$$\begin{aligned}
&A\sqrt{\left|\mathrm{Aut}(G^{(1)})\right|\left|\mathrm{Aut}(G^{(2)})\right|} \\
&\leq 2\sum_{\mathbf{M}\in\mathcal{M}^\star\setminus\mathcal{M}_{\mathrm{PM}}} n^{\frac{|U^{(1)}|+|U^{(2)}|}{2}}\left[\left(\frac{\sqrt{n}}{K^{|U^{(1)}|+|U^{(2)}|}}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^{|E_\Delta|}\right)\wedge\left(\frac{\lambda^{|E_\Delta|/2}}{K^{(|U^{(1)}|+|U^{(2)}|)/2}}\right)\right]2^{|\mathbf{M}_{\mathrm{SM}}|} \\
&= 2\sum_{\mathbf{M}\in\mathcal{M}^\star\setminus\mathcal{M}_{\mathrm{PM}}}\left[\left(\frac{\sqrt{n}}{K}\left(\frac{\lambda}{\sqrt{\bar{q}}}\right)^r\right)\wedge\left(\frac{\lambda^{r/2}}{\sqrt{K/n}}\right)\right]^U 2^{|\mathbf{M}_{\mathrm{SM}}|}\ ,
\end{aligned}$$

where $U = U(\mathbf{M}) = |U^{(1)}| + |U^{(2)}|$ and $r = r(\mathbf{M}) = |E_\Delta|/U$. Write $A(\mathbf{M})$ for the summand in the last line.

Step 3: Bounding of A by summing over shadows. Recall the definition of shadows and of $\mathcal{M}_{\mathrm{shadow}}$ in Section C.2. We now regroup the sum inside A by enumerating all possible matching that are compatible with a shadow. We get

$$A \leq \frac{2}{\sqrt{|\mathrm{Aut}(G^{(1)})||\mathrm{Aut}(G^{(2)})|}}\sum_{\substack{W^{(1)}\subset V^{(1)},W^{(2)}\subset V^{(2)}\\ \underline{\mathbf{M}}\in\mathcal{M}\setminus\mathcal{M}_{\mathrm{PM}}}}\ \sum_{\mathbf{M}\in\mathcal{M}_{\mathrm{shadow}}(W^{(1)},W^{(2)},\underline{\mathbf{M}})} A(\mathbf{M})\ .$$

Remark that $A(\mathbf{M})$ is the same for all $\mathbf{M} \in \mathcal{M}_{\mathrm{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$ and only depends on $U = |U^{(1)}| + |U^{(2)}| = |W^{(1)}| + |W^{(2)}|$, and $\underline{\mathbf{M}}$. Besides, we have $\mathbf{M}_{\mathrm{SM}} = \underline{\mathbf{M}}$. We have the following control for the cardinality of $\mathcal{M}_{\mathrm{shadow}}$.

Lemma 28 *For any $W^{(1)}$, $W^{(2)}$, and $\underline{\mathbf{M}}$, we have*

$$|\mathcal{M}_{\mathrm{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})| \leq \min(|\mathrm{Aut}(G^{(1)})|, |\mathrm{Aut}(G^{(2)})|)\ .$$

Observe that two matchings $\mathbf{M}$ and $\mathbf{M}'$ that belong $\mathcal{M}_{\mathrm{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$ have isomorphic symmetric difference graph G_Δ and have a common value of $|\mathbf{M}_{\mathrm{SM}}|$. Hence

$$A \leq 2\sum_{W^{(1)}\subset V^{(1)},W^{(2)}\subset V^{(2)},\underline{\mathbf{M}}\in\mathcal{M}\setminus\mathcal{M}_{\mathrm{PM}}} A(\mathbf{M})\ , \tag{86}$$

where $\mathbf{M}$ is any matching in $\mathcal{M}_{\text{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$.

Step 4: Bounding $A(\mathbf{M})$. Recall that $|\mathbf{M}_{\text{SM}}| + |U^{(1)}| + |U^{(2)}| = |V_\Delta|$. For $\mathbf{M} \in \mathcal{M}^\star$, we have $|\mathbf{M}_{\text{SM}}| \geq \#\text{CC}_\Delta$. Since $|E_\Delta| \geq |V_\Delta| - \#\text{CC}_\Delta$, we deduce that $|E_\Delta| \geq |U|$ which, in turn, implies that $r(\mathbf{M}) \geq 1$. Also, by definition of the edit distance, we have $|E_\Delta| \geq d(G^{(1)}, G^{(2)}) \vee 1$. Finally, we have $|E_\Delta| \geq |\mathbf{M}_{\text{SM}}|/2$ as any semi-matched node is connected to an edge in G_Δ. Gathering these three lower bounds, we get

$$A(\mathbf{M}) \leq 2^{|\mathbf{M}_{\text{SM}}|} D^{-8c_{\text{s}}|E_\Delta|} \leq 2^{|\mathbf{M}_{\text{SM}}|} D^{-2c_{\text{s}}[|U| + |\mathbf{M}_{\text{SM}}| + d(G^{(1)}, G^{(2)}) \vee 1]} \ .$$

Step 5: Final bound on A. Plugging this bound on $A(\mathbf{M})$ back in Equation (86) we get

$$A \leq 2 \sum_{W^{(1)} \subset V^{(1)}, W^{(2)} \subset V^{(2)}, \underline{\mathbf{M}} \in \mathcal{M} \setminus \mathcal{M}_{\text{PM}}} 2^{|\mathbf{M}_{\text{SM}}|} D^{-2c_{\text{s}}[|W^{(1)}| + |W^{(2)}| + |\mathbf{M}_{\text{SM}}| + d(G^{(1)}, G^{(2)}) \vee 1]}, \quad (87)$$

since for $\mathbf{M} \in \mathcal{M}_{\text{shadow}}(W^{(1)}, W^{(2)}, \underline{\mathbf{M}})$, we have $|U^{(1)}| = |W^{(1)}|$ and $|U^{(2)}| = W^{(2)}$. So when we enumerate over all possible sets $W^{(1)}, W^{(2)}, \underline{\mathbf{M}}$ that have respective cardinality u_1, u_2, and m, and since these sets have bounded cardinalities resp. by $(2D)^{u_1}$, $(2D)^{u_2}$ and $(2D)^{2m}$, we have

$$A \leq 2 \sum_{u_1, u_2, m \geq 0} 2^m (2D)^{u_1 + u_2 + 2m} D^{-2c_{\text{s}}\left[d(G^{(1)}, G^{(2)}) \vee 1 + u_1 + u_2 + m\right]} \leq 2D^{-c_{\text{s}} d(G^{(1)}, G^{(2)}) \vee 1},$$

since $c_{\text{s}} \geq 5$ and $D \geq 2$. Together with (85), this concludes the proof. ■

Proof [Proof of Proposition 25] Since the operator norm of a symmetric matrix is bounded by the maximum L_1 norm of its rows, we have

$$\|\Gamma - \text{Id}\|_{op} \leq \max_{G^{(1)}} \left|\Gamma_{G^{(1)}, G^{(1)}} - 1\right| + \sum_{G^{(2)} \in \mathcal{G}_{\leq D}, G^{(2)} \neq G^{(1)}} \left|\Gamma_{G^{(1)}, G^{(2)}}\right|$$

To bound the latter sum, we use that for a fixed template $G^{(1)}$, the number of templates $G^{(2)} \in \mathcal{G}_{\leq D}$ such that $d(G^{(1)}, G^{(2)}) = u$ is bounded by $(u + D)^{2u}$. We also use that if $G^{(2)} \neq G^{(1)}$, then $d(G^{(1)}, G^{(2)}) \geq 1$ as they are not isomorphic. It then follows from Proposition 24 that

$$\begin{aligned}
\sum_{G^{(2)} \in \mathcal{G}_{\leq D}, G^{(2)} \neq G^{(1)}} \left|\Gamma_{G^{(1)}, G^{(2)}}\right| &\leq \sum_{G^{(2)} \in \mathcal{G}_{\leq D}, G^{(2)} \neq G^{(1)}} 3D^{-c_{\text{s}} d(G^{(1)}, G^{(2)})} \\
&\leq \sum_{2D \geq u \geq 1} |\{G^{(2)} : d(G^{(1)}, G^{(2)}) = u\}| 3D^{-c_{\text{s}} u} \\
&\leq \sum_{2D \geq u \geq 1} 3(u + D)^{2u} D^{-c_{\text{s}} u} \leq \sum_{2D \geq u \geq 1} 3D^{-(c_{\text{s}} - 6)u} \leq 3D^{-c_{\text{s}}/2} \ ,
\end{aligned}$$

since $D \geq 2$ provided we have $c_{\text{s}} \geq 14$. Using again Proposition 24, to bound $\left|\Gamma_{G^{(1)}, G^{(1)}} - 1\right|$ concludes the proof. ■

Proof [Proof of Lemma 28] The proof is a straighforward variation of that of Lemma A.5 in Carpentier et al. (2025a), the twist being that $|\text{Aut}(G)|$ is restrict to automorphisms that let v_1 and v_2 fixed, whereas all matchings $\mathbf{M}$ of $(G^{(1)}, G^{(2)})$ contain $(v_1^{(1)}, v_1^{(2)})$ and $(v_2^{(1)}, v_2^{(2)})$. ■

Appendix E. Proof of Proposition 22

Consider two templates $G^{(1)}, G^{(2)}$, some node matching $\mathbf{M} \in \mathcal{M}^*$ and two injections $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$. The expectation $\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}} P_{G^{(2)},\pi^{(2)}}\right]$ only depends on the graphs $G^{(1)}\left[\pi^{(1)}\right]$ and $G^{(2)}\left[\pi^{(2)}\right]$. To avoid clumsy notations, and since the injections $(\pi^{(1)}, \pi^{(2)}) \in \Pi(\mathbf{M})$ are fixed, we directly work in this proof with the graphs $G^{(i)}\left[\pi^{(i)}\right] =: (V^{(i)}, E^{(i)})$, with $i = 1, 2$. Also, we shall introduce new notations to account for the case where v_1 or v_2 are isolated in $G^{(1)}$ or in $G^{(2)}$. For $i = 1, 2$, we define $G^{\star(i)} = (V^{\star(i)}, E^{(i)})$ by removing isolated nodes. We recall the notation $E_\cap = E^{(1)} \cap E^{(2)}$ and $E_\Delta = E^{(1)} \Delta E^{(2)}$. We partition E_Δ according to these connected components $E_\Delta = \cup_{\ell=1}^{\#\mathrm{CC}_\Delta} E_{\Delta,\ell}$. Since we work with the pruned graph, we shall redefine the node set. In particular, we define $V^\star_\cap = V^{\star(1)} \cap V^{\star(2)}$, we partition $V^\star_\cap$ into the sets $V^\star_{PM}$ and $V^\star_{SM}$ of perfectly and semi-matched nodes. Besides we define $U^{\star(1)} = V^{\star(1)} \setminus U^{\star(2)}$, $U^{\star(2)} = V^{\star(2)} \setminus U^{\star(1)}$, and $U^\star = U^{\star(1)} \cup U^{\star(2)}$. Obviously, we have $|U| = |U^\star|$ when neither v_1 nor v_2 is isolated in both $G^{(1)}$ and $G^{(2)}$. In general, we have $|U^\star| \in [|U|, |U| + 2]$. Given a subset W of $V^\star_\cap$, we define $E_\cap[W]$ as the subset of edges in $E_\cap$ that connects nodes in W. We first focus on proving (74).

