Recovery, detection and confidence sets of communities in a sparse stochastic block model

B. J. K. Kleijn[1] and J. van Waaij[2]

[1] *Korteweg-de Vries Institute for Mathematics, University of Amsterdam*

[2] *Department of Mathematics, Humboldt-Universität zu Berlin*

October 2018

Abstract

Posterior distributions for community assignment in the planted bi-section model are shown to achieve frequentist exact recovery and detection under sharp lower bounds on sparsity. Assuming posterior recovery (or detection), one may interpret credible sets (or enlarged credible sets) as consistent confidence sets. If credible levels grow to one quickly enough, credible sets can be interpreted as frequentist confidence sets without conditions on the parameters. In the regime where within-class and between-class edge-probabilities are very close, credible sets may be enlarged to achieve frequentist asymptotic coverage. The 'diameters' of credible sets are controlled and match rates of posterior convergence.

1 Communities in random graphs

The stochastic block model is a generalization of the Erdös-Rényi random graph model (Erdös and Rényi, 1959) where one studies a version X^n of the complete graph between n vertices under percolation, with edge probability $p_n \in [0, 1]$. Stochastic block models (Holland, Laskey, and Leinhardt, 1983) are similar but concern random graphs with vertices that belong to one of several classes and edge probabilities that depend on those classes. If we think of the graph X^n as data and the class assignments of the vertices as unobserved, an interesting statistical challenge presents itself regarding estimation of (and other forms of inference on) the vertices' class assignments, a task referred to as *community detection* (Girvan and Newman, 2002). The stochastic block model and its generalizations have applications in physics, biology, sociology, image processing, genetics, medicine, logistics, *etcetera* and are widely employed as a canonical model to study clustering and community detection (Fortunato, 2010; Abbe, 2018); (Zhang and Zhou, 2016) even state that, *"Community detection for the stochastic block model is probably the most studied topic in network analysis."*

In an asymptotic sense, one may wonder under which conditions on edge probabilities community detection can be done in a 'statistically consistent' way as the number of vertices n grows, that is, whether it is possible to estimate the true class assignments correctly (or correctly for a fraction of the vertices that goes to one) with high probability. Note that already in the Erdös-Rényi model, asymptotic behaviour is very rich: consider, for example, *connectedness* of the Erdös-Rényi graphs X^n with a sequence of edge probabilities (p_n) that becomes *sparse*: $p_n \to 0$. A sharp transition exists: $p_n \geq (A/n)\log n$ leads to a connected graph with high probability, if and only if $A > 1$. In even more sparse circumstances, there is a single *giant component* (a connected component of size $O(n)$) with high probability, if and only if $p_n \geq C/n$ with $C > 1$. Below the $1/n$-threshold, the graph X^n fragments into many disconnected sub-graphs of order no larger than $O(\log(n))$ with high probability. At the boundaries $1/n$ and $\log(n)/n$, the Erdös-Rényi graph is said to undergo *phase transitions* (Bollobás, Janson, and Riordan, 2007), from the *fragmented phase* to the sparse *Kesten-Stigum phase*, and then to the less sparse *Chernoff-Hellinger phase*.

Here and in (Abbe, Bandeira, and Hall, 2016; Massoulié, 2014; Mossel, Neeman, and Sly, 2016b), the community detection problem is studied in the simplest context, that of the so-called *planted bi-section model*, which is a stochastic block model with two classes, each of n vertices and edge probabilities p_n (within-class) and q_n (between-class). A famous sufficient condition for so-called *exact recovery* of the class assignments in the planted bi-section model comes from (Dyer and Frieze, 1989): if there exists a constant $A > 0$ such that,

$$p_n - q_n \geq A\frac{\log n}{n}, \tag{1}$$

then community detection by simple minimization of the number of edges between estimated classes constitutes an estimator for the class assignments that concentrates on the true class assignment with high probability. Restriction (1) is expressed most naturally in the Chernoff-Hellinger phase and excludes the Kesten-Stigum phase.

Detection of the class assignment poses the weaker requirement that the *fraction* of vertices that are classified correctly goes to one with high probability (see definition 2.1): it was conjectured in (Decelle et al., 2011b; Decelle et al., 2011a) to be possible in block models, if, with c_n and d_n such that $p_n = c_n/n$ and $q_n = d_n/n$.

$$(c_n - d_n)^2 > 2(c_n + d_n). \tag{2}$$

Essentially Decelle *et al.* argue that in the Kesten-Stigum phase random graphs like that of the planted bi-section model allow estimation of the underlying class assignment, only if their distribution is sufficiently dissimilar from that of a Erdös-Rényi graph ($p_n = q_n$ makes class assignments indistinguishable, and correspondingly, small values for $p_n - q_n$ define a regime in which inference on the class assignment is relatively difficult, see (Banerjee, 2018)). An analogous claim in the Chernoff-Hellinger phase was first considered more rigorously in (Massoulié, 2014) and later confirmed, both from a probabilistic/statistical perspective in (Mossel, Neeman, and Sly, 2015; Mossel, Neeman, and Sly, 2016b), and independently from an information theoretic perspective in (Abbe, Bandeira, and Hall, 2016). Defining a_n and b_n by $np_n = a_n \log(n)$ and $nq_n = b_n \log(n)$ and assuming that $C^{-1} \leq a_n, b_n \leq C$ for all but finitely many $n \geq 1$, the class assignment in the planted bi-section model can be recovered exactly (Mossel, Neeman, and Sly, 2016b), if and only if,

$$(a_n + b_n - 2\sqrt{a_n b_n} - 1) \log n + \tfrac{1}{2} \log \log n \to \infty. \tag{3}$$

(Mossel, Neeman, and Sly, 2016b) also find a sharp condition for detection in the Kesten-Stigum phase. To summarize, the sparse Erdös-Rényi phases as well as the proximity of the Erdös-Rényi sub-model in the parameter space play a role in the statistical perspective on communities in the stochastic block model.

Estimation methods used for the community detection problem include spectral clustering (see (Krzakala et al., 2013) and many others), maximization of the likelihood and other modularities (Girvan and Newman, 2002; Bickel and Chen, 2009; Choi, Wolfe, and Airoldi, 2012; Amini et al., 2013), semi-definite programming (Hajek, Wu, and Xu, 2016; Guédon and Vershynin, 2016), and penalized ML detection of communities with minimax optimal misclassification ratio (Zhang and Zhou, 2016; Gao et al., 2017). More generally, we refer to (Abbe, 2018) and the very informative introduction of (Gao et al., 2017) for extensive bibliographies and a more comprehensive discussion. Bayesian methods have been popular throughout, *e.g.* the original work by Snijders and Nowicki (2001) (Nowicki and Snijders, 2001), the work of (Decelle et al., 2011b; Decelle et al., 2011a) and more recently, (Suwan et al., 2016), based on an empirical prior choice, and (Mossel, Neeman, and Sly, 2016a).

Interest in the stochastic block model has generated a wealth of algorithms that estimate the class assignment. Naturally, great emphasis has been placed on computational efficiency of these algorithms. Methods that maximize the likelihood or other modularities (Girvan and Newman, 2002; Bickel and Chen, 2009) do not compare well in this respect. For example, (Gao et al., 2017) argue along these lines (and favour a localized method over global (penalized) ML-estimation) (Zhang and Zhou, 2016). Broadly speaking, MCMC methods to simulate from posteriors are comparable to the relevant maximization methods as far as computational burdens are concerned.

Given that algorithms exists with much more favourable computational properties, calculation or simulation of posteriors seems unnecessarily laborious. As long as point-estimation is the only goal there is little to argue, but estimation is only the *first* step in statistical inference: when a consistent estimator has been found, immediate questions regarding (limiting) accuracy and reliability arise. From a Bayesian perspective, the posterior provides estimates of accuracy and credibility without further process, but to the algorithmic frequentist relying on a point-estimator, more detailed inferential questions concerning the sampling distributions of point-estimators are often (prohibitively) hard to analyse. In the stochastic block model, questions concerning accuracy have been addressed in (Zhang and Zhou, 2016), but, to the best of the authors' knowledge, frequentist uncertainty quantification with *confidence sets for class assignment* has not been addressed in the literature.

In this paper, our goal is to explore the limits of what is possible from the statistical point of view, similar to what Mossel *et al.* do from the probabilistic point of view and Abbe *et al.* from the information theoretic point of view. In section 3, it is shown that posteriors detect and recover underlying class assignments under conditions on $(p_n), (q_n)$ that are sharp. In section 4, it is demonstrated that Bayesian credible sets can be converted to asymptotically consistent confidence sets in various ways (Kleijn, 2016). If we assume posterior consistency in the form of exact recovery (or detection), credible sets (or enlarged credible sets) are consistent confidence sets. Moreover, if credible levels grow to one quickly enough, credible sets can be interpreted as frequentist confidence sets without conditions on the parameters. When,

$$n|p_n - q_n| \to 0, \quad n\bigl|p_n^{1/2} - q_n^{1/2}\bigr| \to \infty,$$

close to the Erdös-Rényi submodel where communities are the hardest to distinguish, credible sets may be enlarged to achieve frequentist asymptotic coverage. The 'diameters' of credible sets can be controlled when posterior convergence has been established, or with the help of the minimax bound for the expected number of misclassified vertices (Zhang and Zhou, 2016).

We conclude that, in the context of the planted bi-section model (and also much wider), the relatively high computational cost of simulating a posterior is justifiable if one is interested in more than estimation, in particular uncertainty quantification concerning the class assignment. Posteriors play a role comparable to that of bootstrap-approximated sampling distributions, leading to approximation of confidence intervals asymptotically. Le Cam's Bernstein-von Mises theorem arrives at the same conclusion (limited to the parametric setting of local-asymptotic normality).

This advantage of Bayesian methods over point-estimators has been noted before. Particularly, in the context of the planted bi-section model, physicists (Decelle et al., 2011a) expressed this point as follows: *"[The Bayesian] method provides [...] the marginal probabilities with which each node belongs to a given community. [...] Physically speaking, the Boltzmann distribution tells us more about a network than [just] the ground state does."*. (In the context of the cavity method/belief propagation the 'Boltzmann distribution' is the posterior distribution, and the 'ground state' signifies the point of maximal posterior density, the MAP-/ML-estimator.)

2 The planted bi-section model

In a stochastic block model, each vertex is assigned to one of $K \geq 2$ classes through an unobserved *class assignment vector* θ'. Each vertex belongs to a class and any edge occurs (independently of others) with a probability depending on the classes of the vertices that it connects. In the *planted bi-section model*, there are only two classes ($K = 2$) and, at the n-th iteration ($n \geq 1$), there are $2n$ vertices (labelled with indices $1 \leq i \leq 2n$), n in each class, with class assignment vector $\theta' \in \Theta'_n$ (with components $\theta'_1, \ldots, \theta'_{2n} \in \{0,1\}$), where Θ'_n is the subset of $\{0,1\}^{2n}$ of all finite binary sequences that contain as many ones as zeroes. Denote that space in which the random graph X^n takes its values by $\mathscr{X}_n$ (*e.g.* represented by its adjacency matrix with entries $\{X_{ij} : 1 \leq i,j \leq 2n\}$). The ($n$-dependent) probability of an edge occuring ($X_{ij} = 1$) between vertices $1 \leq i,j \leq 2n$ *within the same class* is denoted $p_n \in (0,1)$; the probability of an edge *between classes* is denoted $q_n \in (0,1)$,

$$Q_{ij}(\theta') := P_{\theta,n}(X_{ij} = 1) = \begin{cases} p_n, & \text{if } \theta'_{n,i} = \theta'_{n,j}, \\ q_n, & \text{if } \theta'_{n,i} \neq \theta'_{n,j}, \end{cases} \tag{4}$$

Note that if $p_n = q_n$, X^n is the Erdös-Rényi graph $G(2n, p_n)$ and the class assignment $\theta_n \in \Theta'_n$ is not identifiable. Another identifiability issue that arises is that the model is invariant under interchange of class labels 0 and 1. This is expressed in the parameter spaces Θ'_n through equivalence relations: $\theta'_1 \sim_n \theta'_2$, if $\theta'_{2,n} = \neg\theta'_{1,n}$ (by componentwise negation). To prevent non-identifiability, we parametrize the model for X^n in terms of a parameter θ_n in a quotient space $\Theta_n = \Theta'_n / \sim_n$, for every $n \geq 1$. For $\theta'_n \in \Theta'_n$ we denote the equivalence class $\{\theta'_n, \neg\theta'_n\}$ by θ_n. Note that the set Θ_n can be identified with the set of partitions of $\{1, \ldots, 2n\}$ consisting of exactly two sets, via the identification

$$\theta_n \longleftrightarrow \left\{ \left\{ i : \theta'_{n,i} = 0 \right\}, \left\{ i : \theta'_{n,i} = 1 \right\} \right\},$$

and note that this is independent of the choice of the representation.

