Topics in mathematical and applied statistics

Pas, S.L. van der

4

Bayesian community detection

Abstract
We introduce a Bayesian estimator of the underlying class structure in the stochastic block model, when the number of classes is known. The estimator is the posterior mode corresponding to a Dirichlet prior on the class proportions, a generalized Bernoulli prior on the class labels, and a beta prior on the edge probabilities. We show that this estimator is strongly consistent when the expected degree is at least of order $\log^2 n$*, where* n *is the number of nodes in the network.*

4.1 Introduction

The stochastic block model (SBM) (Holland et al., 1983) is a model for network data in which individual nodes are considered members of classes or communities, and the probability of a connection occurring between two individuals depends solely on their class membership. It has been applied to social, biological and communication networks, for example in Park and Bader (2012), Bickel and Chen (2009) and Snijders and Nowicki (1997) amongst many others. There are many extensions of the SBM for various applications, including the degree-corrected SBM (Karrer and Newman, 2011; Zhao et al., 2012) which accounts for possible heterogeneity among nodes within the same class, and the mixed-membership SBM (Airoldi et al., 2008), in which the assumption that the classes are disjoint is removed. These extensions allow for additional modelling flexibility.

Two main SBM research directions are the recovery of the class labels (*community*

This chapter has been submitted as: S.L. van der Pas and A.W. van der Vaart. Bayesian community detection. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

detection) and recovery of the remaining model parameters, consisting of the probability vector generating the class labels, and the class-dependent probabilities of creating an edge between nodes. In this paper, we focus on community detection, noting that once strong consistency of a community detection method has been established, consistency of the natural plug-in estimators for the remaining parameters follows directly by results in (Channarond et al., 2012).

A large number of methods for recovering the class labels has been proposed. Those most closely related to this work are the modularities. Newman and Girvan (2004) introduced the term *modularity* for 'a measure of the quality of a particular division of a network'. They described one such measure for models in which edges are more likely to occur within classes than between classes, in which case there is a community structure in the colloquial sense, although the SBM does not require this assumption. Bickel and Chen (2009) studied more general modularities, defining them as functions of the number of connections between all combinations of classes and the proportion of nodes placed in each class. They introduced the likelihood modularity, and provided general conditions under which modularities are consistent. Their method and theory was extended to the degree-corrected SBM by Zhao et al. (2012).

Spectral methods for community detection have gained in popularity, and refined results on error bounds are now available for the SBM and extensions of the SBM, as evidenced in Rohe et al. (2011), Jin (2015), Sarkar and Bickel (2015) and Lei and Rinaldo (2015) for example. Many other algorithms have been introduced, most of them currently lacking formal proofs of consistency. A notable exception is the Largest Gaps algorithm (Channarond et al., 2012), which only takes the degree of each node as its input, and is strongly consistent under a separability condition.

A Bayesian approach towards recovering the class assignments in the SBM was first suggested by Snijders and Nowicki (1997), motivated by computational advantages of Gibbs sampling over maximum likelihood estimation. They considered two classes and proposed uniform priors on the class proportions and the edge probabilities. This approach was extended in (Nowicki and Snijders, 2001) to allow for more classes, with a Dirichlet prior on the class proportions and beta priors on the edge probabilities. Hofman and Wiggins (2008) described a similar Bayesian approach for a special case of the SBM and suggested a variational approach to overcome the computational issues associated with maximizing over all possible class assignments.

Bayesian methods for the SBM have barely been studied from a theoretical point of view, although recent results for parameter recovery by Pati and Bhattacharya (2015), for detecting the number of communites by Hayashi et al. (2016) and for an empirical Bayes approach to community detection by Suwan et al. (2016) are encouraging. In this work, we provide theoretical results on community detection, establishing that the Bayesian posterior mode is strongly consistent for the class labels if the expected degree is at least of order $\log^2 n$, where n is the number of nodes. This is proven by relating the posterior mode to the maximizer of the likelihood modularity of Bickel and Chen (2009). The likelihood modularity has been claimed to be strongly consistent under the weaker assumption that the expected degree is of larger order than $\log n$ (Bickel and Chen, 2009; Bickel et al., 2015; Zhao et al., 2012). However, their proof assumes that the likelihood modularity is globally Lipschitz, while it is only locally so. The Bayesian method is based on a combination

of likelihood and prior, and for this reason the proof of our main theorem, Theorem 4.3, runs into a similar problem. We were able to resolve this only under the slightly stronger assumption that the expected degree is of larger order than $(\log n)^2$. The literature on other methods for community detection shows that the order $\log n$ is sufficient for consistent detection. However, these results are usually obtained under additional assumptions such asăa restriction to two classes or an ordering of the connection probabilities, and their implications for the likelihood or Bayesian modularities is unclear. We discuss this and the relevant literature further following the statement of our main result in Section 4.3.5.

This paper is organized as follows. We introduce the SBM and the associated notation in Section 4.2. Our main results are in Section 4.3, where we describe the prior and the link with the likelihood modularity, present the consistency results and discuss the underlying assumptions, especially those on the expected degree. The method is illustrated on a data set in Section 4.4, and we conclude with a Discussion in Section 4.5. All proofs are given in the Appendix.

4.2 The stochastic block model

We introduce the notation and generative model for the SBM with $K \in \{1, 2, \ldots\}$ classes. Consider an undirected random graph with n nodes, numbered $1, 2, \ldots, n$, and edges encoded by the $n \times n$ symmetric adjacency matrix (A_{ij}), with entries in $\{0, 1\}$. Thus $A_{ij} = A_{ji}$ is equal to 1 or 0 if the nodes i and j are or are not connected by an edge, respectively. Self-loops are not allowed, so $A_{ii} = 0$ for $i = 1, \ldots, n$. The generative model for the random graph is:

1. The nodes are randomly labeled with i.i.d. variables $Z_1, \ldots, Z_n$, taking values in a finite set $\{1, \ldots, K\}$, according to probabilities $\pi = (\pi_1, \ldots, \pi_K)$.

2. Given $Z = (Z_1, \ldots, Z_n)$, the edges are independently generated as Bernoulli variables with $\mathbb{P}(A_{ij} = 1 \mid Z) = P_{Z_i, Z_j}$, for $i < j$, for a given $K \times K$ symmetric matrix $P = (P_{ab})$.

The probability vector π is considered fixed, but unknown. Although this is not visible in the notation, the matrix P may change with n, a case of particular interest being that P tends to zero, which gives a sparse graph. The order of magnitude of $\|P\|_\infty = \max_{a,b} P_{ab}$ is the same as the order of magnitude of $\rho_n = \sum_{a,b} \pi_a \pi_b P_{ab}$, the probability of there being an edge between two randomly selected nodes. The *expected degree* of a randomly selected node is $\lambda_n = (n-1)\rho_n$, and twice the expected total number of edges in the network is $\mu_n = n(n-1)\rho_n$.

The likelihood for the model is given by

$$\prod_{i<j} P_{Z_i Z_j}^{A_{ij}} (1 - P_{Z_i Z_j})^{1 - A_{ij}} \prod_i \pi_{Z_i} = \prod_{a \le b} P_{ab}^{O_{ab}(Z)} (1 - P_{ab})^{n_{ab}(Z) - O_{ab}(Z)} \prod_a \pi_a^{n_a(Z)}, \quad (4.1)$$

where $O_{ab}(Z)$ is the number of edges between nodes labelled a and b by the labelling Z, $n_{ab}(Z)$ is the maximum number of edges that can be created between nodes labelled a and b, and $n_a(Z)$ is the number of nodes labelled a, and a and b range over $\{1, 2, \ldots, K\}$.

More formally, for a given labelling $e = (e_1, \dots, e_n) \in \{1, \dots, K\}^n$ of nodes, and class labels $a, b \in \{1, \dots, K\}$, we define

$$O_{ab}(e) = \begin{cases} \sum_{i,j} A_{ij} \mathbf{1}_{\{e_i=a, e_j=b\}}, & a \neq b, \\ \sum_{i<j} A_{ij} \mathbf{1}_{\{e_i=a, e_j=b\}}, & a = b, \end{cases}$$

$$n_{ab}(e) = \begin{cases} n_a(e) n_b(e), & a \neq b, \\ \frac{1}{2} n_a(e)(n_a(e) - 1), & a = b, \end{cases}$$

$$n_a(e) = \sum_{i=1}^{n} \mathbf{1}_{\{e_i=a\}}.$$

Since the matrix A is symmetric with zero diagonal by assumption, for $a \neq b$ the variable $O_{ab}(e)$ can also be written as $\sum_{i<j} A_{ij}[\mathbf{1}_{\{e_i=a, e_j=b\}} + \mathbf{1}_{\{e_j=a, e_i=b\}}]$, which explains the different appearances of the diagonal and off-diagonal entries. The numbers $n_{ab}(e)$ are equal to the numbers $O_{ab}(e)$ when all A_{ij} are equal to 1. We collect the variables $O_{ab}(e)$ and $n_{ab}(e)$ in $K \times K$ matrices $O(e)$ and $n(e)$.

Now consider the $K \times K$ probability matrix $R(e, c)$ and K probability vector $f(e)$ with entries

$$R_{ab}(e, c) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{e_i=a, c_i=b\}}, \qquad f_a(e) = \frac{n_a(e)}{n}. \tag{4.2}$$

The row sums of $R(e, c)$ are equal to $R(e, c)\mathbf{1} = f(e)$, while the column sums are equal to $\mathbf{1}^T R(e, c) = f(c)^T$. Thus, the matrix $R(e, c)$ can be seen as a coupling of the marginal probability vectors $f(e)$ and $f(c)$. If $e = c$, then it is diagonal with diagonal $f(c) = f(e)$. More generally, the matrix can be viewed as measuring the discrepancy between labellings e and c. This can be precisely measured as half the L_1-distance of $R(e, c)$ to its diagonal, as evidenced by Lemma 4.1, which is noted in Bickel and Chen (2009).

For a vector v we denote by $\mathrm{Diag}(v)$ the diagonal matrix with diagonal v, and for a matrix M we denote its diagonal by $\mathrm{diag}\,(M)$.