We have $\mathbb{E}\left[Y_{ij}|X\right] = \lambda X_{ij} = \lambda \mathbf{1}_{z_i=z_j}$ and

$$\mathbb{E}\left[Y_{ij}^2|X\right] = \bar{q} + \lambda X_{ij}(1-2q) = \bar{q}\left(\frac{\bar{p}}{\bar{q}}\right)^{\mathbf{1}_{z_i=z_j}},$$

with $\bar{p} = \bar{q} + \lambda(1-2q) \geq \bar{q}$ for $q \leq 1/2$. Since the Y_{ij} are conditionally independent given X, and since for $\mathbf{M} \in \mathcal{M}^\star$, no connected component of G_Δ is only composed of nodes from $U^{\star(j)}$, we have

$$\begin{aligned}
\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}} P_{G^{(2)},\pi^{(2)}}\right] &= \mathbb{E}\left[\prod_{(i,j)\in E_\Delta} Y_{ij} \prod_{(i,j)\in E_\cap} Y_{ij}^2\right] \\
&= \mathbb{E}\left[\prod_{(i,j)\in E_\Delta} (\lambda \mathbf{1}_{z_i=z_j}) \prod_{(i,j)\in E_\cap} \bar{q}\left(\frac{\bar{p}}{\bar{q}}\right)^{\mathbf{1}_{z_i=z_j}}\right] \qquad (88)\\
&= \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\mathbb{E}\left[\prod_{\ell=1}^{\#\mathrm{CC}_\Delta} \prod_{(i,j)\in E_{\Delta,\ell}} \mathbf{1}_{z_i=z_j} \prod_{k=1}^{K} \prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j=k}} \left(\frac{\bar{p}}{\bar{q}}\right)\right] \\
&= \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM} \in [K]^{V_{PM}}} \mathbb{P}\left[z \sim (\varphi_\Delta, \varphi_{PM})\right] \prod_{k=1}^{K} \left(\frac{\bar{p}}{\bar{q}}\right)^{\left|E_\cap\left[\varphi_{SM}^{-1}(\{k\}) \cup \varphi_{PM}^{-1}(\{k\})\right]\right|},
\end{aligned}$$

where, for a given $\varphi_\Delta : [\#\mathrm{CC}_\Delta] \to [K]$, the assignment $\varphi_{SM} : V^\star_{SM} \to [K]$ is defined by

$$\varphi_{SM}(i) = \varphi_\Delta(\ell) \text{ for } i \in V^\star_{SM} \cap V^\star_{\Delta,\ell}, \qquad (89)$$

and where

$$\begin{aligned}
\mathbb{P}\left[z \sim (\varphi_\Delta, \varphi_{PM})\right] &:= \mathbb{P}\left[z_i = \varphi_\Delta(\ell), \text{for all } i \in V^\star_{\Delta,\ell} \text{ and } z_i = \varphi_{PM}(i), \text{ for all } i \in V^\star_{PM}\right] \\
&= K^{-|V^\star_\Delta| - |V^\star_{PM}|} = K^{-|V^\star_\cup|}.
\end{aligned}$$

So

$$\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right] = \frac{\lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}}{K^{|V_\cup^\star|}} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V_{PM}^\star}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|}, \tag{90}$$

We have in particular for $j = 1, 2$

$$\mathbb{E}\left[P^2_{G^{(j)},\pi^{(j)}}\right] = \frac{\bar{q}^{|E^{(j)}|}}{K^{|V^{\star(j)}|}} \sum_{\varphi_j\in[K]^{V^{\star(j)}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E^{(j)}\left[\varphi_j^{-1}(\{k\})\right]\right|}. \tag{91}$$

To lower bound $\mathbb{E}[P^2_{G^{(j)},\pi^{(j)}}]$, we consider two different subfamilies of assignments $\varphi_j : V^{\star(j)} \to [K]$ in the sum of (91). For some given $\varphi_\Delta : [\#\mathrm{CC}_\Delta] \to [K]$ and some $\varphi_{PM} : V^\star_{PM} \to [K]$, we consider

1. φ_j defined by $\varphi_j(i) = \varphi_\Delta(\ell)$, for all $i \in V^\star_{\Delta,\ell}$; $\varphi_j(i) = \varphi_{PM}(i)$, for all $i \in V^\star_{PM}$;
2. φ_j defined by $\varphi_j(i) = \varphi_\Delta(\ell)$, for all $i \in V^\star_{\Delta,\ell} \cap V^\star_{SM}$; $\varphi_j(i) = \varphi_{PM}(i)$, for all $i \in V^\star_{PM}$, and $\varphi_j(i) = \varphi_{U^{\star(j)}}(i)$ for $i \in U^{\star(j)}$, where $\varphi_{U^{\star(j)}} \in [K]^{U^{\star(j)}}$.

First lower bound. Let us consider some $\varphi_\Delta : [\#\mathrm{CC}_\Delta] \to [K]$ and $\varphi_{PM} : V^\star_{PM} \to [K]$. Let us restrict to assignments $\varphi_j : V^{\star(j)} \to [K]$ defined by $\varphi_j(i) = \varphi_\Delta(\ell)$, for all $i \in V^\star_{\Delta,\ell}$; and $\varphi_j(i) = \varphi_{PM}(i)$, for all $i \in V^\star_{PM}$. We observe that $\varphi_j^{-1}(\{k\}) \cap (V^\star_{SM} \cup V^\star_{PM}) = \varphi_{SM}^{-1}(\{k\}) \cup \varphi_{PM}^{-1}(\{k\})$, so

$$\left|E^{(j)}\left[\varphi_j^{-1}(\{k\})\right]\right| = \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\}) \cup \varphi_{PM}^{-1}(\{k\})\right]\right| + \left|\left(E_\Delta \cap E^{(j)}\right)\left[\varphi_j^{-1}(\{k\}) \cap \left(V^\star_{SM} \cup U^{\star(j)}\right)\right]\right|, \tag{92}$$

with

$$\begin{aligned}\sum_{k=1}^K \left|\left(E_\Delta \cap E^{(j)}\right)\left[\varphi_j^{-1}(\{k\}) \cap \left(V^\star_{SM} \cup U^{\star(j)}\right)\right]\right| &= \sum_{\ell=1}^{\#\mathrm{CC}_\Delta}\sum_{k=1}^K \left|\left(E_{\Delta,\ell} \cap E^{(j)}\right)\left[\varphi_j^{-1}(\{k\}) \cap \left(V^\star_{SM} \cup U^{\star(j)}\right)\right]\right| \\ &= \sum_{\ell=1}^{\#\mathrm{CC}_\Delta} \left|\left(E_{\Delta,\ell} \cap E^{(j)}\right)\right| = \left|E_\Delta \cap E^{(j)}\right|,\end{aligned} \tag{93}$$

since φ_j is constant on each $V^\star_{\Delta,\ell}$. So, we can lower-bound (91) by

$$\mathbb{E}\left[P^2_{G^{(j)},\pi^{(j)}}\right] \geq \frac{\bar{q}^{|E^{(j)}|}}{K^{|V^{\star(j)}|}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\left|E_\Delta \cap E^{(j)}\right|} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V^\star_{PM}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|}.$$

Since $|E_\Delta| = |E_\Delta \cap E^{(1)}| + |E_\Delta \cap E^{(2)}|$ and $2|E_\cap| = |E^{(1)}| + |E^{(2)}| - |E_\Delta \cap E^{(1)}| - |E_\Delta \cap E^{(2)}|$, we get a first lower-bound

$$\prod_{j=1,2} \mathbb{E}\left[P^2_{G^{(j)},\pi^{(j)}}\right]^{1/2} \geq \frac{\bar{q}^{|E_\cap|}\,\bar{p}^{|E_\Delta|/2}}{K^{\left(|V^{\star(1)}|+|V^{\star(2)}|\right)/2}} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V^\star_{PM}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|}$$
$$= \frac{\bar{q}^{|E_\cap|}\,\bar{p}^{|E_\Delta|/2}}{K^{|V^\star_\cup|-\left(|U^{\star(1)}|+|U^{\star(2)}|\right)/2}} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V^\star_{PM}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|}. \tag{94}$$

Second lower bound. Let us consider again some $\varphi_\Delta : [\#\mathrm{CC}_\Delta] \to [K]$, $\varphi_{PM} : V^\star_{PM} \to [K]$ and $\varphi_{U^{\star(j)}} : U^{\star(j)} \to [K]$. We now restrict to assignments $\varphi_j : V^{*(j)} \to [K]$ defined by $\varphi_j(i) = \varphi_\Delta(\ell)$, for all $i \in V^*_{\Delta,\ell} \cap V^\star_{SM}$; $\varphi_j(i) = \varphi_{PM}(i)$, for all $i \in V^\star_{PM}$, and $\varphi_j(i) = \varphi_{U^{\star(j)}}(i)$ for $i \in U^{\star(j)}$. We still have the decomposition (92), but we do not have the identity (93) anymore. Yet, we have $\left|E^{(j)}\left[\varphi_j^{-1}(\{k\})\right]\right| \geq \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\}) \cup \varphi_{PM}^{-1}(\{k\})\right]\right|$. Since $\mathbf{M} \in \mathcal{M}^\star$, no connected component of G_Δ is only composed of nodes from $U^{\star(j)}$. Since $\bar{p} \geq \bar{q}$ when $q \leq 1/2$, we can lower-bound (91) by

$$\mathbb{E}\left[P^2_{G^{(j)},\pi^{(j)}}\right] \geq \frac{\bar{q}^{|E^{(j)}|}}{K^{|V^{\star(j)}|}} \sum_{\varphi_{U^{\star(j)}}\in[K]^{U^{\star(j)}}} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V^\star_{PM}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|}$$
$$\geq \frac{\bar{q}^{|E^{(j)}|}}{K^{|V^{\star(j)}|-|U^{\star(j)}|}} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V^\star_{PM}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|},$$

and get a second lower bound

$$\prod_{j=1,2} \mathbb{E}\left[P^2_{G^{(j)},\pi^{(j)}}\right]^{1/2} \geq \frac{\bar{q}^{(|E^{(1)}|+|E^{(2)}|)/2}}{K^{\left(|V^{\star(1)}|+|V^{\star(2)}|-|U^{\star(1)}|-|U^{\star(2)}|\right)/2}}$$
$$\times \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V^\star_{PM}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|}$$
$$= \frac{\bar{q}^{|E_\cap|+|E_\Delta|/2}}{K^{|V^\star_\cup|-|U^{\star(1)}|-|U^{\star(2)}|}} \sum_{\varphi_\Delta \in [K]^{\#\mathrm{CC}_\Delta}} \sum_{\varphi_{PM}\in[K]^{V^\star_{PM}}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\sum_{k=1}^K \left|E_\cap\left[\varphi_{SM}^{-1}(\{k\})\cup\varphi_{PM}^{-1}(\{k\})\right]\right|}. \tag{95}$$

Conclusion. Combining (90), (94) and (95), we then conclude that

$$\frac{\mathbb{E}\left[P_{G^{(1)},\pi^{(1)}}P_{G^{(2)},\pi^{(2)}}\right]}{\prod_{j=1,2}\mathbb{E}\left[P^2_{G^{(j)},\pi^{(j)}}\right]^{1/2}} \leq \left(\frac{\lambda^{|E_\Delta|}}{\bar{p}^{|E_\Delta|/2}K^{(|U^{\star(1)}|+|U^{\star(2)}|)/2}}\right) \wedge \left(\frac{\lambda^{|E_\Delta|}}{\bar{q}^{|E_\Delta|/2}K^{|U^{\star(1)}|+|U^{\star(2)}|}}\right)$$
$$\leq \left(\frac{\lambda^{|E_\Delta|/2}}{K^{(|U^{\star(1)}|+|U^{\star(2)}|)/2}}\right) \wedge \left(\frac{\lambda^{|E_\Delta|}}{\bar{q}^{|E_\Delta|/2}K^{|U^{\star(1)}|+|U^{\star(2)}|}}\right),$$

where the last line follows from $\bar{p} = \lambda + q(1-q-2\lambda) \geq \lambda$ when $q + 2\lambda \leq 1$. Since $|U^{\star(1)}| + |U^{\star(2)}| \geq |U^{(1)}| + |U^{(2)}|$, this concludes the proof of (74).

Let us turn to (75). Since G is connected, neither v_1 nor v_2 are isolated. We know that

$$\mathbb{E}[P_{G,\pi}] = \frac{\lambda^{|E|}}{K^{|V|-1}} .$$

Then, we lower bound the second moment $\mathbb{E}[P_{G,\pi}^2]$ by relying on (91) and considering two subfamilies of assignments as previously: one subfamily where the z_i's are identical on the vertices of G and one where the z_i's are let arbitrary. This respectively leads us to $\mathbb{E}[P_{G,\pi}^2] \geq \bar{p}^{|E|}K^{-|V|+1}$ and $\mathbb{E}[P_{G,\pi}^2] \geq \bar{q}^{|E|}$. Hence

$$\begin{aligned} 0 \leq \frac{\mathbb{E}\left[P_{G,\pi}\right]}{\mathbb{E}\left[P_{G,\pi}^2\right]^{1/2}} &\leq \left(\frac{\lambda^{|E|}}{\bar{p}^{|E|/2}K^{(|V|-1)/2}}\right) \wedge \left(\frac{\lambda^{|E|}}{\bar{q}^{|E|/2}K^{|V|-1}}\right) \\ &\leq \left(\frac{\lambda^{|E|/2}}{K^{(|V|-1)/2}}\right) \wedge \left(\frac{\lambda^{|E|}}{\bar{q}^{|E|/2}K^{(|V|-1)}}\right) \quad , \end{aligned}$$

where we used again that $\bar{p} \geq \lambda$. This concludes the proof of Proposition 22.