The probability measure for the graph X^n corresponding to parameter θ is denoted $P_{\theta,n}$. The likelihood is given by,

$$p_{\theta,n}(X^n) = \prod_{i<j} Q_{i,j}(\theta)^{X_{ij}} (1 - Q_{i,j}(\theta))^{1-X_{ij}}.$$

For the *sparse versions* of the planted bi-section model, we also define edge probabilities that vanish with growing n: take (a_n) and (b_n) such that $a_n \log n = np_n$ and $b_n \log n = nq_n$ for the Chernoff-Hellinger phase; take (c_n) and (d_n) such that $c_n = np_n$ and $d_n = nq_n$ for the Kesten-Stigum phase. The fact that we do not allow loops (edges that connect vertices with themselves) leaves room for $2 \cdot \frac{1}{2}n(n-1) + n^2 = 2n^2 - n = \frac{1}{2} \cdot (2n)(2n-1)$ possible edges in the random graph X^n observed at iteration n.

The statistical question of interest in this model is to reconstruct the unobserved class assignment vectors θ_n *consistently*, that is, correctly with probability growing to one as $n \to \infty$. This can be stated in a strong and in a weak version, defined below.

Definition 2.1 *Let $\theta_{0,n} \in \Theta_n$ be given. An estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ is said to* recover the class assignment $\theta_{0,n}$ exactly *if,*

$$P_{\theta_{0,n}}\big(\hat{\theta}_n(X^n) = \theta_{0,n}\big) \to 1,$$

that is, if $\hat{\theta}_n$ indicates the correct partition assignment with high probability.

Based on the definition of k (just before eq. (7) below) we also relax this consistency requirement somewhat in the form of the following definition, *c.f.* (Mossel, Neeman, and Sly, 2016b) and others.

Definition 2.2 *Let $\theta_{0,n} \in \Theta_n$ be given. An estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ is said to* detect the class assignment $\theta_{0,n}$ *if,*

$$\frac{1}{2n}\left|\sum_{i=1}^{2n}(-1)^{\hat{\theta}_{n,i}}\,(-1)^{\theta_{0,n,i}}\right| \xrightarrow{P_{\theta_{0,n}}} 1,$$

that is, if the fraction of correct assignments in $\hat{\theta}_n$ grows to one with high probability.

Below, we specialize to the Bayesian approach: we choose prior distributions π_n for all Θ_n, $(n \geq 1)$ and calculate the posterior: denoting the likelihood by $p_{\theta,n}(X^n)$, the posterior for the parameter θ_n is written as a fraction of sums, for all $A \subset \Theta_n$,

$$\Pi(A|X^n) = \sum_{\theta_n \in A} p_{\theta,n}(X^n)\,\pi_n(\theta_n) \Bigg/ \sum_{\theta_n \in \Theta_n} p_{\theta,n}(X^n)\,\pi_n(\theta_n),$$

where $\pi_n : \Theta_n \to [0,1]$ is the probability mass function prior Π_n. Here, we only consider uniform priors (Π_n) for $\theta_n \in \Theta_n$, so for all $n \geq 1$ and $\theta_n \in \Theta_n$, $\pi(\theta_n) = \pi_n := (|\Theta_n|)^{-1}$.

3 Exact recovery and detection with posteriors

Consider the sequence of experiments in which we observe random graphs $X^n \in \mathscr{X}_n$ generated by the planted bi-section model of definition eq. (4). We assume, for every $n \geq 1$, that the prior is the uniform distribution over Θ_n: since we can choose n vertices from a total of $2n$ in $\binom{2n}{n}$ ways and $\theta \sim \neg\theta$, $\pi_n = (\frac{1}{2}\binom{2n}{n})^{-1}$.

Given true parameters $\theta_{0,n} \in \Theta_n$ $(n \geq 1)$, choose representations $\theta'_{0,n} \in \Theta'_n$ and define $Z_n(\theta'_0) \subset \{1, \ldots, 2n\}$ to be *class zero* (the set of all those i such that $\theta'_{0,i} = 0$) and call the complement $Z_n^c(\theta'_0)$ *class one*. For the questions concerning exact recovery and detection, we are interested in the sets $V'_{n,k} \subset \Theta'_n$, defined to contain all those θ'_n that differ from $\theta'_{0,n}$ by exactly k exchanges of pairs: for $\theta'_n \in \Theta'_n$ we have $\theta'_n \in V'_{n,k}$, if the set of vertices in class zero *c.f.* $\theta'_{0,n}$, $Z(\theta'_{0,n}) = \{1 \le i \le 2n : \theta'_{0,n,i} = 0\}$, from which we leave out the set of vertices in class zero *c.f.* θ'_n, $Z(\theta'_n) = \{1 \le i \le 2n : \theta'_{n,i} = 0\}$, has k elements. Conversely, for any $\theta'_{1,n}$ and $\theta'_{2,n}$ in Θ'_n, we denote the minimal number of pair-exchanges necessary to take $\theta'_{1,n}$ into $\theta'_{2,n}$ by $k'(\theta'_{1,n}, \theta'_{2,n})$. Note that $k'(\theta'_{1,n}, \neg\theta'_{2,n}) = n - k'(\theta'_{1,n}, \theta'_{2,n})$, which leads to the distance measure between two representation classes

$$k(\theta_{1,n}, \theta_{2,n}) = k'(\theta'_{1,n}, \theta'_{2,n}) \wedge k'(\theta'_{1,n}, \neg\theta'_{2,n}) \tag{5}$$

and note that this is independent of choice of the representations and that this function k takes values in $\{0, \dots, \lfloor n/2 \rfloor\}$. Now define,

$$V_{n,k} = V_{n,k}(\theta_{0,n}) = \{\theta_n : k(\theta_n, \theta_{0,n}) = k\} = \left\{\theta_n : \theta'_n \in V'_{n,k}\right\}, \tag{6}$$

for $k \in \{1, \dots, \lfloor n/2 \rfloor\}$. Given some sequence (k_n) of positive integers we then define V_n as the disjoint union,

$$V_n = \bigcup_{k=k_n}^{\lfloor n/2 \rfloor} V_{n,k} \tag{7}$$

Since we can choose two subsets of k elements from two sets of size n in $\binom{n}{k}^2$ ways, the cardinal of $V_{n,k}$ is $\binom{n}{k}^2$, when $k < n/2$ and $\frac{1}{2}\binom{n}{n/2}^2$ when n is even and $k = n/2$. In both cases the number of elements in $V_{n,k}$ is therefore bounded by $\binom{n}{k}^2$.

According to lemma 2.2 in (Kleijn, 2016) (with $B_n = \{\theta_{0,n}\}$), for any test sequences $\phi_{k,n} : \mathscr{X}_n \to [0,1]$ $(k \ge 1, n \ge 1)$, we have,

$$\begin{aligned} P_{\theta_0,n}\Pi(V_n|X^n) &= \sum_{k=k_n}^{\lfloor n/2 \rfloor} P_{\theta_0,n}\Pi(V_{n,k}|X^n) \\ &\le \sum_{k=k_n}^{\lfloor n/2 \rfloor} \Big(P_{\theta_0,n}\phi_{k,n}(X^n) + \sum_{\theta_n \in V_{n,k}} P_{\theta_n,n}(1 - \phi_{k,n}(X^n)) \Big) \end{aligned}$$

for every $n \ge 1$. Suppose that for any $k \ge 1$ there exists a sequence $(a_{n,k})_{n \ge 1}$, $a_{n,k} \downarrow 0$ and, for any $\theta_n \in V_{n,k}$, a test function $\phi_{\theta_n,n}$ that distinguishes $\theta_{0,n}$ from θ_n as follows,

$$P_{\theta_0,n}\phi_{\theta_n,n}(X^n) + P_{\theta_n,n}(1 - \phi_{\theta_n,n}(X^n)) \le a_{n,k}, \tag{8}$$

for all $n \ge 1$. Then using test functions $\phi_{k,n}(X^n) = \max\{\phi_{\theta_n,n}(X^n) : \theta_n \in V_{n,k}\}$, as well as the fact that,

$$P_{\theta_0,n}\phi_{k,n}(X^n) \le \sum_{\theta_n \in V_{n,k}} P_{\theta_0,n}\phi_{\theta_n,n}(X^n),$$

we see that,

$$\begin{aligned} P_{\theta_0,n}\Pi(V_n|X^n) &\le \sum_{k=k_n}^{\lfloor n/2 \rfloor} \sum_{\theta_n \in V_{n,k}} \Big(P_{\theta_{0,n},n}\phi_{\theta_n,n}(X^n) + P_{\theta_n,n}(1 - \phi_{\theta_n,n}(X^n)) \Big) \\ &\le \sum_{k=k_n}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 a_{k,n}. \end{aligned}$$

This inequality forms the basis for the results on exact recovery and detection in the next two subsections.

3.1 Posterior consistency: exact recovery

For the case of exact recovery, we are interested in the expected posterior masses of subsets of Θ_n of the form:

$$V_n = \{\theta_n \in \Theta_n : \theta_n \neq \theta_{0,n}\} = \bigcup_{k=1}^{\lfloor n/2 \rfloor} V_{n,k}.$$

The theorem states a sufficient condition for (p_n) and (q_n), which is related to requirement eq. (3) in the Chernoff-Hellinger phase.

THEOREM 3.1 *For some $\theta_{0,n} \in \Theta_n$, assume that $X^n \sim P_{\theta_{0,n}}$, for every $n \geq 1$. If we equip every Θ_n with its uniform prior and (p_n) and (q_n) are such that,*

$$\Bigl(1 + \bigl(1 - p_n - q_n + 2p_n\, q_n + 2\sqrt{p_n(1-p_n)q_n(1-q_n)}\bigr)^{n/2}\Bigr)^{2n} \to 1, \tag{9}$$

then the posterior succeeds in exact recovery, i.e.

$$\Pi(\theta_n = \theta_{0,n}|X^n) \xrightarrow{P_{\theta_{0,n}}} 1, \tag{10}$$

as $n \to \infty$.

