Supplementary materials: Bayesian community detection with unknown number of communities

January 21, 2018

A. ALGORITHM 1

We present the details of the Gibbs sampling algorithm mentioned in §3.1 of the main document.

Algorithm 1 Collapsed sampler for MFM-SBM

procedure C-MFM-SBM

Initialize $z = (z_1, \ldots, z_n)$ and $Q = (Q_{rs})$.

for each iter $= 1$ to M **do**

Update $Q = (Q_{rs})$ conditional on z in a closed form as

$$p(Q_{rs} \mid A) \sim \text{Beta}(\bar{A}_{[rs]} + a, n_{rs} - \bar{A}_{[rs]} + b)$$

Where $\bar{A}_{[rs]} = \sum_{z_i=r, z_j=s, i\neq j} A_{ij}$, $n_{rs} = \sum_{i\neq j} I(z_i = r, z_j = s), r = 1, \ldots, k; s = 1, \ldots, k$. Here k is the number of clusters formed by current z.

Update $z = (z_1, \ldots, z_n)$ conditional on $Q = (Q_{rs})$, for each i in $(1, ..., n)$, we can get a closed form expression for $P(z_i = c \mid z_{-i}, A, Q)$:

$$\propto \begin{cases} [|c| + \gamma][\prod_{j>i} Q_{cz_j}^{A_{ij}}(1 - Q_{cz_j})^{(1-A_{ij})}][\prod_{k<i} Q_{z_k c}^{A_{ki}}(1 - Q_{z_k c})^{(1-A_{ki})}] & \text{at an existing table c} \\ \frac{V_n(|\mathcal{C}_{-i}|+1)}{V_n(|\mathcal{C}_{-i}|)}\gamma m(A_i) & \text{if c is a new table} \end{cases}$$

where $\mathcal{C}_{-i}$ denotes the partition obtained by removing z_i and

$$m(A_i) = \prod_{t=1}^{|\mathcal{C}_{-i}|} \big[\text{Beta}(a, b)\big]^{-1} \text{Beta}\Big[\sum_{j\in\mathcal{C}_t, j>i} A_{ij} + \sum_{j\in\mathcal{C}_t, j<i} A_{ji} + a, |\mathcal{C}_t| - \sum_{j\in\mathcal{C}_t, j>i} A_{ij} - \sum_{j\in\mathcal{C}_t, j<i} A_{ji} + b\Big].$$

end for

end procedure

B. ESTIMATION PERFORMANCE UNDER MODEL MISSPECIFICATION

As mentioned at the end of § 5.1 of the main document, we investigate the robustness of MFM-SBM to deviations from the block model assumption. To this end, we generate data from a degree-corrected block model

$$A_{ij} \sim \text{Bernoulli}(\theta_{ij}), \quad \theta_{ij} = w_i w_j Q_{z_i z_j}, \quad 1 \le i < j \le n, \tag{A.1}$$

with node specific weights w_is. If all w_is are one, this reduces to the usual block model. We randomly set 30% of the w_is to 0.8 and the remaining to one. We generate 100 datasets for the same choices of (n, K, p) as in § 5.1. Performance in estimating the number of communities is summarized in Figures 1 and 2, while Table 1 reports estimation accuracy of the cluster configurations. As in § 5.1 of the main document, MFM-SBM continues to have superior performance when the block structure is vague. These simulations indicate that MFM-SBM can handle mild deviations from the block model assumption without degrading performance, though certainly there will be a breakdown point if the true model is very different from an SBM.

(k, p)	**MFM-SBM**	**LEM**	**HMM**	**MH-MCMC**
$k = 2, p = 0.50$	0.89 (1.00)	1.00 (1.00)	0.99 (1.00)	1.00 (1.00)
$k = 2, p = 0.24$	0.93 (0.75)	0.21 (0.73)	NA (NA)	0.54 (0.57)
$k = 3, p = 0.50$	0.96 (0.99)	0.75 (0.94)	1.00 (0.99)	0.87 (0.99)
$k = 3, p = 0.33$	0.93 (0.88)	0.78 (0.73)	0.47 (0.80)	0.38 (0.82)

Table 1: *Cluster membership estimation under degree-corrected model. The value outside the parenthesis denotes the proportion of correct estimation of the number of clusters out of 100 replicates. The value inside the parenthesis denotes the average Rand index value when the estimated number of clusters is true. NA's indicate no correct estimation of the number of clusters out of all replicates.*

C. CONVERGENCE DIAGNOSTICS

Our first set of simulations investigate the algorithmic performance of MFM-SBM relative to other available Bayesian methods for different choices of the number of nodes n, number of communities K, the within-community edge probability p, and the relative community sizes.

Figures 3 – 7 show average value of $\mathrm{RI}(z, z_0)$ for the first 300 MCMC iterations from 100 randomly chosen starting configurations for the MFM-SBM algorithm. In each figure, the block structure gets increasingly vague as one moves from the left to the right. It can be readily seen from

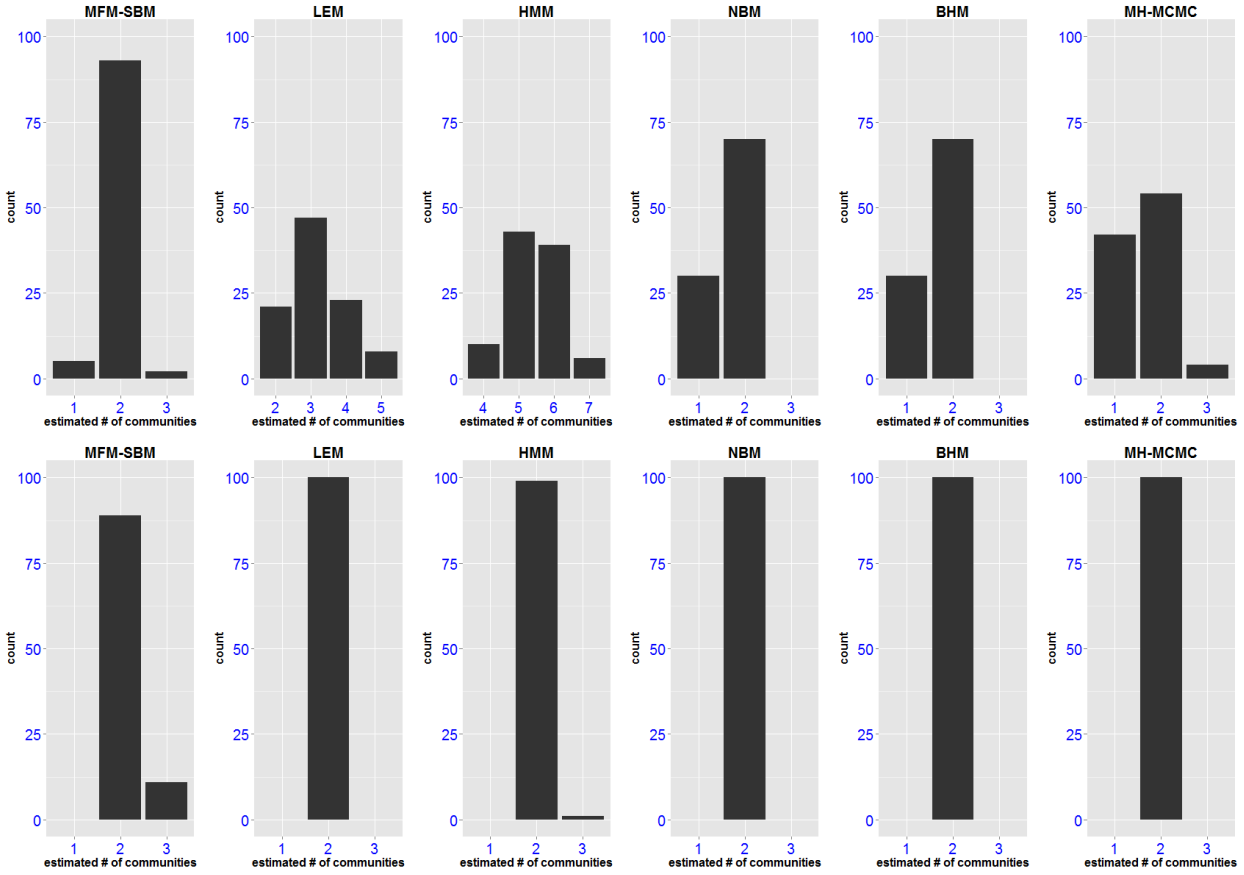

Figure 1: *Balanced* **degree-corrected** *network with* 100 *nodes and* 2 *communities. Histograms of estimated number of communities across* 100 *replicates. The lower panel is the case when the community structure in the network is prominent (p = 0.5); the top panel is for a vague block structure (p = 0.24). From left to right: our method (MFM-SBM), leading eigenvector method (LEM), hierarchical modularity measure (HMM), non back-tracking matrix (NBM) , Bethe Hessian matrix (BHM) & MH-MCMC.*

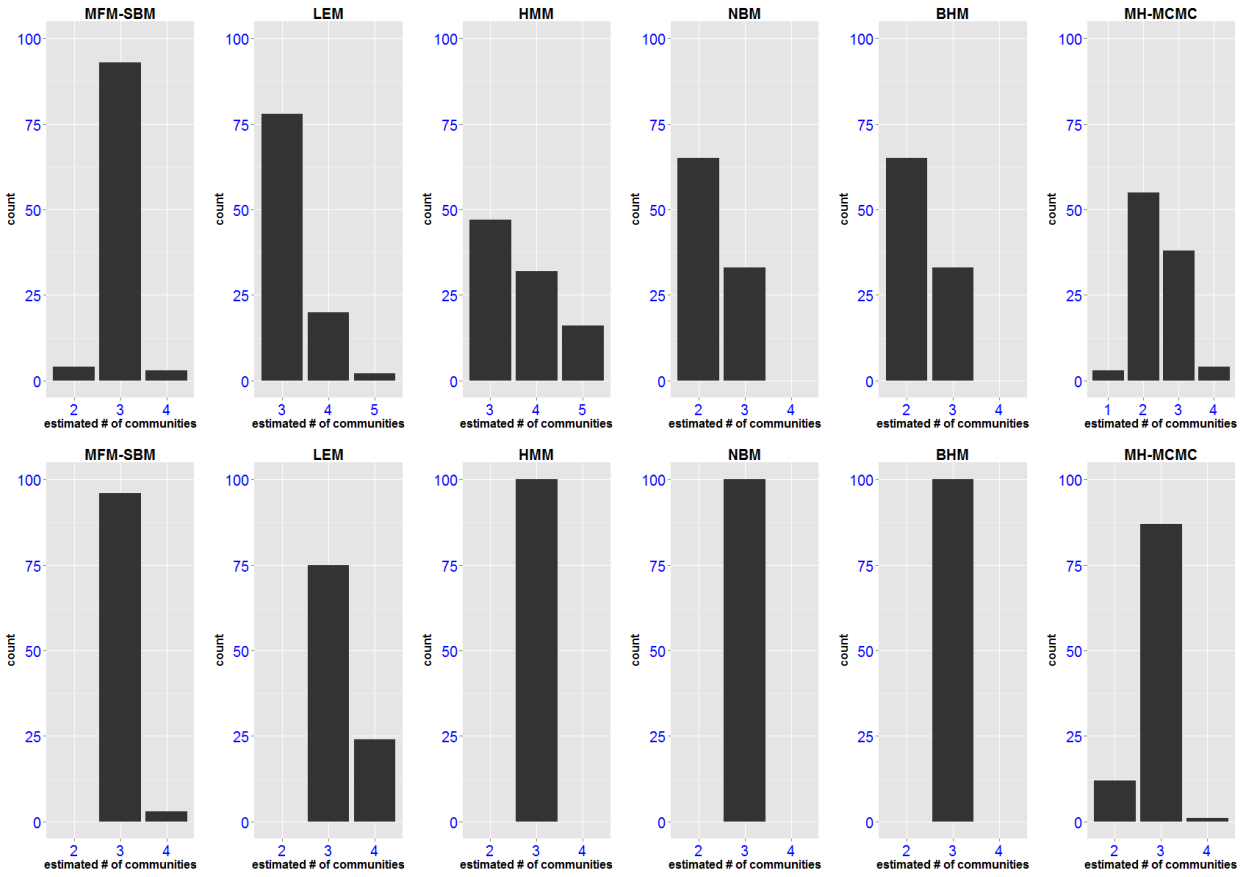

Figure 2: *Balanced* **degree-corrected** *network with* 100 *nodes and* 3 *communities. Histograms of estimated number of communities across* 100 *replicates. The lower panel is the case when the community structure in the network is prominent (p = 0.5); the top panel is for a vague block structure (p = 0.33). From left to right: our method (MFM-SBM), leading eigenvector method (LEM), hierarchical modularity measure (HMM), non back-tracking matrix (NBM) , Bethe Hessian matrix (BHM) & Bayesian competitor (MH-MCMC).*

Figures 3 and 6 that for balanced networks with sufficient number of nodes per community, the Rand index rapidly converges to 1 or very close to 1 within 300 MCMC iterates, indicating rapid mixing and convergence of the chain. The convergence is somewhat slowed down if the network

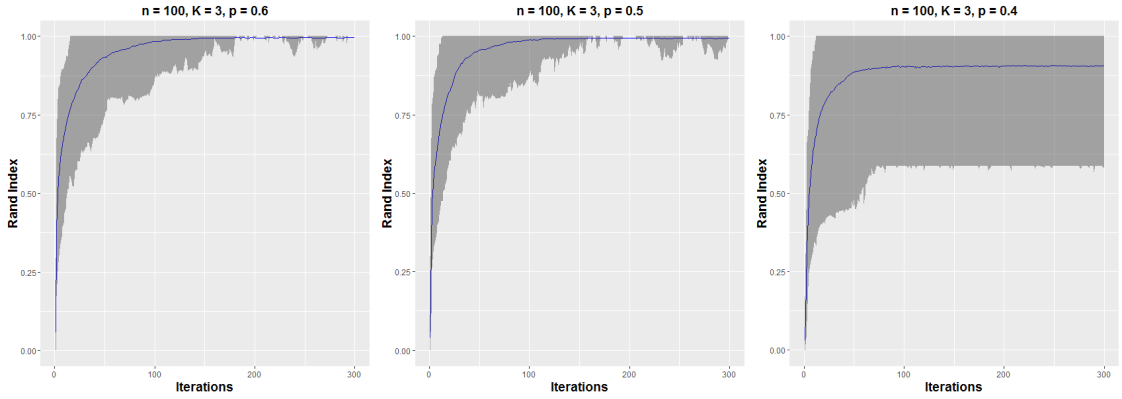

Figure 3: *Average Rand index (solid blue line) vs. MCMC iteration for MFM-SBM for 100 different starting configurations in a balanced network.* $n = 100$ *nodes in* $K = 3$ *communities of sizes 33, 33 and 34. The shaded regions correspond to the variation of the Rand index obtained from MFM-SBM due to random initializations.*