Appendix F. Expectation and variance of m-clique counting: proof of Lemma 8

Let us prove Lemma 8. Without loss of generality, we assume henceforth that $(i,j) = (1,2)$. We recall that $G = (V,E)$ is a clique on $V = \{v_1, \ldots, v_m\}$, where we have removed the edge (v_1, v_2), and

$$S_{12} = \sum_{\pi \in \Pi_{1,2}} P_{G,\pi}(Y),$$

where $\Pi_{1,2}$ is the set of injections π from V to $\{1 \ldots, n\}$ such that $\pi(v_1) = 1$ and $\pi(v_2) = 2$. We will repeatedly use the identity

$$|E| = \frac{m(m-1)}{2} - 1 = \frac{(m+1)(m-2)}{2}.$$

We first compute the mean and upper bound the variance under $\mathbb{P}_{12} = \mathbb{P}\left[\cdot | z_1 = z_2\right]$, and then under $\mathbb{P}_{\not{1}2} = \mathbb{P}\left[\cdot | z_1 \neq z_2\right]$.

Mean under $\mathbb{P}_{12}$. We have

$$\begin{aligned} \mathbb{E}_{12}\left[P_{G,\pi}(Y)\right] &= \lambda^{|E|}\, \mathbb{E}_{12}\left[\prod_{(i,j)\in\pi(E)} \mathbf{1}_{z_i = z_j}\right] \\ &= \lambda^{\frac{(m+1)(m-2)}{2}}\, \mathbb{P}_{12}\left[z_i = z_1 : i = 3, \ldots, m\right] = \frac{\lambda^{\frac{(m+1)(m-2)}{2}}}{K^{m-2}}. \end{aligned}$$

Hence

$$\mathbb{E}_{12}\left[S_{12}\right] = \frac{(n-2)!}{(n-m)!}\left(\frac{\lambda^{\frac{m+1}{2}}}{K}\right)^{m-2} .$$

Variance under $\mathbb{P}_{12}$. Let $\pi^{(1)}, \pi^{(2)} \in \Pi_{12}$ such that $|\text{range}(\pi^{(1)}) \cap \text{range}(\pi^{(2)})| = 2 + u$, with $u \in \{0, \ldots, m-2\}$. Following (88), we have

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] = \lambda^{|E_\Delta|} \bar{q}^{|E_\cap|} \mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta} \mathbf{1}_{z_i = z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i = z_j}} \frac{\bar{p}}{\bar{q}}\right].$$

Case $u = 0$. For $u = 0$, we have $E_\cap = \emptyset$. We also have $|V_\Delta| = |V_\cup| = 2m-2$ and $|E_\Delta| = 2|E| = (m+1)(m-2)$, so, when $u = 0$,

$$\begin{aligned} \mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] &= \lambda^{(m+1)(m-2)} \mathbb{P}_{12}\left[z_i = z_1 : i = 3, \ldots, 2m-2\right] \\ &= \frac{\lambda^{(m+1)(m-2)}}{K^{2m-4}} = \mathbb{E}_{12}\left[P_{G,\pi^{(1)}}\right] \mathbb{E}_{12}\left[P_{G,\pi^{(2)}}\right]. \end{aligned} \tag{96}$$

Case $1 \le u \le m-3$. We now have $|V_\Delta| = |V_\cup| = 2m - 2 - u$, $|E_\cap| = (u+3)u/2$, and $|E_\Delta| = 2(|E| - |E_\cap|) = (m+1)(m-2) - (u+3)u$. Hence

$$\begin{aligned} \mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] &= \lambda^{(m+1)(m-2)-(u+3)u} \bar{q}^{\frac{(u+3)u}{2}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\frac{(u+3)u}{2}} \mathbb{P}_{12}\left[z_i = z_1 : i = 3, \ldots, 2m-2-u\right] \\ &= \frac{\lambda^{(m+1)(m-2)}}{K^{2m-4}} \left(\frac{\bar{p}}{\lambda^2}\right)^{\frac{(u+3)u}{2}} K^u \le \frac{\bar{p}^{(m+1)(m-2) - \frac{(u+3)u}{2}}}{K^{2m-4}} K^u\ , \end{aligned} \tag{97}$$

since $\bar{p} = \bar{q} + \lambda(1-2q) \ge \lambda$ when $2\lambda + q \le 1$.
Case $u = m-2$. When $u = m-2$ then $\text{range}(\pi^{(1)}) = \text{range}(\pi^{(2)})$ and $E_\Delta = \emptyset$. We define $V_k(z) := \{i \in \{1, \ldots, m\} : z_i = k\}$. We have

$$\begin{aligned} \mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] &= \bar{q}^{|E|} \mathbb{E}_{12}\left[\prod_{k=1}^{K} \prod_{\substack{(i,j)\in\pi^{(1)}(E) \\ z_i = z_j = k}} \frac{\bar{p}}{\bar{q}}\right] = \bar{q}^{\frac{(m+1)(m-2)}{2}} \mathbb{E}_{12}\left[\prod_{k=1}^{K} \prod_{\substack{1\le i<j\le m \\ (i,j)\ne(1,2) \\ z_i = z_j = k}} \frac{\bar{p}}{\bar{q}}\right] \\ &= \bar{q}^{\frac{(m+1)(m-2)}{2}} \mathbb{E}_{12}\left[\left(\frac{\bar{p}}{\bar{q}}\right)^{\frac{1}{2}\sum_{k=1}^{K} |V_k(z)|(|V_k(z)|-1) - 1}\right] \\ &= \bar{q}^{\frac{(m+1)(m-2)}{2}} \mathbb{E}_{12}\left[\left(\frac{\bar{p}}{\bar{q}}\right)^{\frac{1}{2}\sum_{k=1}^{K} |V_k(z)|^2 - \frac{m+2}{2}}\right]. \end{aligned}$$

Let $\ell(z) := |\{k : V_k(z) \neq \emptyset\}|$. We have $\ell(z) \le m-1$ under $\mathbb{P}_{12}$, since $z_1 = z_2$ a.s. For $\ell(z) = \ell^\star$, the sum $\sum_{k=1}^{K} |V_k(z)|^2$ is maximized when $\ell^\star - 1$ groups have 1 node and the remaining group has $m - (\ell^\star - 1)$ nodes, so

$$\sum_{k=1}^{K} |V_k(z)|^2 \le \ell(z) - 1 + (m - (\ell(z) - 1))^2 = m^2 - (\ell(z) - 1)(2m - \ell(z)).$$

Since $\bar{p} \geq \bar{q}$ (since $q \leq 1/2$), we deduce

$$\begin{aligned}
\mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] &\leq \bar{q}^{\frac{(m+1)(m-2)}{2}} \left(\frac{\bar{p}}{\bar{q}}\right)^{\frac{(m+1)(m-2)}{2}} \mathbb{E}_{12}\left[\left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{(\ell(z)-1)(2m-\ell(z))}{2}}\right] \\
&\leq \bar{p}^{\frac{(m+1)(m-2)}{2}} \mathbb{E}_{12}\left[\left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell(z)-1)\frac{(m+1)}{2}}\right] \\
&\leq \bar{p}^{\frac{(m+1)(m-2)}{2}} \sum_{\ell=1}^{m-1} \left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell-1)\frac{(m+1)}{2}} \mathbb{P}_{12}\left[\ell(z)=\ell\right]. \qquad (98)
\end{aligned}$$

We observe that, for $\ell \leq m-1 \leq K$,

$$\mathbb{P}_{12}\left[\ell(z)=\ell\right] \leq \binom{m-2}{\ell-1} \frac{K(K-1)\cdots(K-\ell+1)}{K^\ell} \left(\frac{\ell}{K}\right)^{m-2-(\ell-1)} \leq \binom{m-2}{\ell-1} \left(\frac{m-2}{K}\right)^{m-2-(\ell-1)},$$

so plugging in (98), we get

$$\begin{aligned}
\mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] &\leq \bar{p}^{\frac{(m+1)(m-2)}{2}} \sum_{\ell=1}^{m-1} \binom{m-2}{\ell-1} \left(\frac{m-2}{K}\right)^{m-2-(\ell-1)} \left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell-1)\frac{(m+1)}{2}} \\
&\leq \bar{p}^{\frac{(m+1)(m-2)}{2}} \left(\frac{m-2}{K} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2}. \qquad (99)
\end{aligned}$$

<u>Combining the three terms.</u> Combining (96),(97), and (99)

$$\begin{aligned}
\mathrm{var}_{12}(S_{12}) &= \sum_{u=0}^{m-2} \sum_{\substack{\pi^{(1)},\pi^{(2)} \in \Pi_{12} \\ |\mathrm{range}(\pi^{(1)}) \cap \mathrm{range}(\pi^{(2)})| = 2+u}} \left(\mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] - \mathbb{E}_{12}\left[P_{G,\pi^{(1)}}\right] \mathbb{E}_{12}\left[P_{G,\pi^{(2)}}\right]\right) \\
&\leq \sum_{u=1}^{m-2} \binom{m-2}{u}^2 \frac{(n-2)!u!}{(n-2m+u+2)!} \mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \\
&\leq \frac{(n-2)!(m-2)!}{(n-m)!} \bar{p}^{\frac{(m+1)(m-2)}{2}} \left(\frac{m-2}{K} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2} \\
&\quad + \sum_{u=1}^{m-3} \binom{m-2}{u}^2 \frac{(n-2)!u!}{(n-2m+u+2)!} \frac{\bar{p}^{(m+1)(m-2) - \frac{(u+3)u}{2}}}{K^{2m-4-u}},
\end{aligned}$$

so

$$\begin{aligned}
\mathrm{var}_{12}(S_{12}) &\leq \frac{(n-2)!(m-2)!}{(n-m)!} \bar{p}^{\frac{(m+1)(m-2)}{2}} \left(\frac{m-2}{K} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2} \\
&\quad + \frac{(n-2)!(m-2)!}{(n-m)!} \frac{\bar{p}^{(m+1)(m-2)}}{K^{2m-4}} \sum_{u=1}^{m-3} \binom{m-2}{u} \frac{(n-m)!}{(m-2-u)!(n-2m+u+2)!} \left(\frac{K}{\bar{p}^{\frac{m}{2}}}\right)^u.
\end{aligned}$$

We have

$$\sum_{u=1}^{m-3}\binom{m-2}{u}\frac{(n-m)!}{(m-2-u)!(n-2m+u+2)!}\left(\frac{K}{\bar{p}^{\frac{m+1}{2}}}\right)^{u} \leq (n-m)^{m-2}\sum_{u=1}^{m-3}\binom{m-2}{u}\left(\frac{K}{(n-m)\bar{p}^{\frac{m+1}{2}}}\right)^{u} \leq \left(n-m+\frac{K}{\bar{p}^{\frac{m+1}{2}}}\right)^{m-2}.$$

So,

$$\mathrm{var}_{12}(S_{12}) \leq \frac{(n-2)!(m-2)!}{(n-m)!}\bar{p}^{\frac{(m+1)(m-2)}{2}}\left(\frac{m-2}{K}+\left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2} + \frac{(n-2)!(m-2)!}{(n-m)!}\frac{\bar{p}^{(m+1)(m-2)}}{K^{2m-4}}\left(n-m+\frac{K}{\bar{p}^{\frac{m+1}{2}}}\right)^{m-2}\bar{p}^{1/2}.$$

We have proved (36).

Mean under $\mathbb{P}_{\not 12}$. We have

$$\mathbb{E}_{\not 12}\left[P_{G,\pi}(Y)\right] = \lambda^{|E|}\,\mathbb{E}_{\not 12}\left[\prod_{(i,j)\in\pi(E)}\mathbf{1}_{z_i=z_j}\right] = 0,$$

since $z_1 \neq z_2$ a.s. under $\mathbb{P}_{\not 12}$.

Variance under $\mathbb{P}_{\not 12}$. Let $\pi^{(1)},\pi^{(2)} \in \Pi_{12}$ such that $|\mathrm{range}(\pi^{(1)}) \cap \mathrm{range}(\pi^{(2)})| = 2+u$, with $u \in \{0,\dots,m-2\}$. Following (88), we have

$$\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] = \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\mathbb{E}_{\not 12}\left[\prod_{(i,j)\in E_\Delta}\mathbf{1}_{z_i=z_j}\prod_{\substack{(i,j)\in E_\cap\\ z_i=z_j}}\frac{\bar{p}}{\bar{q}}\right].$$

<u>Case $0 \leq u \leq m-3$.</u> We have $1,2 \in V_\Delta$ and $|E_\Delta| = (m+1)(m-2)-(u+3)u \geq 4$, so $\prod_{(i,j)\in E_\Delta}\mathbf{1}_{z_i=z_j} = 0$ a.s. under $\mathbb{P}_{\not 12}$ since $z_1 \neq z_2$. Hence $\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] = 0$ in this case.

