J. Appl. Prob. **55,** 325–352 (2018)
*doi:*10.1017/jpr.2018.22

RECOVERING A HIDDEN COMMUNITY BEYOND THE KESTEN–STIGUM THRESHOLD IN $O(|E|\log^*|V|)$ TIME

BRUCE HAJEK,* *University of Illinois at Urbana-Champaign*
YIHONG WU,** *Yale University*
JIAMING XU,*** *Purdue University*

Abstract

Community detection is considered for a stochastic block model graph of n vertices, with K vertices in the planted community, edge probability p for pairs of vertices both in the community, and edge probability q for other pairs of vertices. The main focus of the paper is on weak recovery of the community based on the graph G, with $o(K)$ misclassified vertices on average, in the sublinear regime $n^{1-o(1)} \le K \le o(n)$. A critical parameter is the effective signal-to-noise ratio $\lambda = K^2(p-q)^2/((n-K)q)$, with $\lambda = 1$ corresponding to the Kesten–Stigum threshold. We show that a belief propagation (BP) algorithm achieves weak recovery if $\lambda > 1/\mathrm{e}$, beyond the Kesten–Stigum threshold by a factor of $1/\mathrm{e}$. The BP algorithm only needs to run for $\log^* n + O(1)$ iterations, with the total time complexity $O(|E|\log^* n)$, where $\log^* n$ is the iterated logarithm of n. Conversely, if $\lambda \le 1/\mathrm{e}$, no local algorithm can asymptotically outperform trivial random guessing. Furthermore, a linear message-passing algorithm that corresponds to applying a power iteration to the nonbacktracking matrix of the graph is shown to attain weak recovery if and only if $\lambda > 1$. In addition, the BP algorithm can be combined with a linear-time voting procedure to achieve the information limit of exact recovery (correctly classify all vertices with high probability) for all $K \ge (n/\log n)(\rho_{\mathrm{BP}} + o(1))$, where ρ_{BP} is a function of p/q.

Keywords: Hidden community; belief propagation; message passing; spectral algorithms; high-dimensional statistics

2010 Mathematics Subject Classification: Primary 62H12
Secondary 62C20

1. Introduction

The problem of finding a densely connected subgraph in a large graph arises in many research disciplines such as theoretical computer science, statistics, and theoretical physics. To study this problem, the stochastic block model [19] for a single dense community is considered.

Definition 1.1. (*Planted dense subgraph model.*) Given $n \ge 1$, $C^* \subset [n]$, and $0 \le q \le p \le 1$, the corresponding *planted dense subgraph model* is a random undirected graph $G = (V, E)$ with $V = [n]$, such that two vertices are connected by an edge with probability p if they are

Received 16 August 2016; revision received 9 January 2018.
* Postal address: Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. Email address: b-hajek@illinois.edu
** Postal address: Department of Statistics and Data Science, Yale University, New Haven, CT 06511, USA.
*** Postal address: Krannert School of Management, Purdue University, West Lafayette, IN 47907, USA.

both in C^*, and with probability q otherwise, with the outcomes being mutually independent for distinct pairs of vertices.

The terminology is motivated by the fact that the subgraph induced by the community C^* is typically denser than the rest of the graph if $p > q$; see [4], [7], [14], [27], and [30]. The problem of interest is to recover C^* based on the graph G.

We consider a sequence of planted dense subgraphs indexed by n and assume that p and q depend on n. For a given n, the set C^* could be deterministic or random. We also introduce $K \geq 1$ depending on n, and assume either that $|C^*| \equiv K$ or $|C^*|/K \to 1$ in probability as $n \to \infty$. Where it matters, we specify which assumption holds. Since the focus of this paper is to understand the fundamental limits of recovering the hidden community in the planted dense subgraph model, we assume the model parameters (K, p, q) are known to the estimators. It remains an open question as to whether this assumption can be relaxed without changing the fundamental limits of recovery. Decelle *et al.* [9] suggested a method for estimating the parameters but it is unclear how to incorporate it into our theorems. For simplicity, we further impose the mild assumptions that K/n is bounded away from 1 and p/q is bounded from above. We primarily focus on two types of recovery guarantees.