Lemma 4.1. *For every labelling c, e in the K-class stochastic block model:*

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{c_i \neq e_i\}} = \tfrac{1}{2} \|\mathrm{Diag}(f(c)) - R(e, c)\|_1.$$

Proof. The diagonal of $R(e, c)$ gives the fractions of labels on which c and e agree. Hence the left side of the lemma is $1 - \sum_a R_{aa}(e, c) = \sum_a (f_a(c) - R_{aa}(c))$. The elements of both $K \times K$ matrices $\mathrm{Diag}(f(c))$ and $R(e, c)$ can be viewed as probabilities that add up to 1. Thus the sum of the differences of the diagonal elements is minus the sum of the differences of the off-diagonal elements. Because $f_a(c) \geq R_{aa}(e, c)$ for every a, we have $\sum_a (f_a(c) - R_{aa}(e, c)) = \sum_a |f_a(c) - R_{aa}(e, c)|$. Similarly the off-diagonal elements of $\mathrm{Diag}(f(c))$, which are zero, are smaller than the off-diagonal elements of $R(e, c)$ and hence we can add absolute values. Thus the sum over the diagonal is half the sum of the absolute values of all terms in $\mathrm{Diag}(f(c)) - R(e, c)$. □

4.3 Bayesian approach to community detection

Our main results are presented in this section. We first discuss the choice of prior in Section 4.3.1, and define the estimator, in Section 4.3.2. The resulting Bayesian modularity is closely related to the likelihood modularity of Bickel and Chen (2009). The relationship is clarified in Section 4.3.3. We briefly consider the issue of identifiability in the SBM in Section 4.3.4, and conclude with our main theorem on the strong consistency of the Bayesian modularity in Section 4.3.5.

4.3.1 The prior

We adopt the Bayesian approach of Nowicki and Snijders (2001). We put prior distributions on the parameters of the stochastic block model with K known, the vector π and the matrix P, yielding a joint probability distribution of (A, Z, π, P). Next we marginalize over π and P as in McDaid et al. (2013), leading to a joint distribution of (A, Z). Finally we "estimate" the unobserved vector Z by the posterior mode of the conditional distribution of Z given A. From a frequentist point of view this means that Z is treated as a parameter of the problem, equipped with a hierarchical prior that chooses first π and then Z. Accordingly we shall change notation from Z to e, reserving Z for the frequentist description of the stochastic block model in Section 4.2.

The prior on π is a Dirichlet, and independently the P_{ab} for $a \leq b$ receive independent beta priors:

$$\begin{aligned} \pi &\sim \mathrm{Dir}(\alpha, \ldots, \alpha), \\ P_{ab} &\overset{i.i.d.}{\sim} \mathrm{Beta}(\beta_1, \beta_2), \quad 1 \leq a \leq b \leq K. \end{aligned}$$

This is essentially the same set-up as in Nowicki and Snijders (2001) and McDaid et al. (2013), except that we use a more flexible $\mathrm{Beta}(\beta_1, \beta_2)$ instead of a uniform prior on the P_{ab}. We assume $\alpha, \beta_1, \beta_2 > 0$.

We complete the Bayesian model by specifying class labels $e = (e_1, \ldots, e_n)$ and edges $A = (A_{ij} : i < j)$ through

$$\begin{aligned} e_i \mid \pi, P &\overset{i.i.d.}{\sim} \pi, \quad 1 \leq i \leq n, \\ A_{ij} \mid \pi, P, e &\overset{ind.}{\sim} \mathrm{Bernoulli}(P_{e_i, e_j}), \quad 1 \leq i < j \leq n. \end{aligned}$$

Abusing notation we write $p(e)$, $p(A \mid e)$ and $p(e \mid A)$ for marginal and conditional probability density functions.

4.3.2 The Bayesian modularity

The Bayesian estimator of the class labels will be the posterior mode, that is:

$$\widehat{e} = \operatorname*{argmax}_e p(e \mid A).$$

The posterior mode can be interpreted as a modularity-based estimator in the sense of Bickel and Chen (2009), in that it maximizes a function that only depends on the $O_{ab}(e)$ and

the $n_a(e)$. This can be seen from the joint density of (A, e), which is found by marginalizing the likelihood (4.1) over π and P. The conjugacy between the multinomial and Dirichlet distributions gives the marginal density of the class assignment e as:

$$p(e) = \int_{S_K} \prod_a \pi_a^{n_a(e)} \frac{\prod_a \pi_a^{\alpha-1}}{D(\alpha)}\, d\pi = \frac{\Gamma(\alpha K)}{\Gamma(\alpha)^K \Gamma(n+\alpha K)} \prod_a \Gamma(n_a(e)+\alpha). \tag{4.3}$$

Here the integral is relative to the Lebesgue measure on the K-dimensional unit simplex and $D(\alpha) = \Gamma(\alpha)^K/\Gamma(K\alpha)$ is the norming constant for the Dirichlet density. Similarly the conjugacy between the Bernoulli and Beta distributions gives the marginal conditional density of A given e as:

$$\begin{aligned} p(A \mid e) &= \int_{[0,1]^{K(K+1)/2}} \prod_{a\le b} P_{ab}^{O_{ab}(e)} (1-P_{ab})^{n_{ab}(e)-O_{ab}(e)} \prod_{a\le b} \frac{P_{ab}^{\beta_1-1}(1-P_{ab})^{\beta_2-1}}{B(\beta_1,\beta_2)}\, dP \\ &= \prod_{a\le b} \frac{1}{B(\beta_1,\beta_2)} B(O_{ab}(e)+\beta_1, n_{ab}(e)-O_{ab}(e)+\beta_2), \end{aligned} \tag{4.4}$$

where $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the beta-function. The joint density of A and e is given by the product of (4.3) and (4.4), and n^{-2} times its logarithm is up to a constant that is free of e equal to

$$Q_B(e) = \frac{1}{n^2} \sum_{1\le a\le b\le K} \log B(O_{ab}(e)+\beta_1, n_{ab}(e)-O_{ab}(e)+\beta_2) + \frac{1}{n^2}\sum_{a=1}^{K} \log\Gamma(n_a(e)+\alpha).$$

This is a modularity in the sense of Bickel and Chen (2009), which we define as the *Bayesian modularity*. As $p(e \mid A)$ is proportional to $p(e, A)$, the posterior mode is equal to the class assignment that maximizes the Bayesian modularity, so the Bayesian estimator is equal to:

$$\widehat{e} = \operatorname*{argmax}_e Q_B(e). \tag{4.5}$$

4.3.3 Similarity to the likelihood modularity

The Bayesian modularity $Q_B(e)$ consists of a two parts, originating from the likelihood and the prior on the classification, respectively. The first part is close to the *likelihood modularity* given by

$$Q_{ML}(e) = \frac{1}{n^2} \sum_{1\le a\le b\le K} n_{ab}(e)\, \tau\Big(\frac{O_{ab}(e)}{n_{ab}(e)}\Big),$$

where $\tau(x) = x\log x + (1-x)\log(1-x)$. This criterion, obtained in Bickel and Chen (2009), results from replacing in the log conditional likelihood of A given e (the logarithm of (4.1) with Z replaced by e and discarding the term involving the parameters π_a) the parameters P_{ab} by their maximum likelihood estimators $\hat{P}_{ab} = O_{ab}(e)/n_{ab}(e)$. In other words, the parameters are *profiled out* rather than integrated out as for the Bayesian modularity. The corresponding estimator

$$\widehat{e}_{ML} = \operatorname*{argmax}_e Q_{ML}(e)$$

is consistent, and hence one may hope that the Bayesian estimator can be proved consistent by showing that the Bayesian and likelihood modularities are close. This will indeed be our line of approach, but the execution must be done with care. For instance, the second, prior part of the Bayesian modularity does play a role in the proof of strong consistency, although it is negligible when proving weak consistency.

The following lemma links the Bayesian and likelihood modularities.

Lemma 4.2. *There exists a constant C such that, for $\mathcal{E} = \{1,\dots,K\}^n$ the set of all possible labellings:*

$$\max_{e\in\mathcal{E}}\Big|Q_B(e) - Q_{ML}(e) - Q_P(e)\Big| \le \frac{C\log n}{n^2},$$

for

$$Q_P(e) = \frac{1}{n^2}\sum_{a:n_a(e)+\lfloor\alpha\rfloor\ge 2} n_a(e)\log(n_a(e)) - \frac{1}{n}.$$

Consequently $\max_{e\in\mathcal{E}}\big|Q_B(e) - Q_{ML}(e)\big| = O\big(\log n/n\big)$.

4.3.4 Identifiability and consistency

A classification $\widehat{e}$ is said to be *weakly consistent* if the fraction of misclassified nodes tends to zero (partial recovery), and *strongly consistent* if the probability of misclassifying any of the nodes tends to zero (exact recovery). In defining consistency in a precise manner, the complication of the possible unidentifiability of the labels needs to be dealt with. From the observed data A we can at best recover the partition of the n nodes in the K classes with equal labels Z_i, but not the values $Z_1,\dots,Z_n$ of the labels, in the set $\{1,2,\dots,K\}$, attached to the classes. Thus consistency will be up to a permutation of labels.

To make this precise define, for a given permutation $(1,\dots,K) \to (\sigma(1),\dots,\sigma(K))$, the *permutation matrix* P_σ as the matrix with rows

$$\begin{matrix} e_{\sigma(1)}^T \\ \vdots \\ e_{\sigma(K)}^T, \end{matrix}$$

for $e_1,\dots,e_K$ the unit vectors in $\mathbb{R}^K$. Then pre-multiplication of a matrix by P_σ permutes the rows, and post-multiplication by P_σ^T the columns: $P_\sigma R$ is the matrix with jth row equal to the $\sigma(j)$th row of R, and RP_σ^T is the matrix with jth column the $\sigma(j)$th column of R. Thus $P_\sigma R(e,Z)$ is the matrix that would result if we would permute the labels of the classes of the assignment e, and $P_\sigma P P_\sigma^T$ and $P_\sigma R(e,Z)P_\sigma^T$ are the matrices that would result if we would relabel the classes throughout. Since we cannot recover the labels, the matrix $P_\sigma R(e,Z)$ is just as good or bad as $R(e,Z)$ for measuring discrepancy between a labelling e and the true labelling Z; furthermore, nothing should change if we choose different names for the classes.

Thus, taking into account the unidentifiability of the labels, by Lemma 4.1, an estimator $\widehat{e}$ is *weakly consistent* if

$$\|P_\sigma R(\widehat{e}, Z) - \mathrm{Diag}(f(Z))\|_1 \to 0,$$

for some permutation matrix P_σ. The classification $\widehat{e}$ is said to be *strongly consistent* if

$$\mathbb{P}(P_\sigma R(\widehat{e}, Z) = \mathrm{Diag}(f(Z))) \to 1,$$

for some permutation matrix P_σ.

The permutation matrix P_σ is for large n uniquely defined: if $\|(P_\sigma)_j R - \mathrm{Diag}(\pi)\|_1 \leq \min_a \pi_a$, for $j = 1, 2$, then $(P_\sigma)_1 = (P_\sigma)_2$. This follows because the assumption implies that $\|(P_\sigma)_1^{-1}\mathrm{Diag}(\pi) - (P_\sigma)_2^{-1}\mathrm{Diag}(\pi)\|_1 \leq 2\min_a \pi_a$, by the triangle inequality and the fact that the L_1-norm is invariant under permutations. Furthermore, for $P_\sigma = (P_\sigma)_2(P_\sigma)_1^{-1}$ the left side is $\|P_\sigma \mathrm{Diag}(\pi) - \mathrm{Diag}(\pi)\|_1$, which is at least two times the sum of the two smallest coordinates of π if $P_\sigma \neq I$.