PROOF According to lemma B.1, for every $n \geq 1$, $k \geq 1$ and given, $\theta_{0,n}$, there exists a test sequence satisfying eq. (8) with $a_{n,k} = (1-\mu_n)^{2k(n-k)}$ and $\mu_n = p_n + q_n - 2p_n\, q_n - 2(p_n(1-p_n)q_n(1-q_n))^{1/2} \in [0,1]$. Therefore, with $z_n = (1-\mu_n)^{n/2}$,

$$\begin{aligned} P_{\theta_{0,n}}\Pi(V_n|X^n) &\leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 (1-\mu_n)^{2k(n-k)} \leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 (1-\mu_n)^{nk} \\ &\leq \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{2n}{2k} (1-\mu_n)^{nk} \leq \sum_{l=1}^{2n} \binom{2n}{l} z_n^l = (1+z_n)^{2n} - 1 \end{aligned}$$

The right-hand side goes to zero if eq. (9) is satisfied. $\square$

EXAMPLE 3.2 Consider eq. (9) in the sparse setting of the Chernoff-Hellinger phase, where $np_n = a_n \log n$, $nq_n = b_n \log n$ with $a_n, b_n = O(1)$. In that case,

$$\begin{aligned} &\Bigl(1+\bigl(1 - p_n - q_n + 2p_n\, q_n + 2\sqrt{p_n(1-p_n)q_n(1-q_n)}\bigr)^{n/2}\Bigr)^{2n} \\ &\quad= \Bigl(1 + \Bigl(1 - \bigl(a_n + b_n - 2\sqrt{a_n b_n} + o(n^{-1}\log n)\bigr)\frac{\log n}{n}\Bigr)^{n/2}\Bigr)^{2n} \\ &\quad\approx \Bigl(1 + n^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_n b_n})}\Bigr)^{2n} = \Bigl(1 + \frac{1}{n} n^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_n b_n}-2)}\Bigr)^{2n} \\ &\quad\approx \exp\bigl(2e^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_n b_n}-2)\log n}\bigr) \end{aligned} \tag{11}$$

Accordingly, in the Chernoff-Hellinger phase eq. (9) amounts to the sufficient condition,

$$(a_n + b_n - 2\sqrt{a_n b_n} - 2) \log n \to \infty, \tag{12}$$

which closely resembles (but is not exactly equal to) eq. (3), the requirement of (Mossel, Neeman, and Sly, 2016b). In fact there is a trade-off: eq. (3) is slightly weaker than eq. (12) but applies only if there exists a $C > 0$ such that $C^{-1} \leq a_n, b_n \leq C$ for large enough n (Mossel, Neeman, and Sly, 2016b; Zhang and Zhou, 2016). More particularly, one of the sequences (a_n) and (b_n) may fade away with growing n or equal zero outright. For example if $b_n = 0$ and $\liminf_n a_n > 2$, edges between classes are completely absent but, separately, the Erdös-Rényi graphs spanned by vertices in $Z_n(\theta_0')$ and $Z_n^c(\theta_0')$ respectively are connected with high probability. Similarly, if $a_n = 0$ and $\liminf_n b_n > 2$, the posterior succeeds in exact recovery: apparently, with b_n above 2, edges between classes are abundant enough to guarantee the existence of a path in X^n that visits all vertices at least once, with high probability. It is tempting to state the following, well-known (Abbe, Bandeira, and Hall, 2016; Mossel, Neeman, and Sly, 2016b) sufficient condition for the sequences $a_n > 0$ and $b_n > 0$:

$$(\sqrt{a_n} - \sqrt{b_n})^2 > c, \textit{ for some } c > 2 \textit{ and } n \textit{ large enough,} \tag{13}$$

(even though it ignores the logarithm in eq. (12)).

Corollary 3.3 *Under the conditions of theorem 3.1, the MAP-/ML-estimator* $\hat{\theta}_n$ *recovers* $\theta_{0,n}$ *exactly:* $P_{\theta_{0,n}}(\hat{\theta}_n(X^n) = \theta_{0,n}) \to 1$.

Proof Due to the uniformity of the prior, for every $n \geq 1$, maximization of the posterior density (with respect to the counting measure) on Θ_n, is the same as maximization of the likelihood. Due to eq. (10), the posterior density in the point $\theta_{0,n}$ in Θ_n converges to one in $P_{\theta_{0,n}}$-probability. Accordingly, the point of maximization is $\theta_{0,n}$ with high probability. □

Remark 3.4 Although (Mossel, Neeman, and Sly, 2016b) formulates seminal results, there are small imperfections. Lemma 5.1 is not correct as stated: in case u and v (in their notation) are connected (the simplest counterexample is the graph with two nodes and one edge), where one node is class one and the other class zero (or -1 in their notation). However, the conclusion of the lemma still holds, if we require in addition that u and v are not connected. In the proof of theorem 2.5, this requirement should be taken into account. Also note that $|E(G) \cap B_{\tau'}| \leq |E(G) \cap B_\tau|$ (in their notation), contrary to (Mossel, Neeman, and Sly, 2016b, proof of lemma 5.1).

3.2 Posterior consistency: detection

For the case of detection, the requirement of convergence is less stringent: we require only that estimated θ_n differ from the true $\theta_{0,n}$ by $o(n)$ differences, rather than matching

$\theta_{0,n}$ exactly. It is well-known that edge-probabilities may be of orders smaller than $o(\log(n)/n)$ and of order $O(n^{-1})$. More precisely, *c.f.* (Mossel, Neeman, and Sly, 2016b, proposition 2.9) we have that

$$\frac{n(p_n - q_n)^2}{p_n + q_n} \to \infty, \tag{14}$$

is a necessary and sufficient condition for detection. So where exact recovery can be achieved only in the Chernoff-Hellinger phase, detection of communities is possible in the Kersten-Stigum phase as well. In this section, we show that the posterior for the uniform prior also 'detects' the true class assignment (when that notion is translated appropriately into a property of the posterior, see also (Mossel, Neeman, and Sly, 2016a)).

We are interested in the expected posterior masses of subsets of Θ_n of the form:

$$W_n = \bigcup_{k=k_n}^{\lfloor n/2 \rfloor} V_{n,k},$$

for a (possibly) divergent sequence k_n of order $o(n)$: the posterior concentrates on class assignments θ_n that differ from $\theta_{0,n}$ by no more than k_n pair exchanges, so the fraction of misclassified vertices becomes negligible in the limit $n \to \infty$.

THEOREM 3.5 *For some $\theta_{0,n} \in \Theta_n$, let $X^n \sim P_{\theta_{0,n}}$ for every $n \geq 1$. If we equip all Θ_n with uniform priors and (p_n), (q_n) are such that,*

$$\frac{n}{k_n}\Bigl(1 - p_n - q_n + 2p_n\, q_n + 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\Bigr)^{n/2} \to 0, \ \text{as } n \to \infty, \tag{15}$$

then,

$$\Pi(W_n|X^n) \xrightarrow{P_0} 0, \ \text{as } n \to \infty, \tag{16}$$

i.e. the posterior detects $\theta_{0,n}$ at rate k_n.

PROOF *(of theorem 3.5)* According to lemma B.1, for every $n \geq 1$, $k \geq 1$ and given, $\theta_{0,n}$, there exists a test sequence satisfying eq. (8) with $a_{n,k} = (1-\mu_n)^{2k(n-k)}$. Therefore, using the inequalities $\binom{2n}{k} \leq \frac{(2n)^k}{k!}$ and $(n+m)! \geq n!m!$, the Stirling lower bound formula, and finally our assumption $n(1-\mu_n)^{n/2}/k_n \to 0$, we see that for big enough n,

$$P_{\theta_{0,n}}\Pi(W_n|X^n) \leq \sum_{k=k_n}^{\lfloor n/2 \rfloor} \binom{n}{k}^2 (1-\mu_n)^{2k(n-k)} \leq \sum_{k=2k_n}^{n} \binom{2n}{k} (1-\mu_n)^{k(n-k/2)}$$

$$\leq \sum_{k=2k_n}^{\infty} \frac{1}{k!}(2n)^k (1-\mu_n)^{kn/2} \leq \frac{\bigl(2n(1-\mu_n)^{n/2}\bigr)^{2k_n}}{(2k_n)!} e^{2n(1-\mu_n)^{n/2}}.$$

Application of Stirling's approximation then leads to,

$$P_{\theta_{0,n}}\Pi(W_n|X^n) \leq \frac{1}{\sqrt{4\pi k_n}}\left(\frac{n(1-\mu_n)^{n/2}}{k_n}\right)^{2k_n} e^{2k_n + 2n(1-\mu_n)^{n/2}}$$

$$\leq \frac{1}{\sqrt{4\pi k_n}}\left(\frac{n(1-\mu_n)^{n/2}}{k_n} e^{1+n(1-\mu_n)^{n/2}/k_n}\right)^{2k_n} \leq \frac{n(1-\mu_n)^{n/2}}{k_n} e^{1+n(1-\mu_n)^{n/2}/k_n}$$

which converges to zero as $n \to \infty$. $\square$

EXAMPLE 3.6 Note that as $p_n, q_n \to 0$, we may expand,

$$\sqrt{p_n} - \sqrt{q_n} = \frac{1}{2\sqrt{\frac{1}{2}(p_n + q_n)}}(p_n - q_n) + O(|p_n - q_n|^2).$$

which means that,

$$\mu_n = (\sqrt{p_n} - \sqrt{q_n})^2 + O(n^{-2}) = \frac{(p_n - q_n)^2}{2(p_n + q_n)} + O(n^{-2})$$

Assuming only condition eq. (2), we would arrive at the conclusion that $n\mu_n > 1 + O(n^{-1})$, which is insufficient in the proof of theorem 3.5. Note that a non-divergent choice $k_n = O(1)$ forces us back into the Chernoff-Hellinger phase where exact recovery is possible.

COROLLARY 3.7 *Under the conditions of theorem 3.5 with (p_n) and (q_n) such that,*

$$n\big(p_n + q_n - 2p_n\,q_n - 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\big) \to \infty, \tag{17}$$

as $n \to \infty$, then, there exists a sequence $k_n = o(n)$ such that,

$$\Pi\big(\, k_n(\theta_n, \theta_{0,n}) \geq k_n \;\big|\; X^n \,\big) \xrightarrow{P_0} 0, \tag{18}$$

$n \to \infty$, i.e. the posterior detects $\theta_{0,n}$.

PROOF Define, for every $\beta \in (0,1)$, $k_{\beta,n} = \beta\, n$. We follow the proof of theorem 3.5 with $k_n = k_{\beta,n}$ and note that,

$$P_{\theta_{0,n}}\Pi\big(\, k_n(\theta_n, \theta_{0,n}) \geq k_{\beta,n} \;\big|\; X^n \,\big) \leq \frac{1}{\beta}(1-\mu_n)^{n/2} e^{1+\beta^{-1}(1-\mu_n)^{n/2}}.$$

Due to eq. (17),

$$(1-\mu_n)^{n/2} = \big(1 - p_n - q_n + 2p_n\,q_n + 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\big)^{n/2} \to 0,$$

so $P_{\theta_{0,n}}\Pi(k_n(\theta_n, \theta_{0,n}) \geq k_{\beta,n} \mid X^n) \to 0$. Let $\beta_m \downarrow 0$ be given; if we let $m(n)$ go to infinity slowly enough, posterior convergence continues to hold with β equal to $\beta_{m(n)}$, that is, for $k_n = k_{\beta_{m(n)},n}$. $\square$

According to (Mossel, Neeman, and Sly, 2016b, proposition 2.9) (see eq. (14)), $n\mu_n \to \infty$ is also a *necessary condition* for the possibility of detection. So corollary 3.7 shows that uniform priors lead to posteriors that detect the truth under the weakest possible condition on the parameter sequences (p_n) and (q_n). This is encouraging to the Bayesian and to the frequentist who uses Bayesian methods in this model and in models like it, *e.g.* the stochastic block model.

4 Uncertainty quantification

The most immediate results on uncertainty quantification are obtained with the help of the results in the previous section: if we know that the sequences (p_n) and (q_n) satisfy requirements like eq. (9) or eq. (15), so that recovery or detection is guaranteed, then a consistent sequence of confidence sets is easily constructed from credible sets, as shown in subsection 4.1 and the sizes of these credible sets as well as the sizes of associated confidence sets are controlled.