Definition 1.2. (*Exact recovery.*) Given an estimator $\hat{C} = \hat{C}(G) \subset [n]$, $\hat{C}$ *exactly recovers* C^* if $\lim_{n\to\infty} \mathbb{P}\{\hat{C} \neq C^*\} = 0$, where the probability is taken with respect to the randomness of G and with respect to possible randomness in C^* and the algorithm for generating $\hat{C}$ from G.

Depending on the application, it may be enough to seek an estimator $\hat{C}$ which almost completely agrees with C^*.

Definition 1.3. (*Weak recovery.*) Given an estimator $\hat{C} = \hat{C}(G) \subset [n]$, $\hat{C}$ *weakly recovers* C^* if, as $n \to \infty$, $(1/K)|\hat{C} \triangle C^*| \to 0$, where the convergence is in probability and $\triangle$ denotes the set difference.

Exact and weak recovery are the same as strong and weak consistency, respectively, as defined in [32]. Clearly, an estimator that exactly recovers C^* also weakly recovers C^*. Also, it is not difficult to show that the existence of an estimator satisfying Definition 1.3 is equivalent to the existence of an estimator such that $\mathbb{E}[|\hat{C} \triangle C^*|] = o(K)$; see [16, Appendix A] for a proof.

Intuitively, if the community size K decreases, or p and q get closer, recovery of the community becomes more difficult. A critical role is played by the parameter

$$\lambda = \frac{K^2(p-q)^2}{(n-K)q}, \tag{1.1}$$

which can be interpreted as the effective signal-to-noise ratio for classifying a vertex according to its degree. It turns out that if the community size scales *linearly* with the network size, optimal recovery can be achieved via degree-thresholding in linear time. For example, if $K \asymp n - K \asymp n$ and p/q is bounded, a naïve degree-thresholding algorithm can attain weak recovery in linear time in the number of edges, provided that $\lambda \to \infty$, which is information-theoretically necessary when p is bounded away from 1. Moreover, one can show that degree-thresholding followed by a linear-time voting procedure achieves exact recovery whenever it is information-theoretically possible in this asymptotic regime; see Appendix A for a proof.

Since it is easy to recover a hidden community of size $K = \Theta(n)$ weakly or exactly up to the information limits, we next turn to the *sublinear* regime where $K = o(n)$. However, detecting and recovering polynomially small communities of size $K = n^{1-\Theta(1)}$ is known (see [14]) to suffer a fundamental computational barrier; see Section 2 for details. In the search for the

critical point where statistical and computational limits depart, the main focus of this paper is in the slightly sublinear regime of $K = n^{1-o(1)}$ and $np = n^{o(1)}$ and analysis of the belief propagation (BP) algorithm for community recovery.

The BP algorithm is an iterative algorithm which aggregates the likelihoods computed in the previous iterations with the observations in the current iteration. Running BP for one iteration and then thresholding the beliefs reduces to degree thresholding. Montanari [30] analyzed the performance of the BP algorithm for community recovery in a different regime with $p = a/n$, $q = b/n$, and $K = \kappa n$, where a, b, and κ are assumed to be fixed as $n \to \infty$. In the limit where first $n \to \infty$, and then $\kappa \to 0$ and $a, b \to \infty$, it was shown that using a local algorithm, namely BP running for a constant number of iterations, $\mathbb{E}[|\hat{C}\Delta C^*|] = o(n)$; conversely, if $\lambda < 1/\mathrm{e}$, for all local algorithms, $\mathbb{E}[|\hat{C}\Delta C^*|] = \Omega(n)$. However, since we focus on $K = o(n)$ and weak recovery demands $\mathbb{E}[|\hat{C}\Delta C^*|] = o(K)$, the following question remains unresolved: is $\lambda > 1/\mathrm{e}$ the performance limit of BP algorithms for weak recovery when $K = o(n)$? With regard to the local algorithm, loosely speaking, an algorithm is t-local if the computations determining the status of any given vertex u depend only on the subgraph induced by vertices whose distance to u is at most t; see [30] for a formal definition. In this paper, t is allowed to slowly grow with n so long as $(2 + np)^t = n^{o(1)}$.

In this paper we answer positively this question by analyzing BP running for $\log^* n + O(1)$ iterations. Here $\log^*(n)$ is the iterated logarithm, defined as the number of times the logarithm function must be iteratively applied to n to obtain a result less than or equal to 1. We show that if $\lambda > 1/\mathrm{e}$, weak recovery can be achieved by a BP algorithm running for $\log^*(n) + O(1)$ iterations, whereas if $\lambda < 1/\mathrm{e}$, all local algorithms including BP cannot asymptotically outperform trivial random guessing without the observation of the graph.