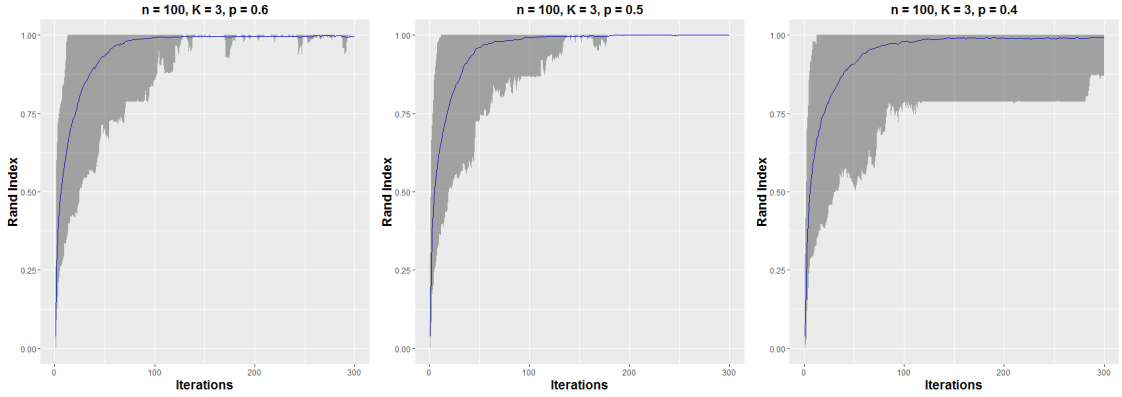

Figure 4: *Average Rand index (solid blue line) vs. MCMC iteration for MFM-SBM with 100 different starting configurations in an unbalanced network.* $n = 100$ *nodes in* $K = 3$ *communities of sizes 22, 33 and 45.*

is unbalanced and the block structure is vague; see for example, the right-most panel of Figures 5. However, with a clearer block structure or more nodes available per community, the convergence improves; see the left two panels of Figures 4 and 5 and the right most panel of Figure 7. We additionally conclude from Figure 5 - 7 that as the number of community increases, we need more nodes per community to get precise recovery of the community memberships.

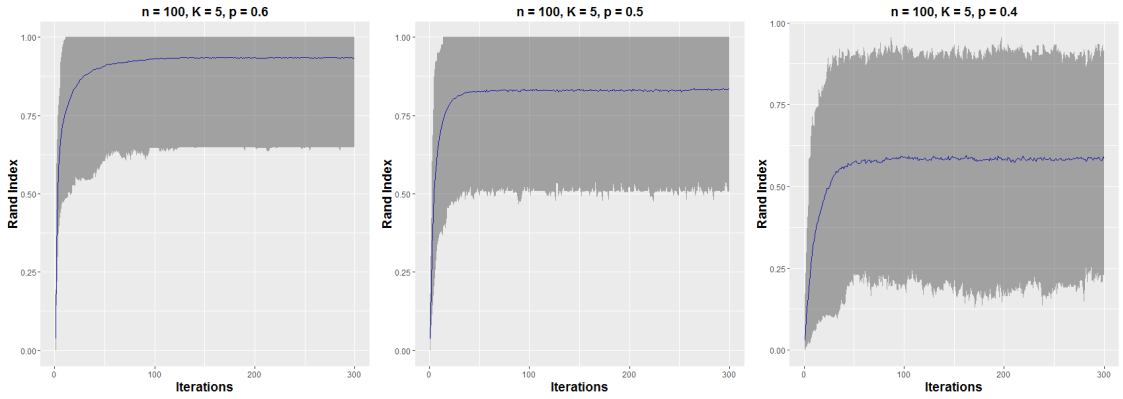

Figure 5: *Average Rand index (soild blue line) vs. MCMC iteration for MFM-SBM with 100 different starting configurations in a balanced network.* $n = 100$ *nodes in* $K = 5$ *communities of size 20 each.*

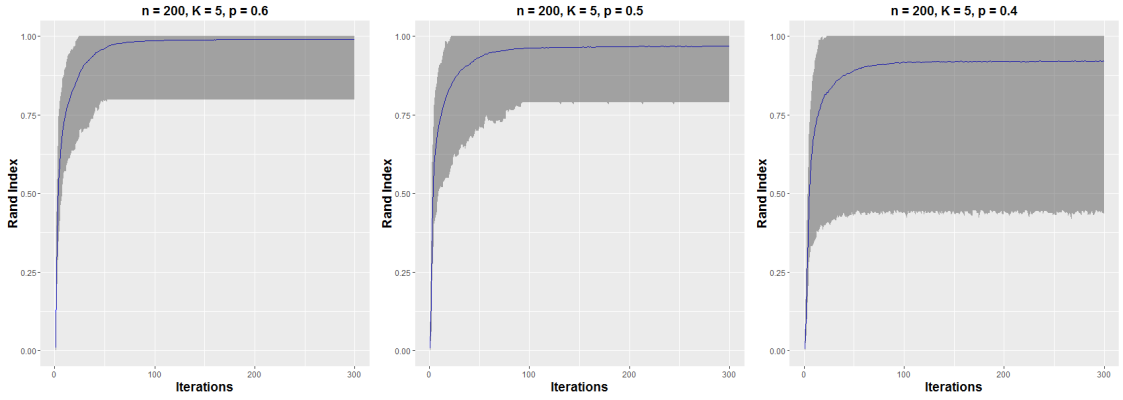

Figure 6: *Average Rand index (solid blue line) vs. MCMC iteration for MFM-SBM with 100 different starting configurations in a balanced network.* $n = 200$ *nodes in* $K = 5$ *communities of size 40 each.*

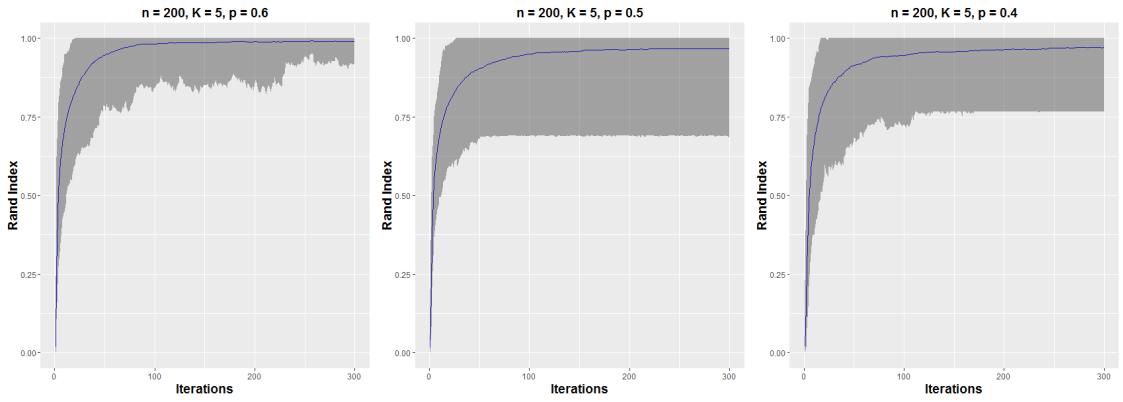

Figure 7: *Average Rand index (solid blue line) vs. MCMC iteration for MFM-SBM with 100 different starting configurations in an unbalanced network.* $n = 200$ *nodes in* $K = 5$ *communities of sizes 20, 30, 40, 50 and 60.*

We also found in the more complicated cases (e.g., right panels of Figure 3), MH-MCMC (Figure 8) does not converge as fast as our approach.

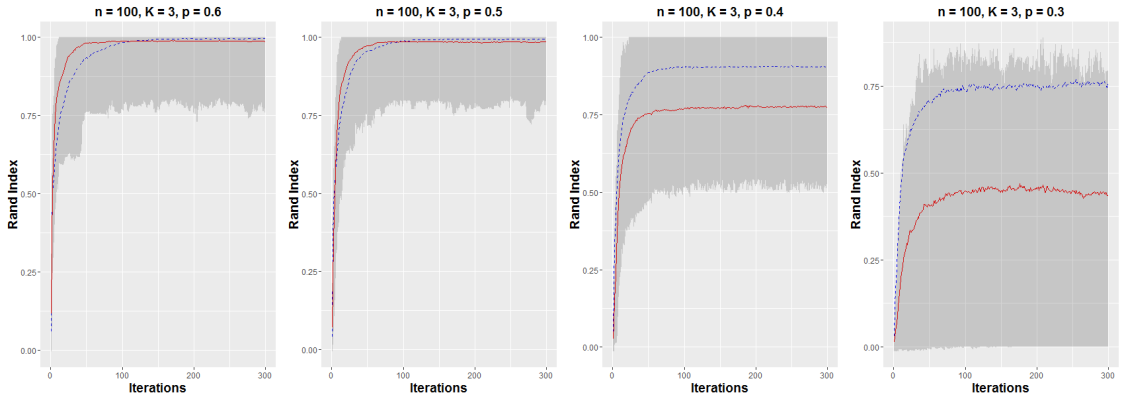

Figure 8: *Average Rand index vs. MCMC iteration for the MH-MCMC of [4] with 100 different starting configurations in a balanced network (solid red line).* $n = 100$ *nodes in* $K = 3$ *communities of sizes 33, 33 and 34. The shaded regions correspond to the variation of the Rand index obtained from MH-MCMC due to random initializations. The average Rand index for MFM-SBM with 100 different starting configurations is additionally provided for comparison (dashed blue line).*

C.1 Mixing of the MCMC chain for Q

We report the results based on the simulated datasets in Figure 3 with 100 nodes, 3 communities in equal sizes and different diagonal values p for Q. The average effective sample sizes for the 250 MCMC iterations (leaving out first 50 MCMC iterations as burn-in) across 100 randomly chosen

starting configurations are 252 for $p = 0.4$; 243 for $p = 0.5$ and 235 for $p = 0.6$. The reported effective sample size here is an average of element-wise effective sample sizes for all terms in matrix θ. The effective sample sizes are very close to the number of MCMC iterations. We also display the trace plots for several representative elements of the matrix θ based on simulated datasets in Figure 3.

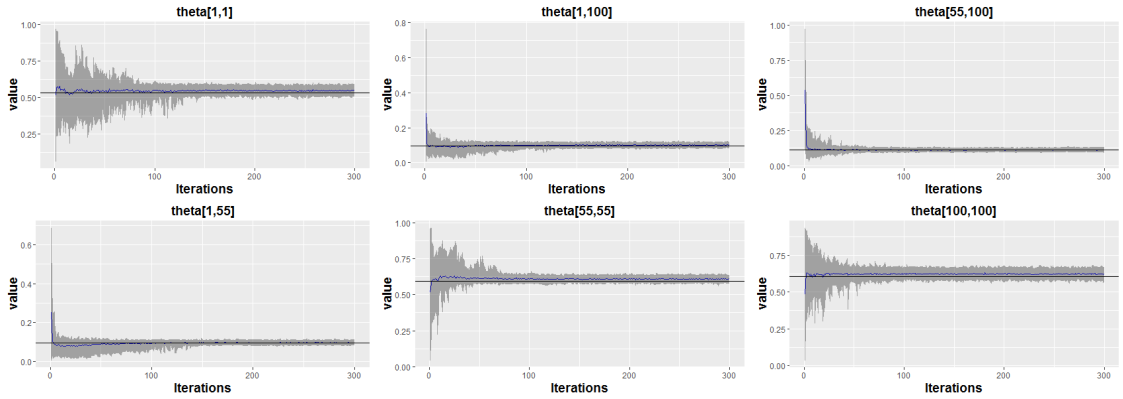

Figure 9: *θ_{ij}'s averaged across 100 different initializations vs. MCMC iteration for MFM-SBM in a balanced network. $n = 100$ nodes in $K = 3$ communities of sizes 33, 33 and 34; $p = 0.6$. The shaded regions correspond to the variation of the MCMC sample due to random initializations.*

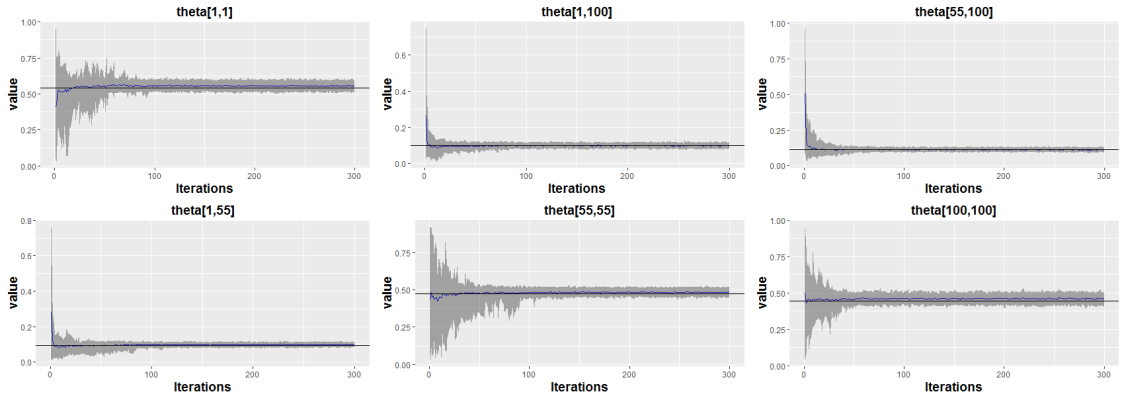

Figure 10: *θ_{ij}'s averaged across 100 different initializations vs. MCMC iteration for MFM-SBM in a balanced network. $n = 100$ nodes in $K = 3$ communities of sizes 33, 33 and 34; $p = 0.5$.*

Figures 9 to 11 depict traceplots for some representative θ_{ij}s averaged over 100 initializations for the first 300 MCMC iterations. The reference line in each subplot is the true value of the representative element based on the true clustering configuration. It is evident that θ_{ij}s rapidly converge to the stationary distributions tightly centered around the true values.

