Supplementary material for "Statistical Guarantees for Consensus Clustering"

This supplement contains the detailed proofs of the results and some extra simulations.

A INCONSISTENCY OF BESTOFK

Using the notation of the present paper, the name "BestOfK" should be "BestOfN". We will use our notation in the following proposition and keep the name "BestOfK".

Proposition 3. *BestOfK is not consistent unless N grows exponentially fast in n.*

Proof. We will prove this proposition by providing a counterexample. Suppose $K = 2$, $1 - \tilde{p} = 0.6$ and $q = 0.4$. Then for a label vector z from the RPM, by the Hoeffding inequality,

$$\mathbb{P}(\text{Mis}(z, z^*) \geq 0.1) \geq 1 - \exp(2(0.4 - 0.1)^2 n) - \exp(2(0.6 - 0.1)^2 n) \geq 1 - 2\exp(-0.18n)$$

where we have accounted for the two permutations in the definition of Mis. Suppose we observe N i.i.d. label vectors $z_1, \ldots, z_N$ from the RPM. Then

$$\mathbb{P}(\min_{i \in [N]} \text{Mis}(z_i, z^*) \geq 0.1) \geq (1 - 2\exp(-0.18n))^N \geq 1 - 2N\exp(-0.18n).$$

This probability (of missing the target) approaches 1 unless N grows exponentially fast in n. □

B RELATIONS AMONG CLUSTERING DISTANCES

Let n_k be the size of the kth cluster of $Z \in \mathcal{E}_K^n$ and, n_ℓ^* the size of the ℓth cluster of $Z^* \in \mathcal{E}_L^n$, and let X and X^* be the corresponding association matrices. The Mirkin distance (34, Eqn (6)) is given by

$$d'_M(Z, Z^*) = \sum_k n_k^2 + \sum_\ell (n_\ell^*)^2 - 2\sum_{k,\ell} n_{k\ell}^2 \tag{13}$$

where $n_{k\ell}$ is the number of objects that are in cluster k according to Z and cluster ℓ according to Z^*. It is not hard to see that $\sum_k n_k^2 = \|X\|_F^2$ and similarly $\sum_\ell (n_\ell^*)^2 = \|X^*\|_F^2$. We also have $Z(Z^*)^T = (n_{k\ell})$, hence, using $\|A\|_F^2 = \text{tr}(AA^T)$,

$$\sum_{k,\ell} n_{k\ell}^2 = \|Z(Z^*)^T\|_F^2 = \text{tr}(Z(Z^*)^T Z^* Z^T) = \text{tr}((Z^*)^T Z^* Z^T Z) = \text{tr}(X^* X).$$

Combining these facts, we obtain the first equality below

$$d'_M(Z, Z^*) = \|X - X^*\|_F^2 = \|X - X^*\|_{\ell_1}. \tag{14}$$

The second equality follows from $X - X^*$ having elements in $\{-1, 0, 1\}$. Here $\|\cdot\|_{\ell_1}$ denotes the ℓ_1 norm of a matrix viewed as a vector. The equality $d'_M(Z, Z^*) = \|X - X^*\|_{\ell_1}$ immediately shows that d'_M is indeed a distance on the space of clusterings. It also connects the Mirkin distance with the Rand index.

To see the connection with the Rand index, let N_{disagree} be the number of pairs of objects for which Z and Z' disagree about their co-clustering, that is, whether the two objects are in the same cluster or not. Similarly, let N_{agree} be the number of pairs of objects for which Z and Z' agree about their co-clustering. We have $N_{\text{disagree}} + N_{\text{agree}} = \binom{n}{2}$. The Rand index is defined as the proportion of the agreements, that is,

$$\text{Rand} = \frac{N_{\text{agree}}}{\binom{n}{2}}.$$

It is easy to see that $\|X - X^*\|_{\ell_1} = 2N_{\text{disagree}}$ where the factor of 2 is due to the double-counting caused by the symmetry of $X - X^*$. This proves the relation

$$\frac{1}{2} d'_M = \binom{n}{2}(1 - \text{Rand}). \tag{15}$$

The symmetric difference distance (SDD) is another name for N_{disagree}, hence $d'_M/2 = \text{SDD}$. The Binder loss is defined as half the expression in (13), that is, $d'_M/2 = \text{Binder}$.

B.1 Consistency in Mis implies consistency in Mirkin distance

Let us now show that the consistency in Mis implies consistency in the *normalized* Mirkin distance defined as $d_M := d'_M/n^2$. See (34, Eqn (9)). It then follows that consistency in Mis implies consistency in the normalized SDD, normalized Binder loss and the Rand index, as discussed above. This claim follows from the following inequality:

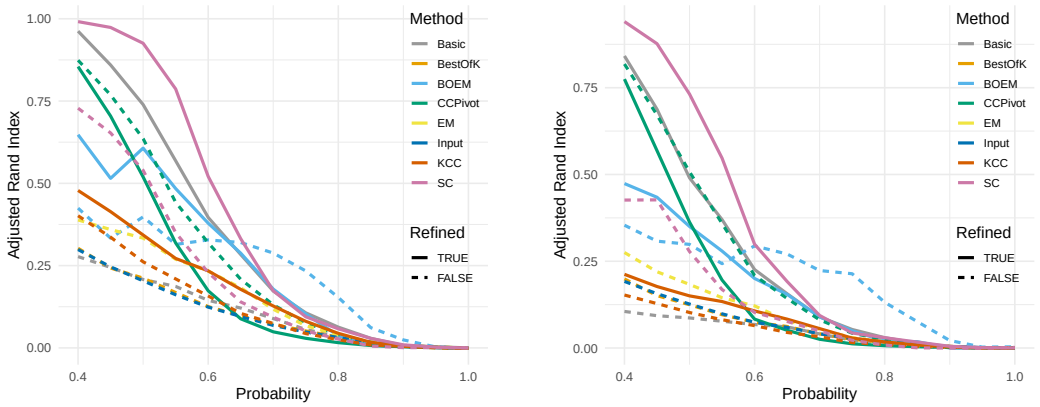
Proposition 4. *We have* $d_M \leq 2 \cdot \text{Mis}$.

Proof. Let X and X^* be the association matrices corresponding to label vectors z and z^*. Then $d'_M(z, z^*) = \|X - X^*\|_{\ell^1}$ as shown in (14). The entries of $X - X^*$ take values in $\{-1, 0, 1\}$. Assume, WLOG, that the optimal permutation between z and z^* is the identity. Then:

1. If the label $z_i = z_i^*$, then the ith row of $X - X^*$ has at most "$n \cdot \text{Mis}$" nonzero entries. There are at most n such rows.
2. If the label $z_i \neq z_i^*$, then the ith row of $X - X^*$ has at most n nonzero entries. There are at most "$n \cdot \text{Mis}$" such rows.