A necessary requirement for consistency is that the classes can be recovered from the likelihood, i.e. the model parameters must be identifiable. If π has strictly positive coordinates, so that all labels will appear in the data eventually, then as explained in Bickel and Chen (2009) an appropriate condition is that P does not have two identical rows. If $\pi_a = 0$ for some a, then class a will never be consumed; the identifiability condition should then be imposed after deleting the ath column from P. Thus, we call the pair (P, π) *identifiable* if the rows of P are different after removing the columns corresponding to zero coordinates of π. Throughout we assume that P is symmetric.

4.3.5 Consistency results and assumptions

We are now ready to present our results on consistency for the Bayesian maximum a posteriori (MAP) estimator (4.5). Theorem 4.3 shows strong consistency of the Bayesian estimator if $\lambda_n \gg (\log n)^2$. The proof rests on a proof of weak consistency under similar conditions, stated in the appendix as Theorem 4.4.

Recall that $\rho_n = \sum_{a,b} \pi_a \pi_b P_{ab}$ is the probability of a new edge, and $\lambda_n = (n-1)\rho_n$ is the expected degree of a node.

Theorem 4.3 (strong consistency). *(i) If (P, π) is fixed and identifiable with $0 < P < 1$ and $\pi > 0$ then the MAP classifier $\widehat{e} = \arg\max_e Q_B(e)$ is strongly consistent.*

(ii) If $P = \rho_n S$, where (S, π) is fixed and identifiable with $S > 0$ and $\pi > 0$, then the MAP classifier $\widehat{e} = \arg\max_e Q_B(e)$ is strongly consistent if $\lambda_n \gg (\log n)^2$.

The theorem distinguishes two cases: i is the *dense* case, while ii is the *sparse* case. The second is the most interesting of the two, as it touches on the question how much information is required to recover the underlying community structure. Much recent research effort has gone into determining detection and computational boundaries, in particular for special cases of the SBM with $K = 2$ (see e.g. Mossel et al. (2012), Chen and Xu (2014), Abbe et al. (2014) and Zhang and Zhou (2015)).

Weakly consistent estimation of the class labels for an arbitrary, but known, number of classes is possible under the assumption $\lambda_n \gg \log n$, as this was shown to hold for

spectral clustering by Lei and Rinaldo (2015). *Strong* consistency of maximum likelihood was shown to hold in the special cases of planted bisection ($K = 2$ and equal community sizes) and planted clustering (equal community sizes and P_{ab} can take two values) by Abbe et al. (2014); Chen and Xu (2014), again under the assumption $\lambda_n \gg \log n$. Gao et al. (2015) and Gao et al. (2016) achieve optimality in different senses, under assumptions on the average within-community and between-community edge probabilities; Gao et al. (2015) introduce a two-stage procedure which achieves the optimal proportion of misclassified nodes in a special case where P_{ab} can only take two values, while Gao et al. (2016) obtain minimax rates for the proportion of misclassified nodes in the degree corrected SBM.

Strong consistency of the likelihood modularity for an arbitrary number of classes K has been claimed under the same assumption $\lambda_n \gg \log n$ (Bickel and Chen, 2009), and those results have been extended to the degree-corrected SBM (Zhao et al., 2012). However, these results were obtained by application of an abstract theorem to the special case of the likelihood modularity, which would require the function $\tau(x) = x \log x + (1 - x) \log(1 - x)$, or the function $\sigma(x) = x \log x$, to be globally Lipschitz. As τ and σ are only locally Lipschitz, it is still unclear whether $\lambda_n \gg \log n$ is a sufficient condition for either weakly or strongly consistent estimation by maximum likelihood. From our proof of Theorem 4.3, which proceeds by comparing the Bayesian modularity to the likelihood modularity, it immediately follows that $\lambda_n \gg (\log n)^2$ is certainly sufficient. Given weak consistency the problem can be reduced to a neighbourhood of the true parameter on which the Lipschitz condition is reasonable. However, it is precisely our proof of weak consistency that needs the additional $\log n$ factor.

The Largest Gaps algorithm of Channarond et al. (2012) is strongly consistent provided that $\min_{a \neq b} |\sum_{k=1}^{K} \alpha_k (P_{ak} - P_{bk})|$ is at least of order $\sqrt{\log n / n}$, implying that at least one of the P_{ab} is of the same order, and thus $\lambda_n \gg \sqrt{n \log n}$. This much stronger condition is not surprising, as the Largest Gaps algorithm only uses the degree of a node and does not take into account any finer information on the group structure, such as the information contained in the O_{ab}.

To the best of our knowledge, for $K > 2$, it remains to be shown that $\lambda \gg \log n$ is sufficient for strong consistency of any community detection method for the general SBM. For the minimax rate for the proportion of misclustered nodes in community detection, when only classes of sizes proportional to n are considered, a phase transition when going from the case $K = 2$ to $K \geq 3$ was observed by Zhang and Zhou (2015). Their results show that if $K = 2$, communities of the same size are most difficult to distinguish, while if $K \geq 3$, small communities are harder to discover. This shift in the nature of the communities that are harder to detect may be what has been preventing a general strong consistency result under the assumption $\lambda_n \gg \log n$ so far.

4.4 Application to the karate club data set

Some options for implementing the Bayesian modularity are given in Section 4.4.1, after which the results of applying the Bayesian and likelihood modularities to the well-studied karate club data of Zachary (1977) are discussed in Section 4.4.2.

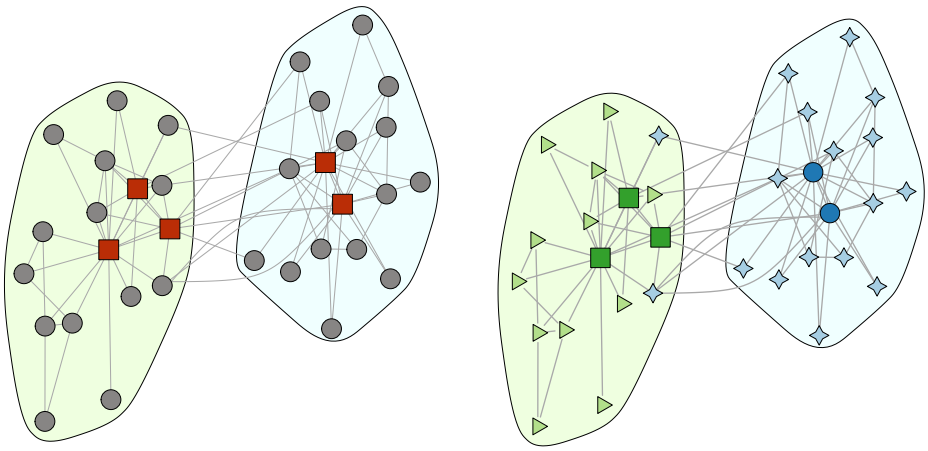

Figure 4.1: Communities detected by the Bayesian modularity when $K = 2$ (left) and $K = 4$ (right), with $\alpha = \beta_1 = \beta_2 = 1/2$. The polygons contain the two groups the karate club was split into; the left one is Mr. Hi's club, the right one is the Officers' club. The shapes of the nodes represent the communities selected by the modularities. Figure made using the igraph package (Csardi and Nepusz, 2006).

4.4.1 Implementation

Two recent works explicitly discuss implementation of Bayesian methods for the SBM. McDaid et al. (2013) followed the approach of Nowicki and Snijders (2001) and added a Poisson prior on K. After marginalizing over π and P, they employ an allocation sampler to sample from the joint density of K and z given A, and use the posterior mode to estimate K. Their algorithm can scale to networks with approximately ten thousand nodes and ten million edges. Côme and Latouche (2014), claiming that the algorithm of McDaid et al. (2013) suffers from poor mixing properties, propose a greedy inference algorithm for the same problem. For the karate club data in Section 4.4.2, the network was small enough that a tabu search (Glover, 1989), run for a number of different initial configurations, yielded good results. We used $\alpha = 1/2$ for the Dirichlet prior, and $\beta_1 = \beta_2 = 1/2$ for the beta prior.

4.4.2 Karate club

Zachary (1977) described a karate club which split into two clubs after a conflict over the price of the karate lessons. The new club was led by Mr. Hi, the karate teacher of the original club, while the remainder of the old club stayed under the former Officers' rule. The data consists of an adjacency matrix for those 34 individuals who interacted with other club members outside club meetings and classes. Each of these individuals' affiliations after the conflict is known.

The communities selected by the Bayesian modularity for $K = 2$ and $K = 4$ are given in Figure 4.1. In both instances, the tabu search led to nearly the same solution for both the

Bayesian and likelihood modularities, only differing at one node for $K = 4$, which is not surprising in light of Lemma 4.2. For $K = 2$, the results of Bickel and Chen (2009) for this data set are recovered. For $K = 4$, the partition in Figure 4.1 yields a higher value of the likelihood modularity than the partition into four classes found by Bickel and Chen (2009), and an even higher value is obtained by switching club member 20 to the second-largest class. This discrepancy is likely due to the heuristic nature of the tabu search algorithm, and for the same reason, it may be the case that improvement over the partitions found by the Bayesian modularity in Figure 4.1 are possible.

For $K = 2$, the communities found by the algorithms do not correspond in the slightest to the two karate clubs, instead grouping the nodes with the highest degrees, corresponding to Mr. Hi, the president of the original club, and their closest supporters, together. Incidentally, this partition is the same as the one returned by the Largest Gaps algorithm of Channarond et al. (2012), which solely uses the degrees of the nodes and discards all other information.

These bad results are no reason to shelve the Bayesian and likelihood modularities, as there is no reason to believe that the two karate clubs form communities in the sense of the stochastic block model. Mr. Hi and the club's president are clear outliers within their groups, and neither of the algorithms were designed to be robust to such a phenomenon. The communities selected by the modularities are communities in the sense that they form connections within and between the groups in a similar fashion. This sense does not correspond to the social notion of a community in this setting.

The results for four classes unify the social and stochastic senses of community. The prominent members of each of the new clubs are placed into two separate, small, communities. The other members are classified nearly perfectly, with two exceptions. However, one of those exceptional individuals is the only person described by Zachary (1977) as being a supporter of the club's president before the split, who joined Mr. Hi's club, making this person's affiliation up for debate. The second is described as only a weak supporter of Mr. Hi. The increased number of communities allows for some outliers within the social communities, and leads to a more detailed understanding of the dynamics within both of the groups. We essentially recover the two communities, each with a core that is more connective than the remainder of the nodes.

4.5 Discussion

An advantage of Bayesian modelling is that it does not solely result in an estimator, but in a full posterior distribution. The posterior mode studied in this paper is but one aspect of the posterior, and its good behaviour in terms of consistency is encouraging. Further study into other aspects in the posterior may prove to be fruitful. One possible research direction would be to use the posterior to *quantify uncertainty* in the estimate of the class labels. A second issue that may be resolved by the Bayesian approach is the question of estimating the number of classes, K. This remains an important open question, as noted by Bickel and Chen (2009), despite recent attempts (e.g. Saldana et al. (2014), Chen and Lei (2014) and Wang and Bickel (2015)). By introducing a prior on K, such as the Poisson-prior suggested by McDaid et al. (2013), the number of communities K can be detected by the posterior.