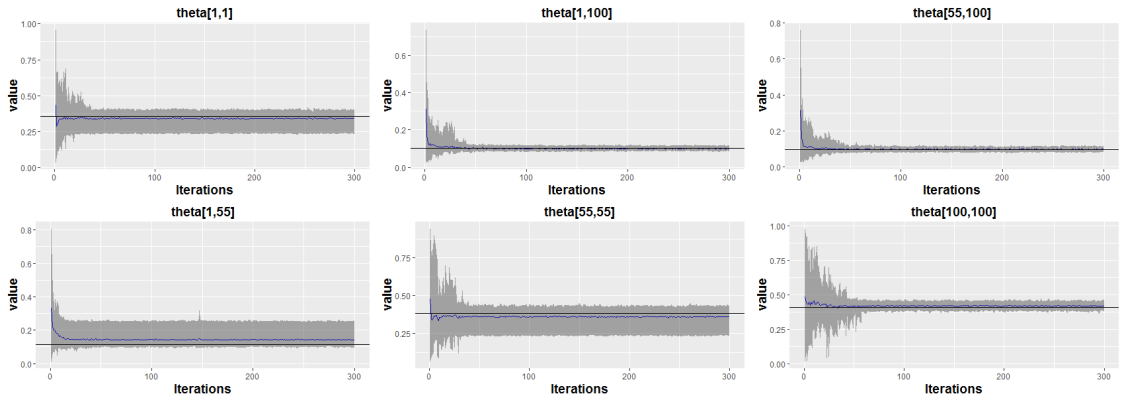

Figure 11: θ_{ij}*'s averaged across 100 different initializations vs. MCMC iteration for MFM-SBM in a balanced network.* $n = 100$ *nodes in* $K = 3$ *communities of sizes 33, 33 and 34;* $p = 0.4$.

Method	**MFM-SBM**	**NBM**	**BHM**	**LEM**	**HMM**	**MH-MCMC**
Number of clusters	5	3	3	4	4	6

Table 2: *Estimated number of clusters for US Politics data*

D. COMMUNITY DETECTION IN BOOKS ABOUT US POLITICS DATA

We now provide details of the second real dataset mentioned in §6 of the main document. We consider a network of books about US politics sold by the online bookseller Amazon.com [3]. In this network the vertices represent 105 recent books on American politics bought from Amazon, and edges join pairs of books that are frequently purchased by the same buyer. Books were divided according to their stated or apparent political alignment, liberal or conservative, except for a small number of books that were explicitly bipartisan or centrist, or had no clear affiliation. This is a undirected network data with 105 nodes.

Results from MFM-SBM is again based on 10,000 MCMC iterations leaving out a burn-in of 4,000, initialized at a randomly generated configuration with 9 clusters. Both Beta$(2, 2)$ and Beta$(1, 1)$ priors on the elements of Q are investigated here. From Table 5 and Table 6, both LEM and HMM find two large clusters consisting of mainly liberal or conservative books respectively (refer to cluster 3&4 in table 5 and cluster 2&4 in table 6). The remaining nodes of the two clusters in these two clustering configurations consist of books from different categories.

Among two prior choices in MFM-SBM, Beta$(2, 2)$ prior on the elements of Q provide a more interpretable result. From Table 3 (MFM-SBM), we find one cluster (cluster 5) consisting of books from different categories. The remaining four clusters form two large clusters consisting of mainly liberal (cluster 1&3) or conservative (cluster 2&4) books respectively. It is also interesting to

MFM-SBM	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
liberal	29	0	9	0	5
conservative	1	8	0	34	6
neutral	2	0	0	3	8

Table 3: *Contingency table of cluster index and book categories using MFM-SBM with Beta*$(2, 2)$ *priors on the elements of* Q

observe "core-periphery" structure [1] in those four clusters. From that heatmap of Q in Figure 13, it is evident that there are two core clusters surrounded by another cluster with sparse within group connections. This structure reveals that the books in the core parts are popular books most frequently purchased by the same buyer; while the books in the peripheral region are more likely to be purchased by the same buyer more specific to his interests. Both MFM-SBM with Beta$(1, 1)$ prior and MH-MCMC reveals 6 clusters with similar "core-periphery" structure.

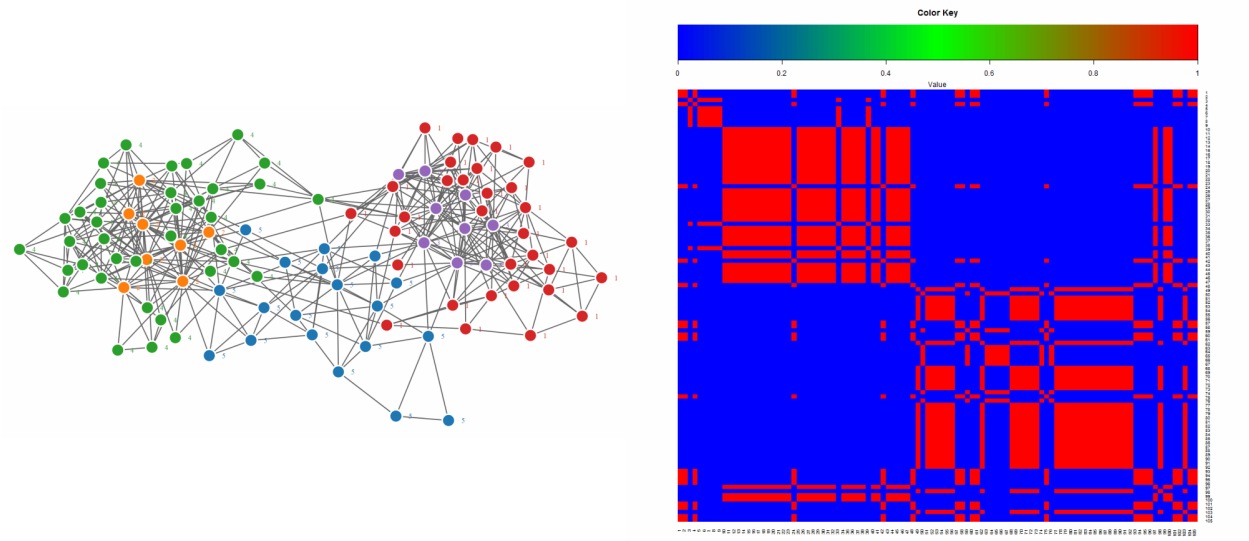

Figure 12: *Estimated configuration for the US Politics books data using MFM-SBM with Beta*$(2, 2)$ *prior on the elements of* Q

The modularity based approaches (LEM and HMM) in the `igraph` package could not find the core-periphery structure as shown in Figure 16 and Figure 17 respectively. The heatmaps in Figures 12, 14, 16, 17 and 18 are obtained after rearranging the nodes in order of the clusters corresponding to conservatives, liberal and neutral.

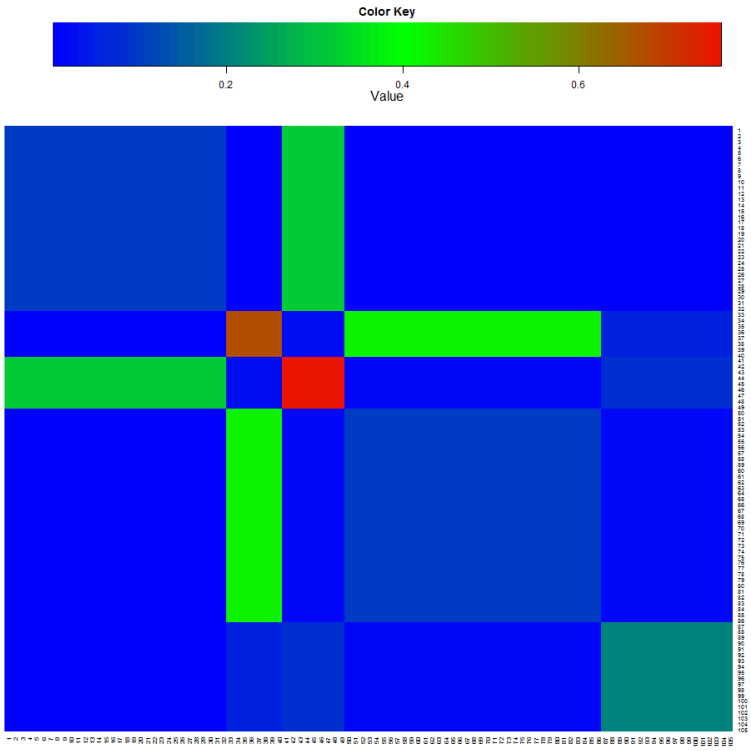

Figure 13: *Heatmap for Q matrix for the US Politics books data using MFM-SBM with Beta$(2, 2)$ prior on the elements of Q*

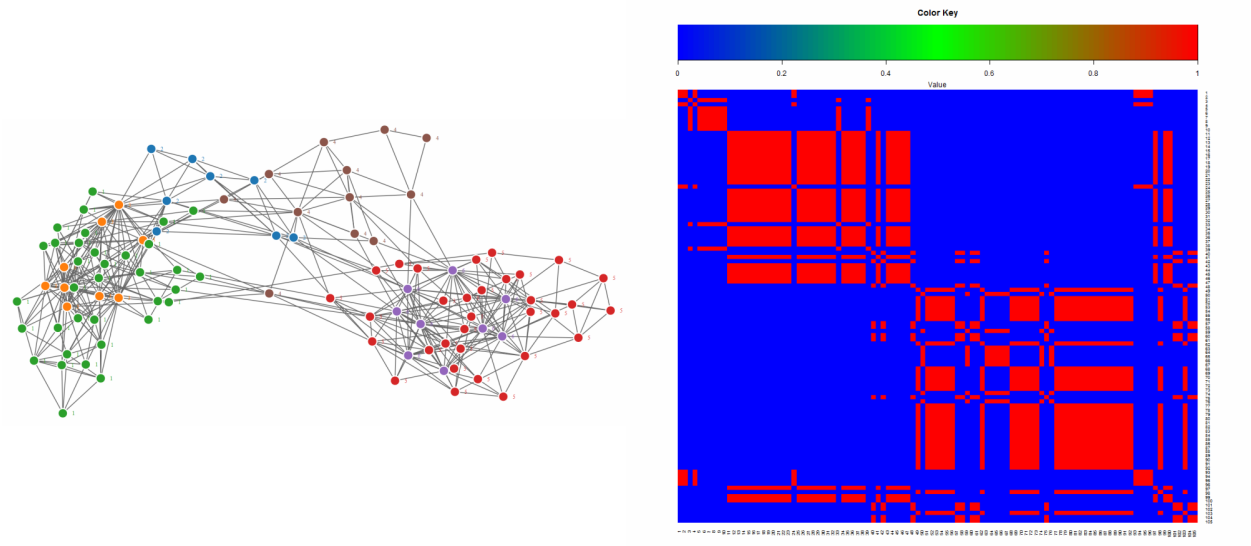

Figure 14: *Estimated configuration for the US Politics books data using MFM-SBM with Beta$(1, 1)$ prior on the elements of Q*

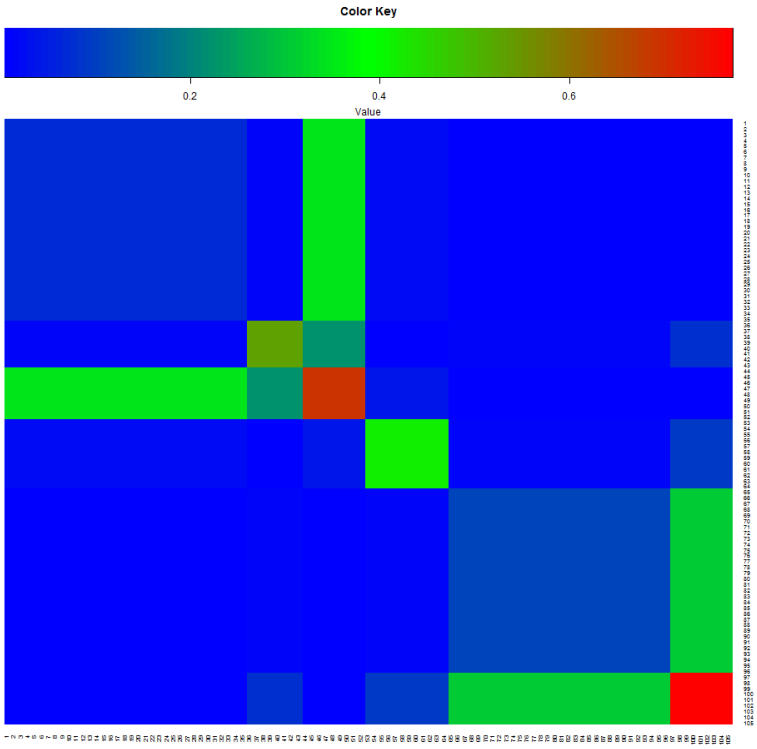

Figure 15: *Heatmap for Q matrix for the US Politics books data using MFM-SBM with Beta$(1, 1)$ prior on the elements of Q*