Therefore, $d'_M = \|X - X^*\|_{\ell^1} \leq n \cdot (n \cdot \text{Mis}) + (n \cdot \text{Mis}) \cdot n = 2n^2 \cdot \text{Mis}$ and the result follows. □

C Extra Simulation Results



(a) $n = 100, N = 20, K = 6, p_1 = 0.8$ (b) $n = 100, N = 20, K = 6, p_1 = 0.9$

Figure 4: Significant improvements due to local refinement in the case of unbalanced cluster sizes.

Figure 4 shows some extra cases of unbalanced cluster sizes (various values of p_1 as defined earlier), showing the significant improvement of the refinement step in such cases. All the results for the unbalanced case (including those in the main text) are averaged over 120 runs.

Tables 1, 2 and 3 show the average ARI in all the eight settings (abbreviated Set in the tables) shown in Figures 2, 3 and 4. The tables show the performance of the methods at noise probabilities $p = 0.45, 0.55$ and 0.65 respectively—corresponding to a cross-section of each plot at a line parallel to the y-axis, crossing the x-axis at the respective value of p. The settings are as follows:

1. Set 1: Balanced, $n = 100$, $N = 20$.
2. Set 2: Balanced, $n = 100$, $N = 200$.
3. Set 3: Balanced, $n = 500$, $N = 20$.
4. Set 4: Balanced, $n = 500$, $N = 200$.
5. Set 5: Unbalanced, $n = 100$, $N = 20$, $p_1 = 0.5$.
6. Set 5: Unbalanced, $n = 100$, $N = 20$, $p_1 = 0.75$.

Method	Refined	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
Basic	FALSE	1.00	1.00	1.00	1.00	0.82	0.35	0.24	0.093
Basic	TRUE	1.00	1.00	1.00	1.00	0.98	0.91	0.86	0.690
BestOfK	FALSE	0.32	0.30	0.29	0.31	0.33	0.28	0.24	0.150
BOEM	FALSE	0.50	0.48	0.57	0.69	0.38	0.41	0.33	0.310
BOEM	TRUE	0.65	0.59	0.94	1.00	0.60	0.70	0.52	0.430
CCPivot	FALSE	0.77	1.00	0.77	1.00	0.81	0.78	0.77	0.670
CCPivot	TRUE	0.84	1.00	0.81	1.00	0.81	0.75	0.70	0.570
EM	FALSE	0.97	0.98	0.97	0.99	0.72	0.41	0.36	0.220
Input	FALSE	0.30	0.30	0.30	0.30	0.33	0.28	0.25	0.160
KCC	FALSE	1.00	1.00	1.00	1.00	0.91	0.44	0.34	0.130
KCC	TRUE	1.00	1.00	1.00	1.00	0.93	0.50	0.41	0.180
SC	FALSE	0.99	1.00	1.00	1.00	0.96	0.65	0.65	0.430
SC	TRUE	1.00	1.00	1.00	1.00	0.99	0.98	0.97	0.880

Table 1: Mean adjusted rand index (ARI) for all settings at noise probability $p = 0.45$.

Method	Refined	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
Basic	FALSE	0.96	1.00	0.98	1.00	0.66	0.26	0.19	0.077
Basic	TRUE	0.97	1.00	0.98	1.00	0.89	0.64	0.57	0.370
BestOfK	FALSE	0.20	0.21	0.20	0.20	0.23	0.19	0.17	0.096
BOEM	FALSE	0.21	0.21	0.23	0.29	0.33	0.39	0.31	0.240
BOEM	TRUE	0.46	0.36	0.56	0.65	0.55	0.62	0.48	0.280
CCPivot	FALSE	0.51	0.99	0.51	0.99	0.57	0.52	0.44	0.360
CCPivot	TRUE	0.61	0.98	0.54	0.94	0.52	0.36	0.32	0.200
EM	FALSE	0.87	0.95	0.91	0.97	0.64	0.33	0.27	0.140
Input	FALSE	0.20	0.20	0.20	0.20	0.23	0.18	0.16	0.098
KCC	FALSE	0.94	1.00	0.98	1.00	0.67	0.28	0.21	0.081
KCC	TRUE	0.97	1.00	0.98	1.00	0.74	0.36	0.27	0.130
SC	FALSE	0.95	1.00	0.98	1.00	0.75	0.41	0.35	0.170
SC	TRUE	0.97	1.00	0.98	1.00	0.95	0.86	0.79	0.550

Table 2: Mean adjusted rand index (ARI) for all settings at noise probability $p = 0.55$.

7. Set 6: Unbalanced, $n = 100$, $N = 20$, $p_1 = 0.8$.
8. Set 7: Unbalanced, $n = 100$, $N = 20$, $p_1 = 0.9$.

D PROOFS

D.1 PROOF OF PROPOSITION 1

For any $Z \in \mathcal{E}_K^n$, we have $\|Z\|_F^2 = \sum_{k,i} Z_{ki}^2 = \sum_{k,i} Z_{ki} = n$. Thus, $\|Z\|_F^2 = \|\widehat{P}_j Z_j\|_F^2 = n$ for all $j \in [N]$. Hence, solving (5) is equivalent to maximizing $f(Z) := \sum_{j=1}^N w_j \operatorname{tr}(Z^T \widehat{P}_j Z_j) = \operatorname{tr}(Z^T \bar{Z})$ over $\mathcal{E}_K^n$, where $\bar{Z} := \sum_j w_j \widehat{P}_j Z_j$. Let $Z = (z_1, \ldots, z_n)$ and $\bar{Z} = (\bar{z}_1, \ldots, \bar{z}_n)$. Maximizing $f(Z) = \sum_{i=1}^n \langle z_i, \bar{z}_i \rangle$ is a separable problem over i, and maximizing $z \mapsto \langle z, \bar{z}_i \rangle$ over $\mathcal{E}_K$ amounts to finding the index of the maximum element of $\bar{z}_i$, that is, the "argmax" of $\bar{z}_i$, as claimed.

D.2 PROOF OF THEOREM 1

We have $Z_j = P_j Z_j'$ where $Z_j' = (z_{j1}', \ldots, z_{jn}')$ and z_{ji}' are i.i.d. draws as in (8). Since the algorithm is invariant to permutations P_j, without loss of generality we assume $P_j = I_n$, hence $Z_j = Z_j'$. We write $X^* = (Z^*)^T Z^*$ for the true association matrix. Let E_n be the all-ones $n \times n$ matrix.