4.6 Proofs

After stating some repeatedly used notation, this appendix starts with the proof of Theorem 4.4, which is a theorem on weak consistency of the Bayesian modularity. It is followed by a number of supporting Lemmas, after which we proceed to the proof of Theorem 4.3, and some additional supporting Lemmas.

We write diag (P) for the diagonal of P if P is a matrix, and Diag(f) for the diagonal matrix with diagonal f if f is a vector.

4.6.1 Weak consistency

The following quantities will be used in the course of multiple proofs. The function H_P, with domain $K \times K$ probability matrices, is given by, for $\tau(u) = u \log u + (1-u)\log(1-u)$,

$$H_P(R) = \frac{1}{2}\sum_{a,b}(R\mathbf{1})_a(R\mathbf{1})_b\,\tau\left(\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\right). \tag{4.6}$$

For $\tau_0(u) = u\log(u) - u$, define

$$G_P(R) = \frac{1}{2}\sum_{a,b}(R\mathbf{1})_a(R\mathbf{1})_b\,\tau_0\Big(\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\Big).$$

The sums defining these functions are over all pairs (a,b) with $1 \le a,b \le K$, unlike the sums defining the modularities Q_B and Q_{ML}, which are restricted to $a \le b$.

Theorem 4.4 (weak consistency)**.** *(i) If (P,π) is fixed and identifiable, then the MAP classifier $\widehat{e} = \arg\max_z Q_B(e)$ is weakly consistent.*

(ii) If $P = \rho_n S$ for $\rho_n \to 0$, and (S,π) is fixed and identifiable, then the MAP classifier $\widehat{e} = \arg\max_z Q_B(e)$ is weakly consistent provided $n\rho_n \gg (\log n)^2$.

Proof. By Lemma 4.2 the Bayesian modularity Q_B is equivalent to the likelihood modularity Q_{ML} up to order $(\log n)/n$. With the notation $\widetilde{O}_{ab}(e) = O_{ab}(e)$ if $a \ne b$, and $\widetilde{O}_{ab}(e) = 2O_{ab}(e)$ if $a = b$, the likelihood modularity is in turn equivalent up to the same order to

$$\mathbb{L}(e) = \frac{1}{2n^2}\sum_{a,b} n_a(e)n_b(e)\,\tau\Big(\frac{\widetilde{O}_{ab}(e)}{n_a(e)n_b(e)}\Big). \tag{4.7}$$

Indeed the terms of $Q_{ML}(e)$ for $a < b$ are identical to the sums of the terms of $\mathbb{L}(e)$ for $a < b$ and $a > b$, while for $a = b$ the terms of $Q_{ML}(e)$ and $\mathbb{L}(e)$ differ only subtly: the first uses $n_{aa}(e) = \frac{1}{2}n_a(e)(n_a(e)-1)$, where the second uses $\frac{1}{2}n_a(e)^2$. Thus the difference is bounded in absolute value by the sum over a of (where e is suppressed from the notation)

$$\Big|\frac{n_a^2}{2n^2}\tau\Big(\frac{\widetilde{O}_{aa}}{n_a^2}\Big) - \frac{n_a\big(n_a-1\big)}{2n^2}\tau\Big(\frac{\widetilde{O}_{aa}}{n_a(n_a-1)}\Big)\Big| \le \frac{1}{2n}\|\tau\|_\infty + \frac{n_a^2}{2n^2}l\Big(\frac{\widetilde{O}_{aa}}{n_a^2(n_a-1)}\Big).$$

where $l(x) = x(1\vee\log(1/x))$, in view of Lemma 4.7. We now use that $n_a l(u/n_a) \lesssim \log n_a \le \log n$, for $0 \le u \le 1$.

Combining the preceding, we conclude that

$$\eta_{n,1} := \max_e |\mathbb{L}(e) - Q_B(e)| = O\left(\frac{\log n}{n}\right).$$

Since $Q_B(\widehat{e}) \geq Q_B(Z)$, by the definition of $\widehat{e}$, it follows that $\mathbb{L}(\widehat{e}) - \mathbb{L}(Z) \geq -2\eta_{n,1}$. The next step is to replace $\mathbb{L}$ in this equality by an asymptotic value.

For x equal to a big multiple of $(\|P\|_\infty^{1/2} \vee n^{-1/2})/n^{1/2}$, the right side of Lemma 4.5 tends to zero and hence $\max_e \left\|\widetilde{O}(e) - \mathbb{E}\big(\widetilde{O}(e) \mid Z\big)\right\|_\infty / n^2$ is of this order in probability. We also have, by Lemma 4.6:

$$\max_e \left\| \frac{1}{n^2} \mathbb{E}\big(\widetilde{O}(e) \mid Z\big) - R(e,Z)PR(e,Z)^T \right\|_\infty = \max_e \frac{1}{n} \left\| \mathrm{Diag}(R(e,Z))\, \mathrm{diag}\,(P) \right\|_\infty \to 0,$$

as each entry of $\mathrm{Diag}(R(e,Z))\,\mathrm{diag}\,(P)$ is bounded above by one. By Lemma 4.7, $\big|v\tau(x/v) - v\tau(y/v)\big| \leq l(|x-y|)$, uniformly in $v \in [0,1]$, where $l(x) = x(1 \vee \log(1/x))$. It follows that

$$\eta_{n,2} := \max_e \big|\mathbb{L}(e) - L(e)\big| = o_P\Big(l\Big(\frac{\|P\|_\infty^{1/2} \vee n^{-1/2}}{n^{1/2}}\Big)\Big),$$

for

$$L(e) = \frac{1}{2} \sum_{a,b} f_a(e) f_b(e)\, \tau\Big(\frac{(R(e,Z)PR(e,Z)^T)_{ab}}{f_a(e) f_b(e)}\Big).$$

Combining this with the preceding paragraph, we conclude that $L(\widehat{e}) \geq L(Z) - 2(\eta_{n,1} + \eta_{n,2})$.

Proof of i. For given $\delta > 0$, let $\mathcal{R}_\delta$ be the set of all probability matrices R with

$$\min_{P_\sigma} \left\| P_\sigma R - \mathrm{Diag}(R^T \mathbf{1}) \right\|_1 \geq \delta, \qquad \text{and} \qquad \min_{a:\pi_a > 0} (R^T \mathbf{1})_a \geq \delta.$$

Here the minimum is taken over the (finite) set of all permutation matrices P_σ on K labels. Furthermore, set

$$\eta := \inf_{R \in \mathcal{R}_\delta} \Big[H_P\Big(\mathrm{Diag}(R^T \mathbf{1})\Big) - H_P(R) \Big],$$

where H_P is as defined in (4.6). Because $\mathcal{R}_\delta$ is compact and the maps $R \mapsto H_P(R)$ and $R \mapsto \mathrm{Diag}(R^T \mathbf{1})$ are continuous, the infimum in the display is assumed for some $R \in \mathcal{R}_\delta$. Because no $R \in \mathcal{R}_\delta$ can be transformed into a diagonal element by permuting rows and every $R \in \mathcal{R}_\delta$ has a nonzero element in every column a with $\pi_a > 0$, Lemma 4.8 shows that $\eta_n > 0$.

Because $L(e) = H_P(R(e,Z))$ for every e, and $R(Z,Z) = \mathrm{Diag}(f(Z)) = \mathrm{Diag}(R(\widehat{e},Z)^T \mathbf{1})$, we conclude that

$$H_P(\mathrm{Diag}(R(\widehat{e},Z)^T \mathbf{1})) - H_P(R(\widehat{e},Z)) \leq 2(\eta_{n,1} + \eta_{n,2}).$$

If $2(\eta_{n,1} + \eta_{n,2})$ is smaller than η_n, then it follows that $R(\widehat{e},Z)$ cannot be contained in $\mathcal{R}_\delta$. Since $R(\widehat{e},Z)^T \mathbf{1} = f(Z) \xrightarrow{P} \pi$, by the law of large numbers, for sufficiently small $\delta > 0$ this must be because $R(\widehat{e},Z)$ fails the first requirement defining $\mathcal{R}_\delta$. That is, $\|P_\sigma R(\widehat{e},Z) -$

$\mathrm{Diag}(f(Z))\|_1 \le \delta$ for some permutation matrix P_σ. As this is true eventually for any $\delta > 0$, it follows that $\min_{P_\sigma} \|P_\sigma R(\widehat{e}, Z) - \mathrm{Diag}(\pi)\|_1 \xrightarrow{P} 0$.

Proof of ii. In view of Lemma 4.9, the number $\eta = \eta_n$, which now depends on n, is now bounded below by ρ_n times a positive number that depends on (S, π). The preceding argument goes through provided $\eta_{n,1} + \eta_{n,2}$ is of smaller order than η_n. This leads to $l\Big(\sqrt{\rho_n/n}\Big) + \log(n)/n \ll \rho_n$, or $(\rho_n/n)\log^2\Big(n/(\rho_n\|S\|_\infty)\Big) \ll \rho_n^2$. □

Lemma 4.5. *Let $\widetilde{O}_{ab}(e) = O_{ab}(e)$ if $a \neq b$, and $\widetilde{O}_{ab}(e) = 2O_{ab}(e)$ if $a = b$. For any $x > 0$,*

$$\mathbb{P}\Big(\max_e \Big\|\widetilde{O}(e) - \mathbb{E}\Big(\widetilde{O}(e) \mid Z\Big)\Big\|_\infty > xn^2\Big) \le 2K^{n+2} e^{-x^2 n^2/(8\|P\|_\infty + 4x/3)}.$$

Proof. This Lemma is adapted from Lemma 1.1 in Bickel and Chen (2009). There are K^n possible values of e and $\|\cdot\|_\infty$ is the maximum of the K^2 entries in the matrix. We use the union bound to pull these maxima out of the probability, giving the factor K^{n+2} on the right. Next it suffices to bound the tail probability of each variable

$$\widetilde{O}_{ab}(e) - \mathbb{E}\Big(\widetilde{O}_{ab}(e) \mid Z\Big) = \sum_{i,j}\Big(A_{ij} - \mathbb{E}(A_{ij} \mid Z)\Big)(\mathbf{1}\{e_i = a, e_j = b\} + \mathbf{1}\{e_i = b, e_j = a\}).$$