<u>Case $u = m-2$.</u> When $u = m-2$ then $\mathrm{range}(\pi^{(1)}) = \mathrm{range}(\pi^{(2)})$ and $E_\Delta = \emptyset$. Following the same lines as for deriving (98), we get

$$\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] = \bar{q}^{\frac{(m+1)(m-2)}{2}}\mathbb{E}_{\not 12}\left[\left(\frac{\bar{p}}{\bar{q}}\right)^{\frac{1}{2}\sum_{k=1}^{K}|V_k(z)|^2-\frac{m}{2}}\right] \leq \bar{p}^{\frac{(m+1)(m-2)}{2}}\,\mathbb{E}_{\not 12}\left[\left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{(\ell(z)-1)(2m-\ell(z))}{2}-1}\right].$$

We have $2 \leq \ell(z) \leq m$ a.s. under $\mathbb{P}_{\not 12}$, since $z_1 \neq z_2$ a.s. Hence

$$\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] \leq \bar{p}^{\frac{(m+1)(m-2)}{2}}\sum_{\ell=2}^{m-1}\left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell-1)\frac{(m+1)}{2}-1}\mathbb{P}_{\not 12}\left[\ell(z)=\ell\right] + \bar{p}^{\frac{(m+1)(m-2)}{2}}\left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{(m-1)m}{2}-1}\mathbb{P}_{\not 12}\left[\ell(z)=m\right]$$

Arguing as in the previous variance bound, we observe that, for $\ell \leq m \leq K$, we have

$$\mathbb{P}_{\not{1}2}\left[\ell(z)=\ell\right] \leq \binom{m-2}{\ell-2}\left(\frac{m-2}{K}\right)^{m-2-(\ell-2)},$$

so we get

$$\begin{aligned}
&\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] \\
&\quad \leq \bar{p}^{\frac{(m+1)(m-2)}{2}} \sum_{\ell=2}^{m-1} \binom{m-2}{\ell-2}\left(\frac{m-2}{K}\right)^{m-2-(\ell-2)} \left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell-1)\frac{(m+1)}{2}-1} + \bar{p}^{\frac{(m+1)(m-2)}{2}} \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{(m-2)(m+1)}{2}} \\
&\quad \leq \bar{p}^{\frac{(m+1)(m-2)}{2}} \left(\frac{m-2}{K} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2} \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m-1}{2}} + \bar{p}^{\frac{(m+1)(m-2)}{2}} \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{(m-2)(m+1)}{2}}.
\end{aligned}$$

<u>Final bound on the variance.</u>

$$\begin{aligned}
\mathrm{var}_{\not{1}2}(S_{12}) &= \sum_{\substack{\pi^{(1)},\pi^{(2)}\in\Pi_{12} \\ \mathrm{range}(\pi^{(1)})=\mathrm{range}(\pi^{(2)})}} \left(\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] - \mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}}\right]\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(2)}}\right]\right) \\
&\leq \frac{(n-2)!(m-2)!}{(n-m)!}\bar{p}^{\frac{(m+1)(m-2)}{2}} \left(\left(\frac{m-2}{K} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m+1}{2}}\right)^{m-2} \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{m-1}{2}} + \left(\frac{\bar{q}}{\bar{p}}\right)^{\frac{(m-2)(m+1)}{2}}\right).
\end{aligned}$$

We have proved (37).

Appendix G. Counting self-avoiding paths: Proof of Lemma 14

G.1. Results under the distribution $\mathbb{P}_{12}$

Mean under $\mathbb{P}_{12}$. We have

$$\mathbb{E}_{12}\left[P_{G,\pi}(Y)\right] = \lambda^{|E|}\,\mathbb{E}_{12}\left[\prod_{(i,j)\in\pi(E)} \mathbf{1}_{z_i=z_j}\right] = \frac{1}{K^{|V|-2}}\lambda^{|E|} = \frac{1}{K^{m-2}}\lambda^{m-1},$$

since $z_1 = z_2$ a.s. under $\mathbb{P}_{12}$.

Variance under $\mathbb{P}_{12}$. As for the proof of Lemma 8, we consider $\pi^{(1)}, \pi^{(2)} \in \Pi_{12}$ such that $|\mathrm{range}(\pi^{(1)}) \cap \mathrm{range}(\pi^{(2)})| = 2 + u$, with $u \in \{0, \dots, m-2\}$. Following (88), we have

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] = \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta} \mathbf{1}_{z_i=z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}}\right].$$

Case $u = 0$. For $u = 0$, we have $E_\cap = \emptyset$. We also have $|V_\Delta| = |V_\cup| = 2m-2$ and $|E_\Delta| = 2(m-1)$, so, when $u = 0$,

$$\begin{aligned}\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] &= \lambda^{2(m-1)}\mathbb{P}_{12}\left[z_i = z_1 : i = 3,\dots,2m-2\right] \\ &= \frac{\lambda^{2(m-1)}}{K^{2m-4}} = \mathbb{E}_{12}\left[P_{G,\pi^{(1)}}\right]\mathbb{E}_{12}\left[P_{G,\pi^{(2)}}\right] \quad . \end{aligned} \tag{100}$$

Case $1 \le u \le m-3$. We now have $|V_\cup| = 2m-2-u$. By definition of u, there are $m-u-2$ unmatched nodes in $\pi^{(2)}(V) \setminus \pi^{(1)}(V)$. Note that all the nodes in $\pi^{(2)}(V) \setminus \pi^{(1)}(V)$ belong to G_Δ. Since the line graph G is connected, all the nodes in $\pi^{(2)}(V) \setminus \pi^{(1)}(V)$ are connected through G_Δ to a node in $\pi^{(1)}(V) \cap \pi^{(2)}(V)$. Define $E_\Delta^{(2)} \subset E_\Delta$ as the subset of edges (i,j) such that either $i \in \pi^{(2)}(V) \setminus \pi^{(1)}(V)$ or $j \in \pi^{(2)}(V) \setminus \pi^{(1)}(V)$. Then, we have

$$\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta^{(2)}} \mathbf{1}_{z_i=z_j} | z_{\pi^{(1)}(V)}\right] \le \frac{1}{K^{m-u-2}} \ ,$$

almost surely. It then follows that

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] \le \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\frac{1}{K^{m-u-2}}\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta\setminus E_\Delta^{(2)}} \mathbf{1}_{z_i=z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}}\right] .$$

Then, we consider $E_\Delta^{(1)} = E_\Delta \cap (\pi^{(1)}(V) \times \pi^{(1)}(V)) \subset E_\Delta \setminus E_\Delta^{(2)}$ as the subset of edges that only arise through the labeling $\pi^{(1)}$ of G. We have

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] \le \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\frac{1}{K^{m-u-2}}\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}}\right] . \tag{101}$$

Note that $E_\Delta^{(1)}$ and $E_\cap$ form a partition of the edges of the labelled line graph on $\pi^{(1)}(V)$. If $E_\cap = \emptyset$, we arrive at

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] \le \lambda^{|E_\Delta|}\frac{1}{K^{2m-u-4}} = \lambda^{2(m-1)}\frac{1}{K^{2m-u-4}} \quad .$$

If $E_\cap \neq \emptyset$, then the graph $G_\cap$ is a collection of $t \ge 1$ connected components with $|V_\cap| = |E_\cap| + t \ge 2$ nodes in total. The term $\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j}$ is not equal to 0 if and only if some constraints on the communities z_i of the nodes are satisfied. Namely, this sets the constraint that the nodes from each connected component of $G_\Delta^{(1)}$ must belong to the same community. Also, by definition of $\mathbb{P}_{12}$, we have $z_1 = z_2$. Furthermore, by invariance of the choice of the communities, we can assume that z_1 is fixed. By considering separately the cases where $V_\cap \cap \{1,2\} = \emptyset, \{1\}, \{2\}, \{1,2\}$, we deduce that $m-2-|V_\cap|+t+1 = m-1-|E_\cap|$ nodes in G have their communities fixed by the constraints, and only $|E_\cap|-1$ can still be assigned freely - namely their label can still be chosen freely without setting the term $\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j}$ to 0. This set of nodes - that we write V_u - is composed of (i) exactly one node pro connected component in the graph $G_\Delta^{(1)}$ that one can choose freely and (ii) of all perfectly matched nodes, i.e. the nodes that are not in $G_\Delta^{(1)}$, except v_1 and v_2. This yields

$$\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}}\right] \leq \left(\frac{1}{K}\right)^{m-1-|E_\cap|} \mathbb{E}_{12}\left[\prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}} \;\Big|\; \prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} = 1\right] .$$

Write $E_\cap^= = \{(i,j) \in E_\cap : z_i = z_j\}$. Now we consider the labels of the $|E_\cap| - 1$ nodes in V_u that have their labels unconstrained. We order them as $(v'_1, \ldots, v'_{|E_\cap|-1})$ according to their order in the path $v_1, v_3, \ldots, v_m, v_2$ so that either (i) node v'_i is connected to a connected component of $G_\Delta^{(1)}$ to which v'_{i-1} belongs, or (ii) v'_i is connected to v'_{i-1} (with the convention $v'_0 = v_1$). The communities of these nodes are sampled independently which leads to the set $E_\cap^=$ of identical edges. If the community of v'_i is the same as that of v'_{i-1} (resp. of v_1 for $i = 1$) - which happens with probability $1/K$ - then one edge is added in $E_\cap^=$ - namely the edge connecting v'_i with either the neighboring connected component represented by v'_{i-1}, or with v'_{i-1} itself. Finally, the last edge in $E_\cap$ belongs to $E_\cap^=$, if $v'_{|E_\cap|-1}$ and v_2 (and therefore v_1) belong to the same community. As a consequence, we can decompose $E_\cap^=$ as $E_\cap^= = N_1 + N_2$ where $N_1 \sim \mathrm{Bin}(|E_\cap| - 1, 1/K)$ and $N_2 \in \{0,1\} = \mathbf{1}\{z_{v'_{|E_\cap|-1}} = z_{v_1}\}$. One easily checks that $\mathbb{P}_{12}[N_2 = 1|N_1 = |E_\cap| - 1] = 1$ and that $\mathbb{P}_{12}[N_2 = 1|N_1 = |E_\cap| - 2] = 0$. Next, for any $a \in \{1, \ldots, |E_\cap| - 3\}$, conditionally to $N_1 = a$, we have $\mathbb{P}_{12}[N_2 = 1|N_1 = |E_\cap| - 2] \leq 1/(K-1)$. Indeed, if the community of second-to last node is the same as that of v_1, then this probability is equal to 0, whereas, if the community of second-to last node differs from that of v_1, this probability is equal to $1/(K-1)$. As a consequence, we get

$$\mathbb{E}_{12}\left[\prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}} \;\Big|\; \prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} = 1\right] = \mathbb{E}_{12}\left[\left(\frac{\bar{p}}{\bar{q}}\right)^{|E_\cap^=|} \;\Big|\; \prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} = 1\right] \leq \mathbb{E}\left[\left(\frac{\bar{p}}{\bar{q}}\right)^{N_1+N_2}\right].$$

We have to bound the exponential moment of $N_1 + N_2$. With probability at most $(K-1)^{-(|E_\cap|-1)}$, $N_1 + N_2$ is equal to $|E_\cap|$. Conditionally to $N_1 < |E_\cap| - 1$, $N_1 + N_2$ is stochastically upper bounded by a Binomial distribution with parameters $|E_\cap|$ and $1/(K-1)$. For $b \in \mathbb{R}$, we therefore derive that $\mathbb{E}[e^{a(N_1+N_2)}] \leq \mathbb{E}[e^{bN_3}] + (K-1)^{-(|E_\cap|-1)}e^{b|E_\cap|}$, where $N_3 \sim \mathrm{Bin}(|E_\cap|, 1/(K-1))$. As a consequence, we obtain

$$\mathbb{E}_{12}\left[\prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}} \;\Big|\; \prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} = 1\right] \leq \left[\left(\frac{1}{K}\right)^{|E_\cap|-1}\left(\frac{\bar{p}}{\bar{q}}\right)^{|E_\cap|} + \exp\left(|E_\cap|\log\left(\frac{\bar{p}}{(K-1)\bar{q}} + 1 - \frac{1}{K-1}\right)\right)\right] .$$

Coming back to (101), we get

$$\begin{aligned}\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] &\leq \lambda^{|E_\Delta|}\bar{p}^{|E_\cap|}\frac{1}{K^{2m-u-4}} + \lambda^{|E_\Delta|}\frac{1}{K^{2m-u-3}}\bar{q}^{|E_\cap|}\left(2\frac{\bar{p}}{\bar{q}} + K\right)^{|E_\cap|} \\ &\leq \frac{\lambda^{2(m-1)}}{K^{2m-4}}\left[K^u\left(\frac{\bar{p}}{\lambda^2}\right)^{|E_\cap|} + K^{u-1}\left[\left(4\frac{\bar{p}}{\lambda^2}\right)^{|E_\cap|} + \left[\frac{2K\bar{q}}{\lambda^2}\right]^{|E_\cap|}\right]\right] .\end{aligned}$$

Next, we observe that, if $u \leq m-3$, we have $|E_\cap| \in [1, u]$, whereas, for $u = m-2$, we have $|E_\cap| \in [1, m-1]$. This yields

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \leq \frac{\lambda^{2(m-1)}}{K^{2m-4}} \left[2\left(4K\frac{\bar{p}}{\lambda^2}\right)^u + \frac{1}{K}\left(\frac{2K^2\bar{q}}{\lambda^2}\right)^u + 3K^u\right] \quad , \tag{102}$$

for $u \leq m-3$ and

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \leq \frac{\lambda^{2(m-1)}}{K^{2m-4}} \left[\frac{2}{K}\left(\frac{4K\bar{p}}{\lambda^2}\right)^{m-1} + \frac{1}{K^2}\left(\frac{2K^2\bar{q}}{\lambda^2}\right)^{m-1} + 3K^{m-2}\right] \quad , \tag{103}$$

for $u = m-2$.