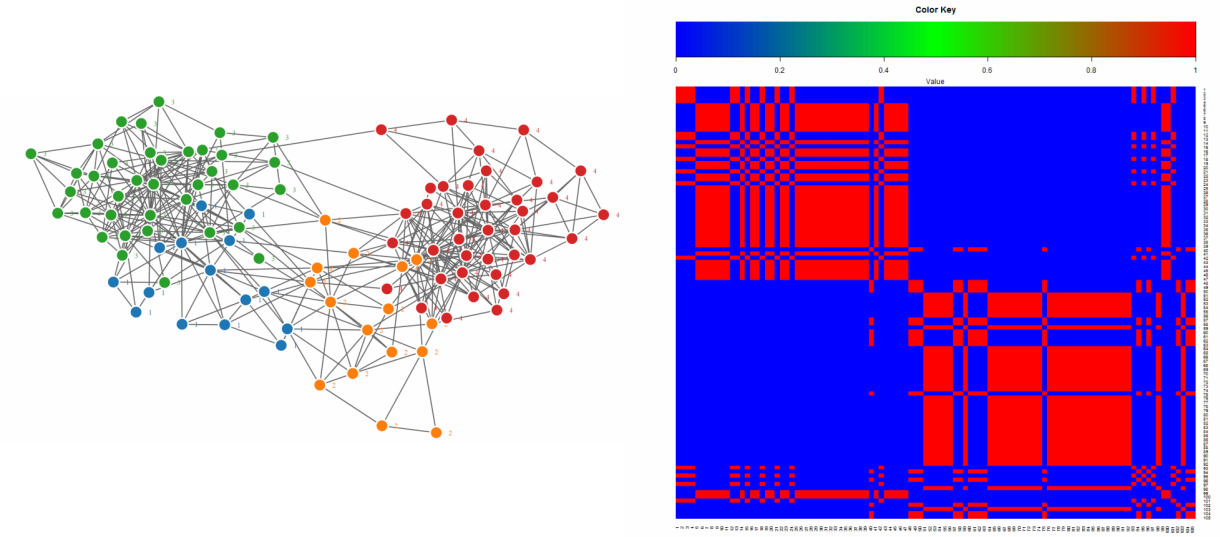

Figure 16: *Estimated configuration for the US Politics books data using LEM*

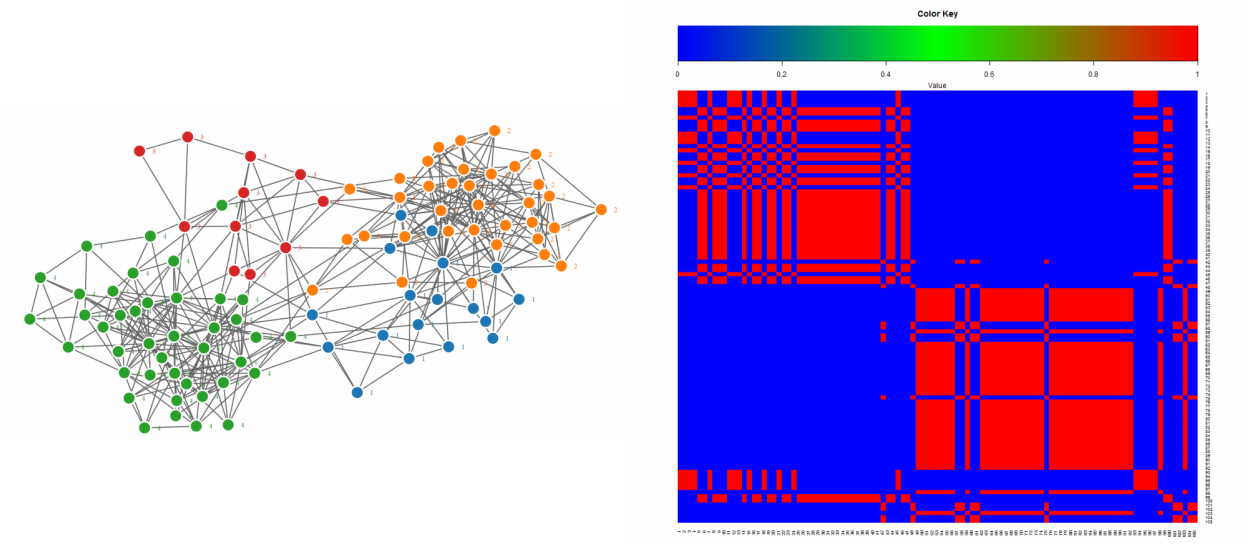

Figure 17: *Estimated configuration for the US Politics books data using HMM*

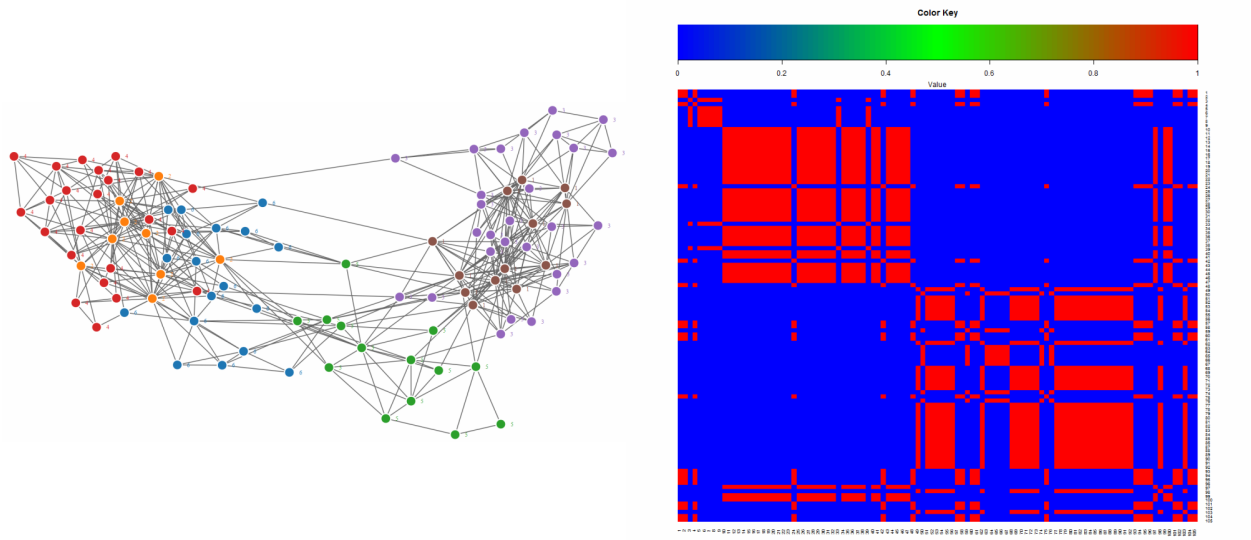

Figure 18: *Estimated configuration for the US Politics books data using MH-MCMC*

MFM-SBM	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
liberal	0	0	0	5	29	9
conservative	32	4	9	3	1	0
neutral	3	4	0	4	2	0

Table 4: *Contingency table of cluster index and book categories using MFM-SBM with Beta*$(1, 1)$ *priors on the elements of* Q

LEM	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
liberal	0	8	0	35
conservative	11	3	35	0
neutral	4	5	2	2

Table 5: *Contingency table of cluster index and book categories using LEM*

HMM	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**
liberal	0	0	5	38
conservative	13	33	2	1
neutral	5	2	4	2

Table 6: *Contingency table of cluster index and book categories using HMM*

MH-MCMC	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
liberal	12	0	26	0	5	0
conservative	0	9	1	22	3	14
neutral	1	0	1	1	6	4

Table 7: *Contingency table of cluster index and book categories using MH-MCMC*

E. PROOF OF THEOREM 4.1

Marginal likelihood approximation and prior-ratio bound:

The posterior expected risk $E[d(z, z_0) \mid \mathcal{A}] = \sum_r r P[d(z, z_0) = r \mid \mathcal{A}]$. Recall that $\mathcal{Z}_{n,K}$ denotes the space of all cluster configurations of n objects into K groups, Π denotes a prior distribution on $\mathcal{Z}_{n,K}$, and z_0 denotes the true configuration. We have

$$P[d(z, z_0) = r \mid \mathcal{A}] = \frac{\sum_{z:d(z,z_0)=r} \mathcal{L}(\mathcal{A} \mid z)\Pi(z)}{\sum_{z\in\mathcal{Z}_{n,K}} \mathcal{L}(\mathcal{A} \mid z)\Pi(z)} = \frac{\sum_{z:d(z,z_0)=r} \exp\{\ell(z) - \ell(z_0) + \Pi_\ell(z, z_0)\}}{\sum_{z\in\mathcal{Z}_{n,K}} \exp\{\ell(z) - \ell(z_0) + \Pi_\ell(z, z_0)\}},$$

where recall $\ell(z) = \log \mathcal{L}(\mathcal{A} \mid z)$ is the log-marginal likelihood of cluster configuration z, and $\Pi_\ell(z, z_0) := \log\{\Pi(z)/\Pi(z_0)\}$. Since $\sum_{z\in\mathcal{Z}_{n,K}} \exp\{\ell(z) - \ell(z_0) + \Pi_\ell(z, z_0)\} \geq 1$, we can bound

$$E[d(z, z_0) \mid \mathcal{A}] \leq \sum_r r \sum_{z:d(z,z_0)=r} \exp\{\ell(z) - \ell(z_0) + \Pi_\ell(z, z_0)\}. \tag{A.2}$$

Next, we approximate the log-marginal likelihood $\ell(z)$ by a more manageable quantity, quantifying the approximation error. Recall the expression for $\mathcal{L}(\mathcal{A} \mid z)$ from (13) in the main document. To handle the combinatorial term, we use the well-known approximation $\log \binom{N}{s} \approx -N\mathcal{H}(s/N)$ (see, e.g., Chapter 1 of [2]), where $\mathcal{H} : [0, 1] \to \mathbb{R}$ is the (negative) Binary entropy function given by $\mathcal{H}(x) = x \log x + (1 - x) \log(1 - x)$.

In fact, using the two-sided Stirling bound $\sqrt{2\pi}N^{N+1/2}e^{-N} \leq N! \leq eN^{N+1/2}e^{-N}$, it is straightforward to verify that

$$\left|\log \binom{N}{s} - \big(-N\mathcal{H}(s/N)\big)\right| \leq C \log N,$$

where C is a global constant independent of s and N. Note that $\mathcal{H}(x) < 0, \mathcal{H}'(x) = \log\{x/(1 - x)\} = \text{logit}(x)$ and $\mathcal{H}''(x) = [x(1 - x)]^{-1}$ for all $x \in (0, 1)$. In particular, the positivity of the second derivative of $\mathcal{H}$ implies that $\mathcal{H}$ is convex over $(0, 1)$, a fact which is crucial to our subsequent derivations.

Using the above approximation and that $n_\uparrow(z), n_\downarrow(z) \leq n^2$, we can write $\ell(z) = \widetilde{\ell}(z) + \ell_R(z)$, where

$$\widetilde{\ell}(z) = n_\uparrow(z)\mathcal{H}\left\{\frac{A_\uparrow(z)}{n_\uparrow(z)}\right\} + n_\downarrow(z)\mathcal{H}\left\{\frac{A_\downarrow(z)}{n_\downarrow(z)}\right\}, \tag{A.3}$$

with the remainder term $|\ell_R(z)| \leq C \log n$ for a global constant C independent of z and n.

Putting together the various approximations, we have from (A.2) that

$$E[d(z, z_0) \mid \mathcal{A}] \leq \sum_r r \sum_{z:d(z,z_0)=r} \exp\{\widetilde{\ell}(z) - \widetilde{\ell}(z_0) + \Delta(z, z_0)\}, \tag{A.4}$$

where $\Delta(z, z_0) = \ell_R(z) - \ell_R(z_0) + \Pi_l(z, z_0)$. Since $|\Pi_l(z, z_0)| \leq CKd(z, z_0)$ by assumption, we have $|\Delta(z, z_0)| \leq C \max\{Kd(z, z_0), \log n\}$ for all z. We subsequently aim to bound $\widetilde{\ell}(z) - \widetilde{\ell}(z_0)$ from above inside a large $\mathbb{P}$-probability set. *The following result is key to our derivations.*

Proposition E.1. *Fix $\nu > 1$. There exists a set $\mathcal{C}$ with $\mathbb{P}(\mathcal{C}) \geq 1 - e^{-C(\log n)^\nu}$, such that for any $\mathcal{A} \in \mathcal{C}$, we have*

$$\widetilde{\ell}(z_0) - \widetilde{\ell}(z) \geq \frac{C\bar{D}(p_0, q_0)\ n\ d(z, z_0)}{K} \tag{A.5}$$

for all $z \in \mathcal{Z}_{n,K}$, where recall that

$$\bar{D}(p_0, q_0) := \frac{(p_0 - q_0)^2}{(p_0 \vee q_0)\{1 - (p_0 \wedge q_0)\}}. \tag{A.6}$$

Proposition E.1 quantifies the difference between the (approximate) log-marginal likelihood of the true configuration $\widetilde{\ell}(z_0)$ and that of any other configuration $\widetilde{\ell}(z)$ in terms of $d(z, z_0)$, the sample size n, the number of communities K, and the quantity $\bar{D}(p_0, q_0)$. The proof of Proposition E.1 is long and hence deferred to the next subsection. Substituting the bound (A.5) from Proposition E.1 in (A.4) and using the crude bound $|\{z \in \mathcal{Z}_{n,K} : d(z, z_0) = r\}| \leq K^r \binom{n}{r}$, we obtain, inside the set $\mathcal{C}$,

$$E[d(z, z_0) \mid r] \leq \sum_r r \binom{n}{r} K^r \exp\left\{ - \frac{C\bar{D}(p_0, q_0)\ nr}{K} + C \max\{Kr, \log n\} \right\} \leq e^{-\frac{C\bar{D}(p_0,q_0)n}{K}},$$

where the second inequality uses the crude bound $\binom{n}{r} \lesssim e^{r \log n}$ and the geometric sum formula. This establishes Theorem 4.1.

Proof of Proposition E.1

We now provide a running proof of Proposition E.1. We break the proof up into several parts which are somewhat independent of each other for improved readability. We first introduce some useful notation and collect some concentration inequalities. The concentration inequalities are used to define the large $\mathbb{P}$-probability set $\mathcal{C}$ in (A.19). The final part of the proof bounds $\widetilde{\ell}(z_0) - \widetilde{\ell}(z)$ inside $\mathcal{C}$. Readers primarily interested in the bound for the log-marginal likelihood difference can skip directly to the final part after familiarizing with the new notations.