The proof is based on analyzing the analogous BP algorithm to classify the root node of a multitype Galton–Watson tree, which is the limit in distribution of the neighborhood of a given vertex in the original graph G. In contrast to the analysis of BP in [30], where the number of iterations was held fixed regardless of the size of graph n, our analysis on the tree and the associated coupling lemmas entail the number of iterations converging slowly to ∞ as the size of the graph increases, in order to guarantee adequate performance of the algorithm in the case that $K = o(n)$. Also, our analysis is based mainly on studying the recursions of exponential moments of beliefs instead of Gaussian approximations as used in [30].

Furthermore, we analyze a linear message-passing algorithm corresponding to applying the power method to the *nonbacktracking matrix* of the graph (see [6] and [25]) whose spectrum has been shown to be more informative than that of the adjacency matrix for the purpose of clustering. It is established that this linear message-passing algorithm followed by thresholding provides weak recovery if $\lambda > 1$ and it does not improve upon trivial random guessing asymptotically if $\lambda < 1$.

As shown in Remark 4.1, the threshold $\lambda = 1$ coincides with the Kesten–Stigum threshold (see [23] and [31]), which originated in the study of phase transitions of limiting offspring distributions of multitype Galton–Watson trees. Since the local neighborhood of a given vertex under stochastic block models is a multitype Galton–Watson tree in the limit, the Kesten–Stigum threshold also plays a critical role in the study of community detection. Dacelle *et al.* [9] first conjectured and later rigorously proved that for stochastic block models with two equal-sized planted communities, recovering a community partition positively correlated with the planted one is efficiently attainable if it is above the Kesten–Stigum threshold (see [6], [26], and [34]), while it is information-theoretically impossible if below the threshold; see [33]. With more than three equal-sized communities, correlated recovery has been shown

to be information-theoretically possible beyond the Kesten–Stigum threshold (see [1] and [5]); however, a conjecture of [9], that no polynomial-time algorithm can succeed in correlated recovery beyond the Kesten–Stigum threshold, still stands. In contrast, we show that in the case of a single hidden community, the BP algorithm achieves weak recovery efficiently beyond the Kesten–Stigum threshold by a factor of e. The problems mentioned above with equal-sized communities are balanced in the sense that the expected degree of a vertex given its community label is the same for all community labels. The single community problem we study is unbalanced; vertex degrees reveal information on vertex community labels. Hence, our results do not disprove that the Kesten–Stigum threshold is the limit for computationally tractable algorithms in the balanced case.

Finally, we address exact recovery. As shown in [16, Theorem 3], if there is an algorithm that can provide weak recovery even if the community size is random and only approximately equal to K, then it can be combined with a linear-time voting procedure to achieve exact recovery whenever it is information-theoretically possible. For $K = o(n)$, we show that both the BP and the linear message-passing algorithms indeed can be upgraded to achieve exact recovery via local voting. Somewhat surprisingly, BP plus voting achieves the information limit of exact recovery if

$$K \geq \frac{n}{\log n}\left(\rho_{\mathrm{BP}}\left(\frac{p}{q}\right) + o(1)\right),$$

where

$$\rho_{\mathrm{BP}}(c) \coloneqq \frac{1}{\mathrm{e}(c-1)^2}\left(1 - \frac{c-1}{\log c}\log\frac{\mathrm{e}\log c}{c-1}\right).$$

2. Related work

The problem of recovering a single community demonstrates a fascinating interplay between statistics and computation and a potential departure between computational and statistical limits.

In the special case of $p = 1$ and $q = \frac{1}{2}$, the problem of finding one community reduces to the classical planted clique problem [21]. If the clique has size $K \leq 2(1-\varepsilon)\log_2 n$ for any $\varepsilon > 0$, then it cannot be uniquely determined; if $K \geq 2(1+\varepsilon)\log_2 n$, an exhaustive search finds the clique with high probability. In contrast, polynomial-time algorithms are only known to find a clique of size $K \geq c\sqrt{n}$ for any constant $c > 0$ (see [2], [3] [10], and [13]), and it was shown in [11] that if $K \geq (1+\varepsilon)\sqrt{n/\mathrm{e}}$, the clique can be found in $O(n^2 \log n)$-time with high probability and $\sqrt{n/\mathrm{e}}$ may be a fundamental limit for solving the planted clique problem in nearly linear time in the number of edges in the graph. Recent work by Metra *et al.* [28] showed that the degree-r sum-of-squares (SOS) relaxation cannot find the clique unless $K \gtrsim (\sqrt{n}/\log n)^{1/r}$; an improved lower bound $K \gtrsim n^{1/3}/\log n$ for the degree-4 SOS was proved in [12]. Further improved lower bounds can be found in [20] and [36].