Combining the three terms. Combining (100), (102), and (103), we obtain

$$\begin{aligned}
\mathrm{var}_{12}(R_{12}) &= \sum_{u=0}^{m-2} \sum_{\substack{\pi^{(1)},\pi^{(2)} \in \Pi_{12} \\ |\mathrm{range}(\pi^{(1)}) \cap \mathrm{range}(\pi^{(2)})| = 2+u}} \left(\mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] - \mathbb{E}_{12}\left[P_{G,\pi^{(1)}}\right] \mathbb{E}_{12}\left[P_{G,\pi^{(2)}}\right]\right) \\
&\leq \sum_{u=1}^{m-2} \binom{m-2}{u}^2 \frac{(n-2)!u!}{(n-2m+u+2)!} \mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \\
&\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \sum_{u=1}^{m-2} (m-2)^{2u} \frac{1}{(n-m)^u} \mathbb{E}_{12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \\
&\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \frac{\lambda^{2(m-1)}}{K^{2m-4}} \left[\sum_{u=1}^{m-3} \left(2\left(4Km^2\frac{\bar{p}}{(n-m)\lambda^2}\right)^u + \frac{1}{K}\left(\frac{2m^2K^2\bar{q}}{(n-m)\lambda^2}\right)^u + 3\left(m^2\frac{K}{n-m}\right)^u\right)\right. \\
&\qquad \left. + m^{2m-4}\left(\frac{2(n-m)}{K}\left(\frac{4K\bar{p}}{(n-m)\lambda^2}\right)^{m-1} + \frac{n-m}{K^2}\left(\frac{2K^2\bar{q}}{(n-m)\lambda^2}\right)^{m-1} + 3\left(\frac{K}{n-m}\right)^{m-2}\right)\right] \\
&\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \frac{\lambda^{2(m-1)}}{K^{2m-4}} m \left[\frac{8Km^2\bar{p}}{(n-m)\lambda^2} + \frac{2m^2K\bar{q}}{(n-m)\lambda^2} + 3\frac{m^2K}{n-m} + 3\left(\frac{m^2K}{n-m}\right)^{m-2}\right. \\
&\qquad \left. + \frac{2nm^2}{K}\left(\frac{4Km^2\bar{p}}{(n-m)\lambda^2}\right)^{m-1} + \frac{nm^2}{K^2}\left(\frac{2m^2K^2\bar{q}}{(n-m)\lambda^2}\right)^{m-1}\right].
\end{aligned}$$

This concludes the proof of (55).

G.2. Results under the distribution $\mathbb{P}_{1\!\!/2}$

Mean under $\mathbb{P}_{1\!\!/2}$. We have

$$\mathbb{E}_{1\!\!/2}\left[P_{G,\pi}(Y)\right] = \lambda^{|E|}\, \mathbb{E}_{1\!\!/2}\left[\prod_{(i,j)\in\pi(E)} \mathbf{1}_{z_i = z_j}\right] = 0,$$

since $z_1 \neq z_2$ a.s. under $\mathbb{P}_{1\!\!/2}$.

Variance under $\mathbb{P}_{\not 12}$. Let $\pi^{(1)}, \pi^{(2)} \in \Pi_{12}$ such that $|\text{range}(\pi^{(1)}) \cap \text{range}(\pi^{(2)})| = 2 + u$, with $u \in \{0, \dots, m-2\}$. Following (88), we again have

$$\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] = \lambda^{|E_\Delta|} \bar{q}^{|E_\cap|} \mathbb{E}_{\not 12}\left[\prod_{(i,j)\in E_\Delta} \mathbf{1}_{z_i = z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i = z_j}} \frac{\bar{p}}{\bar{q}}\right].$$

If $u = 0$, we have $E_\cap = \emptyset$ and since $z_1 \neq z_2$, we deduce that $\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] = 0$.

Case $1 \leq u \leq m-2$. First, if $E_\cap = \emptyset$, we easily derive as above that $\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] = 0$. Hence, we assume henceforth that $|E_\emptyset| \geq 1$. Arguing as for $\mathbb{E}_{12}$ in the previous subsection, we derive from (101) that

$$\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \leq \lambda^{|E_\Delta|} \bar{q}^{|E_\cap|} \frac{1}{K^{m-u-2}} \mathbb{E}_{\not 12}\left[\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i = z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i = z_j}} \frac{\bar{p}}{\bar{q}}\right],$$

and we also have in a similar way as for $\mathbb{E}_{12}$ in the previous subsection

$$\mathbb{E}_{\not 12}\left[\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i = z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i = z_j}} \frac{\bar{p}}{\bar{q}}\right] \leq \left(\frac{1}{K}\right)^{m-1-|E_\cap|} \mathbb{E}\left[\left(\frac{\bar{p}}{\bar{q}}\right)^{N_1' + N_2'}\right],$$

where $N_1' \sim \text{Bin}(|E_\cap| - 1, 1/K)$ and N_2' satisfies $\mathbb{P}[N_2' = 1 | N_1 = |E_\cap| - 1] = 0$ (as now v_1 and v_2 do not have the same labels) and $\mathbb{P}[N_2' = 1 | N_1' < |E_\cap| - 1] \leq \frac{1}{K-1}$, so that $N_1' + N_2'$ is stochastically dominated by Binomial distribution $\text{Bin}(|E_\cap|, 1/(K-1))$. This yields

$$\begin{aligned}
\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] &\leq \lambda^{|E_\Delta|} \frac{1}{K^{2m-u-3}} \bar{q}^{|E_\cap|} \left(2\frac{\bar{p}}{\bar{q}} + K\right)^{|E_\cap|} \\
&\leq \frac{\lambda^{2(m-1)}}{K^{2m-4}} \cdot K^{u-1} \left[\left(4\frac{\bar{p}}{\lambda^2}\right)^{|E_\cap|} + \left[\frac{2K\bar{q}}{\lambda^2}\right]^{|E_\cap|}\right].
\end{aligned}$$

Recall that $|E_\cap| \in [1, u]$ if $u \leq m-3$ and $|E_\cap| \in [1, u+1]$ for $u = m-2$. This yields

$$\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \leq \frac{\lambda^{2(m-1)}}{K^{2m-4}} \left[\frac{1}{K}\left(4K\frac{\bar{p}}{\lambda^2}\right)^u + \frac{1}{K}\left(\frac{2K^2\bar{q}}{\lambda^2}\right)^u + 2K^{u-1}\right], \tag{104}$$

for $u \leq m-3$ and

$$\mathbb{E}_{\not 12}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \leq \frac{\lambda^{2(m-1)}}{K^{2m-4}} \left[\frac{1}{K^2}\left(\frac{4K\bar{p}}{\lambda^2}\right)^{m-1} + \frac{1}{K^2}\left(\frac{2K^2\bar{q}}{\lambda^2}\right)^{m-1} + 2K^{m-2}\right], \tag{105}$$

for $u = m-2$.

Final bound on the variance. Combining (104) and (105), we obtain

$$\begin{aligned}
\mathrm{var}_{\not{1}2}(T_{12}) &= \sum_{u=0}^{m-2} \sum_{\substack{\pi^{(1)},\pi^{(2)}\in\Pi_{12}\\ |\mathrm{range}(\pi^{(1)})\cap\mathrm{range}(\pi^{(2)})|=2+u}} \left(\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] - \mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}}\right]\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(2)}}\right]\right)\\
&\leq \sum_{u=1}^{m-2}\binom{m-2}{u}^2 \frac{(n-2)!u!}{(n-2m+u+2)!}\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right]\\
&\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \frac{\lambda^{2(m-1)}}{K^{2m-4}}\left[\sum_{u=1}^{m-3}\left(\frac{1}{K}\left(4Km^2\frac{\bar{p}}{(n-m)\lambda^2}\right)^u + \frac{1}{K}\left(\frac{2m^2K^2\bar{q}}{(n-m)\lambda^2}\right)^u + 2\left(m^2\frac{K}{n-m}\right)^u\right)\right.\\
&\qquad\left. + m^{2m-4}\left(\frac{n}{K^2}\left(\frac{4K\bar{p}}{(n-m)\lambda^2}\right)^{m-1} + \frac{n}{K^2}\left(\frac{2K^2\bar{q}}{(n-m)\lambda^2}\right)^{m-1} + 2\left(\frac{K}{n-m}\right)^{m-2}\right)\right]\\
&\leq \left(\frac{(n-2)!}{(n-m)!}\right)^2 \frac{\lambda^{2(m-1)}}{K^{2m-4}} m\left[\frac{4m^2\bar{p}}{(n-m)\lambda^2} + \frac{2m^2K\bar{q}}{(n-m)\lambda^2}\right.\\
&\qquad\left. + 2m^2\frac{K}{n-m} + 2\left(\frac{mK}{n-m}\right)^{m-2} + \frac{nm^2}{K}\left(\frac{4Km^2\bar{p}}{(n-m)\lambda^2}\right)^{m-1} + \frac{nm^2}{K^2}\left(\frac{2m^2K^2\bar{q}}{(n-m)\lambda^2}\right)^{m-1}\right].
\end{aligned}$$

We have proved (56).

Appendix H. Proof of Proposition 5

In this proof, we need additional notation. Let $V_{\mathrm{cyc},0}$ denote the set of cycle nodes having the same community as v_1 and v_2. For $k = 1, \ldots, I$ let $V_{\mathrm{cyc},k}$ be the set of cycle nodes of community k. For short, we write $V_{\mathrm{cyc},\neq 0} = \cup_{k=1}^{I} V_{\mathrm{cyc},k}$.

We consider separately cycle edges and fastener edges. For $k = 0, \ldots, I$, we write $E^{\uparrow}_{\mathrm{cyc},k} \subset E_{\mathrm{cyc}}$ for the collection of cycle edges that are incident to a node in $V_{\mathrm{cyc},k}$ and a node from a different community — these edges will be called *boundary* edges. For $k = 0, \ldots, I$, we define $V_{\mathrm{fst},k} := V_{\mathrm{cyc},k} \cap V_{\mathrm{fst}}$, the subset of fastener nodes within community k. Note that $|V_{\mathrm{fst},k}|$ also corresponds to the number of fastener edges within community k. For short, we also write $V_{\mathrm{fst},\neq 0} = \cup_{k=1}^{I} V_{\mathrm{fst},k}$.

First, we consider the specific and trivial case where $I = 1$ and $V_{\mathrm{cyc},1} = V_{\mathrm{cyc}}$. Then, we obviously have

$$|E^{\neq}| = |V_{\mathrm{fst}}| = a\kappa\gamma \geq \gamma + 1 \geq (\gamma + a)I\ ,$$

since $\kappa \geq 2/a$. We assume henceforth $V_{\mathrm{cyc},k} \neq V_{\mathrm{cyc}}$ for any k.

The proof mainly divides into the two following lemmas. The first one states that the boundary edges of any community that does not contain v_1 or v_2 — namely the sets of edges $E^{\uparrow}_{\mathrm{cyc},k}$ for $k \in \{1, \ldots, I\}$ — is of size at least 2γ.

Lemma 29 *If $|V_{\mathrm{cyc},k}| < \kappa\gamma$, i.e. $V_{\mathrm{cyc},k} \neq V_{\mathrm{cyc}}$, we have*

$$|E^{\uparrow}_{\mathrm{cyc},k}| \geq 2\gamma\ .$$

The second lemma lower bounds the sum of twice the number of fasteners $V_{\mathrm{fst},\neq 0}$ that do not belong to the same community as v_1 or v_2, plus the number of boundary edges for the community of v_1, v_2, namely $E^{\uparrow}_{\mathrm{cyc},0}$. If the nodes $V_{\mathrm{cyc},\neq 0}$ were sampled uniformly, then $|V_{\mathrm{fst},\neq 0}|$ would be of the order of $a|V_{\mathrm{cyc},\neq 0}|$. However, the number of fasteners in $V_{\mathrm{fst},\neq 0}$ can be much smaller for unfavorable configurations such that fastener nodes are in the same community as v_1 or v_2. Nevertheless, if this happens, then the number of boundary edges in $E^{\uparrow}_{\mathrm{cyc},0}$ must be high accordingly to compensate this drop in the number of fasteners in $V_{\mathrm{fst},\neq 0}$.