If the sequences (p_n) and (q_n) are unknown, or if we require explicit confidence levels, confidence sets can still be constructed from credible sets under conditions requiring that credible levels grow to one quickly enough. Enlargement of credible sets may be used to mitigate this condition, whenever we are close to the Erdös-Rényi submodel, as discussed in subsection 4.2.

Regarding the sizes of credible sets, the most natural way to compile a minimal-order credible set $E_n(X^n)$ in a discrete space like Θ_n, is to calculate the posterior weights $\Pi(\{\theta_n\}|X^n)$ of all $\theta_n \in \Theta_n$, order Θ_n by decreasing posterior weight into a finite sequence $\{\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,|\Theta_n|}\}$ and define $E_n(X^n) = \{\theta_{n,1}, \ldots, \theta_{n,m}\}$, for the smallest $m \geq 1$ such that $\Pi(\{\theta_{n,1}, \ldots, \theta_{n,m}\}|X^n)$ is greater than or equal to the required credible level. To provide guarantees regarding the sizes of credible sets, one would like to show that these $E_n(X^n)$ are of an order that is upper bounded with high probability. (Although it is not so clear what the upper bound should be, ideally.)

Here we shall follow a different path based on the smallest number $k(\theta_n, \eta_n)$ of pair-exchanges between two representations θ'_n and η'_n in Θ'_n of θ_n and η_n respectively, see eq. (5). The maps $k : \Theta_n \times \Theta_n \to \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ are interpreted in a role similar to that of a metric on larger parameter spaces: the *diameter* $\mathrm{diam}_n(C)$ of a subset $C \subset \Theta_n$ is,

$$\mathrm{diam}_n(C) = \max\bigl\{k(\theta_n, \eta_n) : \theta_n, \eta_n \in C\bigr\}.$$

by definition.

4.1 Posterior recovery/detection and confidence sets

If the posteriors concentrate amounts of mass on $\{\theta_{0,n}\}$ arbitrarily close to one with growing n, then a sequence of credible sets of a certain, fixed level contains $\theta_{0,n}$ for large enough n. If such posterior concentration occurs with high $P_{\theta_{0,n}}$-probability, then the sequence of credible sets is also an asymptotically consistent sequence of confidence sets.

PROPOSITION 4.1 *Let $c_n \in [0, 1]$ be given, with $c_n > \epsilon > 0$ for large enough n. Suppose* *that the posterior recovers the communities exactly,*

$$\Pi(\theta = \theta_{0,n}|X^n) \xrightarrow{P_{\theta_{0,n}}} 1. \tag{19}$$

Then any sequence (D_n) of (P_n^{Π}-almost-sure) credible sets of levels c_n satisfies,

$$P_{\theta_{0,n}}\big(\,\theta_{0,n} \in D_n(X^n)\,\big) \to 1,$$

i.e. (D_n) is a consistent sequence of confidence sets. Credible sets of minimal order/diameter equal $\{\theta_0\}$ with high $P_{\theta_{0,n}}$-probability.

Proof Note that with the uniform priors Π_n, $P_{\theta_{0,n}} \ll P_n^{\Pi}$ for all $n \geq 1$, so that P_n^{Π}-almost-surely defined credible sets D_n of credible level at least ϵ, also satisfy,

$$P_{\theta_{0,n}}\big(\,\Pi(D_n(X^n)|X^n) \geq \epsilon\,\big) = 1.$$

So if, in addition,

$$P_{\theta_{0,n}}\big(\,\Pi(\{\theta_{0,n}\}|X^n) > 1-\epsilon\,\big) \to 1,$$

then $\theta_{0,n} \in D_n(X^n)$ with high $P_{\theta_{0,n}}$-probability. Since all posterior mass is concentrated at $\theta_{0,n}$ with high probability, the $\{\theta_{0,n}\}$ form a sequence of unique credible sets of minimal order (or minimal diameter $k_n = 0$) with confidence levels greater than $\epsilon > 0$ for large enough n. □

In the Kesten-Stigum phase, enlargement of credible sets is sufficient to obtain confidence sets. Recall the definition of the $V_{n,k}(\theta_n)$ in eq. (6) (with $\theta_{0,n}$ replaced by θ_n). Given some fixed underlying $\theta_{0,n} \in \Theta_n$, we write $V_{n,k}$ for $V_{n,k}(\theta_{0,n})$. Making a certain choice for the upper bounds $k_n \geq 1$, we arrive at,

$$B_n(\theta_n) = \bigcup_{k=0}^{k_n} V_{n,k}(\theta_n), \tag{20}$$

for every $n \geq 1$ and $\theta_n \in \Theta_n$. Similar as for $V_{n,k}$ we write B_n for $B_n(\theta_{0,n})$. Given a subset D_n of Θ_n, the set $C_n \subset \Theta_n$ associated with D_n under $B_n(\theta_n)$ (see definition C.3) then is the set of $\theta_n \in \Theta_n$ whose k-distance from some element of D_n is at most k_n,

$$C_n = \{\theta_n \in \Theta_n : \exists_{\eta_n \in D_n},\, k(\eta_n, \theta_n) \leq k_n\},$$

the k_n-enlargement of D_n. If we know that the sequences (p_n) and (q_n) satisfy requirement eq. (15), posterior concentration occurs around $\{\theta_{0,n}\}$ in 'balls' of diameters $2k_n$ with growing n, and there exist credible sets D'_n of levels greater than $1/2$ and of diameters $2k_n$ centred on $\theta_{0,n}$. The credible sets D_n of *minimal diameters* of any level greater than $1/2$ must intersect D_n. Then the k_n-enlargements C_n of the D_n contain $\theta_{0,n}$.

Theorem 4.2 *Let $c_n \in [0,1]$ be given, with $c_n > \epsilon > 0$ for large enough n. Suppose that the posterior detects communities with rate (k_n),*

$$\Pi\big(\, k(\theta_n, \theta_{0,n}) \leq k_n \;\big|\; X^n \,\big) \xrightarrow{P_{\theta_{0,n}}} 1.$$

Let (B_n) denote a sequence of (P_n^{Π}-almost-sure) credible sets of levels c_n of minimal diameters. *Then* $\mathrm{diam}_n(B_n) \leq 2k_n$ *with high $P_{\theta_{0,n}}$-probability and the k_n-enlargements C_n of the B_n satisfy,*

$$P_{\theta_{0,n}}\big(\, \theta_{0,n} \in C_n(X^n) \,\big) \to 1,$$

i.e. the k_n-enlargements (C_n) form a consistent sequence of confidence sets.

Proof As in the proof of proposition proposition 4.1, P_n^{Π}-almost-surely defined credible sets D_n of credible level at least c_n also satisfy,

$$P_{\theta_{0,n}}\big(\, \Pi(D_n(X^n)|X^n) \geq c_n \,\big) = 1.$$

Now fix $n \geq 1$. For every $\theta_n \in \Theta_n$ and every $x^n \in \Theta_n$, let $k_n(\theta_n, x^n)$ denote the radius of the smallest ball in Θ_n centred on θ_n of posterior mass (at least) c_n. Let $\hat{\theta}_n(x^n) \in \Theta_n$ be such that,

$$k_n(\hat{\theta}_n(x^n)) = \min\big\{ k_n(\theta_n, x^n) : \theta_n \in \Theta_n \big\},$$

i.e. the centre point of a *smallest* k_n-ball in Θ_n. Convergence of the posterior implies that the balls $B_n(\theta_{0,n})$ of radii k_n centred on $\theta_{0,n}$ contain a fraction of the posterior mass arbitrarily close to one, so assuming that n is large enough, we may assume that $c_n > \epsilon > 0$ and $\Pi(B_n(\theta_{0,n})|X^n) > 1 - \epsilon$ with high $P_{\theta_{0,n}}$-probability. Conclude that,

$$B_n(\theta_{0,n}) \cap B_n(\hat{\theta}_n(X^n)) \neq \varnothing,$$

with high $P_{\theta_{0,n}}$-probability, which amounts to asymptotic coverage of $\theta_{0,n}$ for the k_n-enlargement $C_n(X^n)$ of $B_n(\hat{\theta}_n(X^n))$. $\square$

Note that the case $k_n = 0$ obtains in the Chernoff-Hellinger phase and the cases $k_n = o(n)$ correspond to the Kesten-Stigum phase. The theorem remains valid in the case $k_n = \beta\, n$ (for some $\beta \in (0,1)$), as in the proof of corollary 3.7.

4.2 Confidence sets directly from credible sets

To use proposition 4.1 or theorem 4.2, the statistician needs to know that the sequences (p_n) and (q_n) satisfy eq. (9) or eq. (15), basically to satisfy the testing condition eq. (8). (Particularly, eq. (17) is *not* strong enough to apply theorem 4.2).) But even if that knowledge is not available and testing cannot serve as a condition, the use of credible sets as confidence sets remains valid, as long as credible levels grow to one fast enough. The following proposition also provides lower bounds for confidence levels of credible sets. (Write $b_n = |\Theta_n|^{-1} = (\frac{1}{2}\binom{2n}{n})^{-1}$.)

Proposition 4.3 *Let $\theta_{0,n}$ in Θ_n with uniform priors Π_n, $n \geq 1$, be given and let D_n be a sequence of credible sets, such that,*

$$\Pi(D_n(X^n)|X^n) \geq 1 - a_n,$$

for some sequence (a_n) with $a_n = o(b_n)$. Then,

$$P_{\theta_0,n}\big(\,\theta_0 \in D_n(X^n)\,\big) \geq 1 - b_n^{-1} a_n.$$

Proof If $\theta_{0,n} \notin D_n(X^n)$ then $\Pi(\{\theta_{0,n}\}|X^n) \leq a_n$, P_n^{Π}-almost-surely. Then,

$$\begin{aligned}
P_{\theta_0,n}\big(\theta_0 \in \Theta \setminus D_n(X^n)\big) &= P_n^{\Pi|\{\theta_0\}}\big(\theta_0 \in \Theta \setminus D_n(X^n)\big) \\
&= b_n^{-1} \int_{\{\theta_{0,n}\}} P_{\theta,n}\big(\theta_0 \in \Theta \setminus D_n(X^n)\big)\, d\Pi_n(\theta) \\
&\leq b_n^{-1} P_n^{\Pi}\big(1\{\theta_0 \in \Theta_n \setminus D_n(X^n)\}\, \Pi(\{\theta_{0,n}\}|X^n)\big) \leq b_n^{-1} a_n,
\end{aligned}$$

by Bayes's Rule eq. (30). □

Theorem C.4 leaves room for mitigation of the lower bound on credible levels if we are willing to use enlarged credible sets. There are two competing influences when enlarging: on the one hand, the prior masses $b_n = \Pi_n(B_n(\theta_{0,n}))$ become larger, relaxing the rate at which credible levels are required to go to one. On the other hand, enlargement leads to likelihood ratios with random fluctuations that take them further away from one (see lemmas C.6 and C.7), thus interfering with notions like contiguity and remote contiguity (see appendix C). Whether proposition 4.3 is useful and whether enlargement of credible sets helps, depends on the sequences (p_n), (q_n), essentially to maintain sufficient distance from the Erdös-Rényi submodel. On the other hand, close to the Erdös-Rényi submodel, fluctuations of likelihood ratios are relatively small, so remote contiguity is achieved relatively easily.