The recent work of Hajek *et al.* [14] focused on the $p = n^{-\alpha}$, $q = cn^{-\alpha}$ case for fixed constants $c < 1$ and $0 < \alpha < 1$, and $K = \Theta(n^{\beta})$ for $0 < \beta < 1$. It was shown that no polynomial-time algorithm can attain the information-theoretic threshold of detecting the planted dense subgraph unless the planted clique problem can be solved in polynomial time; see [14, Hypothesis 1] for the precise statement. For exact recovery, the maximum likelihood estimator succeeds with high probability if $\alpha < \beta < \frac{1}{2} + \frac{1}{4}\alpha$; however, no randomized polynomial-time solver exists, conditioned on the same planted clique hardness hypothesis.

In sharp contrast to the computational barriers discussed in the previous two paragraphs, in the regime $p = a\log n/n$ and $q = b\log n/n$ for fixed a and b, and $K = \rho n$ for a fixed constant $0 < \rho < 1$, recent work by Hajek *et al.* [15] derived a function $\rho^*(a, b)$ such that if $\rho > \rho^*$,

exact recovery is achievable in polynomial time via semidefinite programming relaxations of the maximum likelihood estimation; if $\rho < \rho^*$, any estimator fails to exactly recover the cluster with probability tending to 1 regardless of the computational costs.

In summary, the previous work revealed that for exact recovery, a significant gap between the information limit and the limit of polynomial-time algorithms emerges as the community size K decreases from $K = \Theta(n)$ to $K = n^\beta$ for $0 < \beta < 1$. In the search of the exact phase transition point where information and computational limits depart, in this paper we focus on the regime $K = n^{1-o(1)}$. In Appendix B, we show that BP plus voting attains the sharp information limit if $K \geq (n/\log n)(\rho_{\mathrm{BP}}(p/q) + o(1))$. However, as soon as $\lim_{n\to\infty} K \log n/n \leq \rho_{\mathrm{BP}}(p/q)$, we observe a gap between the information limit and the necessary condition of local algorithms, given by $\lambda > 1/\mathrm{e}$; see Figure 1 for an illustration. For weak recovery, as soon as $K = o(n)$, a gap between the information limit and the necessary condition of local algorithms emerges.

3. Main results

As mentioned above, in the search for the critical point where statistical and computational limits depart, we focus on the regime where K is slightly sublinear in n and invoke the following assumption.

Assumption 3.1. *As $n \to \infty$, $p \geq q$, $p/q = O(1)$, $n^{1-o(1)} \leq K \leq o(n)$, and the signal-to-noise ratio λ is a positive constant; see 1.1.*

3.1. Upper and lower bounds for BP

Let $\sigma \in \{0, 1\}^n$ denote the indicator vector of C^* and A denote the adjacency matrix of the graph G. To detect whether a given vertex i is in the community, a natural approach is to compare the log-likelihood ratio $\log(\mathbb{P}\{G \mid \sigma_i = 1\}/\mathbb{P}\{G \mid \sigma_i = 0\})$ to a certain threshold. However, it is often computationally expensive to evaluate the log-likelihood ratio. As we show in this paper, when the average degree scales as $n^{o(1)}$, the neighborhood of vertex i is tree-like with high probability as long as the radius t of the neighborhood satisfies $(2 + np)^t = n^{o(1)}$; moreover, on the tree, the log-likelihoods can be exactly computed in a finite recursion via BP. These two observations together suggest the following BP algorithm for approximately computing the log-likelihoods for the community recovery problem; see Lemma 4.1 for the derivation of the BP algorithm on a tree. Let ∂i denote the set of neighbors of i in G and