Lemma 30 *We have*

$$|E^{\uparrow}_{\mathrm{cyc},0}| + 2|V_{\mathrm{fst},\neq 0}| \geq 2a|V_{\mathrm{cyc},\neq 0}| \geq 2aI \ . \tag{106}$$

Let us explain how Proposition 5 follows from Lemmas 5 and 29. First, we observe that an edge between two communities is either a fastener edge and is therefore incident to $V_{\mathrm{fst},\neq 0}$ or is a cycle edge and therefore both arises in $E^{\uparrow}_{\mathrm{cyc},k}$ and $E^{\uparrow}_{\mathrm{cyc},k'}$ for $k \neq k'$ being the two communities to which the nodes of the edge belong. Hence, we have the following decomposition

$$2|E^{\neq}| \geq \sum_{k=0}^{I} |E^{\uparrow}_{\mathrm{cyc},k}| + 2|V_{\mathrm{fst},\neq 0}| \ ,$$

and Lemmas 29 and 30 ensure that

$$2|E^{\neq}| \geq 2\gamma I + 2aI \ ,$$

which concludes the proof of the Proposition 5.

Proof [Proof of Lemma 29] We consider three cases, (i) at least a layer does not intersect $V_{\mathrm{cyc},k}$; (ii) each layer intersects $V_{\mathrm{cyc},k}$ and $|V_{\mathrm{cyc},k}| \leq (\kappa-1)\gamma$; (iii) each layer intersects $V_{\mathrm{cyc},k}$ and $|V_{\mathrm{cyc},k}| > (\kappa-1)\gamma$.

Case (i): at least a layer does not intersect $V_{\mathrm{cyc},k}$. In this case there exist two layers i_1 and i_2 (not necessarily distinct) such that

$$L_{i_1} \cap V_{\mathrm{cyc},k} \neq \emptyset; \quad L_{i_1-1} \cap V_{\mathrm{cyc},k} = \emptyset; \quad L_{i_2} \cap V_{\mathrm{cyc},k} \neq \emptyset; \quad L_{i_2+1} \cap V_{\mathrm{cyc},k} = \emptyset,$$

with the convention that $L_0 = L_\kappa$ and $L_{\kappa+1} = L_1$. There are γ edges between L_{i_1-1} and one node of $V_{\mathrm{cyc},k}$ in L_{i_1}. Similarly, there are γ edges between L_{i_2+1} and one node of $V_{\mathrm{cyc},k}$ in L_{i_2}. Since $\kappa \geq 3$, these edges are distinct and we deduce $|E^{\uparrow}_{\mathrm{cyc},k}| \geq 2\gamma$.

Case (ii): each layer intersects $V_{\mathrm{cyc},k}$, with $|V_{\mathrm{cyc},k}| \leq (\kappa-1)\gamma$. Since in this case each layer intersects $V_{\mathrm{cyc},k}$, it follows from the definition of the graph that each node v in $V_{\mathrm{cyc}} \setminus V_{\mathrm{cyc},k}$ is connected to $V_{\mathrm{cyc},k}$ by at least two edges corresponding to nodes of $V_{\mathrm{cyc},k}$ in the previous layer and in the following layer of that of v. As a consequence,

$$|E^{\uparrow}_{\mathrm{cyc},k}| \geq 2|V_{\mathrm{cyc}} \setminus V_{\mathrm{cyc},k}| = 2[\kappa\gamma - |V_{\mathrm{cyc},k}|] \geq 2\gamma \ ,$$

as soon as $|V_{\mathrm{cyc},k}| \leq (\kappa-1)\gamma$.

Case (iii): each layer intersects $V_{\mathrm{cyc},k}$, with $|V_{\mathrm{cyc},k}| > (\kappa-1)\gamma$. It remains to consider the case where $|V_{\mathrm{cyc},k}| \in [(\kappa-1)\gamma+1, \kappa\gamma-1]$ which is non empty only if $\gamma > 1$. Then, we consider the complementary set $V' = V_{\mathrm{cyc}} \setminus V_{\mathrm{cyc},k}$ which satisfies $|V'| \leq \gamma - 1 \leq \kappa(\gamma-1)$. Since $V_{\mathrm{cyc},k}$

and V' have a symmetric role for counting $|E^{\uparrow}_{\mathrm{cyc},k}|$, we can apply the same arguments as in Case (i) and (ii), switching the role of $V_{\mathrm{cyc},k}$ and V'. Accordingly, $|E^{\uparrow}_{\mathrm{cyc},k}| \geq 2\gamma$, and the proof of Lemma 29 is complete.

■

Proof [Proof of Lemma 30]

Write $\mathcal{L}_0 = \{\omega \in [\kappa] : L_\omega \cap V_{\mathrm{cyc},0} \neq \emptyset\}$ the collection of layers that intersect $V_{\mathrm{cyc},0}$. To start with, we consider three cases (a) $\mathcal{L}_0 = \emptyset$, (b) $\mathcal{L}_0 = [\kappa]$, and (c) $|\mathcal{L}_0| \in [2; \kappa - 1]$.

Case (a): $\mathcal{L}_0 = \emptyset$. In this case no fastener node belongs to the community of v_1, v_2, therefore we have

$$|V_{\mathrm{fst},\neq 0}| = |V_{\mathrm{fst}}| = a|V_{\mathrm{cyc}}| \geq a|V_{\mathrm{cyc},\neq 0}| \ ,$$

and (106) holds.

Case (b): $\mathcal{L}_0 = [\kappa]$. In this case we observe as in the proof of Lemma 29 (case (ii)), that any node v in $V_{\mathrm{cyc}} \setminus V_{\mathrm{cyc},0}$ is connected to two nodes in $V_{\mathrm{cyc},0}$, one in the previous layer and one in the following layer. Thus, we get

$$|E^{\uparrow}_{\mathrm{cyc},0}| \geq 2\sum_{k=1}^{I} |V_{\mathrm{cyc},k}| = 2|V_{\mathrm{cyc},\neq 0}| \geq 2a|V_{\mathrm{cyc},\neq 0}| \ ,$$

and (106) holds.

Case (c): $|\mathcal{L}_0| \in [2; \kappa - 1]$. This is the intermediary case between cases (a) and (b). Consider $[\kappa]$ as the nodes of a cycle graph and $\mathcal{L}_0$ as a subset of nodes of this graph. Write $\#\mathrm{CC}(\mathcal{L}_0) \geq 1$ for the number of connected components of the subgraph induced by $\mathcal{L}_0$.

Step 1: lower bound of $|E^{\uparrow}_{\mathrm{cyc},0}|$. We partition $\mathcal{L}_0$ into $\mathcal{L}_{0,\mathrm{in}}$, $\mathcal{L}_{0,\mathrm{bd}}$, and $\mathcal{L}_{0,\mathrm{sg}}$ where $\mathcal{L}_{0,\mathrm{in}}$ are internal layers of $\mathcal{L}_0$ that is, they are not at the boundary of a connected component of $\mathcal{L}_0$, $\mathcal{L}_{0,\mathrm{sg}}$ correspond to connected components that are singletons, and $\mathcal{L}_{0,\mathrm{bd}}$ stands for all the the remaining boundary layers.

Fix any $\omega \in \mathcal{L}_{0,\mathrm{in}}$ and any node $v \in L_\omega \setminus V_{\mathrm{cyc},0}$. This node v is connected by at least two edges to $V_{\mathrm{cyc},0}$ (at least one node in the previous layer and one node in the following layer). Fix any $\omega \in \mathcal{L}_{0,\mathrm{bd}}$. For any node $v \in L_\omega \setminus V_{\mathrm{cyc},0}$, we also observe that v is connected by at least one edge to $V_{\mathrm{cyc},0}$. Also, for any $v \in L_\omega \cap V_{\mathrm{cyc},0}$, there are γ edges between v and nodes in a layer that does not belong to $\mathcal{L}_0$. Finally, we consider any $\omega \in \mathcal{L}_{0,\mathrm{sg}}$ and any $v \in L_\omega \cap V_{\mathrm{cyc},0}$. This node is connected to 2γ nodes outside the layers in $\mathcal{L}_0$. Gathering these bounds, we get

$$\begin{aligned} |E^{\uparrow}_{\mathrm{cyc},0}| &\geq 2\sum_{\omega\in\mathcal{L}_{0,\mathrm{in}}} |L_\omega \setminus V_{\mathrm{cyc},0}| + \sum_{\omega\in\mathcal{L}_{0,\mathrm{bd}}} [|L_\omega \setminus V_{\mathrm{cyc},0}| + \gamma] + 2\sum_{\omega\in\mathcal{L}_{0,\mathrm{sg}}} \gamma \\ &\geq 2\sum_{\omega\in\mathcal{L}_0} |L_\omega \setminus V_{\mathrm{cyc},0}| + 2\#\mathrm{CC}(\mathcal{L}_0) \ , \end{aligned} \tag{107}$$

where we used that, for any $\omega \in \mathcal{L}_0$, $|L_\omega \setminus V_{\mathrm{cyc},0}| \leq \gamma - 1$, and where we recall that $\#\mathrm{CC}(\mathcal{L}_0)$ is the number of connected components of $\mathcal{L}_0$.

Step 2: lower bound of $|V_{\mathrm{fst},\neq 0}|$. We now focus on the set $[\kappa] \setminus \mathcal{L}_0$ of layers that do not intersect $V_{\mathrm{cyc},0}$. We start from

$$|V_{\mathrm{fst},\neq 0}| \geq \sum_{\omega\in[\kappa]\setminus\mathcal{L}_0} |V_{\mathrm{fst}} \cap L_\omega|.$$

Although $|V_{\text{fst}} \cap L_\omega| \in [\lfloor a\gamma \rfloor, \lceil a\gamma \rceil]$ for each ω, we have to be slightly more careful to control $|V_{\text{fst},\neq 0}|$. The set $[\kappa] \setminus \mathcal{L}_0$ decomposes into $\#\text{CC}(\mathcal{L}_0)$ connected components, since the set $\mathcal{L}_0$ decomposes in $\#\text{CC}(\mathcal{L}_0)$ connected components. We write $r_1, \ldots, r_{\#\text{CC}(\mathcal{L}_0)}$ for their respective number of layers. Into the i-th connected components, there are at least $\lfloor a r_i \gamma \rfloor$ fasteners. From this and from $\sum_i r_i \geq \kappa - |\mathcal{L}_0|$, we deduce that

$$\begin{aligned} |V_{\text{fst},\neq 0}| &\geq \sum_{i=1}^{\#\text{CC}(\mathcal{L}_0)} \lfloor a r_i \gamma \rfloor \geq a \left[\kappa - |\mathcal{L}_0|\right] \gamma - \#\text{CC}(\mathcal{L}_0) \\ &= a \sum_{\omega \in [\kappa] \setminus \mathcal{L}_0} |L_\omega \setminus V_{\text{cyc},0}| - \#\text{CC}(\mathcal{L}_0) \ . \end{aligned} \tag{108}$$

Combining (107) and (108), we obtain

$$\begin{aligned} |E^{\uparrow}_{\text{cyc},0}| + 2|V_{\text{fst},\neq 0}| &\geq 2a \left[\sum_{\omega \in [\kappa]} |L_\omega \setminus V_{\text{cyc},0}| \right] = 2a \left[\sum_{\omega \in [\kappa]} |L_\omega \cap V_{\text{cyc},\neq 0}| \right] \\ &\geq 2a |V_{\text{cyc},\neq 0}| \ , \end{aligned}$$

which concludes the proof of Lemma 30. ■

Appendix I. Proof of Proposition 15

This proposition is a straightforward consequence of the following result.