We shall consider the 'statistical phase' where distinctions between within-class and between-class edges become less-and-less pronounced:

$$p_n - q_n = o\big(n^{-1}\big), \tag{21}$$

while satisfying also the condition that,

$$p_n^{1/2}(1-p_n)^{1/2} + q_n^{1/2}(1-q_n)^{1/2} = o\big(n|p_n - q_n|\big). \tag{22}$$

In this regime, $p_n, q_n \to 0$ or $p_n, q_n \to 1$. If $p_n, q_n \to 0$ as in the sparse phases, eq. (22) amounts to,

$$n(p_n^{1/2} - q_n^{1/2}) \to \infty, \tag{23}$$

so differences between p_n and q_n may not converge to zero too fast. (Note however that extreme sparsity levels of order $p_n, q_n \propto n^{-\gamma}$ with $1 < \gamma < 2$ are allowed.) For the following lemma we define,

$$\rho_n = \min\Bigl\{ \Bigl(\frac{1-p_n}{p_n}\,\frac{q_n}{1-q_n}\Bigr), \Bigl(\frac{p_n}{1-p_n}\,\frac{1-q_n}{q_n}\Bigr) \Bigr\} = e^{-|\lambda_n|}.$$

where $\lambda_n := \log(1-p_n) - \log(p_n) + \log(q_n) - \log(1-q_n)$, and,

$$\alpha_n = \int 2k(\theta_{0,n},\theta_n)(n-k(\theta_{0,n},\theta_n))\,d\Pi_n(\theta_n|B_n) = \frac{1}{|B_n|}\sum_{k=0}^{k_n}\binom{n}{k}^2 2k(n-k)$$

with the following rate for remote contiguity:

$$d_n = \rho_n^{C\alpha_n|p_n-q_n|}, \tag{24}$$

for some choice of $C > 1$.

Let (k_n) and $\theta_{0,n} \in \Theta_n$ be given for all $n \geq 1$. It is shown in lemma C.7 that, if eq. (22) holds, then,

$$P_{\theta_{0,n}} \lhd d_n^{-1} P_n^{\Pi|B_n},$$

with $B_n = B_n(\theta_{0,n})$ like in eq. (20). This argument amounts to a proof for the following theorem (immediate from theorem C.4).

Theorem 4.4 *Let (k_n) be given and assume that eq. (21) and eq. (22) hold. Let $\theta_{0,n}$ in Θ_n with uniform priors Π_n be given and let D_n be a sequence of credible sets of credible levels $1-a_n$, for some sequence (a_n) such that $b_n^{-1}a_n = o(d_n)$. Then the sets C_n, associated with D_n under B_n as in eq. (20) satisfy,*

$$P_{\theta_0,n}\bigl(\,\theta_0 \in C_n(X^n)\,\bigr) \to 1,$$

i.e. the C_n are asymptotic confidence sets.

Consider the possible choices for (a_n) if we assume $k_n = \beta\, n$ for some fixed $\beta \in (0,1)$ (as in the proof of corollary 3.7), which leads to the type of exponential correction factor in the prior mass sequence b_n that is required to move the restriction on the credible levels $1-a_n$ substantially. First of all, Stirling's approximation gives rise to the following approximate lower bound on the factor between prior mass and prior mass without enlargement:

$$\frac{\Pi_n(B_n)}{\Pi_n(\{\theta_{0,n}\})} = \sum_{k=0}^{k_n}\binom{n}{k}^2 \geq \binom{n}{k_n}^2 \geq \frac{1}{2\pi n}\,\frac{1}{\beta(1-\beta)}f(\beta)^n,$$

where $f:(0,1)\to(1,4]$ is given by,

$$f(\beta) = (1-\beta)^{-2(1-\beta)}\beta^{-2\beta}.$$

Approximating $\alpha_n \approx 2k_n(n-k_n)$ for large n and using eq. (21), we also have,

$$d_n = \rho_n^{C\alpha_n|p_n-q_n|} \approx \rho_n^{2Cn^2\beta(1-\beta)|p_n-q_n|} = e^{-|\lambda_n|o(n)}.$$

So if we assume that $\lambda_n = O(1)$, d_n is sub-exponential and does not play a role for the improvement factor.

Conclude as follows: (let $a_n = o(|\Theta_n|^{-1}) \approx o(4^{-n})$ denote the rates appropriate in proposition 4.3 and assume $\lambda_n = O(1)$) if we have credible sets $D_n(X^n)$ of credible levels $1 - a_n f(\beta)^{n(1+o(1))}$, then the sequence of enlarged confidence sets $(C_n(X^n))$, associated with $D_n(X^n)$ through B_n with $k_n = \beta\, n$, covers the true value of the class assignment parameter with high probability. Credible levels that had to be of order $1 - a_n \approx 1 - o(4^{-n})$ previously, can be of approximate order $1 - o(c^{-n})$ for any $1 < c < 4$ by enlargement by B_n if conditions eqs. (21) and (22) hold; the closer $0 < \beta < 1/2$ is to $1/2$, the closer c is to 1.

To control the sizes of credible and confidence sets obtained in this way, the *existence* of a point-estimator sequence $(\hat{\theta}_n)$ known to converge at certain rates, together with a different condition on the sequences (p_n) and (q_n), can also be used to guarantee that there exist credible/confidence sets of controlled diameters, thus by-passing testing requirements. Arguments from (Kleijn, 2016) will be used to show that, with high $P_{\theta_0,n}$-probability, there exist credible sets $C_n(X^n)$ of controlled diameters δ_n, that are also asymptotic confidence sets, for sequences (δ_n) related to the minimax rate of misclassification (Zhang and Zhou, 2016). We emphasize that we do not expect the bound in the following theorem to be sharp (see remark 4.6).

THEOREM 4.5 *Assume that $X^n \sim P_{\theta_0,n}$ for some $\theta_{0,n} \in \Theta_n$. Let (p_n), (q_n) and (δ_n) be given, then, with high $P_{\theta_0,n}$-probability there exist credible sets $C_n(X^n)$ of diameter smaller than or equal to δ_n such that,*

$$P_{\theta_0,n}\bigl(\,\theta_0 \notin C_n(X^n)\,\bigr) \leq 2 \cdot 4^n \left(\frac{\sqrt{p_n}\sqrt{q_n} + \sqrt{1-p_n}\sqrt{1-q_n}}{(\pi n \delta_n)^{1/n}}\right)^{n/2}. \tag{25}$$

PROOF Theorem 1.1 of (Zhang and Zhou, 2016), applied specifically to the planted bisection model and written in our notation, provides the following exponential minimax bound for the fraction of misclassified vertices: there exists an estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ such that,

$$P_{\theta,n}\left(\frac{k(\theta_n, \hat{\theta}_n(X^n))}{n}\right) \leq \bigl(\sqrt{p_n}\sqrt{q_n} + \sqrt{1-p_n}\sqrt{1-q_n}\bigr)^{n(1+o(1))}, \tag{26}$$

for every $n \geq 1$ and all $\theta_n \in \Theta_n$. Using Markov's inequality,

$$P_{\theta,n}\left(\frac{k(\theta_n, \hat{\theta}_n(X^n))}{n} \geq \delta_n\right) \leq \delta_n^{-1}\bigl(\sqrt{p_n}\sqrt{q_n} + \sqrt{1-p_n}\sqrt{1-q_n}\bigr)^{n(1+o(1))}, \tag{27}$$

which fulfils condition eq. (36) with $a_n = 2\delta_n^{-1/2}(\sqrt{p_n}\sqrt{q_n} + \sqrt{1-p_n}\sqrt{1-q_n})^{n/2}$. Comparing with $b_n = |\Theta_n|^{-1} = 2\binom{2n}{n}^{-1} \approx 4^{-n}\sqrt{\pi n}$, we find that,

$$
\begin{aligned}
b_n^{-1} a_n &\leq 2 \cdot 4^n (\pi n \delta_n)^{-1/2} \big(\sqrt{p_n}\sqrt{q_n} + \sqrt{1-p_n}\sqrt{1-q_n}\big)^{n/2} \\
&= 2 \cdot 4^n \left(\frac{\sqrt{p_n}\sqrt{q_n} + \sqrt{1-p_n}\sqrt{1-q_n}}{(\pi n \delta_n)^{1/n}} \right)^{n/2},
\end{aligned}
$$

which completes the proof. □

REMARK 4.6 The formulation of the upper-bound on confidence levels gives rise to a condition on (p_n), (q_n) and (δ_n):

$$
(\pi n \delta_n)^{1/n} \big(\sqrt{p_n}\sqrt{q_n} + \sqrt{1-p_n}\sqrt{1-q_n}\big) < \frac{1}{16},
$$

for large enough n. Admittedly, the amount of parameter-space left by this restriction is rather small: credible sets $C_n(X^n)$ have confidence levels growing to one, only if the difference between p_n and q_n stays above bounds that do not vanish in the limit. Thus we exclude any of the sparse phases, as well as a sizeable portion of the dense phase. There does not appear any inherent reason why such a restriction should exist: given the type of convergence of eq. (10) one would expect more direct control in the Chernoff-Hellinger phase and also in the Kesten-Stigum phase. It seems likely that the bound achieved here is far from sharp. The reason for this is probably the transition from eq. (26) to eq. (27): the use of Markov's inequality here is crude and it is conjectured that, either, *(i.)* one can demonstrate that the penalized MLE's $(\hat{\theta}_n)$ of (Zhang and Zhou, 2016) and (Gao et al., 2017) satisfy a much sharper version of eq. (27), or, *(ii.)* one can show the existence of estimators $(\tilde{\theta}_n)$ satisfying much sharper versions of eq. (27). Either way, a much sharper version of eq. (25) would result.

Acknowledgement

The authors thank E. Mossel and J. Neeman for a helpful discussion on necessary conditions for exact recovery.

A Definitions and conventions

Because we take the perspective of a frequentist using Bayesian methods, we are obliged to demonstrate that Bayesian definitions continue to make sense under the assumption that the data X is distributed according to a true, underlying P_0.

REMARK A.1 We assume given for every $n \geq 1$, a random graph X^n taking values in the (finite) space $\mathscr{X}_n$ of all undirected graphs with n vertices. We denote the powerset of $\mathscr{X}_n$ by $\mathscr{B}_n$ and regard it as the domain for probability distributions $P_n : \mathscr{B}_n \to [0,1]$

a model $\mathscr{P}_n$ parametrized by $\Theta_n \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$ with finite parameter spaces Θ_n (with powerset $\mathscr{G}_n$) and uniform priors Π_n on Θ_n. As frequentists, we assume that there exists a 'true, underlying distribution for the data; in this case, that means that for every $n \geq 1$, there exists a $\theta_{0,n} \in \Theta_n$ and corresponding $P_{\theta_{0,n}}$ from which the n-th graph X^n is drawn. □

DEFINITION A.2 *Given $n \geq 1$ and a prior probability measure Π_n on Θ_n, define the n-th* prior predictive distribution *as:*

$$P_n^{\Pi}(A) = \int_{\Theta} P_{\theta,n}(A)\, d\Pi_n(\theta), \tag{28}$$

for all $A \in \mathscr{B}_n$. For any $B_n \in \mathscr{G}$ with $\Pi_n(B_n) > 0$, define also the n-th local prior predictive distribution*,*

$$P_n^{\Pi|B}(A) = \frac{1}{\Pi_n(B_n)} \int_{B_n} P_{\theta,n}(A)\, d\Pi_n(\theta), \tag{29}$$

as the predictive distribution on $\mathscr{X}_n$ that results from the prior Π_n when conditioned on B_n.

The prior predictive distribution P_n^{Π} is the marginal distribution for X^n in the Bayesian perspective that considers parameter and sample jointly $(\theta, X^n) \in \Theta \times \mathscr{X}_n$ as the random quantity of interest.