The $n_{ab}(e)$ variables in this sum are conditionally independent given Z, take values in $[-2, 2]$, and have conditional mean zero given Z and conditional variance bounded by $4\,\mathrm{var}(A_{ij} \mid Z) \le 4P_{Z_i Z_j}(1 - P_{Z_i Z_j}) \le 4\|P\|_\infty$. Thus we can apply Bernstein's inequality to find that

$$\mathbb{P}\Big(\Big|\widetilde{O}_{ab}(e) - \mathbb{E}\Big(\widetilde{O}_{ab}(e) \mid Z\Big)\Big| > xn^2\Big) \le 2e^{-x^2 n^4/(8n_{ab}(e)\|P\|_\infty + 4xn^2/3)}.$$

Finally we use the crude bound $n_{ab}(e) \le n^2$ and cancel one factor n^2. □

Lemma 4.6. *Define $\widetilde{O}_{ab}(e) = O_{ab}(e)$ if $a \neq b$, and $\widetilde{O}_{ab}(e) = 2O_{ab}(e)$ if $a = b$. Then, for $R(e, Z)$ as defined in* (4.2)*,*

$$\mathbb{E}(\widetilde{O}_{ab} \mid Z) = n^2 R(e, Z) P R(e, Z)^T - n\mathrm{Diag}(R(e, Z)\,\mathrm{diag}\,(P)).$$

Proof. A similar expression, not taking into account the absence of self-loops, appears in Bickel and Chen (2009).

$$\begin{aligned}
\mathbb{E}(\widetilde{O}_{ab}(e) \mid Z = c) &= \sum_{i \neq j} P_{c_i c_j} \mathbf{1}\{e_i = a, e_j = b\} \\
&= \sum_{a', b'} P_{a'b'} \sum_{i \neq j} \mathbf{1}\{c_i = a', c_j = b'\}\mathbf{1}\{e_i = a, e_j = b\} \\
&= \sum_{a', b'} P_{a'b'} \sum_{i,j} \mathbf{1}\{c_i = a', c_j = b'\}\mathbf{1}\{e_i = a, e_j = b\} - \delta_{ab} \sum_{a'} P_{a'a'} \mathbf{1}\{c_i = a'\}\mathbf{1}\{e_i = a\} \\
&= n^2 \sum_{a', b'} P_{a'b'} R_{aa'}(e, c) R_{bb'}(e, c) - \delta_{ab} n \sum_{a'} P_{a'a'} R_{aa'}(e, c).
\end{aligned}$$

□

Lemma 4.7. *The function $\tau : [0,1] \to \mathbb{R}$ satisfies $|\tau(x) - \tau(y)| \le l(|x-y|)$, for $l(x) = 2x(1 \vee \log(1/x))$.*

Proof. Write the difference between $x \log x$ and $y \log y$ as $|\int_x^y (1 + \log s)\, ds|$. The function $s \mapsto 1 + \log s$ is strictly increasing on $[0,1]$ from $-\infty$ to 1 and changes sign at $s = e^{-1}$. Therefore the absolute integral is bounded above by the maximum of

$$-\int_0^{|x-y| \wedge e^{-1}} (1 + \log s)\, ds = -(|x-y| \wedge e^{-1}) \log |x-y| \wedge e^{-1}$$

and

$$\int_{1-|x-y| \vee e^{-1}}^1 (1 + \log s)\, ds \le |x-y|.$$

□

Proof of Lemma 4.2

Proof. The second assertion of the lemma follows from the first and the fact that $\max_e Q_P(e) \lesssim (\log n)/n$. It suffices to prove the first assertion.

Recall that the Bayesian modularity is given by

$$n^2 Q_B(e) = \sum_{a \le b} \log B\left(O_{ab}(e) + \tfrac{1}{2}, n_{ab}(e) - O_{ab}(e) + \tfrac{1}{2}\right) + \sum_a \log \Gamma(n_a(e) + \alpha). \tag{4.8}$$

We shall show that the first sum on the right is equivalent to $Q_{ML}(e)$, and the second sum is equivalent to $Q_P(e)$. We show this by comparing the sums defining the various modularities term by term. For clarity we shall suppress the argument e. We will repeatedly use the following bound from (Robbins, 1955): for $n \in \mathbb{N}_{\ge 1}$,

$$\Gamma(n+1) = \sqrt{2\pi} n^{n+1/2} e^{-n} e^{a_n}, \tag{4.9}$$

with $(12n+1)^{-1} \le a_n \le (12n)^{-1}$, as well as the fact that $\Gamma(s)$ is monotone increasing for $s \ge 3/2$. In addition, we will bound remainder terms by using the inequality $x \log((x+c)/x) \le c$ for $c \ge 0$ and the fact that $x \log((x-1)/x)$ is bounded for $x > 1$.

First sum of (4.8)**.**

Upper bound, case 1: $O_{ab} \ne 0$ and $n_{ab} \ne O_{ab}$

We apply (4.9):

$$\begin{aligned}
\log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) &\le \log \frac{\Gamma(O_{ab} + \lfloor \beta_1 \rfloor + 1)\Gamma(n_{ab} - O_{ab} + \lfloor \beta_2 \rfloor + 1)}{\Gamma(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor)} \\
&= O_{ab} \log\left(\frac{O_{ab} + \lfloor \beta_1 \rfloor}{n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor - 1}\right) + (n_{ab} - O_{ab}) \log\left(\frac{n_{ab} - O_{ab} + \lfloor \beta_2 \rfloor}{n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor - 1}\right) \\
&\quad + (\lfloor \beta_1 \rfloor + 1/2) \log(O_{ab} + \lfloor \beta_1 \rfloor) + (\lfloor \beta_2 \rfloor + 1/2) \log(n_{ab} - O_{ab} + \lfloor \beta_2 \rfloor) \\
&\quad - (\lfloor \beta_1 + \beta_2 \rfloor - 1/2) \log(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor - 1) + \log \sqrt{2\pi} - \lfloor \beta_1 \rfloor - \lfloor \beta_2 \rfloor + \lfloor \beta_1 + \beta_2 \rfloor - 1 \\
&\quad + \alpha_{ab} + \beta_{ab} - \gamma_{ab},
\end{aligned}$$

where α_{ab}, β_{ab} and γ_{ab} are bounded by constants. By the inequality $x \log((x+c)/x) \leq c$ for $c \geq 0$, and the fact that $x \log((x-1)/x)$ is bounded for $x > 1$, we find the upper bound:

$$\log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) \leq n_{ab} \tau \left(\frac{O_{ab}}{n_{ab}} \right) + O(\log n_{ab}).$$

Upper bound, case 2: $n_{ab} = 1$ *and* $O_{ab} = 0$ *or* $n_{ab} = O_{ab}$*, or* $n_{ab} = 0$
In both cases, the corresponding term of the likelihood modularity vanishes, whereas the contribution of the Bayesian modularity is either $\log B(1 + \beta_1, \beta_2)$, $\log(\beta_1, 1 + \beta_2)$, or $\log B(\beta_1, \beta_2)$.

Upper bound, case 3: $n_{ab} \geq 2$ *and* $O_{ab} = 0$ *or* $n_{ab} = O_{ab}$
Again, the corresponding term of the likelihood modularity vanishes. We show the computations for the case $n_{ab} = O_{ab}$; for the case $O_{ab} = 0$, switch β_1 and β_2. By (4.9):

$$\begin{aligned} \log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) &= \log B(n_{ab} + \beta_1, \beta_2) \leq \log \frac{\Gamma(n_{ab} + \lfloor \beta_1 \rfloor + 1)\Gamma(\beta_2)}{\Gamma(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor)} \\ &= (n_{ab} + \lfloor \beta_1 \rfloor) \log \left(\frac{n_{ab} + \lfloor \beta_1 \rfloor}{n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor} \right) + (1/2) \log(n_{ab} + \lfloor \beta_1 \rfloor) \\ &\quad - (\lfloor \beta_1 + \beta_2 \rfloor + 1/2) \log(n_{ab} + \lfloor \beta_1 + \beta_2 \rfloor) + \log \Gamma(\beta_2) + \lfloor \beta_1 + \beta_2 \rfloor - 1 + \delta_{ab} - \epsilon_{ab}, \end{aligned}$$

where δ_{ab} and ϵ_{ab} are bounded by constants. Arguing as before, the first term is bounded, while the remainder is of order $\log(n_{ab})$. A lower bound is found analogously.

Lower bound The computations for the lower bound are completely analogous, except that we require $O_{ab} + \beta_1 \geq 2$ and $n_{ab} - O_{ab} + \beta_2 \geq 2$. We study four cases. The cases (1) $O_{ab} \geq 2$ and $n_{ab} - O_{ab} \geq 2$, (2) $n_{ab} = 0$ and (3) $n_{ab} > 0$ and $n_{ab} = O_{ab}$ or $O_{ab} = 0$ are similar to cases 1, 2 and 3 respectively of the upper bound. The fourth case is $n_{ab} - O_{ab} = 1$ and $O_{ab} \geq 2$, or $O_{ab} = 1$ and $n_{ab} - O_{ab} \geq 1$. In both instances, the likelihood modularity is equality to a bounded term minus $\log n_{ab}$. By similar calculations as before, the Bayesian modularity is of the order $\log n_{ab}$ as well.

Conclusion We find:

$$\sum_{a \leq b} \log B(O_{ab} + \beta_1, n_{ab} - O_{ab} + \beta_2) = \sum_{a \leq b} n_{ab} \tau \left(\frac{O_{ab}}{n_{ab}} \right) + O(\log n).$$

Second sum of (4.8)**.**
We consider three cases. If $n_a + \lfloor \alpha \rfloor = 0$, then $\alpha > 0$, implies $n_a = 0$, in which case $\log \Gamma(n_a + \alpha) = \log \Gamma(\alpha)$, which is bounded. In case $n_a + \lfloor \alpha \rfloor = 1$, the term $\log \Gamma(n_a + \alpha)$ is equal to either $\log \Gamma(1 + \alpha)$ or $\log \Gamma(\alpha)$ and thus bounded as well. For the case $n_a + \lfloor \alpha \rfloor \geq 2$, we study the upper bound $\Gamma(n_a + \alpha) \leq \Gamma(n_a + \lfloor \alpha \rfloor + 1)$ and the lower bound $\Gamma(n_a + \alpha) \geq \Gamma(n_a + \lfloor \alpha \rfloor)$. By applying (4.9) in both cases, we conclude:

$$\sum_a \log \Gamma(n_a + \alpha) = \sum_{a : n_a + \lfloor \alpha \rfloor \geq 2} n_a \log n_a - n + O(\log n).$$

□

Lemma 4.8. *For any probability matrix R,*

$$H_P(R) \leq H_P(\mathrm{Diag}(R^T \mathbf{1})\Big). \tag{4.10}$$

Furthermore, if (P, π) is identifiable and the columns of R corresponding to positive coordinates of π are not identically zero, then the inequality is strict unless $P_\sigma R$ is a diagonal matrix for some permutation matrix P_σ.

Proof. This Lemma is related to the proof that the likelihood modularity is consistent given in Bickel and Chen (2009). This proof however rests on their incorrect Lemma 3.1, and thus we provide full details on how the argument can be adapted to avoid the use of their Lemma 3.1 altogether.