Proposition 31 *Assume that $q \leq 1/4$, $q + 2\lambda \leq 1$, $n \geq 2\kappa\gamma + 4$, and set $\bar{p} = \bar{q} + \lambda(1 - 2q)$. We also assume that $a\kappa\gamma$ is an even integer and $\kappa \geq 3 \vee 2/a$. Then, for any $1 \leq i < j \leq n$, we have*

$$\mathbb{E}_{ij}\left[R_{ij}\right] = \frac{(n-2)!}{(n-\kappa\gamma-2)!} \left(\frac{\lambda^{\gamma+a}}{K}\right)^{\kappa\gamma} , \tag{109}$$

$$\mathbb{E}_{i\not{j}}\left[R_{ij}\right] = 0, \tag{110}$$

$$\begin{aligned} \text{var}_{i\not{j}}(R_{ij}) \vee \text{var}_{ij}(R_{ij}) \leq \mathbb{E}^2_{12}[R_{12}](\kappa\gamma)^2 \Bigg[&\frac{2(\kappa\gamma)^3 K^2}{n} \left(\frac{\bar{q}}{\lambda^2} \vee \frac{\bar{q}}{\bar{p}}\right)^{\gamma+a} + \frac{2(\kappa\gamma)^3 K}{n} \left(\frac{\bar{p}}{\lambda^2} \vee 1\right)^{\gamma+a} \\ &+ \left[\frac{2(\kappa\gamma)^3 K^2}{n} \left(\frac{\bar{q}}{\lambda^2} \vee \frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right]^{\kappa\gamma} + \left[\frac{2(\kappa\gamma)^3 K}{n} \left(\frac{\bar{p}}{\lambda^2} \vee 1\right)^{\gamma+a}\right]^{\kappa\gamma} \Bigg]. \end{aligned} \tag{111}$$

I.1. Notation and preliminaries

To control the expectation and the variance of R_{ij}, we need to introduce some graph notation as in Appendix A of Carpentier et al. (2025b). For a motif $G = (V, E)$ and a labeling $\pi \in \Pi_{ij}$, we define the labeled graph $\pi(G)$ as the graph with node set $\{\pi(v) : v \in V\}$ and edge set $\{(\pi(v), \pi(v')) : (v, v') \in E\}$. Given two labelings $\pi^{(1)}$ and $\pi^{(2)} \in \Pi_{ij}$, the labeled merged graph $G_\cup = (V_\cup, E_\cup)$ is defined as the union of $\pi^{(1)}(G^{(1)})$ and $\pi^{(2)}(G^{(2)})$, with the convention that two same edges are merged into a single edge. Similarly, we define the intersection graph $G_\cap = (V_\cap, E_\cap)$ and the symmetric difference graph $G_\Delta = (V_\Delta, E_\Delta)$ so that $E_\Delta = E_\cup \setminus E_\cap$. Here, $V_\cap$ (resp. V_Δ) is the set of nodes induced by the edges $E_\cap$ (resp. E_Δ) so that $G_\cap$ (resp. G_Δ) does not contain any isolated node.

The two following expressions are the starting point of our analysis. They easily derive from the definition (7) of Y:

$$\mathbb{E}\left[P_{G,\pi}\right] = \mathbb{E}\left[\prod_{(v,v')\in E} Y_{\pi(v)\pi(v')}\right] = \mathbb{E}\left[\prod_{(v,v')\in E} (\lambda \mathbf{1}_{z_{\pi(v)}=z_{\pi(v')}})\right], \tag{112}$$

$$\mathbb{E}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] = \mathbb{E}\left[\prod_{(i,j)\in E_\Delta} Y_{ij} \prod_{(i,j)\in E_\cap} Y_{ij}^2\right] = \mathbb{E}\left[\prod_{(i,j)\in E_\Delta} (\lambda \mathbf{1}_{z_i=z_j}) \prod_{(i,j)\in E_\cap} \bar{q}\left(\frac{\bar{p}}{\bar{q}}\right)^{\mathbf{1}_{z_i=z_j}}\right].$$

Without loss of generality, we assume in the following that $(i,j) = (1,2)$.

I.2. Results under the distribution $\mathbb{P}_{12}$

In what follows, we consider the graph $G := G_{\kappa,\gamma,a} = (V,E)$. We will study respectively the expectation and variance of the associated polynomial, both under $\mathbb{P}_{12}$ and under $\mathbb{P}_{\not{1}\not{2}}$.

Mean under $\mathbb{P}_{12}$. Since the graph G is connected, we have from Equation (112)

$$\mathbb{E}_{12}\left[P_{G,\pi}(Y)\right] = \lambda^{|E|}\,\mathbb{E}_{12}\left[\prod_{(i,j)\in\pi(E)} \mathbf{1}_{z_i=z_j}\right] = \frac{1}{K^{|V|-2}}\lambda^{|E|} = \frac{1}{K^{\kappa\gamma}}\lambda^{(\gamma+a)\kappa\gamma}, \tag{113}$$

since $z_1 = z_2$ a.s. under $\mathbb{P}_{12}$. The identity (109) follows.

Variance under $\mathbb{P}_{12}$. Let us turn to proving (111), which is the main part of the proof. Fix $\pi^{(1)}, \pi^{(2)} \in \Pi_{12}$. We start by controlling $\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right]$. Let $u \in \{0,\dots,|V_{\text{cyc}}|\}$ be the integer so that

$$|\text{range}(\pi^{(1)}) \cap \text{range}(\pi^{(2)})| = 2 + u\ .$$

Recall that $V_\cap, V_\Delta, E_\cap, E_\Delta$ stand for the common and symmetric-difference vertex- and edge-sets. Following (88), we have

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] = \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\,\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta} \mathbf{1}_{z_i=z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}}\right].$$

<u>Case $u=0$.</u> Then, we have $E_\cap = \emptyset$, $|V_\Delta| = |V_\cup| = 2\kappa\gamma + 2$ and $|E_\Delta| = 2\kappa\gamma(\gamma+a)$. Therefore

$$\begin{aligned}
\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] &= \lambda^{2\kappa\gamma(\gamma+a)}\,\mathbb{P}_{12}\left[z_i = z_1 \text{ for } i \in V_\Delta\right] \\
&= \lambda^{2\kappa\gamma(\gamma+a)} K^{-2\kappa\gamma} \\
&= \mathbb{E}_{12}\left[P_{G,\pi^{(1)}}\right]\mathbb{E}_{12}\left[P_{G,\pi^{(2)}}\right] \ ,
\end{aligned} \tag{114}$$

by (113).
<u>Case $1 \leq u \leq \kappa\gamma$.</u> In this case $|V_\cup| = 2\kappa\gamma + 2 - u$ and there are $\kappa\gamma - u$ vertices that appear only in $\pi^{(2)}(V) \setminus \pi^{(1)}(V)$. All these vertices belong to G_Δ and each such vertex is connected (within G_Δ) to at least one vertex in the intersection $\pi^{(1)}(V) \cap \pi^{(2)}(V)$ because G is connected. Let $E_\Delta^{(2)} \subset E_\Delta$ be the set of edges of G_Δ that arise from $\pi^{(2)}[G]$. In particular, $E_\Delta^{(2)}$ contains all the edges that have at least one endpoint in $\pi^{(2)}(V) \setminus \pi^{(1)}(V)$. Conditionally on the communities of the vertices of $\pi^{(1)}(V)$, the product of indicators on $E_\Delta^{(2)}$ enforces the community of the $\kappa\gamma - u$ vertices in $\pi^{(2)}(V) \setminus \pi^{(1)}(V)$. Hence, we have

$$\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta^{(2)}} \mathbf{1}_{z_i=z_j} \middle| z_{\pi^{(1)}(V)}\right] \leq \frac{1}{K^{\kappa\gamma-u}} \qquad \text{a.s.}$$

Therefore (conditioning and removing $E_\Delta^{(2)}$), we get

$$\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] \leq \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\frac{1}{K^{\kappa\gamma-u}}\mathbb{E}_{12}\left[\prod_{(i,j)\in E_\Delta\setminus E_\Delta^{(2)}} \mathbf{1}_{z_i=z_j} \prod_{\substack{(i,j)\in E_\cap \\ z_i=z_j}} \frac{\bar{p}}{\bar{q}}\right].$$

Now, define $E_\Delta^{(1)} := E_\Delta \setminus E_\Delta^{(2)}$. Observe that $E_\Delta^{(1)}$ together with $E_\cap$ form a partition of the edges of the labeled copy of G coming from $\pi^{(1)}$. Given the community assignments $(z_i)_{i\in\pi^{(1)}(V)}$, denote $\ell(z)$ the number of distinct communities in $\pi^{(1)}(V)$, and $|E^{\neq}|$ the number of edges in the graph between distinct communities. We have

$$\begin{aligned}\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] &\leq \lambda^{|E_\Delta|}\bar{p}^{|E_\cap|}\frac{1}{K^{\kappa\gamma-u}}\ \mathbb{E}_{12}\left[\left(\frac{\bar{q}}{\bar{p}}\right)^{|E^{\neq}|}\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j}\right] \\ &\leq \lambda^{|E_\Delta|}\bar{p}^{|E_\cap|}\frac{1}{K^{\kappa\gamma-u}}\ \mathbb{E}_{12}\left[\left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell(z)-1)(\gamma+a)}\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j}\right], \end{aligned} \tag{115}$$

where, importantly, we apply Proposition 5 in the second line. Consider the graph $\overline{\mathcal{G}}_\Delta^{(1)} = (\pi^{(1)}(V), E_\Delta^{(1)})$. In Equation (115) above, the condition $\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} = 1$ enforces that the nodes of each of the connected components of $\overline{\mathcal{G}}_\Delta^{(1)}$ are in the same community. Let cc denote the number of connected components of $\overline{\mathcal{G}}_\Delta^{(1)}$ when we remove those containing $\pi^{(1)}(v_1) = 1$ or $\pi^{(1)}(v_2) = 2$. Then, conditionally to $\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} = 1$, we have $(\ell(z) - 1) \in [0, \mathrm{cc}]$ and for any $x \in [0, \mathrm{cc}]$, we have

$$\mathbb{P}_{12}\left[\ell(z) - 1 = x \middle| \prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j} = 1\right] \leq \mathrm{cc}^{\mathrm{cc}}K^{-(\mathrm{cc}-x)}\,. \tag{116}$$

Since the condition $\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j}=1$ enforces that the nodes of each of the connected components of $\overline{\mathcal{G}}_\Delta^{(1)}$ are in the same community, we get

$$\mathbb{P}_{12}\left[\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j}=1\right] \leq \frac{1}{K^{\kappa\gamma+1-cc-1}} = \frac{1}{K^{\kappa\gamma-cc}}$$

and, combining with (116)

$$\begin{aligned}\mathbb{E}_{12}\left[\mathbf{1}_{l(z)-1=x}\left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell(z)-1)(\gamma+a)}\prod_{(i,j)\in E_\Delta^{(1)}} \mathbf{1}_{z_i=z_j}\right] &\leq \frac{1}{K^{\kappa\gamma-\mathrm{cc}}}\cdot \mathrm{cc}^{\mathrm{cc}}K^{-(\mathrm{cc}-x)}\left(\frac{\bar{q}}{\bar{p}}\right)^{x(\gamma+a)}\\ &= \frac{1}{K^{\kappa\gamma}}\cdot \mathrm{cc}^{\mathrm{cc}}K^{x}\left(\frac{\bar{q}}{\bar{p}}\right)^{x(\gamma+a)} .\end{aligned}$$

Since G is connected, each connected component of $\overline{\mathcal{G}}_\Delta^{(1)}$ must contain at least one node in $\pi^{(1)}(V)\cap \pi^{(2)}(V)$. As a consequence, we have $\mathrm{cc}\leq u$. Combining Equation (115) with the last equation, and using $|E_\Delta|+2|E_\cap|=2|E|$, we get

$$\begin{aligned}\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] &\leq \lambda^{|E_\Delta|}\bar{p}^{|E_\cap|}\frac{1}{K^{2\kappa\gamma-u}}u^u\sum_{x=0}^{u}\left(K\left(\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right)^x\\ &\leq \lambda^{2|E|}\left(\frac{\bar{p}}{\lambda^2}\right)^{|E_\cap|}\frac{1}{K^{2\kappa\gamma-u}}u^u\sum_{x=0}^{u}\left(K\left(\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right)^x .\end{aligned}$$

Lemma 32 *We have*

$$|E_\cap|\leq(\gamma+a)u.$$

Proof [Proof of Lemma 32] This result is a consequence of Proposition 5. Indeed, consider a community assignment such that the $2+u$ nodes in $\pi^{(1)}(V)\cap\pi^{(2)}(V)$ belong to same community, whereas the $\kappa\gamma-u$ other nodes in $\pi^{(1)}(V)\setminus\pi^{(2)}(V)$ belong to distinct communities. Hence, for that assignment, we have $\kappa\gamma-u+1$ communities and $|E^{\neq}|=|E|-|E_\cap|$. Then, Proposition 5 states that

$$|E|-|E_\cap|\geq(\gamma+a)[\kappa\gamma-u] .$$

Since $|E|=(\gamma+a)\kappa\gamma$, Lemma 32 holds. ■

We deduce from the previous lemma that

$$\begin{aligned}\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] &\leq \mathbb{E}_{12}^2\left[P_{G,\pi}(Y)\right]\left[K\left(\frac{\bar{p}}{\lambda^2}\vee 1\right)^{\gamma+a}\right]^u u^u\sum_{x=0}^{u}\left(K\left(\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right)^x\\ &\leq \mathbb{E}_{12}^2\left[P_{G,\pi}(Y)\right]u^{u+1}\left[\left[K^2\left(\frac{\bar{q}}{\lambda^2}\vee\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right]^u+\left[K\left(\frac{\bar{p}}{\lambda^2}\vee 1\right)^{\gamma+a}\right]^u\right] .\end{aligned} \tag{117}$$