Method	Refined	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8
Basic	FALSE	0.790	1.000	0.88	1.00	0.47	0.17	0.120	0.061
Basic	TRUE	0.810	1.000	0.89	1.00	0.60	0.33	0.280	0.160
BestOfK	FALSE	0.120	0.130	0.12	0.12	0.13	0.11	0.098	0.062
BOEM	FALSE	0.085	0.098	0.12	0.11	0.24	0.29	0.320	0.270
BOEM	TRUE	0.130	0.150	0.21	0.22	0.36	0.33	0.290	0.150
CCPivot	FALSE	0.250	0.690	0.22	0.67	0.28	0.24	0.210	0.140
CCPivot	TRUE	0.280	0.540	0.21	0.37	0.22	0.12	0.088	0.051
EM	FALSE	0.600	0.920	0.74	0.95	0.46	0.23	0.180	0.078
Input	FALSE	0.120	0.120	0.12	0.12	0.14	0.11	0.096	0.059
KCC	FALSE	0.590	1.000	0.89	1.00	0.39	0.16	0.100	0.045
KCC	TRUE	0.740	1.000	0.89	1.00	0.52	0.24	0.180	0.083
SC	FALSE	0.750	1.000	0.88	1.00	0.47	0.19	0.140	0.076
SC	TRUE	0.810	1.000	0.89	1.00	0.65	0.40	0.330	0.190

Table 3: Mean adjusted rand index (ARI) for all settings at noise probability $p = 0.65$.

Lemma 1. *Let $Z \sim \mathcal{L}(Z^*, p)$ and let $X = Z^T Z$ be the corresponding association matrix. Then,*

$$M := \mathbb{E}[X] = (1-\xi)X^* + \xi\Big(\frac{1}{K}E_n + (1-\frac{1}{K})I_n\Big) \tag{16}$$

where $\xi = p(2-p)$.

Proof of Lemma 1. We have $X_{ij} = (Z^T Z)_{ij} = \langle z_i, z_j\rangle$ and $\mathbb{E}[z_i] = (1-p)z_i^* + p\frac{1}{K}1_K$. For $i \neq j$, z_i and z_j are independent, hence

$$\begin{aligned}\mathbb{E}X_{ij} = \langle \mathbb{E}z_i, \mathbb{E}z_j\rangle &= \langle (1-p)z_i^* + p\frac{1}{K}1_K, (1-p)z_j^* + p\frac{1}{K}1_K\rangle \\ &= (1-p)^2\langle z_i^*, z_j^*\rangle + 2p(1-p)\frac{1}{K} + p^2\frac{1}{K}\end{aligned}$$

For $i = j$, we have $\mathbb{E}[X_{ii}] = 1$. The above shows that

$$\mathbb{E}[X] = (1-p)^2X^* + p(2-p)\frac{1}{K}E_n + p(2-p)\Big(1-\frac{1}{K}\Big)I_n$$

which simplifies to the desired expression. □

Let $Z_1, \ldots, Z_N, Z \sim \mathcal{L}(Z^*, p)$ be independent draws, and let $X_j = Z_j^T Z_j$ and $X = Z^T Z$ be the associated association matrices. Setting $\bar{X} = \frac{1}{N}\sum_{t=1}^N X_t$, we obtain

$$\mathbb{E}\|\bar{X} - M\|_F^2 = \sum_{ij}\mathbb{E}(\bar{X}_{ij} - M_{ij})^2 = \sum_{ij}\mathrm{var}(\bar{X}_{ij}) = \frac{1}{N}\sum_{ij}\mathrm{var}(X_{ij}).$$

We have $\mathrm{var}(X_{ij}) = 0$ for $i = j$. For $i \neq j$, one has $X_{ij} \sim \mathrm{Ber}((1-\xi)X_{ij}^* + \xi/K)$, hence

$$\begin{aligned}\mathrm{var}(X_{ij}) &= (1-\xi)X_{ij}^* + \frac{\xi}{K} - \Big((1-\xi)^2X_{ij}^* + 2\frac{\xi}{K}(1-\xi)X_{ij}^* + \frac{\xi^2}{K^2}\Big) \\ &= \psi(\xi)\Big(1-\frac{2}{K}\Big)X_{ij}^* + \psi(\xi/K)\end{aligned}$$

where $\psi(x) = x(1-x)$. Note that $\xi = p(2-p) \in (0,1)$. It follows that

$$N\cdot\mathbb{E}\|\bar{X} - M\|_F^2 \leq \psi(\xi)\Big(1-\frac{2}{K}\Big)\sum_{ij}X_{ij}^* + n^2\psi(\xi/K)$$

where the inequality is due to bounding $\mathrm{var}(X_{ii})$ by the same formula used for $\mathrm{var}(X_{ij}), i \neq j$. Let n_k^* be the number of entities in cluster k of Z^*, that is, $n_k^* = (Z^*1_n)_k$. We have $\sum_{ij}X_{ij}^* = \|Z^*1_n\|^2 = \sum_k (n_k^*)^2$. Using the assumption $n_k^* \leq \beta n/K$, we have

$$N\cdot\mathbb{E}\|\bar{X} - M\|_F^2 \leq \psi(\xi)\Big(1-\frac{2}{K}\Big)\frac{\beta^2n^2}{K} + n^2\psi(\xi/K).$$

Calculating the center separations. Let $\widetilde{M} = (1-\xi)X^* + (\xi/K)E_n$. We note that $M - \widetilde{M}$ is diagonal and

$$\|M - \widetilde{M}\|_F^2 = \|\xi(1-1/K)I_n\|_F^2 = \xi^2(1-1/K)^2 n \le \xi^2 n.$$

It follows that

$$\begin{aligned}\mathbb{E}\|\bar{X} - \widetilde{M}\|_F^2 &= \mathbb{E}\sum_{i\neq j}(\bar{X}_{ij} - \widetilde{M}_{ij})^2 + \mathbb{E}\sum_i (\bar{X}_{ii} - \widetilde{M}_{ii})^2 \\ &\le \mathbb{E}\|\bar{X} - M\|_F^2 + \xi^2 n.\end{aligned}$$

We obtain

$$\frac{1}{n^2}\mathbb{E}\|\bar{X} - \widetilde{M}\|_F^2 \le \frac{2}{N}\Big[\psi(\xi)\Big(1 - \frac{2}{K}\Big)\frac{\beta^2}{K} + \psi(\xi/K)\Big] + \frac{2\xi^2}{n}.$$

The matrix $\widetilde{M}$ is a K-means matrix with K distinct rows. If $z_i = r \neq k = z_{i'}$, then

$$\|\widetilde{M}_{i*} - M_{i'*}\|^2 = (1-\xi)^2\|X^*_{i*} - X^*_{i'*}\|^2 = (1-\xi)^2(n_r^* + n_k^*) \ge 2(1-\xi)^2\frac{n}{\beta K}$$

using $n_k^* \ge n/(\beta K)$, which holds by assumption (9). We have $n_r\delta_r^2 \ge 2(1-\xi)^2(\frac{n}{\beta K})^2$ which gives the following bound, using (49, Proposition 1),

$$\mathbb{E}[\mathrm{Mis}_r] \lesssim \frac{1}{N(1-\xi)^2}\Big[\psi(\xi)(K-2)\beta^4 + \beta^2K^2\psi(\xi/K)\Big] + \frac{\xi^2}{(1-\xi)^2}\frac{\beta^2K^2}{n}.$$