Additional Notation:

For $z, z' \in \mathcal{Z}_{n,K}$, define

$$n_{\uparrow\uparrow}(z,z') = \sum_{i<j} \mathbb{1}(z_i = z_j, z'_i = z'_j), \quad A_{\uparrow\uparrow}(z,z') = \sum_{i<j} a_{ij}\mathbb{1}(z_i = z_j, z'_i = z'_j),$$
$$n_{\uparrow\downarrow}(z,z') = \sum_{i<j} \mathbb{1}(z_i = z_j, z'_i \neq z'_j), \quad A_{\uparrow\downarrow}(z,z') = \sum_{i<j} a_{ij}\mathbb{1}(z_i = z_j, z'_i \neq z'_j),$$
$$n_{\downarrow\uparrow}(z,z') = \sum_{i<j} \mathbb{1}(z_i \neq z_j, z'_i = z'_j), \quad A_{\downarrow\uparrow}(z,z') = \sum_{i<j} a_{ij}\mathbb{1}(z_i \neq z_j, z'_i = z'_j),$$
$$n_{\downarrow\downarrow}(z,z') = \sum_{i<j} \mathbb{1}(z_i \neq z_j, z'_i \neq z'_j), \quad A_{\downarrow\downarrow}(z,z') = \sum_{i<j} a_{ij}\mathbb{1}(z_i \neq z_j, z'_i \neq z'_j).$$

To simplify notation, we shall subsequently use $\dagger$ and $\dagger'$ as dummy variables taking values in the set $\{\uparrow, \downarrow\}$.[1] With this notation, $n_\dagger(z) = \sum_{\dagger,\dagger'} n_{\dagger\dagger'}(z,z')$ and $A_\dagger(z) = \sum_{\dagger,\dagger'} A_{\dagger\dagger'}(z,z')$ for any $z, z' \in \mathcal{Z}_{n,K}$. Denoting $\xi_\uparrow = p_0$ and $\xi_\downarrow = q_0$, we have

$$A_{\dagger\dagger'}(z, z_0) \sim \text{Binomial}(n_{\dagger\dagger'}(z, z_0), \xi_{\dagger'}), \tag{A.7}$$

independently across $\dagger, \dagger'$. For any $\dagger, \dagger'$, additionally denote

$$X_\dagger = \frac{A_\dagger(z)}{n_\dagger(z)}, \quad Y_\dagger = \frac{A_\dagger(z_0)}{n_\dagger(z_0)}, \quad W_{\dagger\dagger'} = \frac{A_{\dagger\dagger'}(z,z_0)}{n_{\dagger\dagger'}(z,z_0)} \tag{A.8}$$
$$\omega_{\dagger\dagger'} = \frac{n_{\dagger\dagger'}(z,z_0)}{n_\dagger(z)}, \quad \widetilde{\omega}_{\dagger\dagger'} = \frac{n_{\dagger\dagger'}(z,z_0)}{n_{\dagger'}(z_0)}. \tag{A.9}$$

It is straightforward to verify that

$$\sum_{\dagger'} \omega_{\dagger\dagger'} = 1, \quad X_\dagger = \sum_{\dagger'} \omega_{\dagger\dagger'} W_{\dagger\dagger'},$$
$$\sum_{\dagger} \widetilde{\omega}_{\dagger\dagger'} = 1, \quad Y_{\dagger'} = \sum_{\dagger} \widetilde{\omega}_{\dagger\dagger'} W_{\dagger\dagger'}.$$

It is evident from (A.7) that $\mathbb{E}W_{\dagger\dagger'} = \xi_{\dagger'}, \mathbb{E}Y_{\dagger'} = \xi_{\dagger'}$ and $\mathbb{E}X_\dagger = \bar{\xi}_\dagger := \sum_{\dagger'} \omega_{\dagger\dagger'}\xi_{\dagger'}$. Further, since the random variables involved are sub-Gaussian, they concentrate around their mean with large probability. We collect some useful concentration bounds next.

Concentration bounds: Fix $z \neq z_0 \in \mathcal{Z}_{n,K}$ with $d(z, z_0) = r$. For a constant $\nu > 1$, let

$$\mathcal{C}_X(z) = \left\{ |X_\dagger - \bar{\xi}_\dagger| \leq \frac{(\log n)^{\nu/2}\sqrt{r}}{\sqrt{n_\dagger(z)}}, \forall\ \dagger \right\} \tag{A.10}$$
$$\mathcal{C}_Y(z) = \left\{ |Y_\dagger - \xi_\dagger| \leq \frac{(\log n)^{\nu/2}\sqrt{r}}{\sqrt{n_\dagger(z_0)}}, \forall\ \dagger \right\}. \tag{A.11}$$

[1] For example, $\sum_\dagger n_\dagger(z)$ is shorthand for $n_\uparrow(z) + n_\downarrow(z)$.

For $T_i \sim \text{Bernoulli}(p_i)$ independently for $i = 1, \ldots, N$, it follows from Hoeffding's inequality that $P(|\bar{T} - \bar{p}| > t) \leq 2e^{-2nt^2}$ for any $t > 0$, where $\bar{p} = N^{-1} \sum_{i=1}^{N} p_i$. Combining with the union bound, it follows that

$$\mathbb{P}\big[\mathcal{C}_X(z) \cap \mathcal{C}_Y(z)\big] \geq 1 - 8\, e^{-r(\log n)^{\nu}}. \tag{A.12}$$

We additionally need control on another set of random variables that appear inside Taylor expansions subsequently. Define, for each $\dagger$,

$$L_\dagger = \sum_{\dagger'} \omega_{\dagger\dagger'} Y_{\dagger'} - X_\dagger = \sum_{\dagger'} \omega_{\dagger\dagger'} (Y_{\dagger'} - W_{\dagger\dagger'}). \tag{A.13}$$

For any $\dagger$, define $\ddagger$ to be the reverse spin of $\dagger$, that is, $\ddagger = \downarrow$ if $\dagger = \uparrow$ and vice versa. With this notation, $Y_{\dagger'} - W_{\dagger\dagger'} = \widetilde{\omega}_{\dagger\dagger'} W_{\dagger\dagger'} + \widetilde{\omega}_{\ddagger\dagger'} W_{\ddagger\dagger'} - W_{\dagger\dagger'} = \widetilde{\omega}_{\ddagger\dagger'} (W_{\ddagger\dagger'} - W_{\dagger\dagger'})$, since $1 - \widetilde{\omega}_{\dagger\dagger'} = \widetilde{\omega}_{\ddagger\dagger'}$. Substituting in (A.13),

$$L_\dagger = \sum_{\dagger'} \omega_{\dagger\dagger'} \widetilde{\omega}_{\ddagger\dagger'} (W_{\ddagger\dagger'} - W_{\dagger\dagger'}). \tag{A.14}$$

Observe that $W_{\ddagger\dagger'}$ and $W_{\dagger\dagger'}$ are independent random variables with $\mathbb{E}W_{\ddagger\dagger'} = \mathbb{E}W_{\dagger\dagger'} = \xi_{\dagger'}$, implying $\mathbb{E}L_\dagger = 0$. Define

$$\mathcal{C}_L(z) = \left\{ |L_\dagger| \leq \frac{C(\log n)^{\nu/2} \sqrt{r}\, \sqrt{\mathbf{n}(z, z_0)}}{n_\dagger(z)},\ \forall\ \dagger \right\}, \tag{A.15}$$

where

$$\mathbf{n}(z, z_0) = \frac{n_{\uparrow\uparrow}(z, z_0) n_{\downarrow\uparrow}(z, z_0)}{n_\uparrow(z_0)} + \frac{n_{\uparrow\downarrow}(z, z_0) n_{\downarrow\downarrow}(z, z_0)}{n_\downarrow(z_0)}. \tag{A.16}$$

Using a sub-Gaussian concentration inequality, we prove below that

$$\mathbb{P}\big[\mathcal{C}_L(z)\big] \geq 1 - 6e^{-r(\log n)^{\nu}}. \tag{A.17}$$

The main idea to establish (A.17) is to recognize $L_\dagger$ as a weighted sum of centered Bernoulli variables in (A.14) and use a rotation invariance property of sub-Gaussian random variables to bound the sub-Gaussian norm of the aforesaid random variable.

Let us recall some useful facts about sub-Gaussian random variables from §5.2.3 of [5]. A mean zero random variable Z is called sub-Gaussian if $E(e^{tZ}) \leq e^{Ct^2 \|Z\|_{\psi_2}^2}$ for all $t \in \mathbb{R}$, where $\|Z\|_{\psi_2} = \sup_{s \geq 1} s^{-1/2} (E|Z|^s)^{1/s}$ is the sub-Gaussian norm of Z and C is an absolute constant. Sub-Gaussian random variables satisfy Gaussian-like tail bounds: $P(|Z| > t) \leq Ce^{-ct^2/\|Z\|_{\psi_2}^2}$,

with $C < 3$. The following rotation invariance property is useful: if $Z_1, \ldots, Z_N$ are independent sub-Gaussian random variables, then $Z = \sum_{i=1}^N a_i Z_i$ is also sub-Gaussian, with

$$\|Z\|_{\psi_2}^2 \leq C \sum_{i=1}^N a_i^2 \, \|Z_i\|_{\psi_2}^2 \,,$$

for some absolute constant C.

Any centered Bernoulli random variable is sub-Gaussian, with sub-Gaussian norm bounded by 1. Since $L_\dagger$ is a weighted sum of Bernoulli random variables, $L_\dagger$ is also sub-Gaussian. Let us attempt to bound the sub-Gaussian norm of $L_\dagger$. First, in (A.14), write $W_{\ddagger\dagger'} - W_{\dagger\dagger'} = (W_{\ddagger\dagger'} - \xi_{\dagger'}) - (W_{\dagger\dagger'} - \xi_{\dagger'})$ as a weighted sum of centered Bernoulli random variables. By rotation invariance,

$$\|W_{\ddagger\dagger'} - W_{\dagger\dagger'}\|_{\psi_2}^2 \leq C\bigg(\frac{1}{n_{\ddagger\dagger'}} + \frac{1}{n_{\dagger\dagger'}}\bigg).$$

Another application of rotation invariance yields,

$$\begin{aligned}\|L_\dagger\|_{\psi_2}^2 &\leq C \sum_{\dagger'} \omega_{\dagger\dagger'}^2 \widetilde{\omega}_{\ddagger\dagger'}^2 \bigg(\frac{1}{n_{\ddagger\dagger'}(z, z_0)} + \frac{1}{n_{\dagger\dagger'}(z, z_0)}\bigg) \\ &= \frac{C}{n_\dagger^2(z)} \sum_{\dagger'} \frac{n_{\dagger\dagger'}(z, z_0)\, n_{\ddagger\dagger'}(z, z_0)}{n_{\dagger'}(z_0)} = \frac{C\mathbf{n}(z, z_0)}{n_\dagger^2(z)},\end{aligned}$$

using the definitions in (A.8) and (A.9) from the first to the second line, and noting that the summation in the penultimate line equals $\mathbf{n}(z, z_0)$ defined in (A.16).

From the general tail bound for sub-Gaussian random variables mentioned previously (see paragraph after equation (A.17)), we have $\mathbb{P}(|L_\dagger| > t) \leq 3e^{-Ct^2/\|L_\dagger\|_{\psi_2}^2}$ for any $t > 0$. Set $t^* = C(\log n)^{\nu/2}\sqrt{r}\,\sqrt{\mathbf{n}(z, z_0)}/n_\dagger(z)$ for an appropriate C and use that $e^{-1/x}$ is increasing in x to obtain $\mathbb{P}(|L_\dagger| > t^*) \leq 3e^{-r(\log n)^\nu}$. The inequality (A.17) follows from an application of the union bound over $\dagger$.

Constructing large probability set:

We use the concentration bounds above to create the large probability set $\mathcal{C}$ in Proposition E.1 within which the log-marginal likelihood differences can be appropriately bounded. Define,

$$\mathcal{C}_r = \cap_{z: d(z, z_0) = r} \big[\mathcal{C}_X(z) \cap \mathcal{C}_Y(z) \cap \mathcal{C}_L(z)\big], \quad \mathcal{C} = \cap_{r=1}^n \mathcal{C}_r. \tag{A.18}$$

We have,

$$\mathbb{P}\big[\mathcal{C}_r^c\big] \leq C\,|z : d(z, z_0) = r|\; e^{-r(\log n)^\nu} \leq C\binom{n}{r} K^r e^{-r(\log n)^\nu} \leq e^{-Cr(\log n)^\nu}.$$

For the first inequality in the above display, we used the union bound to (A.12) and (A.17). The second inequality uses the crude upper bound $|z : d(z, z_0) = r| \leq \binom{n}{r} K^r$, whereas the last inequality uses the bound $\binom{n}{r} \leq e^{r \log n}$ and the fact that $\nu > 1$. Another application of the union bound yields

$$\mathbb{P}\big(\mathcal{C}\big) \geq 1 - e^{-C(\log n)^{\nu}}. \tag{A.19}$$

Bounding the log-marginal likelihood differences:

Fix z with $d(z, z_0) = r$. Recall the approximation $\widetilde{\ell}(\cdot)$ to the log-marginal likelihood from (A.3). We now proceed to bound $\widetilde{\ell}(z_0) - \widetilde{\ell}(z)$ from below inside the set $\mathcal{C}$. Using the notation introduced in (A.8) and (A.9), we can write

$$\widetilde{\ell}(z) = \sum_{\dagger} n_{\dagger}(z) \mathcal{H}(X_{\dagger}),$$

and

$$\widetilde{\ell}(z_0) = \sum_{\dagger'} n_{\dagger'}(z_0) \mathcal{H}(Y_{\dagger'}) = \sum_{\dagger'} \sum_{\dagger} n_{\dagger\dagger'}(z, z_0) \mathcal{H}(Y_{\dagger'}) = \sum_{\dagger} n_{\dagger}(z) \Big[\sum_{\dagger'} \omega_{\dagger\dagger'} \mathcal{H}(Y_{\dagger'}) \Big].$$

Thus, $\widetilde{\ell}(z_0) - \widetilde{\ell}(z) = \sum_{\dagger} n_{\dagger}(z) \big[\sum_{\dagger'} \omega_{\dagger\dagger'} \mathcal{H}(Y_{\dagger'}) - \mathcal{H}(X_{\dagger}) \big]$. To tackle the inner sum, we perform a Taylor expansion of each $\mathcal{H}(Y_{\dagger'})$ around $\mathcal{H}(X_{\dagger})$. After some cancellations since $\sum_{\dagger'} \omega_{\dagger\dagger'} = 1$, we obtain

$$\widetilde{\ell}(z_0) - \widetilde{\ell}(z) = \sum_{\dagger} n_{\dagger}(z) \Big[\sum_{\dagger'} \omega_{\dagger\dagger'} \Big\{ (Y_{\dagger'} - X_{\dagger}) \mathcal{H}'(X_{\dagger}) + \frac{(Y_{\dagger'} - X_{\dagger})^2}{2} \mathcal{H}''(U_{\dagger'\dagger}) \Big\} \Big], \tag{A.20}$$

where $U_{\dagger'\dagger}$ lies between $Y_{\dagger'}$ and $X_{\dagger}$.