For R a diagonal matrix the numbers $(RPR^T)_{ab}/(R\mathbf{1})_a(R\mathbf{1})_b$ reduce to P_{ab}. Consequently, by the definition of H_P,

$$H_P\Big(\mathrm{Diag}(f)\Big) = \sum_{a,b} f_a f_b \, \tau(P_{ab}). \tag{4.11}$$

For a general matrix R, by inserting the definition of τ,

$$\begin{aligned} H_P(R) = &\sum_{a,b} (RPR^T)_{ab} \log \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b} \\ &+ \sum_{a,b} \Big((R\mathbf{1})_a (R\mathbf{1})_b - (RPR^T)_{ab} \Big) \log\Big(1 - \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b}\Big). \end{aligned}$$

Because $(R\mathbf{1})_a(R\mathbf{1})_b - (RPR^T)_{ab} = (R(1-P)R^T)_{ab}$, with 1 the $(K \times K)$-matrix with all coordinates equal to 1, we can rewrite this as

$$\sum_{a,b} \sum_{a',b'} R_{aa'} R_{bb'} \left[P_{a'b'} \log \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b} + (1 - P_{a'b'}) \log\Big(1 - \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a (R\mathbf{1})_b}\Big) \right].$$

By the information inequality for two-point measures, the expressions in square brackets becomes bigger when $(RPR^T)_{ab}/(R\mathbf{1})_a(R\mathbf{1})_b$ is replaced by $P_{a'b'}$, with a strict increase unless these two numbers are equal. After making this substitution the terms in square brackets becomes $\tau(P_{a'b'})$, and we can exchange the order of the two (double) sums and perform the sum on (a, b) to write the resulting expression as

$$\sum_{a',b'} (R^T\mathbf{1})_{a'} (R^T\mathbf{1})_{b'} \tau(P_{a'b'}) = H_P\Big(\mathrm{Diag}(R^T\mathbf{1})\Big).$$

This proves the first assertion (4.10) of the lemma.

If R attains equality, then also for every permutation matrix P_σ, by the equality $H_P(P_\sigma R) = H_P(R)$ and the fact that $(P_\sigma R)^T\mathbf{1} = R^T\mathbf{1}$, we have

$$H_P(P_\sigma R) = H_P\Big(\mathrm{Diag}((P_\sigma R)^T\mathbf{1})\Big). \tag{4.12}$$

We shall show that if R satisfies this equality and $P_\sigma R$ has a positive diagonal, then $P_\sigma R$ is in fact diagonal. Furthermore, we shall show that there exists P_σ such that $P_\sigma R$ has a positive diagonal.

Fix some $(P_\sigma)_m$ that maximizes the number of positive diagonal elements of $P_\sigma R$ over all permutation matrices P_σ, and denote $\bar{R} = (P_\sigma)_m R$. Because the information inequality is strict, the preceding argument shows that (4.12) can be true for $P_\sigma = (P_\sigma)_m$ (giving $P_\sigma R = \bar{R}$) only if

$$P_{a'b'} = \frac{(\bar{R}P\bar{R}^T)_{ab}}{(\bar{R}\mathbf{1})_a(\bar{R}\mathbf{1})_b}, \qquad \text{whenever } \bar{R}_{aa'}\bar{R}_{bb'} > 0. \tag{4.13}$$

Denote the matrix on the right of the equality by Q.

If $\bar{R}$ has a completely positive diagonal, then we can choose $a = a'$ and $b = b'$ and find from equation (4.13), that $P_{ab} = Q_{ab}$, for every a, b. If also $\bar{R}_{aa'} > 0$, then we can also choose $b = b'$ and find that $P_{a'b} = Q_{ab}$, for every b. Thus the ath and a'th rows of P are identical. Since all rows of P are different by assumption, it follows that no $a \neq a'$ with $\bar{R}_{aa'} > 0$ exists.

If $\bar{R}$ does not have a fully positive diagonal, then the submatrix of $\bar{R}$ obtained by deleting the rows and columns corresponding to positive diagonal elements must be the zero matrix, since otherwise we might permute the remaining rows and create an additional nonzero diagonal element, contradicting that $(P_\sigma)_m$ already maximized this number. If I and I^c are the sets of indices of zero and nonzero diagonal elements, then the preceding observation is that $\bar{R}_{ij}$ is zero for every $i, j \in I$. If $\pi > 0$, then we need to consider only R with nonzero columns. For $i \in I$ a nonzero element in the ith column of $\bar{R}$ must be located in the rows with label in I^c: for every $i \in I$ there exists $k_i \in I^c$ with $\bar{R}_{k_i i} > 0$. Then, for $i, j \in I$,

(1) for $a = k_i$, $b = k_j$, $a' = i$, $b' = j$, equation (4.13) implies $Q_{k_i k_j} = P_{ij}$.

(2) for $a = k_i$, $b \in I^c$, $a' = i$, $b' = b$, equation (4.13) implies $Q_{k_i b} = P_{ib}$.

(3) for $a = k_i$, $b \in I^c$, $a' = k_i$, $b' = b$, equation (4.13) implies $Q_{k_i b} = P_{k_i b}$.

We combine these three assertions to conclude that, for $a, i \in I$ and $b \in I^c$,

$$\begin{aligned} P_{ai} = P_{ia} &\overset{(1)}{=} Q_{k_i k_a} \overset{(2)}{=} P_{ik_a} = P_{k_a i}, \\ P_{ab} &\overset{(2)}{=} Q_{k_a b} \overset{(3)}{=} P_{k_a b}. \end{aligned}$$

Together these imply that the ath and the k_ath row of P are equal. Since by assumption they are not (if $\pi > 0$), this case can actually not exist (i.e. $k = 0$).

Finally if $\pi_a = 0$ for some a, then we follow the same argument, but we match only every column $i \in I$ with $\pi_i > 0$ to a row $k_i \in I^c$. By the assumption on R such k_i exist, and the construction results in two rows of P that are identical in the coordinates with $\pi_a > 0$. □

Lemma 4.9. *For any fixed $(K \times K)$-matrix P with elements in $[0,1]$, uniformly in probability matrices R, as $\rho_n \to 0$,*

$$\frac{1}{\rho_n}\Big(H_{\rho_n P}(\mathrm{Diag}(R^T \mathbf{1})) - H_{\rho_n P}(R)\Big) \to G_P(\mathrm{Diag}(R^T \mathbf{1})) - G_P(R). \tag{4.14}$$

Furthermore, if (P,π) is identifiable and the columns of R corresponding to positive coordinates of π are not identically zero, then the right side is strictly positive unless SR is a diagonal matrix for some permutation matrix S.

Proof. From the fact that $|(1-u)\log(1-u)+u| \le u^2$, for $0 \le u \le 1$, it can be verified that, $\left|\rho_n^{-1}\tau(\rho_n u) - \Big(u\log\rho_n + \tau_0(u)\Big)\right| \le \rho_n \to 0$, uniformly in $0 \le u \le 1$. It follows that, uniformly in R,

$$\frac{1}{\rho_n}H_{\rho_n P}(R) = \log\rho_n \sum_{a,b}(RPR^T)_{ab} + \sum_{a,b}(R\mathbf{1})_a(R\mathbf{1})_b\tau_0\Big(\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\Big) + O(\rho_n).$$

The first term on the right is equal to $\log\rho_n(R^T\mathbf{1})^T P(R^T\mathbf{1})$, and hence is the same for R and $\mathrm{Diag}(R^T\mathbf{1})$. Thus this term cancels on taking the difference to form the left side of (4.14), and hence (4.14) follows.

The right side of (4.14) is nonnegative, because the left side is, by Lemma 4.8. This fact can also be proved directly along the lines of the proof of Lemma 4.8, as follows. Write

$$G_P(R) = \sum_{a,b}\sum_{a',b'} R_{aa'}R_{bb'}\Big[P_{a'b'}\log\frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b} - \frac{(RPR^T)_{ab}}{(R\mathbf{1})_a(R\mathbf{1})_b}\Big].$$

By the information inequality for two Poisson distributions the term in square brackets becomes bigger if $(RPR^T)_{ab}/(R\mathbf{1})_a(R\mathbf{1})_b$ is replaced by $P_{a'b'}$. It then becomes $\tau_0(P_{a'b'})$ and the double sum on (a,b) can be executed to see that the resulting bound is $G_P\Big(\mathrm{Diag}(R^T\mathbf{1})\Big)$. Furthermore, the inequality is strictly unless (4.13) holds, with $\bar{R} = R$. Since also $G_P(P_\sigma R) = G_P(R)$, for every permutation matrix P_σ, the final assertion of the lemma is proved by copying the proof of Lemma 4.8. □

4.6.2 Strong consistency

We need slightly adapted versions of the function H_P, given by, with δ_{ab} equal to 1 or 0 if $a = b$ or not,

$$H_{P,n}(R) = \frac{1}{2}\sum_{a,b}(R\mathbf{1})_a\Big((R\mathbf{1})_b - \delta_{ab}/n\Big)\,\tau\Big(\frac{(RPR^T)_{ab} - \delta_{ab}\sum_k P_{kk}R_{ka}/n}{(R\mathbf{1})_a\Big((R\mathbf{1})_b - \delta_{ab}/n\Big)}\Big). \tag{4.15}$$

For given functions $t_{ab} : [0,1] \to \mathbb{R}$, let $X(e)$ be the $K \times K$ matrix with entries

$$X_{ab}(e) = t_{ab}\Big(\frac{\widetilde{O}_{ab}(e)}{n^2}\Big) - t_{ab}\Big(\frac{\mathbb{E}(\widetilde{O}_{ab}(e) \mid Z)}{n^2}\Big). \tag{4.16}$$

Proof of Theorem 4.3 [strong consistency]

Proof. i. By Theorem 4.4, $\widehat{e}$ is weakly consistent, and hence with probability tending to one it belongs to the set of classifications e such that the fractions $f(e)$ are close to π, and the matrices $R(e,Z)$ are close to $\mathrm{Diag}(\pi)$ after the appropriate permutation of the labels (that is, of rows of $R(e,Z)$). Therefore, it is no loss of generality to assume that $\widehat{e}$ is restricted to this set. By Lemmas 4.5 and 4.6, the matrices $\widetilde{O}(e)/n^2$ are then close to

$R(e,Z)PR(e,Z)^T \to \mathrm{Diag}(\pi)P\mathrm{Diag}(\pi)$, and hence are bounded away from zero and one if P has this property.