Here, Mis_r is the misclassification rate over true cluster r. The dependence on β of the first term is $O(\beta^2)$ when $K = 2$ and $O(\beta^4)$ when $K > 2$. Ignoring this difference, we can simplify the bound, by noting that $K^2\psi(\xi/K) = K\xi(1-\xi/K) \le K\xi$ and $\beta^2 \le \beta^4$. Then,

$$\mathbb{E}[\mathrm{Mis}_r] \lesssim \frac{\xi}{(1-\xi)^2}\frac{2K\beta^4}{N} + \frac{\xi^2}{(1-\xi)^2}\frac{\beta^2K^2}{n},$$

from which the bound in the theorem follows since $\mathrm{Mis} = \sum_r (n_r^*/n)\,\mathrm{Mis}_r$.

D.3 Proof sketch for Theorem 2

For the benefit of the readers, we first give a proof sketch for Theorem 2 and its key lemma. A detailed proof is given in Appendix D.4. The proof of Theorem 2 relies on the following key lemma:

Lemma 2. *Let $B(\delta)$ denote the set of label matrices Z with at most $n\delta$ labels different from Z^*, and let $\widehat{Z}(Z)$ be the output of Algorithm 3 with initial label matrix Z. If $n_{min}p(1\wedge I)/K \to \infty$ and $\frac{\log K}{NI} \to 0$, then*

$$\mathbb{P}\big(\exists Z \in B(\delta) \text{ such that } \widehat{z}_i(Z) \neq z_i^*\big) \le e^{-(1-\eta')NI + \frac{3Kn\delta N}{2pn_{min}}} \tag{17}$$

for some $\eta' = o(1)$.

The first step is to prove the case $\delta = 0$ in Lemma 2, corresponding to the initial label matrix in Algorithm 3 being Z^*. If $z_i^* = e_1$, then the algorithm fails to recover z_i^* if there exists $k \neq 1$ such that $Y_k := n_1b_k - n_kb_1 \ge 0$. Y_k is the average of i.i.d. samples, where each sample follows a mixture model depending on which events among $z_i = e_1$, $z_i = e_k$, or $z_i \notin \{e_1, e_k\}$ happens. We compute the MGF of Y_k and obtain the bound

$$\mathbb{E}\big[\exp(tNY_k/(n_1n_2(1-p)))\big] \le \big[(1-\tilde{p})\,e^{-t(1+o(1))} + q\,e^{t(1+o(1))} + (K-2)q\,e^{\frac{2qt^2}{n_{\min}(1-p)^2}}\big]^N.$$

The choice of t has little affect on the last term since $n_{\min}$ is large, so we set $t = \frac{1}{2}\log[(1-\tilde{p})/q]$ to minimize $(1-\tilde{p})e^{-t} + qe^t$. Under the regularity conditions of the lemma and the definition of I in (11), we have

$$\mathbb{E}\big[\exp(tY_k/(n_1n_2(1-p)))\big] \;\le\; \big[2\sqrt{(1-\tilde{p})q} + (K-2)q\big]^{(1-o(1))N} = e^{-(1-o(1))NI}.$$

Applying the Chernoff inequality, it follows that

$$\mathbb{P}(\widehat{z}_i(Z^*) \neq z_i^*) \leq \sum_{k=2}^{K} \mathbb{P}\left(Y_k \geq 0\right) \leq \sum_{k=2}^{K} \mathbb{E}\big[\exp(tY_k/(n_1 n_2(1-p)))\big] \leq (K-1)e^{-(1-o(1))NI}.$$

Using the assumption $\frac{\log K}{NI} \to 0$, we obtain $\mathbb{P}(\widehat{z}_i(Z^*) \neq z_i^*) \leq e^{-(1-\eta)NI}$. This proves the case $\delta = 0$. Now we compare Y_k's obtained from Algorithm 3 initialized with label matrices Z^* and Z, and denoted by $Y_k(Z^*)$ and $Y_k(Z)$, respectively. For all $Z \in B(\delta)$, we show that $|Y_k(Z^*) - Y_k(Z)| \leq 3(n_1 \vee n_k)\, n\delta$ if $z_i^* = e_1$, giving

$$\mathbb{P}\big(\exists Z \in B(\delta) \text{ such that } \widehat{z}_i(Z) \neq z_i^*\big) \leq \sum_{k=2}^{K} \mathbb{P}\big(Y_k \geq -3(n_1 \vee n_k)n\delta\big).$$

We apply the Chernoff inequality with the same choice of t to obtain (17). We arrive at the proof of Theorem 2. Let $\widehat{Z}(Z)$ be the output of Algorithm 3 with initial label matrix Z. Consider the event $\mathcal{A}_\delta = \{\mathrm{Mis}(\widetilde{Z}, Z^*) \leq \delta\}$. For any $\varepsilon > 0$,

$$\mathbb{P}\big(\mathrm{Mis}(\widehat{Z}(\widetilde{Z}), Z^*) > \varepsilon\big) \leq \mathbb{P}\big(\mathcal{A}_\delta^c\big) + \mathbb{P}\big(\exists Z \in B(\delta'),\ \mathrm{Mis}(\widehat{Z}(Z), Z^*) > \varepsilon\big).$$

We have $\mathbb{P}\big(\mathcal{A}_\delta^c\big) = o(1)$ under assumption (b1). Letting $\varepsilon = NIe^{-(1-\eta')NI + \frac{3Kn\delta N}{2pn_{\min}}}$, one can verify that the second probability also converges to 0 under the conditions of the theorem and $\varepsilon = e^{-(1-o(1))NI}$. This proves (20) under assumption (b1). For the proof under assumption (b2), please see Appendix D.4

D.4 Detailed proof of Theorem 2

Let $\widehat{Z}(Z)$ be the output of Algorithm 3 with initial label matrix Z. Consider the event $\mathcal{A}_{\delta'} = \{\mathrm{Mis}(\widetilde{Z}, Z^*) \leq \delta'\}$. For any $\varepsilon > 0$,

$$\mathbb{P}\big(\mathrm{Mis}(\widehat{Z}(\widetilde{Z}), Z^*) > \varepsilon\big) \leq \mathbb{P}\big(\mathcal{A}_{\delta'}^c\big) + \mathbb{P}\big(\exists Z \in B(\delta'),\ \mathrm{Mis}(\widehat{Z}(Z), Z^*) > \varepsilon\big). \tag{18}$$

If assumption (b1) holds, then let $\delta' = \delta$ so that $\mathbb{P}(\mathcal{A}_\delta^c) = o(1)$. If assumption (b2) holds, Then we let $\delta' = \sqrt{n_{\min} pI\delta/(Kn)}$ so that

$$\frac{Kn\delta'}{n_{\min}pI} = \sqrt{\frac{Kn\delta}{n_{\min}pI}} = o(1)$$

and by Markov's inequality,

$$\mathbb{P}(\mathrm{Mis}(\widetilde{Z}, Z^*) > \delta') \leq \frac{1}{\delta'}\mathbb{E}[\mathrm{Mis}(\widetilde{Z}, Z^*)] \leq \frac{\delta}{\delta'} = \sqrt{\frac{Kn\delta}{n_{\min}pI}} = o(1).$$

Then, (b1) is satisfied with $\delta = \delta'$. Therefore, it is enough to only consider assumption (b1) and let $\delta' = \delta$ for the rest of the proof.