Since $\mathcal{H}$ is convex, the quadratic term in (A.20) is positive. We show below that the quadratic term is the dominant term and the linear term is of smaller order. To that end, we first bound the magnitude of the linear term inside $\mathcal{C}$. Since from (A.10), $X_{\dagger}$ concentrates around $\bar{\xi}_{\dagger}$, and $\bar{\xi}_{\dagger}$ lies between p_0 and q_0, $|\mathcal{H}'(X_{\dagger})|$ can be bounded by a constant inside $\mathcal{C}$. Hence, inside $\mathcal{C}$,

$$\Big| \sum_{\dagger} n_{\dagger}(z) \sum_{\dagger'} \omega_{\dagger\dagger'} (Y_{\dagger'} - X_{\dagger}) \mathcal{H}'(X_{\dagger}) \Big| \leq C \sum_{\dagger} n_{\dagger}(z) \, |L_{\dagger}| \leq C (\log n)^{\nu/2} \sqrt{r} \, \sqrt{\mathbf{n}(z, z_0)}, \tag{A.21}$$

where recall from (A.13) that $L_{\dagger} = \sum_{\dagger'} \omega_{\dagger\dagger'} (Y_{\dagger'} - X_{\dagger})$. From the second to third step, we used the bound on $|L_{\dagger}|$ inside $\mathcal{C}$ from (A.15).

Next, we bound from below the quadratic term in (A.20). Since $U_{\dagger'\dagger}$ lies between $Y_{\dagger'}$ and $X_\dagger$ which in turn concentrate around their respective means inside $\mathcal{C}$, we can bound $H''(U_{\dagger'\dagger})$ from below as follows:

$$\mathcal{H}''(U_{\dagger'\dagger}) = \frac{1}{U_{\dagger'\dagger}(1-U_{\dagger'\dagger})} \geq \frac{1}{(p_0 \vee q_0)\{1-(p_0 \wedge q_0)\}},$$

where $\vee$ and $\wedge$ respectively denote the maximum and minimum. Thus,

$$\sum_\dagger n_\dagger(z) \sum_{\dagger'} \omega_{\dagger\dagger'} \frac{(Y_{\dagger'} - X_\dagger)^2}{2} \mathcal{H}''(U_{\dagger'\dagger}) \geq \frac{\sum_\dagger \sum_{\dagger'} n_{\dagger\dagger'}(z,z_0)(Y_{\dagger'} - X_\dagger)^2}{(p_0 \vee q_0)\{1-(p_0 \wedge q_0)\}}.$$

Write

$$(Y_{\dagger'} - X_\dagger) = (\xi_{\dagger'} - \bar{\xi}_\dagger) + (Y_{\dagger'} - \xi_{\dagger'}) + (X_\dagger - \bar{\xi}_\dagger).$$

The bounds on $|Y_{\dagger'} - \xi_{\dagger'}|$ and $|X_\dagger - \bar{\xi}_\dagger|$ from (A.11) and (A.10) imply that $(\xi_{\dagger'} - \bar{\xi}_\dagger)$ is the leading term in the above display. Since we can bound $(a+b)^2 \geq a^2/2$ if $|b| = o(|a|)$, we obtain, inside $\mathcal{C}$,

$$\sum_\dagger \sum_{\dagger'} n_{\dagger\dagger'}(z,z_0)(Y_{\dagger'} - X_\dagger)^2 \geq \frac{1}{2} \sum_\dagger \sum_{\dagger'} n_{\dagger\dagger'}(z,z_0)(\xi_{\dagger'} - \bar{\xi}_\dagger)^2. \tag{A.22}$$

We have $(\xi_{\dagger'} - \bar{\xi}_\dagger) = (\xi_{\dagger'} - \omega_{\dagger\dagger'}\xi_{\dagger'} - \omega_{\dagger\ddagger'}\xi_{\ddagger'}) = \omega_{\dagger\ddagger'}(\xi_{\dagger'} - \xi_{\ddagger'})$, since $\omega_{\dagger\ddagger'} = 1 - \omega_{\dagger\dagger'}$. Also, $|\xi_{\dagger'} - \xi_{\ddagger'}| = |p_0 - q_0|$. Hence

$$\begin{aligned}
&\sum_\dagger \sum_{\dagger'} n_{\dagger\dagger'}(z,z_0)(\xi_{\dagger'} - \bar{\xi}_\dagger)^2 \qquad (A.23)\\
&= \sum_\dagger \sum_{\dagger'} n_{\dagger\dagger'}(z,z_0)\omega_{\dagger\ddagger'}^2 (p_0 - q_0)^2 \\
&= (p_0 - q_0)^2 \sum_\dagger \sum_{\dagger'} n_{\dagger\dagger'}(z,z_0) \frac{n_{\dagger\ddagger'}^2(z,z_0)}{n_\dagger^2(z)} \\
&= (p_0 - q_0)^2 \sum_\dagger \frac{n_{\dagger\uparrow}(z,z_0) n_{\dagger\downarrow}(z,z_0)}{n_\dagger(z)}, \qquad (A.24)
\end{aligned}$$

since

$$\sum_{\dagger'} n_{\dagger\dagger'}(z,z_0) n_{\dagger\ddagger'}^2(z,z_0) = n_{\dagger\uparrow}(z,z_0) n_{\dagger\downarrow}^2(z,z_0) + n_{\dagger\downarrow} n_{\dagger\uparrow}^2(z,z_0) = n_{\dagger\uparrow}(z,z_0) n_{\dagger\downarrow}(z,z_0) n_\dagger(z).$$

Define

$$\widetilde{\mathbf{n}}(z,z_0) = \sum_\dagger \frac{n_{\dagger\uparrow}(z,z_0) n_{\dagger\downarrow}(z,z_0)}{n_\dagger(z)} = \frac{n_{\uparrow\uparrow}(z,z_0) n_{\uparrow\downarrow}(z,z_0)}{n_\uparrow(z)} + \frac{n_{\downarrow\uparrow}(z,z_0) n_{\downarrow\downarrow}(z,z_0)}{n_\downarrow(z)} \tag{A.25}$$

We then have, from (A.24), (A.22), and (A.21), that inside $\mathcal{C}$,

$$\widetilde{\ell}(z_0) - \widetilde{\ell}(z) \geq \frac{(p_0 - q_0)^2}{2(p_0 \vee q_0)\{1-(p_0 \wedge q_0)\}} \widetilde{\mathbf{n}}(z,z_0) - C'(\log n)^{\nu/2} \sqrt{r}\, \sqrt{\mathbf{n}(z,z_0)}. \tag{A.26}$$

We now state a Lemma to bound $\widetilde{\mathbf{n}}(z,z_0)$ and $\mathbf{n}(z,z_0)$ in appropriate directions.

Lemma E.1. *Suppose $K \geq 2$ and $d(z, z_0) = r$. Then, $\widetilde{\mathbf{n}}(z, z_0) \geq min\{Crn/K, Cn^2/K^2\}$ and $\mathbf{n}(z, z_0) \leq C\{nr/K + r^2\}$ for some constant $C > 0$, where $\widetilde{\mathbf{n}}(z, z_0)$ and $\mathbf{n}(z, z_0)$ are defined in* (A.25) *and* (A.16) *respectively.*

The proof of Lemma E.1 is provided in the Appendix G. Substituting the inequalities in Lemma E.1 to (A.26) delivers the bound (A.5) in Proposition E.1.

F. PROOF OF THEOREM 4.3

We first introduce a few notations. Since d_H is not defined between two configurations with different values of k, we instead work with the Rand-Index (R) in the subsequent developments. Define

$$n_{\alpha\beta} = |i : z_i = \alpha, z_i^0 = \beta|,\ \alpha = 1, \ldots, k, \beta = 1, 2; \quad n_\alpha = |i : z_i = \alpha|,\ \alpha = 1, \ldots, k,$$
$$B = 2\sum_{\alpha=1}^{k} n_{\alpha 1} n_{\alpha 2}, \quad R = \frac{n_{\uparrow\uparrow}(z, z_0) + n_{\downarrow\downarrow}(z, z_0)}{\binom{n}{2}}.$$

Clearly $0 \leq R \leq 1$ and $R = 1$ indicates perfect concordance between the configurations z and z_0. To find a lower bound to $\Pi(K \mid \mathcal{A})$, it is enough to find an upper bound to the Bayes factor $\mathcal{L}(\mathcal{A} \mid k)/\mathcal{L}(\mathcal{A} \mid K)$. Observe that

$$\frac{\mathcal{L}(\mathcal{A} \mid k)}{\mathcal{L}(\mathcal{A} \mid K)} \leq \sum_{z \in Z_{n,k}} \frac{\mathcal{L}(A \mid z, k)}{\mathcal{L}(\mathcal{A} \mid z_0, K)} \frac{\Pi(z \mid k)}{\Pi(z_0 \mid K)}. \tag{A.27}$$

Straightforward calculations yield, for the Dirichlet-multinomial prior with Dirichlet concentration parameter γ,

$$\frac{\Pi(z \mid k = 3)}{\Pi(z_0 \mid K = 2)} \leq c_1 e^{nc_2}, \quad \frac{\Pi(z \mid k = 2)}{\Pi(z_0 \mid K = 3)} \leq c_3 e^{c_4 n \log n}. \tag{A.28}$$

Since the analysis leading up to (A.26) does not depend on whether or not z and z_0 share the same k, we have

$$\frac{\mathcal{L}(A \mid z, k)}{\mathcal{L}(\mathcal{A} \mid z_0, K)} \leq \exp\{C' t_n \sqrt{\mathbf{n}(z, z_0)} - \bar{D}(p_0, q_0)\widetilde{\mathbf{n}}(z, z_0)\} \tag{A.29}$$

with probability $1 - e^{-Ct_n^2}$. Denote by $\mathcal{C}$ the set corresponding to the high-probability event in (A.29). In the following, we derive a lower bound for $\mathbf{n}(z, z_0)$ respectively for the following two cases. In both the cases, the upper bound for $\mathbf{n}(z, z_0)$ follows trivially.

1. Overfitted case ($K = 2$ and the model is fitted with $k = 3$): Since the true model is contained in the fitted model, a value of R close to 1 impedes the concentration of k around $K = 2$. We derive lower bound for $\widetilde{\mathbf{n}}(z, z_0)$ in terms of the Rand-Index R and investigate the bounds for different

regimes of R. $R \asymp 1$ corresponds to the case when the separation between the log-marginal likelihoods is relatively weak, but strong enough to offset the model complexity and the prior. In this case $\widetilde{\mathbf{n}}(z, z_0)$ and $\mathbf{n}(z, z_0)$ both are of the order n; however the number of such configurations is polynomial in n, so that the posterior concentrates at $K = 2$ with a rate e^{-Cn}.

2. Underfitted case ($K = 3$ and the model is fitted with $k = 2$): In the underfitted case, R can never approach 1 which makes separation between the log-marginal likelihoods stronger. In this case both $\mathbf{n}(z, z_0)$ and $\widetilde{\mathbf{n}}(z, z_0)$ are of the order n^2 which is enough to offset the model complexity leading to a posterior concentration rate of e^{-n^2}.

In the following, we analyze the above two cases separately.

1. Overfitted case: Here $K = 2$ and $m = n/2$ and

$$\begin{aligned} n_{\uparrow\uparrow}(z, z_0) &= \frac{\sum_{\alpha=1}^{k}(n_{\alpha 1}^2 + n_{\alpha 2}^2)}{2} - m, \quad n_{\uparrow\downarrow}(z, z_0) = \sum_{\alpha=1}^{k} n_{\alpha 1} n_{\alpha 2} \\ n_{\downarrow\uparrow}(z, z_0) &= m^2 - \frac{\sum_{\alpha=1}^{k}(n_{\alpha 1}^2 + n_{\alpha 2}^2)}{2}, \quad n_{\downarrow\downarrow}(z, z_0) = m^2 - \sum_{\alpha=1}^{k} n_{\alpha 1} n_{\alpha 2}. \end{aligned}$$

We express $\mathbf{n}(z, z_0)$ and $\widetilde{\mathbf{n}}(z, z_0)$ in terms of R as

$$\mathbf{n}(z, z_0) = \frac{n_{\uparrow\uparrow}(z, z_0) n_{\downarrow\downarrow}(z, z_0)(1 - R) + n_{\uparrow\downarrow}(z, z_0) n_{\downarrow\uparrow}(z, z_0) R}{(m^2 - m) m^2 / \binom{n}{2}}, \tag{A.30}$$

$$\widetilde{\mathbf{n}}(z, z_0) = \frac{n_{\uparrow\uparrow}(z, z_0) n_{\downarrow\downarrow}(z, z_0)(1 - R) + n_{\uparrow\downarrow}(z, z_0) n_{\downarrow\uparrow}(z, z_0) R}{n_{\uparrow}(z) n_{\downarrow}(z) / \binom{n}{2}}. \tag{A.31}$$

Lemma F.1 derives upper and lower bounds for $\mathbf{n}(z, z_0)$ and $\widetilde{\mathbf{n}}(z, z_0)$ depending on 5 possible range of values for R. For cases 1 and 2, $t_n \sqrt{\mathbf{n}(z, z_0)} - \bar{D}(p_0, q_0)\widetilde{\mathbf{n}}(z, z_0) \leq n t_n - \bar{D}(p_0, q_0) n^2$. For Cases 3 and 4, the bounds are $\{n t_n \sqrt{\eta_n} - \bar{D}(p_0, q_0) n^2 \eta_n\}$ and $\{t_n \sqrt{n} - \bar{D}(p_0, q_0) n^2 \eta_n\}$ respectively.

Thus for each of the cases 1-4, the bound for the ratio of the marginal likelihood in (A.29) is faster than exponential. For Case 5, the bound is $C\{t_n \sqrt{n} - \bar{D}(p_0, q_0) n\}$. Note that this means the ratio of the marginal likelihood in (A.29) can be at the minimum e^{-Cn} for Case 5. However, for z satisfying Case 5, one can improve on the bound of the prior ratio in (A.28) as

$$\frac{\Pi(z \mid k = 3)}{\Pi(z_0 \mid K = 2)} \leq C \sqrt{n} (n + 2)^3. \tag{A.32}$$

The proof of (A.32) is appended with the proof of Lemma F.1. Thus for Case 5, we have

$$\frac{\mathcal{L}(A \mid z, k = 3)}{\mathcal{L}(\mathcal{A} \mid z_0, K = 2)} \frac{\Pi(z \mid k = 3)}{\Pi(z_0 \mid K = 2)} \leq e^{-Cn}.$$

Instead of a global bound on the model complexity, we separately analyze the complexity of configurations corresponding to Cases 1-4 and 5. From the proof of Lemma F.1, configurations

corresponding to Case 5 satisfy the following: choose a constant a from m observations in cluster one and a constant value b from cluster two, then randomly place $a+b$ nodes into three clusters. The number such configurations is at most polynomial in n, say n^κ for some $\kappa > 0$.