DEFINITION A.3 *Given $n \geq 1$, a* (version of) the posterior *is any map $\Pi(\,\cdot\,|X^n = \,\cdot\,) : \mathscr{G}_n \times \mathscr{X}_n \to [0,1]$ such that,*

1. *for $B \in \mathscr{G}_n$, the map $\mathscr{X}_n \to [0,1] : x^n \mapsto \Pi(B|X^n = x^n)$ is $\mathscr{B}_n$-measurable,*
2. *for all $A \in \mathscr{B}_n$ and $V \in \mathscr{G}_n$,*

$$\int_A \Pi(V|X^n)\, dP_n^{\Pi} = \int_V P_{\theta,n}(A)\, d\Pi_n(\theta). \tag{30}$$

Bayes's Rule is expressed through equality eq. (30) and is sometimes referred to as a 'disintegration' (of the joint distribution of (θ, X^n)). Because the models $\mathscr{P}_n$ are dominated (denote the density of $P_{\theta,n}$ by $p_{\theta,n}$), the fraction of integrated likelihoods,

$$\Pi(V|X^n) = \int_V p_{\theta,n}(X^n)\, d\Pi_n(\theta) \Bigg/ \int_{\Theta_n} p_{\theta,n}(X^n)\, d\Pi_n(\theta), \tag{31}$$

for $V \in \mathscr{G}_n$, $n \geq 1$ defines a regular version of the posterior distribution. (Note also that there is no room in definition eq. (30) for X^n-dependence of the prior, so 'empirical Bayes' methods must be based on data Y^n independent of X^n, *e.g.* obtained by sample-splitting.)

Notation and conventions

Asymptotic statements that end in "... with high probability" indicate that said statements are true with probabilities that grow to one. The abbreviations *l.h.s.* and *r.h.s.* refer to "left-" and "right-hand sides" respectively. For given probability measures P, Q on a measurable space $(\Omega, \mathscr{F})$, we define the Radon-Nikodym derivative $dP/dQ : \Omega \to [0, \infty)$, P-almost-surely, referring *only* to the Q-dominated component of P, following (Le Cam, 1986). We also *define* $(dP/dQ)^{-1} : \Omega \to (0, \infty] : \omega \mapsto 1/(dP/dQ(\omega))$, Q-almost-surely. Given random variables $Z_n \sim P_n$, weak convergence to a random variable Z is denoted by $Z_n \xrightarrow{P_n\text{-w.}} Z$, convergence in probability by $Z_n \xrightarrow{P_n} Z$ and almost-sure convergence (with coupling P^∞) by $Z_n \xrightarrow{P^\infty\text{-a.s.}} Z$. The integral of a real-valued, integrable random variable X with respect to a probability measure P is denoted PX, while integrals over the model with respect to priors and posteriors are always written out in Leibniz's or sum notation. The cardinal of a set B is denoted $|B|$.

B Existence of suitable tests

Given $n \geq 1$ and two class assignment vectors $\theta_{0,n}, \theta_n \in \Theta_n$, we are interested in calculation of the likelihood ratio $dP_{\theta,n}/dP_{\theta_0,n}$, because it determines testing power as well as the various forms of remote contiguity that play a role.

Choose a representation θ_0' of θ_0 and a representation θ' of θ so that $k'(\theta_0', \theta') = k(\theta_0, \theta)$, where k and k' are as in section 3. Recall that, $Z_n(\theta_0') \subset \{1, \ldots, 2n\}$ is class zero and the complement $Z_n^c(\theta_0')$ class one. For the sake of presentation (in fig. 1 below), re-label the vertices such that $Z(\theta_0') = \{1, \ldots, n\}$ and $Z^c(\theta_0') = \{n+1, \ldots, 2n\}$. In the case $n = 4$, fig. 1 shows edge probabilities in the familiar block arrangement.

Recall that the likelihood under θ_0 is given by,

$$p_{\theta_0,n}(X^n) = \prod_{i<j} Q_{i,j}(\theta_0)^{X_{ij}} (1 - Q_{i,j}(\theta_0))^{1-X_{ij}}.$$

If we assume that $\theta_{0,n}'$ and θ_n' differ by k pair-exchanges among respective members of the zero- and one-classes, then a look at fig. 1 reveals that the likelihood-ratio depends only on the edges for which exactly one of its end-points changes class. Define,

$$\begin{aligned} A_n &= \{(i,j) \in \{1, \ldots, 2n\} : i < j,\, \theta_{0,n,i}' = \theta_{0,n,j}',\, \theta_{n,i}' \neq \theta_{n,j}'\}, \\ B_n &= \{(i,j) \in \{1, \ldots, 2n\} : i < j,\, \theta_{0,n,i}' \neq \theta_{0,n,j}',\, \theta_{n,i}' = \theta_{n,j}'\}. \end{aligned}$$

Also define,

$$(S_n, T_n) := \Big(\sum\{X_{ij} : (i,j) \in A_n\}, \sum\{X_{ij} : (i,j) \in B_n\}\Big),$$

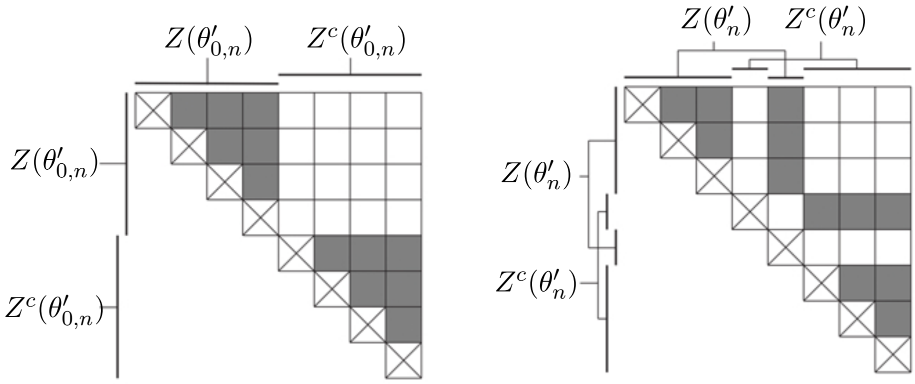

Figure 1: Class assignments and edge probabilitites according to $\theta'_{0,n}$ and to θ'_n for $n = 4$ and $k = 1$. Vertex sets $Z(\cdot)$ and $Z^c(\cdot)$ correspond to zero- and one-classes for the given class assignment. Dark squares correspond to edges that occur with (within-class) probability p_n, and light squares to edges that occur with (between-class) probability q_n.

and note that the likelihood ratio can be written as,

$$\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n) = \left(\frac{1-p_n}{p_n}\,\frac{q_n}{1-q_n}\right)^{S_n - T_n} \tag{32}$$

where,

$$(S_n, T_n) \sim \begin{cases} \mathrm{Bin}(2k(n-k), p_n) \times \mathrm{Bin}(2k(n-k), q_n), & \text{if } X^n \sim P_{\theta_0,n}, \\ \mathrm{Bin}(2k(n-k), q_n) \times \mathrm{Bin}(2k(n-k), p_n), & \text{if } X^n \sim P_{\theta,n}. \end{cases} \tag{33}$$

Based on that, we derive the following lemma.

Lemma B.1 *Let $n \geq 1$, $\theta_{0,n}, \theta_n \in \Theta_n$ be given. Assume that $\theta_{0,n}$ and θ_n differ by k pair-exchanges. Then there exists a test function $\phi_n : \mathscr{X}_n \to [0,1]$ such that,*

$$P_{\theta_0,n}\phi_n(X^n) + P_{\theta,n}(1 - \phi_n(X^n)) \leq a_{n,k},$$

with testing power,

$$a_{n,k} = \bigl(1 - p_n - q_n + 2p_n\, q_n + 2\sqrt{p_n(1-p_n)}\sqrt{q_n(1-q_n)}\bigr)^{2k(n-k)}.$$

Proof The likelihood ratio test $\phi_n(X^n)$ has testing power bounded by the so-called Hellinger transform,

$$P_{\theta_0,n}\phi_n(X^n) + P_{\theta,n}(1 - \phi_n(X^n)) \leq \inf_{0 \leq \alpha \leq 1} P_{\theta_0,n}\Bigl(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\Bigr)^{\alpha},$$

(see, *e.g.* (Le Cam, 1986) and proposition 2.6 in (Kleijn, 2016)). Using $\alpha = 1/2$ (which is the minimum), we find that,

$$P_{\theta_0,n}\Bigl(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\Bigr)^{1/2} = P_{\theta_0,n}\Bigl(\frac{p_n}{1-p_n}\,\frac{1-q_n}{q_n}\Bigr)^{\frac{1}{2}(T_n-S_n)} = Pe^{\frac{1}{2}\lambda_n S_n}\,Pe^{-\frac{1}{2}\lambda_n T_n}$$

where $\lambda_n := \log(1-p_n) - \log(p_n) + \log(q_n) - \log(1-q_n)$ and (S_n, T_n) are distributed binomially, as in the first part of eq. (33). Using the moment-generating function of the binomial distribution, we conclude that,

$$\begin{aligned}
&P_{\theta_0,n}\Bigl(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\Bigr)^{1/2}\\
&\quad= \biggl(\Bigl(1-p_n+p_n\Bigl(\frac{1-p_n}{p_n}\,\frac{q_n}{1-q_n}\Bigr)^{1/2}\Bigr)\Bigl(1-q_n+q_n\Bigl(\frac{p_n}{1-p_n}\,\frac{1-q_n}{q_n}\Bigr)^{1/2}\Bigr)\biggr)^{2k(n-k)}\\
&\quad= \biggl(\Bigl((1-p_n)+p_n^{1/2}q_n^{1/2}\Bigl(\frac{1-p_n}{1-q_n}\Bigr)^{1/2}\Bigr)\Bigl((1-q_n)+p_n^{1/2}q_n^{1/2}\Bigl(\frac{1-q_n}{1-p_n}\Bigr)^{1/2}\Bigr)\biggr)^{2k(n-k)}\\
&\quad= \Bigl((1-p_n)(1-q_n)+2p_n^{1/2}q_n^{1/2}(1-p_n)^{1/2}(1-q_n)^{1/2}+p_nq_n\Bigr)^{2k(n-k)},
\end{aligned}$$

which proves the assertion. □

C Remote contiguity and credibile/confidence sets

Bayesian asymptotics has seen a great deal of development over recent decades, but the essence of the mathematical theory remains that of Schwartz's theorem: a balance between testing power and a minimum of prior mass 'locally', leads to a controlled limit for the posterior distribution with a frequentist interpretation. It has also become clear that the same notion of 'locality' allows conversion of sequences of credible sets to asymptotic confidence sets and that is the purpose of this paper as well. 'Locality' in the above sense is defined through local prior predictive distributions based on a weakened form of contiguity called remote contiguity (Kleijn, 2016).

Asymptotic relations between credible sets and confidence sets are governed by the following definitions and theorem.

Definition C.1 *Let $(\Theta_n, \mathscr{G}_n)$ with priors Π_n be given, denote the sequence of posteriors by $\Pi(\cdot|\cdot) : \mathscr{G}_n \times \mathscr{X}_n \to [0,1]$. Let $\mathscr{D}_n$ denote a collection of measurable subsets of Θ_n. A* sequence of credible sets (D_n) *of* credible levels $1 - a_n$ *(where $0 \le a_n \le 1$, $a_n \downarrow 0$) is a sequence of set-valued maps $D_n : \mathscr{X}_n \to \mathscr{D}_n$ such that $\Pi(\Theta_n \setminus D_n(x^n)|x^n) \le a_n$.*

Note the following: the dependence of D_n on the observed graph X^n may be defined P_n^Π-almost-surely or with high P_n^Π-probability. Whenever the difference is of importance in the main text, it is mentioned explicitly.

Definition C.2 *For $0 \leq a \leq 1$, a set-valued map $x \mapsto C(x)$ defined on $\mathscr{X}$ such that, for all $\theta \in \Theta$, $P_\theta(\theta \notin C(X)) \leq a$, is called a confidence set of level $1-a$. If the levels $1-a_n$ of a sequence of confidence sets $C_n(X^n)$ go to 1 as $n \to \infty$, the $C_n(X^n)$ are said to be asymptotically consistent.*

If measurability of the set C is not guaranteed, interpret definition C.2 in outer measure.