Let π^* be the permutation corresponding to $\mathrm{Mis}(\widetilde{Z}, Z^*)$ in assumption (b1), that is, $\pi^* = \mathrm{argmin}_\pi \sum_{i=1}^n 1\{\tilde{z}_i \neq \pi(z_i^*)\}$. Since we can always assume $\pi^*(z^*)$ to be the true label, without loss of generality, we can assume $\pi^* =$ identity. Writing T_2 for the second term in (18),

$$T_2 \leq \mathbb{P}\Big(\exists Z \in B(\delta),\ \sum_{i=1}^n 1\{\widehat{z}_i(Z) \neq z_i^*\} > n\varepsilon\Big) \leq \mathbb{P}\Big(\sum_{i=1}^n 1\big\{\exists Z \in B(\delta), \widehat{z}_i(Z) \neq z_i^*\big\} > n\varepsilon\Big).$$

By Markov's inequality, we obtain

$$T_2 \leq \frac{1}{n\varepsilon}\sum_{i=1}^n \mathbb{P}\big(\exists Z \in B(\delta),\ \widehat{z}_i(Z) \neq z_i^*\big) \ \leq\ \frac{1}{\varepsilon}e^{-(1-\eta')NI + \frac{3Kn\delta N}{2pn_{\min}}}. \tag{19}$$

where the second inequality follows from Lemma 2, given assumption (a) of the theorem.

Assumption (a) of the theorem also implies $\frac{3Kn\delta N}{2pn_{\min}} = o(NI)$, so this term can be absorbed into $\eta' NI$, giving $T_2 \le \frac{1}{\varepsilon} e^{-(1-\eta'')NI}$ for some $\eta'' = o(1)$. Let

$$\varepsilon = NIe^{-(1-\eta'')NI} = e^{-(1-\eta)NI},$$

where $\eta = \eta'' + \frac{\log(NI)}{NI} = o(1)$. It follows from (19) that

$$T_2 \le \frac{1}{\varepsilon} e^{-(1-\eta'')NI} = \frac{1}{NI} = o(1).$$

Hence, we obtain (12) as desired.

D.4.1 AN AUXILIARY LEMMA

We state the case $\delta = 0$ in Lemma 2 as a separate lemma and prove it first. Recall that $n_{\min} = \min_{k\in[K]} n_k$ where n_k is the size of the kth cluster. We have the following lemma.

Lemma 3 (Local refinement with Z^*)**.** *Suppose the initial label matrix in Algorithm 3 is Z^*, and assume $n_{min}p(1 \wedge I)/K \to \infty$ and $\frac{\log K}{NI} \to 0$, then*

$$\mathbb{P}(\widehat{z}_i \neq z_i^*) = e^{-(1-\eta)NI} \tag{20}$$

for some $\eta = o(1)$. As a direct consequence, $\mathbb{E}[\mathrm{Mis}(\widehat{Z}, Z^*)] \le e^{-(1-\eta)NI}$.

Proof of Lemma 3. Let $q := p/K$ and $\tilde{p} := (K-1)q := p - q$. We first focus on the probability $\mathbb{P}(\widehat{z}_1 \neq z_1^*)$. Let $\mathcal{C}_k^* = \{i \ge 2 : \ z_i^* = e_k\}$. We have $b_k = \sum_{i\in\mathcal{C}_k^*}\langle z_i, z_1\rangle$. Since $z_1^* = e_1$ by assumption, z_1 takes values e_1 and any of $e_\ell, \ell \neq 1$ w.p. $1-\tilde{p}$ and q. For $i \in \mathcal{C}_1^*$, z_i has the same distribution as z_1. For $i \in \mathcal{C}_2^*$, z_i takes values e_2 and any of $e_\ell, \ell \neq 2$ w.p. $1-\tilde{p}$ and q respectively.

Note that (b_1, b_2) is independent of z_1. It follows that

$$(b_1, b_2) \mid z_1 \sim \begin{cases} \mathrm{Bin}(n_1, 1-\tilde{p}) \otimes \mathrm{Bin}(n_2, q), & \text{if } z_1 = e_1 \\ \mathrm{Bin}(n_1, q) \otimes \mathrm{Bin}(n_2, 1-\tilde{p}), & \text{if } z_1 = e_2 \\ \mathrm{Bin}(n_1, q) \otimes \mathrm{Bin}(n_2, q), & \text{if } z_1 \notin \{e_1, e_2\} \end{cases} \tag{21}$$

where $\otimes$ is the notation for the product measure, that is, b_1 and b_2 are independent in each case. The three possibilities above hold with probability $1-\tilde{p}$, q and $(K-2)q$ respectively. Let $Y = n_1b_2 - n_2b_1$ and let $M_Y(\lambda)$ be the moment-generating function (MGF) of Y.

Let $\psi(\lambda; p) = 1 - p + pe^\lambda$ be the MGF of a $\mathrm{Ber}(p)$ variable. Then, the MGF of $\mathrm{Bin}(n, p)$ is $\psi(\lambda; p)^n$ and hence

$$\begin{aligned} \mathbb{E}[e^{\lambda Y} \mid z_1] &= \mathbb{E}[e^{\lambda n_1 b_2} \mid z_1] \cdot \mathbb{E}e^{-\lambda n_2 b_1} \mid z_1] \\ &= \begin{cases} \psi(\lambda n_1; q)^{n_2} \cdot \psi(-\lambda n_2; 1-\tilde{p})^{n_1} & \text{if } z_1 = e_1 \\ \psi(\lambda n_1; 1-\tilde{p})^{n_2} \cdot \psi(-\lambda n_2; q)^{n_1} & \text{if } z_1 = e_2 \\ \psi(\lambda n_1; q)^{n_2} \cdot \psi(-\lambda n_2; q)^{n_1} & \text{if } z_1 \notin \{e_1, e_2\}. \end{cases} \end{aligned}$$