For Cases 1-4, choose $t_n = o(n\sqrt{\eta_n})$ with $3^n e^{-Ct_n^2} \to 0$. For Case 5, choose $t_n = o(\sqrt{n})$ with $n^\kappa e^{-Ct_n^2} \to 0$. Then $\mathbb{P}(\mathcal{C}^c) \to 0$. Hence the right hand side of (A.27) can be bounded by $3^n \exp\{-Cn^2\eta_n\} + n^\kappa \exp\{-Cn\}$ which can be upper bounded by $\exp\{-Cn\}$.

Lemma F.1. *1. If $1-2R \asymp \beta_n$ or $1-2R \asymp Cm^{-1}$ with $\beta_n \to 0$ and $m\beta_n \to 0$, $\mathbf{n}(z,z_0) \le Cn^2$ and $\widetilde{\mathbf{n}}(z,z_0) \ge Cn^2$.*

2. If either $1-R$ or $1-2R$ are constants, $\mathbf{n}(z,z_0) \le Cn^2$ and $\widetilde{\mathbf{n}}(z,z_0) \ge Cn^2$.

3. If $1-R \asymp \eta_n$ with $\eta_n \to 0$ and $m\eta_n \to \infty$, $\mathbf{n}(z,z_0) \le Cn^2\eta_n$ and $\widetilde{\mathbf{n}}(z,z_0) \ge Cn^2\eta_n$.

4. When $1-R = C/m$ and $B/m \to \infty$ and $B/(m^2\eta_n) \to C$, then $\mathbf{n}(z,z_0) \le Cn$, $\widetilde{\mathbf{n}}(z,z_0) \ge n^2\eta_n$.

5. When $1-R = C/m$ for some constant $C > 0$, and $B = Cm$, then $\mathbf{n}(z,z_0) \le Cn$ and $\widetilde{\mathbf{n}}(z,z_0) \ge Cn$

2. Underfitted case: Assume $K = 3$ and $m = n/3$. Then $B = 2\sum_{\alpha=1}^k (n_{\alpha 1}n_{\alpha 2} + n_{\alpha 1}n_{\alpha 3} + n_{\alpha 2}n_{\alpha 3})$. Also, note that

$$\begin{aligned} n_{\uparrow\uparrow}(z,z_0) &= \frac{\sum_{\alpha=1}^k (n_{\alpha 1}^2 + n_{\alpha 2}^2 + n_{\alpha 3}^2)}{2} - \frac{3m}{2}, \quad n_{\uparrow\downarrow}(z,z_0) = \sum_{\alpha=1}^k (n_{\alpha 1}n_{\alpha 2} + n_{\alpha 1}n_{\alpha 3} + n_{\alpha 2}n_{\alpha 3}) \\ n_{\downarrow\uparrow}(z,z_0) &= \frac{3m^2}{2} - \frac{\sum_{\alpha=1}^k (n_{\alpha 1}^2 + n_{\alpha 2}^2 + +n_{\alpha 3}^2)}{2}, \quad n_{\downarrow\downarrow}(z,z_0) = 3m^2 - n_{\downarrow\uparrow}(z,z_0). \end{aligned}$$

It is straightforward to show $n_{\uparrow\uparrow}(z,z_0) \ge Cn^2$. Also,

$$B = \frac{n_1^2+n_2^2}{2} + \left(3 - \frac{9}{2}R\right)m^2 + \left(\frac{3}{2}R - \frac{3}{2}\right)m \ge \frac{9}{4}m^2 + \left(3 - \frac{9}{2}R\right)m^2 + \left(\frac{3}{2}R - \frac{3}{2}\right)m = Cn^2. \tag{A.33}$$

The first inequality in (A.33) follows because $n_1^2 + n_2^2 \ge 2(n/2)^2$ and $n = 3m$. The last equality in (A.33) follows since $0 \le R \le 1$. Hence $n_{\uparrow\uparrow}(z,z_0)n_{\uparrow\downarrow}(z,z_0)/n_\uparrow(z) \ge Cn^2$ and thus $\widetilde{\mathbf{n}}(z,z_0) \ge Cn^2$. Choosing $t_n = o(n)$ concludes the proof.

G. PROOF OF A FEW AUXILIARY LEMMATA

G.1 **Proof of Lemma E.1**

We introduce some additional notations to analyze the terms $n_{\uparrow\downarrow}(z,z_0)$ and $n_{\downarrow\uparrow}(z,z_0)$. Set $m = n/K$ and define $a_k = |\{i : z_i \neq k, z_i^0 = k\}|$, $b_k = |\{i : z_i = k, z_i^0 \neq k\}|$, $n_k = |\{i : z_i = k\}|$ and $n_k^0 = |\{i : z_i^0 = k\}| = m$ for all $k = 1, \ldots, K$. Clearly, $\sum_{k=1}^K a_k = \sum_{k=1}^K b_k = r$ and $n_k^0 - a_k = n_k - b_k$. Fix z with $d(z,z_0) = r$. Then $0 \leq r \leq n - m$. Defining $n_{\uparrow\downarrow}^{(k)}(z,z_0) = |\{i : z_i = z_j = k, z_i^0 \neq z_j^0\}|$ and $n_{\downarrow\uparrow}^{(k)}(z,z_0) = |\{i : z_i^0 = z_j^0 = k, z_i \neq z_j\}|$, we write

$$n_{\uparrow\downarrow}(z,z_0) = \sum_{k=1}^K n_{\uparrow\downarrow}^{(k)}(z,z_0), \quad n_{\downarrow\uparrow}(z,z_0) = \sum_{k=1}^K n_{\downarrow\uparrow}^{(k)}(z,z_0).$$

Observe that,

$$\begin{aligned} n_{\uparrow\downarrow}^{(k)}(z,z_0) &\geq \big|\{i : z_i = k, z_i^0 = k\}\big|\,\big|\{i : z_i = k, z_i^0 \neq k\}\big| = (n_k - b_k)b_k \\ n_{\downarrow\uparrow}^{(k)}(z,z_0) &\geq \big|\{i : z_i = k, z_i^0 = k\}\big|\,\big|\{i : z_i \neq k, z_i^0 = k\}\big| = (n_k^0 - a_k)a_k. \end{aligned}$$

Proof of lower bound on $\widetilde{\mathbf{n}}(z,z_0)$:

Note that

$$\begin{aligned} \sum_{\dagger} \frac{\prod_{\dagger'} n_{\dagger\dagger'}(z,z_0)}{n_{\dagger}(z)} &= \frac{n_{\uparrow\uparrow}(z,z_0)n_{\uparrow\downarrow}(z,z_0)}{n_{\uparrow\uparrow}(z,z_0) + n_{\uparrow\downarrow}(z,z_0)} + \frac{n_{\downarrow\downarrow}(z,z_0)n_{\downarrow\uparrow}(z,z_0)}{n_{\downarrow\downarrow}(z,z_0) + n_{\downarrow\uparrow}(z,z_0)} \\ &= \frac{n_{\uparrow\downarrow}(z,z_0)}{1 + \frac{n_{\uparrow\downarrow}(z,z_0)}{n_{\uparrow\uparrow}(z,z_0)}} + \frac{n_{\downarrow\uparrow}(z,z_0)}{1 + \frac{n_{\downarrow\uparrow}(z,z_0)}{n_{\downarrow\downarrow}(z,z_0)}} := T_1 + T_2. \end{aligned}$$

The proof is based on the following three inequalities:

$$n_{\uparrow\downarrow}(z,z_0) + n_{\downarrow\uparrow}(z,z_0) \geq Crm, \tag{A.34}$$

$$n_{\uparrow\uparrow}(z,z_0) \geq Cm^2, \tag{A.35}$$

$$n_{\downarrow\downarrow}(z,z_0) \geq 2n_{\downarrow\uparrow}(z,z_0) - n_{\uparrow\downarrow}(z,z_0). \tag{A.36}$$

Hence $C > 0$ denotes a generic constant. By (A.34), either $n_{\uparrow\downarrow}(z,z_0) \geq Crm/2$ or $n_{\downarrow\uparrow}(z,z_0) \geq Crm/2$. If $n_{\uparrow\downarrow}(z,z_0) \geq Crm/2$, then $T_1 \geq Crm$ since $n_{\uparrow\uparrow}(z,z_0) \geq Cm^2$ by (A.35). If $n_{\downarrow\uparrow}(z,z_0) \geq Crm/2$ and $n_{\uparrow\downarrow}(z,z_0) < Crm/2$, $n_{\downarrow\uparrow}(z,z_0)/n_{\uparrow\downarrow}(z,z_0) > 1$. Then by (A.36), $n_{\downarrow\downarrow}(z,z_0) \geq 2n_{\downarrow\uparrow}(z,z_0) - n_{\uparrow\downarrow}(z,z_0)$ and hence

$$\frac{n_{\downarrow\uparrow}(z,z_0)}{1 + \frac{n_{\downarrow\uparrow}(z,z_0)}{n_{\downarrow\downarrow}(z,z_0)}} \geq \frac{n_{\downarrow\uparrow}(z,z_0)}{1 + \frac{n_{\downarrow\uparrow}(z,z_0)}{2n_{\downarrow\uparrow}(z,z_0) - n_{\uparrow\downarrow}(z,z_0)}} > \frac{n_{\downarrow\uparrow}(z,z_0)}{2}.$$

Thus $T_2 \geq Crm$. The lower bound on $\widetilde{\mathbf{n}}(z,z_0)$ then follows immediately.

We next turn our attention to proving (A.34) - (A.36). We first show (A.35). Defining $n_{\uparrow\uparrow}^{(k)}(z, z_0) = |\{(i,j) : z_i = z_j = k, z_i^0 = z_j^0\}|$, observe that

$$\begin{aligned}
n_{\uparrow\uparrow}(z, z_0) &= \sum_{k=1}^{K} n_{\uparrow\uparrow}^{(k)}(z, z_0) \\
&\geq \sum_{k=1}^{K} \binom{n_k - b_k}{2} = \sum_{k=1}^{K} (n_k^0 - a_k) \frac{n_k - b_k - 1}{2} = \frac{n}{K} \sum_{k=1}^{K} \frac{n_k - b_k - 1}{2} - \sum_{k=1}^{K} a_k \frac{n_k^0 - a_k - 1}{2} \\
&= \frac{n^2}{2K} - \frac{nr}{2K} - \frac{n}{2} - \frac{nr}{2K} + \frac{r}{2} + \sum_{k=1}^{K} \frac{a_k^2}{2} \\
&\geq \frac{n^2}{2K} - \frac{nr}{K} - \frac{n}{2} + \frac{r}{2} + \frac{r^2}{2K} = \frac{(n-r)^2}{2K} + \frac{r-n}{2} = Cm^2 \qquad \text{(A.37)}
\end{aligned}$$

for some constant $C > 0$. The inequality in (A.37) follows since $\sum a_k^2$ is minimized at $a_k = r/K$.

Next, we show (A.36). Observe that

$$\begin{aligned}
n_{\downarrow\downarrow}(z, z_0) &= \left|\{(i,j) : z_i \neq z_j, z_i^0 \neq z_j^0\}\right| = \left|\{(i,j) : z_i^0 \neq z_j^0\}\right| - n_{\uparrow\downarrow}(z, z_0) \\
&= \binom{n}{2} - K\binom{m}{2} - n_{\uparrow\downarrow}(z, z_0) = \frac{(K-1)K}{2} m^2 - n_{\uparrow\downarrow}(z, z_0).
\end{aligned}$$

The conclusion will then follow if we can show $2n_{\downarrow\uparrow}(z, z_0) \leq \frac{(K-1)K}{2} m^2$. We denote $a_{kt} = |\{(i,j) : z_i = t, z_i^0 = k\}|$, and we fix $a_{kk} = 0$ for all $k = 1, ...K$. Then $\sum_{t=1}^{K} a_{kt} = a_k$ and there are $K - 1$ non-zero terms.

$$\begin{aligned}
n_{\downarrow\uparrow}(z, z_0) &= \sum_{k=1}^{K} \{n_{\downarrow\uparrow}^{(k)}(z, z_0)\} \\
&= \sum_{k=1}^{K} \left\{ (n_k^0 - a_k) a_k + \binom{a_k}{2} - \sum_{t=1}^{K} \binom{a_{kt}}{2} \right\} \\
&= mr + \sum_{k=1}^{K} (-\frac{a_k^2}{2} - \sum_{t=1}^{K} \frac{a_{kt}^2}{2}) \\
&\leq mr + \sum_{k=1}^{K} \left\{ -\frac{a_k^2}{2} - \frac{a_k^2}{2(K-1)} \right\} = mr - \frac{K}{2(K-1)} \sum_{k=1}^{K} a_k^2 \qquad \text{(A.38)} \\
&\leq mr - \frac{r^2}{2(K-1)}. \qquad \text{(A.39)}
\end{aligned}$$

(A.38) follows since $\sum_{t=1}^{K} a_{kt}^2/2$ is minimized at $a_{kt} = a_k/(K-1)$ for $t = 1, ..., K$ and $t \neq k$. (A.39) follows since $\sum_{k=1}^{K} a_k^2$ is minimized at $a_k = r/K$ for $k = 1, ..., K$. Observe that $r \mapsto mr - r^2/2(K-1)$ is maximized at $r = (K-1)m$. Then the upper bound in (A.39) becomes

$$m(K-1)m - \frac{(K-1)m^2}{2} = \frac{(K-1)m^2}{2}.$$

It is easy to see that $2n_{\downarrow\uparrow}(z,z_0) \le (K-1)m^2 \le \frac{(K-1)K}{2}m^2$ when $K \ge 2$.

We finally prove (A.34). We split the proof into two cases.