If $\widehat{e}$ and Z differ at m nodes, then $\widehat{e}$ belongs to the set of e with $\|R(Z,Z) - R(e,Z)\|_1 = m(2/n)$, by Lemma 4.1. In that case $Q_B(e) \geq Q_B(Z)$, for some e in this set, and hence by Lemma 4.2 $Q_{ML}(e) - Q_{ML}(Z) + Q_P(e) - Q_P(Z) \geq -\eta_n$, for some η_n of order $(\log n)/n^2$. It follows that:

$$\begin{aligned}\Big[Q_{ML}(e) - H_{P,n}\Big(R(e,Z)\Big)\Big] - \Big[Q_{ML}(Z) - H_{P,n}\Big(R(Z,Z)\Big)\Big] \\ \geq H_{P,n}\Big(R(Z,Z)\Big) - H_{P,n}\Big(R(e,Z)\Big) - |Q_P(e) - Q_P(Z)| - \eta_n. \end{aligned} \tag{4.17}$$

The first term on the right is bounded below by a multiple of m/n, by Lemmas 4.10 and 4.1. Because $(x+\alpha)\log x - (y+\alpha)\log y = \int_x^y (\log s + (s+\alpha)/s)\,ds$ is bounded in absolute value by a multiple of $|x-y|\log(x \vee y)$, if $\alpha \geq 0$ and $x, y > 0$, the second term $-|Q_P(e) - Q_P(Z)|$ is bounded below by a multiple of $m(\log n)/n^2$, for some positive constant C_2, which is of smaller order than m/n. We conclude that the left side of (4.17) is bounded below by $C_1 m/n$. The left side is $\sum_{a,b}\Big(X_{ab}(e) - X_{ab}(Z)\Big)$, for X defined in (4.16) and t the function with coordinates $t_{ab}(o) = f_a(e)\Big(f_b(e) - \delta_{ab}/n\Big)\tau\Big(o/f_a(e)\Big(f_b(e) - \delta_{ab}/n\Big)\Big)$. Because we restrict e to classifications such that $O_{ab}(e)/n_{ab}(e)$ and $f_a(e)f_b(e)$ are bounded away from zero and one, only the values of the function τ on an open interval strictly within $(0,1)$ matter. On any such interval τ has uniformly bounded derivatives, and hence the bound of Lemma 4.13 is valid. Thus we find that

$$\begin{aligned}\Pr\Big(\#(i : \widehat{e}_i \neq Z_i) = m\Big) &\leq \Pr\Big(\sup_{e:\#(i:e_i \neq Z_i)\leq m} \Big\|X(e) - X(Z)\Big\|_\infty \geq \frac{C_1 m}{n}\Big) \\ &\lesssim K^m \binom{n}{m} e^{-cm^2/(m\|P\|_\infty/n + m/n)} \\ &\leq e^{m\log(Kne/m) - c_1 mn}. \end{aligned}$$

The sum of the right side over $m = 1, \ldots, n$ tends to zero.

ii. We follow the proof for i, but in (4.17) use that $H_{P,n}\Big(R(Z,Z)\Big) - H_{P,n}\Big(R(e,Z)\Big) \geq \rho_n C\|R(Z,Z) - R(e,Z)\|_1 \geq \rho_n C 2m/n$, by Lemma 4.12. Since $\rho_n \gg (\log n)/n$ by assumption, we have that the contribution $m(\log n)/n^2$ of $Q_P(e) - Q_P(Z)$ is still negligible and hence $\rho_n C 2m/n$ is a lower bound for the left side of (4.17). As a bound on the left side of the preceding display, we then obtain

$$\sum_{m=1}^n K^m \binom{n}{m} e^{-c_2\rho_n^2 m^2/(m\rho_n/n + \rho_n m/n)} \leq \sum_{m=1}^n e^{m\log(Kne/m) - c_3\rho_n mn}.$$

This sum tends to zero provided that $n\rho_n \gg \log n$. □

Lemma 4.10. *If P is fixed and symmetric and every pair of rows of P is different and $0 < P < 1$ and $\pi > 0$, then, for sufficiently small $\delta > 0$,*

$$\liminf_{n\to\infty} \inf_{0 < \|R - \mathrm{Diag}(\pi)\| < \delta} \frac{H_{P,n}\Big(\mathrm{Diag}(R^T \mathbf{1})\Big) - H_{P,n}(R)}{\|\mathrm{Diag}(R^T \mathbf{1}) - R\|} > 0. \tag{4.18}$$

Proof. We can reparametrize the $K \times K$ matrices R by the pairs $(R^T\mathbf{1}, R - \mathrm{Diag}(R^T\mathbf{1}))$, consisting of the K vector $f = R^T\mathbf{1}$ and the $K \times K$ matrix $R - \mathrm{Diag}(R^T\mathbf{1})$. The latter matrix is characterized by having nonnegative off-diagonal elements and zero column sums, and can be represented in the basis consisting of all $K \times K$ matrices $\Delta_{bb'}$, for $b \neq b'$, defined by: $(\Delta_{bb'})_{b'b'} = -1$, $(\Delta_{bb'})_{bb'} = 1$ and $(\Delta_{bb'})_{aa'} = 0$, for all other entries (a, a'), i.e. the b'th column of $\Delta_{bb'}$ has a 1 in the bth coordinate and a -1 on the b'th coordinate and all its other columns are zero. Given any matrix $R \geq 0$ the matrix $R - \mathrm{Diag}(R^T\mathbf{1})$ can be decomposed as

$$R - \mathrm{Diag}(R^T\mathbf{1}) = \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'},$$

for $\lambda_{bb'} = R_{bb'} \geq 0$. Since every $\Delta_{bb'}$ has exactly one nonzero off-diagonal element, which is equal to 1, and in a different location for each $b \neq b$, the sum of the off-diagonal elements of the matrix on the right side is $\sum_{b,b'} \lambda_{bb'}$. Because the sum of all its elements is zero, it follows that its sum of absolute elements is given by $\|R - \mathrm{Diag}(R^T\mathbf{1})\|_1 = 2\sum_{b \neq b'} \lambda_{bb'}$.

Thus we obtain a further reparametrization $R \leftrightarrow (f, \lambda)$, in which $R = \mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}$. For given P, f and n, define the function

$$G(\lambda) = H_{P,n}\Big(\mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}\Big).$$

Then we would like to show that there exists C such that

$$\frac{H_{P,n}(\mathrm{Diag}(R^T\mathbf{1})) - H_{P,n}(R)}{\|R - \mathrm{Diag}(R^T\mathbf{1})\|_1} = \frac{G(0) - G(\lambda)}{2\sum_{b \neq b'} \lambda_{bb'}} \geq C > 0,$$

for every f in a neighbourhood of π, λ in a neighbourhood of 0 intersected with $\{\lambda : \lambda \geq 0\}$, and every sufficiently large n. The numerator in the quotient is $f(0) - f(1)$ for the function $f(s) = G(s\lambda)$. Writing this difference in the form $-f'(0) - \int_0^1 \big(f'(s) - f'(0)\big)\, ds$ gives that the numerator is equal to

$$-\nabla G(0)^T \lambda - \int_0^1 \Big(\nabla G(s\lambda) - \nabla G(0)\Big)\Big)^T\, ds\, \lambda. \tag{4.19}$$

It suffices to show that the first term is bounded below by a multiple of $\|\lambda\|_1$ and that the second is negligible relative to the first, as $n \to \infty$, uniformly in f in a neighbourhood of π and λ in a neighbourhood of 0 intersected with $\{\lambda : \lambda \geq 0\}$. Thus it is sufficient to show first that for every coordinate $\lambda_{bb'}$ of λ minus the partial derivative of G at $\lambda = 0$ with respect to $\lambda_{bb'}$ is bounded away from 0, as $n \to \infty$ uniformly in f, and second that every partial derivative is equicontinuous at $\lambda = 0$ uniformly in f and large n.

We have

$$G(\lambda) = \frac{1}{2} \sum_{a,a'} f_a(\lambda)\Big(f_{a'}(\lambda) - \delta_{aa'}/n\Big)\, \tau\Big(\frac{\Big(R(\lambda)PR(\lambda)^T\Big)_{aa'} - \delta_{aa'} e_a(\lambda)/n}{f_a(\lambda)\Big(f_{a'}(\lambda) - \delta_{aa'}/n\Big)}\Big), \tag{4.20}$$

for

$$f(\lambda) = f + \sum_{bb'} \lambda_{bb'}(\Delta_{bb'}\mathbf{1}),$$

$$R(\lambda) = \mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'},$$
$$e_a(\lambda) = \sum_k P_{kk} R_{ak}(\lambda) = P_{aa} f_a + \sum_{b \neq b'} P_{b'b'} \lambda_{bb'} (\delta_{ab} - \delta_{ab'}).$$

By a lengthy calculation, given in Lemma 4.11,

$$\frac{\partial}{\partial \lambda_{bb'}} G(\lambda)_{|\lambda=0} = -\sum_a f_a K(P_{ab'} \| P_{ab}) + \frac{1}{2n} K(P_{b'b'} \| P_{bb}), \tag{4.21}$$

for $K(p\|q) = p \log(p/q) + (1-p)\log\Big((1-p)/(1-q)\Big)$ the Kullback-Leibler divergence between the Bernoulli distributions with success probabilities p and q. The numbers f_a are bounded away from zero for f sufficiently close to π, and hence so is $\sum_a f_a K(P_{ab'} \| P_{ab})$, unless the bth and b'th column of P are identical. The whole expression is bounded below by the minimum over (b, b') of these numbers minus $(2n)^{-1}$ times the maximum of the numbers $K(P_{b'b'} \| P_{bb})$, and hence is positive and bounded away from zero for sufficiently large n.

To verify the equicontinuity of the partial derivatives we can compute these explicitly at λ and take their limit as $n \to \infty$. We omit the details of this calculation. However, we note that every term of $G(\lambda)$ is a fixed function of the quadratic forms in λ

$$\Big(f_a + \sum_{bb'} \lambda_{bb'} (\Delta_{bb'} \mathbf{1})_a\Big)\Big(f_{a'} + \sum_{bb'} \lambda_{bb'} (\Delta_{bb'} \mathbf{1})_{a'} - \delta_{aa'}/n\Big), \tag{4.22}$$
$$\Big(\Big(\mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}\Big) P \Big(\mathrm{Diag}(f) + \sum_{b \neq b'} \lambda_{bb'} \Delta_{bb'}^T\Big)\Big)_{aa'}$$
$$- \frac{\delta_{aa'}}{2n}\Big(P_{aa} f_a + \sum_{b \neq b'} P_{b'b'} \lambda_{bb'} (\delta_{ab} - \delta_{ab'})\Big). \tag{4.23}$$

These forms are obviously smooth in λ, and their dependence and that of their derivatives on n is seen to vanish as $n \to \infty$. For f and λ restricted to neighbourhoods of π and 0, the values of the quadratic forms are restricted to a domain in which the transformation mapping them into $G(\lambda)$ is continuously differentiable. Thus the desired equicontinuity follows by the chain rule. □

Lemma 4.11. *The partial derivatives of the function G at 0 defined by* (4.20) *are given by* (4.21).