Let $\phi(\lambda; \mu) = \exp(\mu(e^\lambda - 1))$ be the MGF of $\mathrm{Poi}(\mu)$ and note that $\psi(\lambda; p)^n \le \phi(\lambda; np)$. Then, for example, we have

$$\mathbb{E}[e^{\lambda Y} \mid z_1 = e_1] \le \phi(\lambda n_1; n_2 q) \cdot \phi(-\lambda n_2; n_1(1-\tilde{p})).$$

Since $\phi(\lambda; \mu) = \exp[\mu(\lambda + o(\lambda))] = \exp[\mu\lambda(1 + o(1))]$ for $\lambda = o(1)$, we obtain

$$\mathbb{E}[e^{\lambda Y} \mid z_1 = e_1] \le \exp\big[n_1 n_2 \lambda q(1 + o(1)) - n_1 n_2 \lambda(1-\tilde{p})\big(1 + o(1)\big)\big]$$

assuming that $\lambda(n_1 + n_2) = o(1)$. Then,

$$\mathbb{E}[e^{\lambda Y} \mid z_1 = e_1] \ \le \ \exp\big[n_1 n_2 \lambda(q - 1 + \tilde{p})\big(1 + o(1)\big)\big]$$

Take $\lambda = t\,[n_1 n_2(1-p)]^{-1}$ for some $t \ge 0$ to be determined below. Noting $q - 1 + \tilde{p} = -(1-p)$,

$$\mathbb{E}[e^{\lambda Y} \mid z_1 = e_1] \ \le \ \exp\big[-t\big(1 + o(1)\big)\big].$$

The case $z_1 = e_2$ is argued similarly and we obtain the bound $\mathbb{E}[e^{\lambda Y} \mid z_1 = e_2] \le \exp\big[t\big(1+o(1)\big)\big]$. For $z_1 \notin \{e_1, e_2\}$, we perform a second-order expansion, assuming $\lambda = o(1)$:

$$\phi(\lambda;\mu) = \exp\big[\mu\big(\lambda + \frac{1}{2}\lambda^2 + o(\lambda^2)\big)\big] \le \exp\big[\mu\big(\lambda+\lambda^2\big)\big]$$

and obtain

$$\psi(\lambda n_1; q)^{n_2} \cdot \psi(-\lambda n_2; q)^{n_1} \le \exp\big[\lambda^2 n_1 n_2 (n_1+n_2) q\big].$$

Let $\gamma := 2q/(1-p)^2$ and let $n_{\text{har}} := 2n_1n_2/(n_1+n_2)$ be the harmonic mean of n_1 and n_2. Note that $n_{\text{har}} \ge n_{\text{min}}$. We have

$$\lambda^2 n_1 n_2 (n_1+n_2) q = \frac{t^2(n_1+n_2)q}{n_1 n_2 (1-p)^2} = \gamma t^2 / n_{\text{har}}.$$

To summarize, the conditional MGF satisfies

$$\mathbb{E}[e^{tY/(n_1n_2(1-p))} \mid z_1] \le \begin{cases} \exp\big[-t\big(1+o(1)\big)\big] & \text{if } z_1 = e_1 \\ \exp\big[t\big(1+o(1)\big)\big] & \text{if } z_1 = e_2 \\ \exp(\gamma t^2/n_{\text{har}}) & \text{if } z_1 \notin \{e_1, e_2\}. \end{cases}$$

Recall that the events $z_1 = e_1, z_1 = e_2$ and $z_1 \notin \{e_1, e_2\}$ happen with probability $1-\tilde{p}$, q and $(K-2)q$ respectively. It follows that

$$M_Y(t/(n_1n_2(1-p))) \le (1-\tilde{p})\, e^{-t(1+o(1))} + q\, e^{t(1+o(1))} + (K-2)q\, e^{\gamma t^2/n_{\text{har}}}. \tag{22}$$

Let us set

$$t = \frac{1}{2}\log((1-\tilde{p})/q) = \frac{1}{2}\log\Big(1 + \frac{K}{p}(1-p)\Big), \tag{23}$$

so that $(1-\tilde{p})e^{-t} = qe^t = \sqrt{(1-\tilde{p})q}$. Then, $t \ge 0$ and since $\log(1+x) \le x$, we have

$$t \le K(1-p)/(2p). \tag{24}$$

The condition $\lambda(n_1+n_2) = o(1)$ is satisfied under assumption $n_{\text{min}}p/K \to \infty$, since

$$\lambda(n_1+n_2) = \frac{(n_1+n_2)t}{n_1n_2(1-p)} = \frac{2t}{n_{\text{har}}(1-p)} \le \frac{K(1-p)/p}{n_{\text{har}}(1-p)} \le \frac{K}{n_{\text{min}}p} = o(1).$$

Recalling that $q = p/K$, the exponent of the last term in (22) satisfies

$$\frac{\gamma t^2}{n_{\text{har}}} \le \frac{\gamma K^2(1-p)^2}{4p^2 n_{\text{har}}} = \frac{2qK^2}{4p^2 n_{\text{har}}} = \frac{K}{2n_{\text{har}}p} \le \frac{K}{2n_{\text{min}}p} = o(I)$$

under the assumption of the lemma. It follows that

$$\begin{aligned} M_Y(t/(n_1n_2(1-p))) &\le 2\sqrt{(1-\tilde{p})q}\, e^{o(t)} + (K-2)qe^{o(I)} \\ &= 2(\sqrt{(1-\tilde{p})q})^{1+o(1)} + (K-2)qe^{o(I)} \\ &= \big[(\sqrt{(1-\tilde{p})q})^{o(1)} \vee e^{o(I)}\big]\, e^{-I} \end{aligned}$$

where the first equality is by $e^{o(t)} = (e^t)^{o(1)} = (\sqrt{(1-\tilde{p})q})^{o(1)}$ for our choice of t, and the second equality by the definition (11) of I. Since $\sqrt{(1-\tilde{p})q} \le e^{-I}$, we have $(\sqrt{(1-\tilde{p})q})^{o(1)} = e^{o(I)}$, hence

$$M_Y(t/(n_1n_2(1-p))) \;\le\; e^{o(I)}e^{-I} = e^{-(1-o(1))I} \;=\; e^{-(1-\eta)I}.$$

Let $Y_1, \dots, Y_N$ be the i.i.d. copies of Y. By Markov's inequality,

$$\mathbb{P}\Big(\sum_{j=1}^N Y_j \ge 0\Big) = \mathbb{P}\Big(e^{\lambda\sum_{j=1}^N Y_j} \ge 1\Big) \le \mathbb{E}e^{\lambda\sum_{j=1}^N Y_j} = M_{Y_1}(\lambda)^N \le e^{-(1-\eta)NI}.$$