Case 1: When $r/m \to 0$ as $m \to \infty$, we want to show that $n_{\uparrow\downarrow}(z,z_0)+n_{\downarrow\uparrow}(z,z_0) \ge Crm$. Observe that

$$\begin{aligned} n_{\uparrow\downarrow}(z,z_0)+n_{\downarrow\uparrow}(z,z_0) &= \sum_{k=1}^{K}\left\{n_{\uparrow\downarrow}^{(k)}(z,z_0)+n_{\downarrow\uparrow}^{(k)}(z,z_0)\right\} \ge \sum_{k=1}^{K}(n_k-b_k)b_k+\sum_{k=1}^{K}(n_k^0-a_k)a_k \\ &= \sum_{k=1}^{K}(n_k^0-a_k)(a_k+b_k) = m\sum_{k=1}^{K}(a_k+b_k)-\sum_{k=1}^{K}(a_k^2+a_kb_k), \end{aligned}$$

which implies

$$\begin{aligned} n_{\uparrow\downarrow}(z,z_0)+n_{\downarrow\uparrow}(z,z_0) &\ge 2mr-\sum_{k=1}^{K}a_k(a_k+b_k) \ge 2mr-\left\{\sum_{k=1}^{K}a_k\right\}\left\{\sum_{k=1}^{K}(a_k+b_k)\right\} \\ &= 2mr-2r^2 = 2rm(1-r/m) \ge Crm. \end{aligned} \tag{A.40}$$

(A.40) follows from the fact that $\sum_{k=1}^{K}a_k(a_k+b_k) < \{\sum_{k=1}^{K}a_k\}\{\sum_{k=1}^{K}(a_k+b_k)\}$.

Case 2: When $r = am$, where a is a constant that satisfies $0 < a \le K-1$,

$$n_{\uparrow\downarrow}(z,z_0)+n_{\downarrow\uparrow}(z,z_0) \ge Crm \tag{A.41}$$

for some $C > 0$. Observe that

$$\begin{aligned} n_{\uparrow\downarrow}(z,z_0)+n_{\downarrow\uparrow}(z,z_0) &= (n_\uparrow(z)-n_{\uparrow\uparrow}(z,z_0))+(n_\uparrow(z_0)-n_{\uparrow\uparrow}(z,z_0)) \\ &= \sum_{k=1}^{K}\binom{n_k}{2}+\sum_{k=1}^{K}\binom{m}{2}-2\sum_{\alpha=1}^{K}\sum_{\beta=1}^{K}\binom{n_{\alpha\beta}}{2} \\ &= \sum_{k=1}^{K}(\frac{n_k^2+n_k^{0\,2}}{2})-\sum_{\alpha=1}^{K}\sum_{\beta=1}^{K}n_{\alpha\beta}^2 \\ &= \sum_{\alpha=1}^{K}\frac{[(\sum_{\beta=1}^{K}n_{\alpha\beta})^2+(\sum_{\beta=1}^{K}n_{\beta\alpha})^2]}{2}-\sum_{\alpha=1}^{K}\sum_{\beta=1}^{K}n_{\alpha\beta}^2 \\ &= \sum_{k=1}^{K}\sum_{a>b}n_{k\alpha}n_{kb}+\sum_{k=1}^{K}\sum_{\alpha>\beta}n_{\alpha k}n_{\beta k}. \end{aligned} \tag{A.42}$$

In the preceding display, $n_{\uparrow\downarrow}(z,z_0)+n_{\downarrow\uparrow}(z,z_0)$ are the sum of squares of all column sums and row sums minus the sum of squares of each term in matrix $N = \{n_{\alpha\beta} : \alpha = 1,\dots,K, \beta = 1,\dots,K\}$. This quantity is essentially the sum of interaction terms within each column and row. The matrix N satisfies the following requirements:

- For diagonal terms of N, we have $\sum_{k=1}^{K}n_{kk} \ge m$.

- For all k in $1, \ldots, K$, $\sum_{\alpha=1}^{K} n_{\alpha k} = m$.

For each column, if there is no term in that column which satisfies $n_{k\alpha} \geq Cm$, from the second requirement above, we can see that there must be at least one term $n_{k\alpha}$ which satisfied $n_{k\alpha} \geq Cm/K$. Then it is straightforward to see for each column k, $\sum_{\alpha>\beta} n_{\alpha k} n_{\beta k} \geq \frac{Cm}{K}(m - \frac{Cm}{K}) \geq Cm^2/K$. When $r = am$, it is easy to show $n_{\uparrow\downarrow}(z, z_0) + n_{\downarrow\uparrow}(z, z_0) \geq \frac{Cm^2}{K} K = Crm$. If there is at least one column or row in which there are more than one term that is Cm (say n_{k1} and n_{k2} are Cm), then from (A.42) and $r = am$, it follows that $n_{\uparrow\downarrow}(z, z_0) + n_{\downarrow\uparrow}(z, z_0) \geq Cm^2 = Crm$. If there is only one term that is Cm in all columns and rows and all other terms are $o(m)$, one can switch labels to make r satisfy $r/m \to 0$ by putting all the Cm terms into diagonal terms of the matrix N. This phenomenon is exemplified in Appendix G.2 for $K = 4$.

Proof of upper bound on $\mathbf{n}(z, z_0)$: From (A.39), $n_{\downarrow\uparrow}(z, z_0) \leq Crm$. In the following, we show that $n_{\uparrow\downarrow}(z, z_0) \leq C\{rm + r^2\}$. We proceed similar to (A.39). Observe that

$$\begin{aligned} n_{\uparrow\downarrow}(z, z_0) &= \sum_{k=1}^{K} \left\{ (n_k^0 - a_k) b_k + \binom{b_k}{2} - \sum_{t=1}^{K} \binom{b_{kt}}{2} \right\} \\ &= mr + \sum_{k=1}^{K} \left\{ -a_k b_k + b_k^2/2 - \sum_{t=1}^{K} b_{kt}^2/2 \right\} \\ &\leq mr + Cr^2 \end{aligned} \tag{A.43}$$

for some constant $C > 0$. Since $n_{\uparrow\uparrow}(z, z_0) \leq n_{\uparrow}(z_0)$ and $n_{\downarrow\downarrow}(z, z_0) \leq n_{\downarrow}(z_0)$, the upper bound for $\mathbf{n}(z, z_0)$ in Lemma E.1 follows.

G.2 Example in the proof of Lemma E.1

Let $N = (n_{\alpha\beta})_{1 \leq \alpha, \beta \leq 4}$ and $n_{11} = Cm$ without loss of generality. A particular instance of occurrence of only Cm term in each of the columns and rows is the following:

$$\begin{bmatrix} Cm & n_{12} & n_{13} & n_{14} \\ n_{21} & n_{22} & Cm & n_{24} \\ n_{31} & n_{32} & n_{33} & Cm \\ n_{41} & Cm & n_{43} & n_{44} \end{bmatrix}$$

in which n_{11}, n_{42}, n_{23} & n_{34} are Cm and all other terms are $O(m)$. Then if we switch the labels as $4 \to 2$, $2 \to 3$ and $3 \to 4$ for z, the matrix N becomes

$$\begin{bmatrix} Cm & n_{12} & n_{13} & n_{14} \\ n_{21} & Cm & n_{23} & n_{24} \\ n_{31} & n_{32} & Cm & n_{34} \\ n_{41} & n_{42} & n_{43} & Cm \end{bmatrix}.$$

Then we have $n_{\uparrow\downarrow}(z,z_0)+n_{\downarrow\uparrow}(z,z_0) \geq \sum_{k=1}^{K} n_{kk}(n_k - n_{kk}) \geq Cm\sum_{k=1}^{K}(n_k - n_{kk}) = Crm$.

G.3 **Proof of Lemma F.1**

Expressing the denominator for (A.31) in terms of B, R and m:

$$\frac{(2R-1)(3-2R)m^4 - (6R-4R^2-1)m^3 + (4-4R)Bm^2 + B(2R-1)m - B^2 + o(m^3)}{2m^2 - m}. \quad \text{(A.44)}$$

(A.44) shows that the denominator is smaller than Cm^2. Since we are interested in finding a lower bound to (A.31), we henceforth assume the denominator to be Cm^2. The numerator for (A.31) is expressed as:

$$(1-R)Bm^2 + (R^2-R)m^3 + (2R-1)(1-R)m^4 - \frac{B^2}{4}.$$

The order of the numerator is decided by the order of B, $1-R$ and $1-2R$. It is straightforward to show $B \leq Cm^2$. Observe that

$$R \;=\; \frac{n_{\uparrow\uparrow}(z,z_0)+n_{\downarrow\downarrow}(z,z_0)}{\binom{n}{2}} = \frac{m^2 - m + \sum_{\alpha=1}^{k}(n_{\alpha 1}-n_{\alpha 2})^2/2}{2m^2-m}.$$

The minimum value for R is achieved when $n_{\alpha 1} = n_{\alpha 2}$ for all α. Then $R_{\text{min}} \asymp 0.5 - 1/4m$. The maximum value of R is achieved when $\sum_{\alpha=1}^{k}(n_{\alpha 1}-n_{\alpha 2})^2$ is the largest. The constraint here is at least one of $n_{\alpha 1}$ and $n_{\alpha 2}$ will be non-zero for all α. Also $\sum_{\alpha=1}^{k} n_{\alpha 1} = \sum_{\alpha=1}^{k} n_{\alpha 2} = m$. Under these constraints, the maximum value will be achieved at $n_{11} = m$, $n_{21} = ... = n_{k1} = 0$, $n_{12} = 0$, $n_{22} = m-(k-2)$ and there are $k-2$ 1's in $n_{\alpha 2}$ for $\alpha > 2$. Then we have

$$R_{\text{max}} \;=\; \frac{m^2 - m + \{(m-k+2)^2 + m^2 + (k-2)\}/2}{2m^2-m} \asymp 1 - \frac{k-2}{2m} + \frac{k^2-3k}{4m^2}.$$

Hence $1-R \geq Ck/m$. Define a sequence $\eta_n \to 0$ and $m\eta_n \to \infty$ as $m \to \infty$. Define another sequence β_n, which satisfies $\beta_n \to 0$ and $m\beta_n \to 0$ as $m \to \infty$. We split into five different cases.
Case 1: If R is close to 0.5 and $1-2R \asymp \beta_n$ or $1-2R \asymp Cm^{-1}$, then we show the lower bound of (A.31) is Cm^2/k. We provide the justification below.

Note that $2B = \sum_{\alpha=1}^{k} n_\alpha^2 - (4R-2)m^2 + (2-2R)m \geq Cm^2/k$. Then observe that the first term of (A.31) can be lower-bounded as

$$\frac{n_{\uparrow\uparrow}(z,z_0)n_{\uparrow\downarrow}(z,z_0)}{n_\uparrow(z)} = \frac{\left\{\frac{\sum_{\alpha=1}^{k}(n_{\alpha 1}^2+n_{\alpha 2}^2)}{2} - m\right\}B/2}{\frac{\sum_{\alpha=1}^{k}(n_{\alpha 1}^2+n_{\alpha 2}^2)}{2} - m + B/2} \geq C\frac{m^2}{k}.$$

Case 2: If R is between 0.5 and 1 and both $1-R$ and $1-2R$ are constants, we provide the justification below.

If $B/m^2 \to 0$ as $m \to \infty$, the numerator for (A.31) is greater than Cm^4. Thus we have the lower bound for (A.31) as Cm^2. If $B/m^2 \to C$ as $m \to \infty$, we have the lower bound of (A.31) to be Cm^2/k from the same justification as in Case 1.

Case 3: If R is close to 1 and $1 - R \asymp \eta_n$, we provide the justification below.

If $\frac{B}{m^2\sqrt{\eta_n}} \to 0$ as $m \to \infty$, the numerator for (A.31) is greater than $C\eta_n m^4$. Thus we have the lower bound for(A.31) as $C\eta_n m^2$. If $\frac{B}{m^2\sqrt{\eta_n}} \to \infty$ as $m \to \infty$, we can have the lower bound of (A.31) to be $C\frac{m^2}{k}$ or $Cm^2\sqrt{\eta_n}$ whichever is smaller, from the same justification in Case 1.

Case 4: If R is close to 1 and $1 - R \asymp Cm^{-1}$, then we show the lower bound of numerator is Cm^{-1}. We provide the justification below.

If $B/m \to \infty$ and $B/(m^2\eta_n) \to C$ as $m \to \infty$, we can have the lower bound of (A.31) to be Cm^2/k or $Cm^2\eta_n$ whichever is smaller from the same justification in case 1. If $B/m \to C$ as $m \to \infty$, we have the lower bound of (A.31) as:

$$\frac{(2R-1)(1-R)m^4}{(2R-1)(3-2R)m^4/\binom{n}{2}} \asymp (1 - \frac{1}{3-2R})m^2 \geq (1 - \frac{1}{1+\frac{k-2}{m}})m^2 \asymp (k-2)m.$$

Case 5: If $1 - R \asymp Cm^{-1}$ when the order of $B/m \to C$ as $m \to \infty$, the lower bound for (A.31) is km. However, the bound for the prior ratio in (A.27) is different. If one of n_i is $n - k + 1$, then $B/m \to \infty$ as $m \to \infty$. If we take a look at the definition of $B = 2\sum_{\alpha=1}^{k} n_{\alpha 1}n_{\alpha 2}$, $n_{\alpha 1}n_{\alpha 2}/m \to C$ or $n_{\alpha 1}n_{\alpha 2}/m \to 0$ for all $\alpha = 1, \ldots, k$. Under the constraint that both $n_{\alpha 1}$ and $n_{\alpha 2}$ are less than m, in order to maximize $n_\alpha = n_{\alpha 1} + n_{\alpha 2}$, one out of $n_{\alpha 1}$ and $n_{\alpha 2}$ has to be $c_1 m - c_2$ and the other one has to be a constant. Then in order to find an upper bound for the prior ratio in the right-most expression of (A.27), there are two n_i's, which are of the form of $n_i = m - c_i$, where c_i is at most of the order of k. Then

$$\begin{aligned} \frac{\Pi(z \mid K=k)}{\Pi(z_0 \mid K=2)} &= \frac{\frac{(k-1)!\prod_{i=1}^{k} n_i!}{(n+k-1)!}}{\frac{m!m!}{(n+1)!}} \asymp \frac{(k-1)!(m-c_1)!(m-c_2)!(n+1)!}{(n+k-1)!m!m!} \\ &\asymp C(k-1)^{k-1/2}2^k\sqrt{n}e^{-c_1 k}(n+k-1)^{c_2-c_3 k}. \end{aligned} \tag{A.45}$$

REFERENCES

[1] Stephen P Borgatti and Martin G Everett. Models of core/periphery structures. *Social networks*, 21(4):375–395, 2000.

[2] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[3] M.E.J. Newman. Modularity and community structure in networks. *proceedings of the national academy of sciences*, 103(23):85778582, 2006.

[4] MEJ Newman and Gesine Reinert. Estimating the number of communities in a network. *arXiv preprint arXiv:1605.02753*, 2016.

[5] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *Arxiv preprint arxiv:1011.3027*, 2010.