Definition C.3 *Let D be a (credible) set in Θ and let $B = \{B(\theta) : \theta \in \Theta\}$ denote a collection of model subsets such that $\theta \in B(\theta)$ for all $\theta \in \Theta$. A model subset C is said to be (a confidence set) associated with D under B, if for all $\theta \in \Theta \setminus C$, $B(\theta) \cap D = \varnothing$.*

The relationship between a credible set D and the model subset C associated with D under B is illustrated in fig. 2 and detailed in the following theorem. (The notation $P_n \lhd d_n^{-1} Q_n$ is explained below, see definition C.5.)

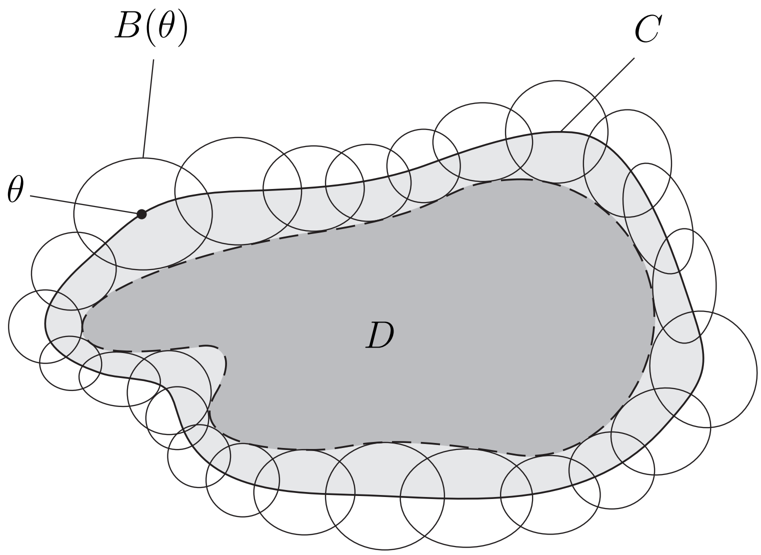

Figure 2: The relation between a credible set D and its associated (minimal) confidence set C under B in Venn diagrams: the extra points θ in the associated confidence set C not included in the credible set D are characterized by non-empty intersection $B(\theta) \cap D \neq \varnothing$.

Theorem C.4 *Let $\theta_{0,n} \in \Theta_n$ $(n \geq 1)$ and $0 \leq a_n \leq 1$, $b_n > 0$ such that $a_n = o(b_n)$ be given. Choose priors Π_n and let D_n denote level-$(1-a_n)$ credible sets in Θ_n. Furthermore, for all $\theta \in \Theta$, let $B_n = \{B_n(\theta_n) \in \mathscr{G}_n : \theta_n \in \Theta_n\}$ and b_n denote sequences such that,* *(i.)* *prior mass is lower bounded,* $\Pi_n(B_n(\theta_0)) \geq b_n$*,*

(ii.) *and for some* $d_n \downarrow 0$ *such that* $b_n^{-1}a_n = o(d_n)$*,* $P_{\theta_0,n} \lhd d_n^{-1}\, P_n^{\Pi|B(\theta_0)}$*.*

Then any confidence sets C_n *associated with the credible sets* D_n *under* B_n *are asymptotically consistent, i.e. for all* $\theta_0 \in \Theta$*,*

$$P_{\theta_0,n}\big(\,\theta_0 \in C_n(X^n)\,\big) \to 1. \tag{34}$$

In most of section 4, the sets B_n are simply,

$$B_n(\theta_n) = \{\theta_n\},$$

for every $n \geq 1$ and every $\theta_n \in \Theta_n$, so that the confidence sets C_n associated with any credible sets $D_n \subset \Theta_n$ under B_n are simply *equal to* D_n. In that case, $P_{\theta_0,n} \lhd c_n^{-1}\, P_n^{\Pi|B(\theta_0)}$ for any rate (c_n), $c_n \downarrow 0$, so all sequences $a_n = o(b_n)$ are permitted. Since the prior mass in $B_n(\theta_{0,n})$ is fixed, theorem C.4 says that, if we have a sequence of credible sets $D_n(X^n) \subset \Theta_n$ of high enough credible levels $1 - a_n$, then these $D_n(X^n)$ are also asymptotically consistent confidence sets (see proposition 4.3).

With an eye on enlarged credible sets, we note that condition *(ii.)* of theorem C.4 says that the sequence $(P_{\theta_0,n})$ is required to be *remotely contiguous* with respect to $P_n^{\Pi|B(\theta_0)}$ at rate $b_n a_n^{-1}$.

Definition C.5 *Given the spaces* $\mathscr{X}_n$*,* $n \geq 1$ *with two sequences* (P_n) *and* (Q_n) *of probability measures and a sequence* $\rho_n \downarrow 0$*, we say that* Q_n *is* ρ_n*-remotely contiguous with respect to* P_n*, notation* $Q_n \lhd \rho_n^{-1} P_n$*, if,*

$$P_n\phi_n(X^n) = o(\rho_n) \quad \Rightarrow \quad Q_n\phi_n(X^n) = o(1), \tag{35}$$

for every sequence of $\mathscr{B}_n$*-measurable* $\phi_n : \mathscr{X}_n \to [0,1]$*.*

Below, we use the following lemma (see section 3 in (Kleijn, 2016)) which shows that uniform tightness of a sequence of re-scaled likelihood ratios is sufficient for remote contiguity.

Lemma C.6 *Given* (P_n)*,* (Q_n)*,* $d_n \downarrow 0$*,* (Q_n) *is* d_n*-remotely contiguous with respect to* (P_n) *if, under* Q_n*, every subsequence of* $(d_n(dP_n/dQ_n)^{-1})$ *has a weakly convergent subsequence.*

According to Prokhorov's theorem, the condition of lemma C.6 is equivalent to uniform tightness: for every $\epsilon > 0$ there exists an $M > 0$ such that,

$$\sup_{n\geq 1} P\Big(d_n\Big(\frac{dP_n}{dQ_n}\Big)^{-1}(X^n) > M\Big) < \epsilon.$$

Based on that, it is no surprise that the following lemma revolves around the central limit. Recall that according to eq. (24), $d_n = \rho_n^{C\alpha_n|p_n-q_n|}$ for some choice of $C > 1$.

Lemma C.7 *Let (k_n) be given and assume that eq. (22) holds. Then for any $\theta_{0,n} \in \Theta_n$,*

$$P_{\theta_{0,n}} \lhd d_n^{-1} P_n^{\Pi|B},$$

with $B_n = B_n(\theta_{0,n})$ like in eq. (20).

Proof Let (k_n) and $\theta_{0,n} \in \Theta_n$ be given. We denote $P_n = P_n^{\Pi|B}$, $Q_n = P_{\theta_{0,n}}$ and apply Jensen's inequality to obtain,

$$\begin{aligned}\frac{dP_n}{dQ_n}(X^n) &= \frac{1}{|B_n|} \sum_{\theta_n \in B_n} \left(\frac{1-p_n}{p_n} \frac{q_n}{1-q_n} \right)^{S_n(\theta_n)-T_n(\theta_n)} \\ &\geq \exp\Bigl(\frac{\lambda_n}{|B_n|} \sum_{\theta_n \in B_n} \bigl(S_n(\theta_n) - T_n(\theta_n) \bigr) \Bigr)\end{aligned}$$

where $(S_n(\theta_n), T_n(\theta_n))$ is distributed as in eq. (33). By invariance of the sum under permutations of the vertices, we re-sum as follows for any $k \geq 1$,

$$\frac{1}{|V_{n,k}|} \sum_{\theta_n \in V_{n,k}} S_n(\theta_n) = \frac{2k(n-k)}{n(n-1)} S_n, \quad \frac{1}{|V_{n,k}|} \sum_{\theta_n \in V_{n,k}} T_n(\theta_n) = \frac{2k(n-k)}{n^2} T_n,$$

where, with the notation $Z_n = Z(\theta'_{0,n}) \subset \{1, \ldots, 2n\}$, for a certain representation $\theta'_{0,n}$ of $\theta_{0,n}$, for the zero elements of $\theta'_{0,n}$,

$$\begin{aligned}S_n &= \sum_{i,j \in Z_n} X_{ij} + \sum_{i,j \in Z_n^c} X_{ij} \sim \mathrm{Bin}(n(n-1), p_n), \\ T_n &= \sum_{i \in Z_n, j \in Z^c} X_{ij} + \sum_{i \in Z_n^c, j \in Z} X_{ij} \sim \mathrm{Bin}(n^2, q_n)\end{aligned}$$

which gives us the upper bound,

$$\frac{dP_n}{dQ_n}(X^n) \geq \rho_n^{\sum_{k=0}^{k_n} 2k(n-k) \frac{|V_{n,k}|}{|B_n|} |\bar{S}_n - \bar{T}_n|} = \rho_n^{\alpha_n |\bar{S}_n - \bar{T}_n|},$$

where $\bar{S}_n = S_n/(n(n-1))$ and $\bar{T}_n = T_n/n^2$. By the central limit theorem,

$$\left(\frac{n(\bar{S}_n - p_n)}{p_n^{1/2}(1-p_n)^{1/2}}, \frac{n(\bar{T}_n - q_n)}{q_n^{1/2}(1-q_n)^{1/2}} \right) \xrightarrow{Q_n\text{-w.}} N(0,1) \times N(0,1),$$

which implies that for every $\epsilon > 0$ there exists an $M > 0$ such that,

$$\sup_{n \geq 1} Q_n \left(\frac{n(\bar{S}_n - p_n)}{p_n^{1/2}(1-p_n)^{1/2}} \vee \frac{n(\bar{T}_n - q_n)}{q_n^{1/2}(1-q_n)^{1/2}} > M \right) < \epsilon$$

Conclude that,

$$\sup_{n \geq 1} Q_n \left(\Bigl(\frac{dP_n}{dQ_n}(X^n) \Bigr)^{-1} \leq \rho_n^{-\alpha_n \left(\frac{M}{n} (p_n^{1/2}(1-p_n)^{1/2} + q_n^{1/2}(1-q_n)^{1/2}) + |p_n - q_n| \right)} \right) \geq 1 - \epsilon.$$

Note that the term in the exponent proportional to M is dominated by $|p_n - q_n|$ by eq. (22). Hence for every $C > 1$ and every $\epsilon > 0$,

$$Q_n \left(\Bigl(\frac{dP_n}{dQ_n}(X^n) \Bigr)^{-1} \leq \rho_n^{-C\alpha_n |p_n - q_n|} \right) \geq 1 - \epsilon,$$

for large enough n. Using the remark following lemma C.6, we see that $P_{\theta_{0,n}} \lhd d_n^{-1} P_n^{\Pi|B}$, with d_n as in eq. (24) $\square$

D Diameters of credible sets

Generalizing from the specific setting of the planted bi-section model, we assume for the moment that the priors Π_n are not necessarily uniform, in fact we assume very little in what follows: we consider general discrete models $\Theta_n \to \mathscr{P}_n$ for sequential data X^n, for all $n \geq 1$. We assume that the (Θ_n, g_n) are metric spaces with priors Π_n (in the planted bi-section model, priors are uniform and $g_n(\hat{\theta}_n, \theta_n) = k(\hat{\theta}_n, \theta_n)/n$). Denote $b_n = \Pi_n(\{\theta_{0,n}\})$.