The above argument shows that $\mathbb{P}\big(\frac{b_2}{n_2} \ge \frac{b_1}{n_1}\big) \le e^{-(1-\eta)NI}$. Repeating the argument for the ith label, it shows that

$$\mathbb{P}(\widehat{z}_i(Z^*) \ne z_i^*) \le \mathbb{P}\left(\max_{k=2,\dots,K} \frac{b_k}{n_k} \ge \frac{b_1}{n_1}\right) \le \sum_{k=2}^K \mathbb{P}\left(\frac{b_k}{n_k} \ge \frac{b_1}{n_1}\right) \le (K-1)e^{-(1-\eta)NI}.$$

If $K = 2$, then we have already obtained (20). If $K > 2$, then

$$(K-1)e^{-(1-\eta)NI} = e^{-(1-\eta)NI + \log(K-1)} = e^{-(1-\eta)NI + o(NI)}.$$

The term $o(NI)$ can be absorbed into ηNI, so we can still obtain (20). □

D.4.2 DETAILED PROOF OF LEMMA 2

Let $i = 1$ without loss of generality, and let $Z \in B(\delta)$, and $\hat{n}_k = n_k(Z)$, the size of the kth cluster in the label matrix Z. Let $b_k(Z^*) = (Z^*_{-1}\bar{X}_1)_k$ and $b_k(Z) = (Z_{-1}\bar{X}_1)_k$ where $\bar{X}_1$ is the first column of $\bar{X}$ in the algorithm. Suppose Z has at most $n\delta$ labels different from Z^*, then

$$\big|[n_1 b_2(Z^*) - n_2 b_1(Z^*)] - [n_1 b_2(Z) - n_2 b_1(Z)]\big| \;\le\; (n_1 \vee n_2)\, n\delta.$$

and

$$\begin{aligned}\big|[n_1 b_2(Z) - n_2 b_1(Z)] - [\hat{n}_1 b_2(Z) - \hat{n}_2 b_1(Z)]\big| &\le |n_1 - \hat{n}_1| \cdot b_2(Z) + |n_2 - \hat{n}_2| \cdot b_1(Z) \\ &\le n\delta\,(n_1 + n_2).\end{aligned}$$

Let $Y(Z^*) = n_1 b_2(Z^*) - n_2 b_1(Z^*)$ and $Y(Z) = \hat{n}_1 b_2(Z) - \hat{n}_2 b_1(Z)$. Combining the two by the triangle inequality, we obtain

$$|Y(Z^*) - Y(Z)| \le 3(n_1 \vee n_2)\, n\delta =: h(\delta).$$

By Markov's inequality, for $\lambda \ge 0$,

$$\begin{aligned}\mathbb{P}\Big(\max_{Z \in B(\delta)} Y(Z) \ge 0\Big) &\le \mathbb{P}\big(Y(Z^*) \ge -h(\delta)\big) \\ &= \mathbb{P}\big(e^{\lambda N Y(Z^*)} \ge e^{-\lambda N h(\delta)}\big) \le e^{\lambda N h(\delta)} \mathbb{E}[e^{\lambda N Y(Z^*)}]\end{aligned}$$

As in the proof of Lemma 3, we take $\lambda = t\,[n_1 n_2 (1-p)]^{-1}$, with t given by (23). Using the upper bound (24), we have

$$\lambda N h(\delta) = \frac{3(n_1 \vee n_2) n N}{n_1 n_2 (1-p)} \delta t \le \frac{3Kn\delta N}{2pn_{\min}}.$$

The result follows as in Lemma 3.

E RPM AND BAYESIAN AGGREGATION

One might ask whether RPM is a useful model in practice. For the applications in which all the label vectors are perturbations of a common true "center", and our goal is to recover this center, RPM is a good first approximation. This is the case for Bayesian label aggregation as we argue below. In such settings, the RPM is like the i.i.d. noise model used in classical regression. Although one can imagine more complex regression models (like those with heteroscedastic noise, or mixtures of regressions, etc.), the i.i.d. setting still provides a lot of insights for understanding the more complex models.

E.1 RPM IS A GOOD MODEL FOR A CONCENTRATED POSTERIOR

Lets us now argue how one can arrive at RPM in the context of Bayesian aggregation, by systematically making some assumptions. First, we note that our goal in the paper is not to prove the "posterior concentration around the truth", also known as *posterior consistency*. This is problem-specific and out of the scope of this work. We assume that we are working with a model for which posterior consistency has already been established. The question that we are trying to answer is then this:

> Given that the posterior concentrates around the true partition, and given that the MCMC has converged—that is, we are sampling from this concentrated posterior—can we obtain a consistent estimate of the center of the posterior (which would be the true partition) based these samples?

For this purpose, it is enough to assume that we are observing samples from the posterior $p(z_1, \ldots, z_n|D)$, where D is the observed data, and this posterior is concentrating around z^* which is the true partition. For simplicity, let us drop D and note that the posterior is some multivariate discrete distribution $p(z_1, \ldots, z_n)$. Then, we proceed in steps:

1. First, we address the label-switching issue. Let z be a sample from the posterior and let permutation π be the minimizer of $H(z^*, \pi(z))$ over all $K!$ permutations, where $H(\cdot, \cdot)$ is the Hamming distance. We note that $\pi(z)$ is an equivalent label vector to z (only the cluster

labels are permuted.) We consider the distribution of $\pi(z)$ as the posterior distribution rather than that of z. This is only to simplify the discussion and is without loss of generality, since our methods are invariant to label-switching. The distribution of $\pi(z)$ will have a single mode at z^* while that of z will have $K!$ modes on all equivalent versions of z^*. Given such simplification, renaming $\pi(z)$ to z from now on, the posterior is a multivariate distribution $p(z_1, \ldots, z_n)$ which is highly concentrated around $z^* = (z_1^*, \ldots, z_n^*)$. This follows from the posterior consistency assumption.

2. We claim that this multivariate distribution can be approximated by the product of its marginals $p(z_1, \ldots, z_n) \approx \prod_{i=1}^n p_i(z_i)$. This is intuitively clear for any concentrated discrete distribution. Alternatively, it is intuitively clear to anyone who has looked at MCMC samples at stationarity. Each node/object i usually has a most likely assignment which is z_i^*, but it occasionally jumps to some other label with a small probability. The fluctuations for different nodes are independent; this is intuitively because the bulk of the labels don't change their clusters all at once; only a few do at any given time.
3. *Non-uniform RPM*: Given the independence assumption across i, the most general form $p_i(z_i)$ can take is a categorical (a.k.a. Multinomial(1,π)) distribution. The bulk of the mass of this categorical variable will be on z_i^*, and the rest distributed among the other labels. For example, for some node, i, the label can jump, say, between $z_i^* = 3$ and 5 with the bulk of the probability on 3. For others it could be that when they jump from z_i^*, they land over a larger collection of labels. Here, we are making the simplifying assumption that for each node, the mass that is put on anything other than z_i^* is uniformly distributed over the label set $[K] \setminus \{z_i^*\}$. This assumption can be removed, and we can work with general categorical variables, at the expense of making the rates and the analysis more complicated.
4. *Inhomogeneous RPM*: Given that the noise in the categorical variable is uniform over $[K] \setminus \{z_i^*\}$, we now assume that the probability of taking any of those values is the same for the all nodes (i.e., independent of i). This is exactly the homogeneous RPM that we consider. This assumption is easy to remove and we can work with the inhomogeneous RPM that allows this probability to depend on i.