Combining the terms. Combining (114) and (117), we get

$$\begin{aligned}
\mathrm{var}_{12}(R_{12}) &= \sum_{u=0}^{\kappa\gamma} \sum_{\substack{\pi^{(1)},\pi^{(2)}\in\Pi_{12}\\ |\mathrm{range}(\pi^{(1)})\cap\mathrm{range}(\pi^{(2)})|=2+u}} \left(\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] - \mathbb{E}_{12}\left[P_{G,\pi^{(1)}}\right]\mathbb{E}_{12}\left[P_{G,\pi^{(2)}}\right]\right)\\
&\leq \sum_{u=1}^{\kappa\gamma}\binom{\kappa\gamma}{u}^2 \frac{(n-2)!u!}{(n-2\kappa\gamma+u-2)!}\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right]\\
&\leq \left(\frac{(n-2)!}{(n-\kappa\gamma-2)!}\right)^2 \sum_{u=1}^{\kappa\gamma}(\kappa\gamma)^{2u}\frac{1}{(n-\kappa\gamma-2)^u}\mathbb{E}_{12}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right]\\
&\leq \left(\frac{(n-2)!}{(n-\kappa\gamma-2)!}\right)^2 \mathbb{E}_{12}^2\left[P_{G,\pi}(Y)\right]\sum_{u=1}^{\kappa\gamma}\frac{u^{u+1}(\kappa\gamma)^{2u}}{(n-\kappa\gamma-2)^u}\left[\left[K^2\left(\frac{\bar{q}}{\lambda^2}\vee\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right]^u + \left[K\left(\frac{\bar{p}}{\lambda^2}\vee 1\right)^{\gamma+a}\right]^u\right]\\
&\leq \mathbb{E}_{12}^2[R_{12}](\kappa\gamma)^2\left[\frac{2(\kappa\gamma)^3K^2}{n}\left(\frac{\bar{q}}{\lambda^2}\vee\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a} + \frac{2(\kappa\gamma)^3K}{n}\left(\frac{\bar{p}}{\lambda^2}\vee 1\right)^{\gamma+a}\right.\\
&\qquad\left. + \left[\frac{2(\kappa\gamma)^3K^2}{n}\left(\frac{\bar{q}}{\lambda^2}\vee\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right]^{\kappa\gamma} + \left[\frac{2(\kappa\gamma)^3K}{n}\left(\frac{\bar{p}}{\lambda^2}\vee 1\right)^{\gamma+a}\right]^{\kappa\gamma}\right].
\end{aligned}$$

This concludes the bound on $\mathrm{var}_{12}(R_{12})$ in Equation (111).

I.3. Results under the distribution $\mathbb{P}_{\not{1}\not{2}}$

Mean under $\mathbb{P}_{\not{1}\not{2}}$. We have from Equation (112)

$$\mathbb{E}_{\not{1}\not{2}}\left[P_{G,\pi}(Y)\right] = \lambda^{|E|}\,\mathbb{E}_{\not{1}\not{2}}\left[\prod_{(i,j)\in\pi(E)}\mathbf{1}_{z_i=z_j}\right] = 0,$$

since $z_1 \neq z_2$ a.s. under $\mathbb{P}_{\not{1}\not{2}}$. Equation (110) follows.

Variance under $\mathbb{P}_{\not{1}\not{2}}$. We follow the same approach as for $\mathbb{P}_{12}$. Let $\pi^{(1)},\pi^{(2)}\in\Pi_{12}$ such that $|\mathrm{range}(\pi^{(1)})\cap\mathrm{range}(\pi^{(2)})| = 2+u$, with $u\in\{0,\ldots,\kappa\gamma\}$. Following (88), we again have

$$\mathbb{E}_{\not{1}\not{2}}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] = \lambda^{|E_\Delta|}\bar{q}^{|E_\cap|}\mathbb{E}_{\not{1}\not{2}}\left[\prod_{(i,j)\in E_\Delta}\mathbf{1}_{z_i=z_j}\prod_{\substack{(i,j)\in E_\cap\\ z_i=z_j}}\frac{\bar{p}}{\bar{q}}\right].$$

If $u=0$, we have $E_\cap=\emptyset$ and since $z_1\neq z_2$, we deduce that $\mathbb{E}_{\not{1}\not{2}}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right]=0$. Then, for $u\in[1,\kappa\gamma]$, we argue as for $\mathbb{E}_{12}$. In particular, the counterpart of (115) still holds up to the difference that under $\mathbb{E}_{\not{1}\not{2}}$, the number of communities that are distinct of z_1 or of z_2 is $\ell(z)-2$. Hence, we have

$$\mathbb{E}_{\not{1}\not{2}}\left[P_{G,\pi^{(1)}}P_{G,\pi^{(2)}}\right] \leq \lambda^{|E_\Delta|}\bar{p}^{|E_\cap|}\frac{1}{K^{\kappa\gamma-u}}\,\mathbb{E}_{\not{1}\not{2}}\left[\left(\frac{\bar{q}}{\bar{p}}\right)^{(\ell(z)-2)_+(\gamma+a)}\prod_{(i,j)\in E_\Delta^{(1)}}\mathbf{1}_{z_i=z_j}\right].$$

Write again cc as the number of connected components of the graph $\overline{\mathcal{G}}_{\Delta}^{(1)} = (\pi^{(1)}(V), E_{\Delta}^{(1)})$ when we remove those containing $\pi^{(1)}(v_1) = 1$ or $\pi^{(1)}(v_2) = 2$. Then, conditionally to $\prod_{(i,j)\in E_{\Delta}^{(1)}} \mathbf{1}_{z_i=z_j} = 1$, we have $(\ell(z) - 2) \in [0, \mathrm{cc}]$ and for any $x \in [0, \mathrm{cc}]$

$$\mathbb{P}_{\not{1}2}\left[\ell(z) - 2 = x \,\middle|\, \prod_{(i,j)\in E_{\Delta}^{(1)}} \mathbf{1}_{z_i=z_j} = 1\right] \leq \mathrm{cc}^{\mathrm{cc}} K^{-(\mathrm{cc}-x)} \ .$$

Hence, we arrive as previously at

$$\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \leq \lambda^{2|E|} \left(\frac{\bar{p}}{\lambda^2}\right)^{|E_{\cap}|} \frac{1}{K^{2\kappa\gamma-u}} u^u \sum_{x=0}^{u} \left(K\left(\frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right)^x ,$$

and the counterpart of (117) holds:

$$\mathbb{E}_{\not{1}2}\left[P_{G,\pi^{(1)}} P_{G,\pi^{(2)}}\right] \leq \mathbb{E}_{12}^2\left[P_{G,\pi}(Y)\right] u^{u+1} \left[\left[K^2\left(\frac{\bar{q}}{\lambda^2} \vee \frac{\bar{q}}{\bar{p}}\right)^{\gamma+a}\right]^u + \left[K\left(\frac{\bar{p}}{\lambda^2} \vee 1\right)^{\gamma+a}\right]^u\right] \ .$$

Then, summing over all $\pi^{(1)}$ and $\pi^{(2)}$, we conclude as previously the bound on $\mathrm{var}_{\not{1}2}(R_{12})$ in Equation (111).

Appendix J. Proof of Lemma 2

The signal condition showing up in the low-degree lower bound is of the form $\lambda \leq D^{-c}\lambda_c$, where

$$\sup_{r\geq 1}\left\{\frac{n\lambda_c^{2r}}{K\lambda_c^r + K^2\bar{q}^r}\right\} = 1 \ . \tag{118}$$

Lemma 2, stated page 4, makes explicit the phase transition in the Condition (118).

Proof of Lemma 2

Let $\rho := q/\lambda$ and

$$\phi(r) := \frac{\lambda^r}{1 + K\rho^r} = \frac{\lambda^{2r}}{\lambda^r + Kq^r}.$$

Let r^* such that $\phi'(r^*) = 0$. We have

$$\rho^{r^*} = \frac{1}{K} \times \frac{\log(1/\lambda)}{\log(\lambda/\rho)} := \frac{u}{K}. \tag{119}$$

We have

$$\phi(r^*) = \frac{\lambda^{r^*}}{1 + K\rho^{r^*}} = \frac{\lambda^{r^*}}{1+u},$$

so if λ_* is such that $\phi(r^*) = K/n$, then

$$\lambda_*^{r^*} = \frac{(1+u)K}{n}. \tag{120}$$

Since $q = \lambda_* \rho$, we then get

$$q^{r^*} = \frac{u(1+u)}{n}, \quad \text{with } u = \frac{\log(1/\lambda_*)}{\log(\lambda_*^2/q)}. \tag{121}$$

Below, we write λ for λ_* We have

$$u = \frac{r^* \log(1/\lambda)}{r^* \log(\lambda/\rho)} = \frac{\log(n/K) - \log(1+u)}{\log(K^2/n) + \log(1+1/u)},$$

and hence

$$u \log(K^2/n) + u \log(1+1/u) + \log(1+u) = \log(n/K). \tag{122}$$

We seek for a solution $u > 0$ and $r^* \geq 1$ to (121) and (122). We denote $K = n^{(1+\delta_n)/2}$.

Case $K \leq (1-\epsilon)\sqrt{n}$. Then $\log(K^2/n) \leq 2\log(1-\epsilon) < 0$ and

$$-2u \log\left(\frac{1}{1-\epsilon}\right) + u \log(1+1/u) + \log(1+u) \geq \frac{1}{2}\log(n)$$

which has no positive solution in u for n large, since the left hand side is bounded from above. In this case, the supremum in (118) is achieved for $r = 1$ and the critical value is the KS threshold

$$\lambda_c = \lambda_{KS} = \frac{K}{2n}\left(1 + \sqrt{1+2nq}\right).$$

Case $(1+\epsilon)\sqrt{n} \leq K \leq n^{1-\epsilon}$. Hence $\frac{2\log(1+\epsilon)}{\log(n)} \leq \delta_n \leq 1 - \epsilon/2$. Then, we have

$$u = \frac{1-\delta}{2\delta} - \frac{u\log(1+1/u) + \log(1+u)}{\log(n)} = \frac{1-\delta}{2\delta} - O\left(\frac{\log(1/\delta)}{\log(n)}\right) = \frac{1-\delta}{2\delta}\left(1 - O\left(\frac{\delta\log(1/\delta)}{\log(n)}\right)\right).$$

We observe that $r^* \geq 1$ iff $q \geq u(1+u)/n$, i.e.

$$q \geq \frac{1-\delta^2}{4\delta^2} \times \frac{1 - O\left(\frac{\delta\log(1/\delta)}{\log(n)}\right)}{n}. \tag{123}$$

Writing $q = n^{-\alpha_n}$ and solving $q^{r^*} = u(1+u)/n$ gives

$$n^{1-\alpha_n r^*} = \frac{1-\delta^2}{4\delta^2} \times \left(1 - O\left(\frac{\delta\log(1/\delta)}{\log(n)}\right)\right)$$

and

$$\begin{aligned}\frac{1}{r^*} &= \alpha_n \left(1 - \frac{\log((1-\delta^2)/4\delta^2) - O(\delta\log(1/\delta)/\log(n))}{\log(n)}\right)^{-1} \\ &= \alpha_n \left(1 + \frac{\log((1-\delta^2)/4\delta^2)}{\log(n)} + O\left(\frac{\delta\log(1/\delta)}{\log(n)^2}\right)\right).\end{aligned} \tag{124}$$

From (120), $q^{r^*} = u(1+u)/n$ and $\log(1+u) = \log(1/\delta) = \log\log(n)$ we get

$$\begin{aligned}
\lambda_c &= \left((1+u)n^{-(1-\delta)/2}\right)^{\alpha_n\left(1+\frac{\log(u(1+u))}{\log(n)}+O\left(\left(\frac{\log\log(n)}{\log(n)}\right)^2\right)\right)} \\
&= n^{-\alpha_n(1-\delta)/2} \times \left(\frac{(1+u)^{1+\delta}}{u^{1-\delta}}\right)^{\alpha_n/2} (1+o(1)) \\
&\sim q^{(1-\delta)/2}\left(\frac{(1+\delta)^{1+\delta}}{(1-\delta)^{1-\delta}(2\delta)^{2\delta}}\right)^{\alpha_n/2}.
\end{aligned}$$