Proof. For given differentiable functions u and v the map $\epsilon \mapsto u(\epsilon)\tau\Big(v(\epsilon)/u(\epsilon)\Big)$ has derivative $v' \log\Big(v/(u-v)\Big) - u' \log\Big(u/(u-v)\Big)$. We apply this for every given pair (a, a') to the functions u and v obtained by taking $\lambda_{bb'}$ in (4.22) and (4.23) equal to ϵ and all other coordinates of λ equal to zero. Then

$$u(0) = f_a(f_{a'} - \delta_{aa'}/n),$$
$$v(0) = f_a(f_{a'} - \delta_{aa'}/n) P_{aa'},$$
$$u'(0) = (\Delta_{bb'} \mathbf{1})_a (f_{a'} - \delta_{aa'}/n) + f_a (\Delta_{bb'} \mathbf{1})_{a'}$$

$$v'(0) = (\Delta_{bb'}P)_{aa'}f_{a'} + f_a(\Delta_{bb'}P)_{a'a} - (\delta_{aa'}/n)P_{b'b'}(\delta_{ab} - \delta_{ab'}).$$

It follows that $v(0)/(u(0) - v(0)) = P_{aa'}/(1 - P_{aa'})$, and $u(0)/(u(0) - v(0)) = 1/(1 - P_{aa'})$. Hence in view of (4.15) the partial derivative in (4.21) is equal to

$$\sum_{a \neq a'} \left[v'(0) \log \frac{P_{aa'}}{1 - P_{aa'}} - u'(0) \log \frac{1}{1 - P_{aa'}} \right].$$

We combine this with the equalities

$$(\Delta_{bb'}\mathbf{1})_a = \begin{cases} 0 & \text{if } a \notin \{b, b'\}, \\ -1 & \text{if } a = b', \\ 1 & \text{if } a = b, \end{cases} \qquad (\Delta_{bb'}P)_{aa'} = \begin{cases} 0 & \text{if } a \notin \{b, b'\}, \\ -P_{b'a'} & \text{if } a = b', \\ P_{b'a'} & \text{if } a = b. \end{cases}$$

□

Lemma 4.12. *If S is fixed and symmetric, every pair of rows of S is different and $S > 0$ and $\pi > 0$ coordinatewise, then there exists $C > 0$ such that, for sufficiently small $\delta > 0$ and any $\rho_n \downarrow 0$,*

$$\liminf_{n \to \infty} \inf_{0 < \|R - \mathrm{Diag}(\pi)\| < \delta} \frac{H_{\rho_n S, n}\left(\mathrm{Diag}(R^T \mathbf{1})\right) - H_{\rho_n S, n}(R)}{\rho_n \|\mathrm{Diag}(R^T \mathbf{1}) - R\|} \geq C.$$

Proof. In the notation of the proof of Lemma 4.10 we must now show that $G(0) - G(\lambda) \geq C\rho_n \|\lambda\|_1$, as $n \to \infty$, uniformly in f in a neighbourhood of π, and λ in a positive neighbourhood of 0. As in that proof we write $G(0) - G(\lambda)$ in the form (4.19) and see that it suffices that the partial derivatives of G at 0 divided by ρ_n tend to negative limits, and that $\left\| \nabla G(\lambda) - \nabla G(0) \right\| / \rho_n$ becomes uniformly small as λ is close enough to zero.

The partial derivative at 0 with respect to $\lambda_{bb'}$ is given in (4.21), where we must replace P by $\rho_n S$. Since the scaled Kullback-Leibler divergence $\rho_n^{-1} K(\rho_n s \| \rho_n t)$ of two Bernoulli laws converges to the Kullback-Leibler divergence $K_0(s\|t) = s \log(s/t) + t - s$ between two Poisson laws of means s and t, as $\rho_n \to 0$, it follows that for $\rho_n \to 0$, uniformly in f,

$$\frac{1}{\rho_n} \frac{\partial}{\partial \lambda_{bb'}} G(\lambda)_{|\lambda = 0} \to - \sum_a f_a K_0(S_{ab'} \| S_{ab}).$$

The right side is strictly negative by the assumption that every pair of rows of S differ in at least one coordinate.

If $P = \rho_n S$, then the function $\lambda \mapsto v(\lambda)$ given in (4.23) takes the form $v = \rho_n v_S$, for v_S defined in the same way but with S replacing P. The function u given in (4.22) does not depend on P or S. Using again that the derivative of the map $\epsilon \mapsto u(\epsilon)\tau\left(v(\epsilon)/u(\epsilon)\right)$ is given by $v' \log\left(v/(u - v)\right) - u' \log\left(u/(u - v)\right)$, we see that the partial derivative with respect to $\lambda_{bb'}$ of the (a, a') term in the sum defining G takes the form

$$\begin{aligned} &\rho_n v_S' \log \frac{\rho_n v_S}{u - \rho v_S} - u' \log \frac{u}{u - \rho_n v_S} \\ &\quad = \rho_n v_S' \log \rho_n - \rho_n v_S' \log(v_S/u) - (\rho_n v_S' - u') \log(1 - \rho_n v_S/u). \end{aligned}$$

Here u and V_S are as in (4.22) and (4.23) (with P replaced by S), and depend on (a, a'). From the fact that the column sums of the matrices $R(\lambda)$ do not depend on λ, we have that

$$\sum_{a,a'}\Big[(R(\lambda)SR(\lambda)^T)_{aa'} - \frac{\delta_{aa'}}{n}\sum_k P_{kk}R(\lambda)_{ak}\Big] = R(\lambda)^T\mathbf{1}SR(\lambda)^T\mathbf{1} - \sum_k P_{kk}\sum_a R(\lambda)_{ak}$$

is constant in λ. This shows that $\sum_{a,a'} v'_S = 0$ and hence the contribution of the term $\rho_n v'_S \log\rho_n$ to the partial derivatives of G vanishes. The term $-(\rho_n v'_S - u')\log(1-\rho_n v_S/u)$ can be expanded as $(\rho_n v'_S - u')\rho_n v_S/u$ up to $O(\rho_n^2)$, uniformly in f and λ. Since these are equicontinuous functions of λ, it follows that $\rho_n^{-1}\big(\nabla G(\lambda) - \nabla G(0)\big)$ becomes arbitrarily small if λ varies in a sufficiently small neighbourhood of 0. □

Lemma 4.13. *There exists a constant $c > 0$ such that for $X(e)$ as in* (4.16)*, for every twice differentiable functions $t_{a,b} : [0,1] \to \mathbb{R}$ with $\|t'_{a,b}\|_\infty \vee \|t''_{a,b}\|_\infty \le 1$, and every $x > 0$,*

$$\Pr\Big(\max_{e:\#(e_i\neq Z_i)\le m}\big\|X(e) - X(Z)\big\|_\infty > x\Big) \qquad \le 6\binom{n}{m}K^{m+2}e^{-\frac{cx^2n^2}{m\|P\|_\infty/n+x}}.$$

Proof. Given Z there are at most $\binom{n}{m}$ groups of m candidate nodes that can be assigned to have $e_i \neq Z_i$, and the label of each node can be chosen in at most $K-1$ ways. Thus conditioning the probability on Z, we can use the union bound to pull out the maximum over e, giving a sum of fewer than $\binom{n}{m}K^m$ terms. Next we pull out the norm giving another factor K^2. It suffices to combine this with a tail bound for a single variable $X_{a,b}(e) - X_{a,b}(Z)$. Write t for $t_{a,b}$.

Assume for simplicity of notation that $e_i = Z_i$, for $i > m$, and decompose

$$\begin{aligned}\frac{1}{n^2}O_{ab}(e) &= \frac{1}{n^2}\Big[\sum_{i\le m \text{ or } j\le m} A_{ij}1_{e_i=a,e_j=b} + \sum_{i>m \text{ and } j>m} A_{ij}1_{e_i=a,e_j=b}\Big]\\ &=: S_1 + S_2.\end{aligned}$$

Let $O_{ab}(Z)/n^2 =: S'_1 + S_2$, with the same variable S_2, be the corresponding decomposition if e is changed to Z, and then decompose, where the expectation signs $\mathbb{E}$ denote conditional expectations given Z,

$$\begin{aligned}&X_{ab}(e) - X_{ab}(Z)\\ &\quad= \Big(t(S_1+S_2) - t(\mathbb{E}S_1 + \mathbb{E}S_2)\Big) - \Big(t(S'_1+S_2) - t(\mathbb{E}S'_1 + \mathbb{E}S_2)\Big)\\ &\quad= t(S_1+S_2) - t(\mathbb{E}S_1 + S_2)\\ &\qquad+ \Big(t(\mathbb{E}S_1+S_2) - t(\mathbb{E}S_1 + \mathbb{E}S_2)\Big) - \Big(t(\mathbb{E}S'_1+S_2) - t(\mathbb{E}S'_1 + \mathbb{E}S_2)\Big)\\ &\qquad+ t(\mathbb{E}S'_1+S_2) - t(S'_1+S_2)\end{aligned}$$

The first and third terms on the far right can be bounded above in absolute value by $\|t'\|_\infty$ times the increment. To estimate the second term we write it as

$$(S_2 - \mathbb{E}S_2)(\mathbb{E}S_1 - \mathbb{E}S'_1)\int_0^1\int_0^1 t''\Big(uS_2 + (1-u)\mathbb{E}S_2 + v\mathbb{E}S_1 + (1-v)\mathbb{E}S'_1\Big)\,du\,dv.$$

Since the first and second derivatives of t are uniformly bounded by 1, it follows that

$$\left|X_{ab}(e) - X_{ab}(Z)\right| \leq |S_1 - \mathbb{E}S_1| + |S_2 - \mathbb{E}S_2|\,|\mathbb{E}S_1 - \mathbb{E}S_1'| + |S_1' - \mathbb{E}S_1'|.$$

The variable $S_1 - \mathbb{E}S_1$ is a sum of fewer than $2mn$ independent variables, each with conditional mean zero, bounded above by $1/n^2$ and of variance bounded above by $\|P\|_\infty/n^4$. Therefore Bernstein's inequality gives that

$$\mathbb{P}\Big(|S_1 - \mathbb{E}S_1| > x\Big) \leq e^{-\frac{1}{2}x^2/(2mn\|P\|_\infty/n^4 + x/(3n^2))}.$$

This is as the exponential factor in the bound given by the lemma, for appropriate c. The variable $S_1' - \mathbb{E}S_1'$ can be bounded similarly. Furthermore $|\mathbb{E}S_1 - \mathbb{E}S_1'| \leq 4mn/n^2 = 4m/n$, and $S_2 - \mathbb{E}S_2$ is the sum of fewer than n^2 variables as before, so that

$$\mathbb{P}\Big(|S_2 - \mathbb{E}S_2|\,|\mathbb{E}S_1 - \mathbb{E}S_1'| > x\Big) \leq e^{-\frac{1}{2}(xn/(4m))^2/(n^2\|P\|_\infty/n^4 + xn/(12mn^2))}.$$

The exponent has a similar form as before, except for an additional factor $n/m \geq 1$. □