We do not lose anything in Step 1. Steps 3 and 4 are simplifying steps that are taken for the ease of understanding and presentation. The main assumption is Step 2, the approximation by the product of marginals. Below we provide some hard evidence that this is very reasonable in the Bayesian setting.

E.2 Hard evidence of near-independence

We consider the problem of recovering the clusters in a stochastic block model (SBM) which is a random network model with latent node clusters. Figure 5 shows the Sequential NMI plot for a Gibbs sampler on a non-parametric SBM (with a Dirichlet Process prior on labels). The sequential NMI means that we compute the NMI of the partition at iteration t relative to that at iteration $t-1$, and the x-axis shows t. The plot suggests that the chain enters stationarity roughly around iteration 100.

We use the MMD-based approach, described in (19; 20; 38), to compare the posterior joint distribution of the labels to its approximate versions:

- Figure 6(a) shows the result for samples from iterations 100 to 1000 of the Gibbs sampler. This is the stationary joint distribution.
- Figure 6(b) shows the result for samples from iterations 1 to 100 of the Gibbs sampler. This is based on the transient samples (effectively, average joint behavior over the transient period).
- Figure 6(c) shows the results on a movie rating dataset with complicated dependent joint distribution (that has nothing to do with SBM).

We refer to (19; 20; 38) for details of how these experiments are performed. The bootstrap MMD serves as the baseline; those methods whose MMD is closer to the bootstrap are better approximations of the joint distribution. In general smaller MMD means a better approximation of the joint. Ind Mult is exactly the approximation by independent multinomials (i.e. product of marginals). The Copula Mult is a good joint model for the discrete multivariate distribution.

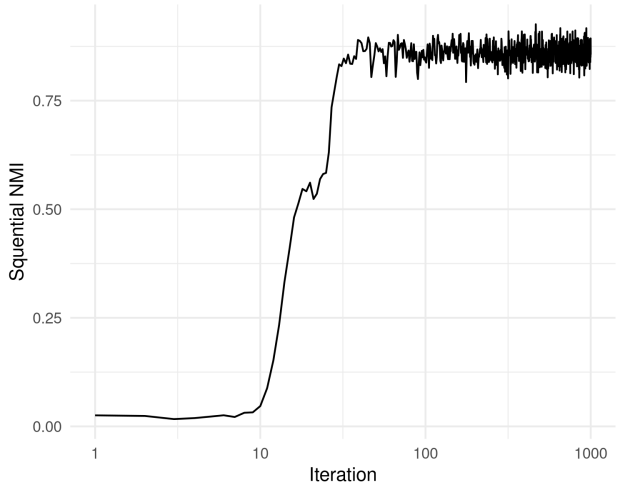
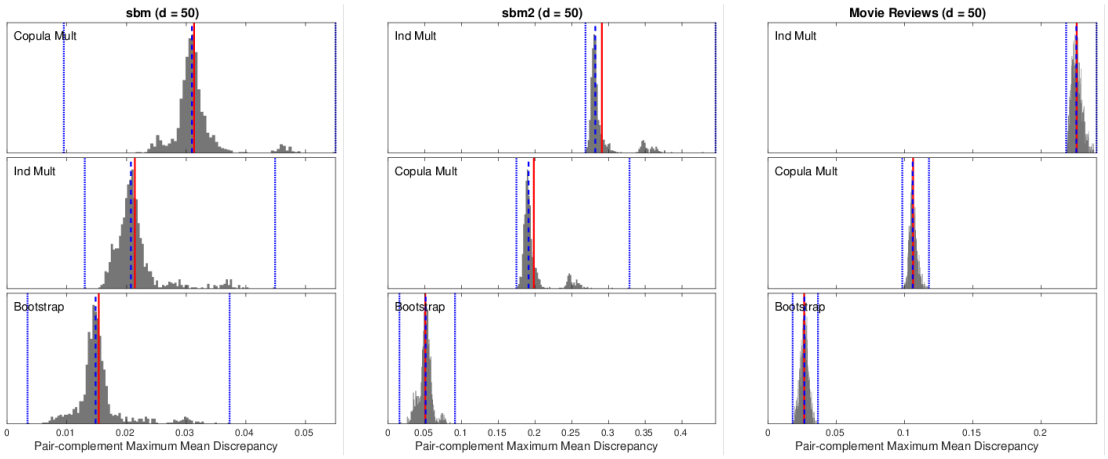


Figure 5: Sequential NMI for the SBM Gibbs sampler



(a) (b) (c)

Figure 6: MMD histograms between the posterior distribution and its various approximations

We see that for the movie rating data and the transient chain, indeed Copula Mult has a much lower MMD, than Ind Mult, showing that there is dependence in the joint that is not captured by Ind Mult. However, for the stationary distribution (Figure 6(a)), the Ind Mult has comparable (and even slightly smaller MMD) relative to the Copula Mult, and is close to bootstrap. This shows that the product of marginals is a good approximation in this case, and justifies Step 2 in our reduction.

F Comparison with Stochastic Block Model

Below we outline some of the similarities and differences between the problem of cluster recovery in SBM and the consensus clustering problem we consider in this paper. Suppose that A is the adjacency matrix generated from the stochastic block model (SBM) and X is the association matrix generated from the RPM.

Similarity: A larger value of X_{ij} increases the likelihood of i and j being in the same cluster in RPM. Similarly, $A_{ij} = 1$, i.e., there is an edge between i and j, increases the likelihood of i and j being in the same community in SBM. Therefore, both X and A can be considered proximity matrices and we can utilize a min-cut algorithm on them to find the clusters. To approximate the

min-cut algorithm, various researchers have proposed the approach of using a good initialization plus a local refinement step. This idea can be applied to many clustering problems, including community detection. In this paper, we show that it can also be applied to consensus clustering.

Difference: The entries of the adjacency matrix A are independent, but the entries of the association matrix X are not. Indeed, the entries on the same row of X have very strong dependence. The likelihood function of X is very different from A, but we can still show that a simple local refinement step outputs optimal labels. The error rate is comparable to the Bayes error rate given by likelihood ratio test.