Theorem D.1 *Let $(\Theta_n, g_n) \to \mathscr{P}_n$ be given, with priors Π_n, for all $n \geq 1$. Let $X^n \sim P_{\theta_0,n}$ for some $\theta_{0,n} \in \Theta_n$. Assume that there exist sequences $a_n, b_n \downarrow 0$, radii (δ_n) and estimators $\hat{\theta}_n$ such that*

$$P_{\theta,n}\big(g(\hat{\theta}_n, \theta_n) \geq \delta_n\big) \leq \tfrac{1}{4}a_n^2 \tag{36}$$

for all Π_n-almost-all $\theta_n \in \Theta_n$. Then, with high $P_{\theta_0,n}$-probability, there exist credible sets $C_n(X^n)$ of diameters $2\delta_n$ such that,

$$P_{\theta_0,n}\big(\,\theta_0 \in C_n(X^n)\,\big) \geq 1 - b_n^{-1}a_n.$$

Remark D.2 The statement of the theorem may seem somewhat strange, given that the point of the assertion appears to be related in rather trivial ways to the condition: if we have such $\hat{\theta}_n$, why bother with posteriors at all? The point of the theorem is that we do not need to calculate (or even be able to calculate) the estimators $\hat{\theta}_n$: merely the *existence* of such estimators implies that the posteriors can be relied upon to provide credible sets of controlled radii δ_n that can serve as confidence sets.

Remark D.3 Regarding condition eq. (36) above, suppose that there exists an estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ such that the collection of random variables $\{\epsilon^{-1}g(\hat{\theta}_n, \theta_n) : n \geq 1, \theta_n \in \Theta_n\}$ satisfies Prokhorov's uniform tightness condition: for all $\epsilon > 0$, there exists an $M > 0$ such that,

$$P_{\theta,n}\big(g(\hat{\theta}_n, \theta_n)) \geq M\,\epsilon_n\big) \leq \epsilon$$

for all $n \geq 1$ and all $\theta_n \in \Theta_n$. (Another way of saying this is that $\hat{\theta}_n$ converges to θ_n at rate ϵ_n, uniformly over Θ_n.) Then there exist functions $\epsilon \mapsto M(\epsilon)$ such that the above display holds. If we pick a sequence $a_n \downarrow 0$ and one such function, we may define $\delta_n = M(\frac{1}{4}a_n^2)\epsilon_n$ and,

$$C_n(X^n) = \{\theta_n \in \Theta_n : g_n(\hat{\theta}_n(X^n), \theta_n) < \delta_n\},$$

then we see that,

$$P_{\theta,n}\big(\theta_n \in C_n(X^n)\big) \geq 1 - \tfrac{1}{4}a_n^2, \tag{37}$$

for all $n \geq 1$ and all $\theta_n \in \Theta_n$. In other words, our assumption implies that *there exist* confidence sets $C_n(X^n)$ of diameter $2\delta_n$ and confidence level $1 - \frac{1}{4}a_n^2$. Our goal below is to show that this implies the existence of credible sets of controlled diameter and credible level high enough to serve also as confidence sets through proposition 4.3.

Proof *(of theorem D.1)*
Fix $n \geq 1$. Integrating eq. (36) with respect to Π_n, Bayes's rule eq. (30) says,

$$\begin{aligned} P_n^\Pi \Pi(\Theta_n \setminus C_n(X^n)|X^n) &= \int_{\mathscr{X}_n} \int_{\Theta_n} 1\{(x^n, \theta_n) : \theta_n \notin C_n(x^n)\}\, d\Pi(\theta_n|X^n = x^n)\, dP_n^\Pi(x^n) \\ &= \int_{\Theta_n} \int_{\mathscr{X}_n} 1\{(x^n, \theta_n) : \theta_n \notin C_n(x^n)\}\, dP_{\theta,n}(x^n)\, d\Pi_n(\theta_n) \\ &= \int_{\Theta_n} P_{\theta,n}\big(\theta_n \notin C_n(X^n)\big)\, d\Pi_n(\theta_n) \leq \tfrac{1}{4} a_n^2, \end{aligned}$$

for every $n \geq 1$. Markov's inequality then implies,

$$P_n^\Pi\big(\Pi(\Theta_n \setminus C_n(X^n)|X^n) \geq \tfrac{1}{2} a_n\big) \leq 2 a_n^{-1} P_n^\Pi \Pi(\Theta_n \setminus C_n(X^n)|X^n) \leq \tfrac{1}{2} a_n.$$

Let F_n denote the subset of $\mathscr{X}_n$ for which $\Pi(\Theta_n \setminus C_n(X^n)|X^n) < \frac{1}{2} a_n$. Then,

$$\begin{aligned} &P_n^\Pi 1\{\theta_0 \in \Theta_n \setminus C_n(X^n)\}\, \Pi(\{\theta_{0,n}\}|X^n) \\ &\qquad \leq P_n^\Pi(\mathscr{X}_n \setminus F_n) + P_n^\Pi 1_{F_n}(X^n) 1\{\theta_0 \in \Theta_n \setminus C_n(X^n)\}\, \Pi(\{\theta_{0,n}\}|X^n). \end{aligned}$$

If $\theta_{0,n} \notin C_n(X^n)$ then $\Pi(\{\theta_{0,n}\}|X^n) < \frac{1}{2} a_n$ since $X^n \in F_n$. Conclude,

$$\begin{aligned} P_{\theta_0,n}\big(\theta_0 \in \Theta \setminus C_n(X^n)\big) &= P_n^{\Pi|\{\theta_0\}}\big(\theta_0 \in \Theta \setminus C_n(X^n)\big) \\ &= b_n^{-1} \int_{\{\theta_{0,n}\}} P_{\theta,n}\big(\theta_0 \in \Theta \setminus C_n(X^n)\big)\, d\Pi_n(\theta) \\ &\leq b_n^{-1} P_n^\Pi 1\{\theta_0 \in \Theta_n \setminus C_n(X^n) \;\; \Pi(\{\theta_{0,n}\}|X^n) \leq b_n^{-1} a_n, \end{aligned}$$

by Bayes's Rule eq. (30). ☐

References

Abbe, E. (2018). "Community Detection and Stochastic Block Models: Recent Developments". In: *Journal of Machine Learning Research* 18.177, pp. 1–86. URL: http://jmlr.org/papers/v18/16-480.html.

Abbe, E., A. S. Bandeira, and G. Hall (2016). "Exact Recovery in the Stochastic Block Model". In: *IEEE: Transactions on Information Theory* 62.1. URL: http://arxiv.org/abs/1405.3267.

Amini, A. A. et al. (2013). "Pseudo-likelihood methods for community detection in large sparse networks". In: *Ann. Statist.* 41.4, pp. 2097–2122. DOI: 10.1214/13-AOS1138.

Banerjee, Debapratim (2018). "Contiguity and non-reconstruction results for planted partition models: the dense case". In: *Electron. J. Probab.* 23, 28 pp. DOI: 10.1214/17-EJP128. URL: https://doi.org/10.1214/17-EJP128.

Bickel, P. J. and A. Chen (2009). "A nonparametric view of network models and Newman-Girvan and other modularities". In: *Proceedings of the National Academy of Sciences* 106.50, pp. 21068–21073. ISSN: 0027-8424. DOI: 10.1073/pnas.0907096106.

Bollobás, B., S. Janson, and O. Riordan (2007). "The phase transition in inhomogeneous random graphs". In: *Random Structures & Algorithms* 31.1, pp. 3–122. DOI: 10.1002/rsa.20168.

Choi, D. S., P. J. Wolfe, and E. M. Airoldi (2012). "Stochastic blockmodels with a growing number of classes". In: *Biometrika* 99.2, pp. 273–284. DOI: 10.1093/biomet/asr053.

Decelle, A. et al. (2011a). "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Phys. Rev. E* 84 (6), p. 066106. DOI: 10.1103/PhysRevE.84.066106.

— (2011b). "Inference and Phase Transitions in the Detection of Modules in Sparse Networks". In: *Phys. Rev. Lett.* 107 (6), p. 065701. DOI: 10.1103/PhysRevLett.107.065701.

Dyer, M. E. and A. M. Frieze (1989). "The solution of some random NP-hard problems in polynomial expected time". In: *Journal of Algorithms* 10.4, pp. 451 –489. ISSN: 0196-6774. DOI: https://doi.org/10.1016/0196-6774(89)90001-1.

Erdös, P. and A. Rényi (1959). "On Random Graphs I". In: *Publicationes Mathematicae* (6).

Fortunato, S. (2010). "Community detection in graphs". In: *Physics Reports* 486.3, pp. 75 –174. ISSN: 0370-1573. DOI: https://doi.org/10.1016/j.physrep.2009.11.002.

Gao, C. et al. (2017). "Achieving Optimal Misclassification Proportion in Stochastic Block Models". In: *Journal of Machine Learning Research* 18.60, pp. 1–45. URL: http://jmlr.org/papers/v18/16-245.html.

Girvan, M. and M. E. J. Newman (2002). "Community structure in social and biological networks". In: *Proceedings of the National Academy of Sciences of the United States of America* 99.12, pp. 7821–7826. DOI: 10.1073/pnas.122653799.

Guédon, O. and R. Vershynin (2016). "Community detection in sparse networks via Grothendieck's inequality". In: *Probability Theory and Related Fields* 165.3, pp. 1025–1049. ISSN: 1432-2064. DOI: 10.1007/s00440-015-0659-z.

Hajek, B., Y. Wu, and J. Xu (2016). "Achieving Exact Cluster Recovery Threshold via Semidefinite Programming". In: *IEEE Trans. Inf. Theor.* 62.5, pp. 2788–2797. ISSN: 0018-9448. DOI: 10.1109/TIT.2016.2546280.

Holland, P. W., K. B. Laskey, and S. Leinhardt (1983). "Stochastic blockmodels: First steps". In: *Social Networks* 5.2, pp. 109 –137. ISSN: 0378-8733. DOI: https://doi.org/10.1016/0378-8733(83)90021-7.

Kleijn, B. J. K. (2016). "On the frequentist validity of Bayesian limits". In: *ArXiv e-prints*. arXiv: 1611.08444 [math.ST].

Krzakala, F. et al. (2013). “Spectral redemption in clustering sparse networks”. In: *Proceedings of the National Academy of Sciences* 110.52, pp. 20935–20940. ISSN: 0027-8424. DOI: 10.1073/pnas.1312486110.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer-Verlag New York. ISBN: 978-1-4612-4946-7. DOI: 10.1007/978-1-4612-4946-7.

Massoulié, L. (2014). “Community detection thresholds and the weak Ramanujan property”. In: *STOC 2014: 46th Annual Symposium on the Theory of Computing*. New York, United States, pp. 1–10. URL: https://hal.archives-ouvertes.fr/hal-00969235.

Mossel, E., J. Neeman, and A. Sly (2015). “Reconstruction and estimation in the planted partition model”. In: *Probability Theory and Related Fields* 162.3, pp. 431–461. ISSN: 1432-2064. DOI: 10.1007/s00440-014-0576-6.

— (Aug. 2016a). “Belief propagation, robust reconstruction and optimal recovery of block models”. In: *Ann. Appl. Probab.* 26.4, pp. 2211–2256. DOI: 10.1214/15-AAP1145. URL: https://doi.org/10.1214/15-AAP1145.

— (2016b). “Consistency thresholds for the planted bisection model”. In: *Electron. J. Probab.* 21, 24 pp. DOI: 10.1214/16-EJP4185.

Nowicki, K. and T. A. B. Snijders (2001). “Estimation and Prediction for Stochastic Blockstructures”. In: *Journal of the American Statistical Association* 96.455, pp. 1077–1087. DOI: 10.1198/016214501753208735.

Suwan, S. et al. (2016). “Empirical Bayes estimation for the stochastic blockmodel”. In: *Electron. J. Statist.* 10.1, pp. 761–782. DOI: 10.1214/16-EJS1115.

Zhang, A. Y. and H. H. Zhou (2016). “Minimax rates of community detection in stochastic block models”. In: *Ann. Statist.* 44.5, pp. 2252–2280. DOI: 10.1214/15-AOS1428.