$$\nu := \log \frac{n - K}{K},$$

which is equal to the log prior ratio $\log(\mathbb{P}\{\sigma_i = 0\}/\mathbb{P}\{\sigma_i = 1\})$. Define the message transmitted from vertex i to its neighbor j at the $(t + 1)$th iteration as

$$R_{i\to j}^{t+1} = -K(p - q) + \sum_{\ell \in \partial i \setminus \{j\}} \log\left(\frac{\exp(R_{\ell\to i}^t - \nu)(p/q) + 1}{\exp(R_{\ell\to i}^t - \nu) + 1}\right) \tag{3.1}$$

for initial conditions $R_{i\to j}^0 = 0$ for all $i \in [n]$ and $j \in \partial i$. Then we approximate

$$\log\left(\frac{\mathbb{P}\{G \mid \sigma_i = 1\}}{\mathbb{P}\{G \mid \sigma_i = 0\}}\right)$$

by the belief of vertex i at the $(t+1)$th iteration R_i^{t+1}, which is determined by combining incoming messages from its neighbors as follows:

$$R_i^{t+1} = -K(p-q) + \sum_{\ell \in \partial i} \log\left(\frac{\exp(R_{\ell \to i}^t - \nu)(p/q) + 1}{\exp(R_{\ell \to i}^t - \nu) + 1}\right). \tag{3.2}$$

Algorithm 3.1. (*BP for weak recovery.*) The algorithm comprises five steps.

(i) Input: $n, K \in \mathbb{N}$. $p > q > 0$, adjacency matrix $A \in \{0,1\}^{n \times n}$, $t_f \in \mathbb{N}$.

(ii) Initialize: set $R_{i \to j}^0 = 0$ for all $i \in [n]$ and $j \in \partial i$.

(iii) Run $t_f - 1$ iterations of BP as in (3.1) to compute $R_{i \to j}^{t_f - 1}$ for all $i \in [n]$ and $j \in \partial i$.

(iv) Compute $R_i^{t_f}$ for all $i \in [n]$ as per (3.2).

(v) Return $\hat{C}$, the set of K indices in $[n]$ with largest values of $R_i^{t_f}$.

Theorem 3.1. *Suppose that Assumption 3.1 holds with $\lambda > 1/\mathrm{e}$ and $(np)^{\log^* \nu} = n^{o(1)}$. Let $t_f = \bar{t}_0 + \log^*(\nu) + 2$, where $\bar{t}_0$ is a constant depending only on λ. Let $\hat{C}$ be produced by Algorithm 3.1. If the planted dense subgraph model (Definition 1.1) is such that $|C^*| \equiv K$ then for any constant $r > 0$, there exists $\nu_0(r)$ such that for all $\nu \geq \nu_0(r)$,*

$$\mathbb{E}[|C^* \triangle \hat{C}|] \leq n^{o(1)} + 2K\mathrm{e}^{-\nu r}. \tag{3.3}$$

If, instead, $|C^*|$ is random with $\mathbb{P}\{||C^*| - K| \geq \sqrt{3K \log n}\} \leq n^{-1/2+o(1)}$ then

$$\mathbb{E}[|C^* \triangle \hat{C}|] \leq n^{1/2+o(1)} + 2K\mathrm{e}^{-\nu r}. \tag{3.4}$$

For either assumption about $|C^*|$, weak recovery is achieved: $\mathbb{E}[|C^* \triangle \hat{C}|] = o(K)$. The running time is $O(|E(G)| \log^* n)$, where $|E(G)|$ is the number of edges in the graph G.

We remark that the same conclusion also holds for the estimator $\hat{C}_o = \{i : R_i^{t_f} \geq \nu\}$, but returning a constant size estimator $\hat{C}$ leads to a simpler analysis of the algorithm for exact recovery.

Next we discuss how to use the BP algorithm to achieve exact recovery. The key idea is to attain exact recovery in two steps. In the first step, we apply BP for weak recovery. In the second step, we use a linear-time local voting procedure to clean-up the residual errors made by BP. In particular, for each vertex i, we count r_i the number of neighbors in the community estimated by BP, and pick the set of K vertices with the largest values of r_i. To facilitate the analysis, we adopt the *successive withholding* method described in [16] and [32] to ensure the first and second steps are independent of each other. In particular, we first randomly partition the set of vertices into a finite number of subsets. One at a time, one subset is withheld to produce a reduced set of vertices, to which BP is applied. The estimate obtained from the reduced set of vertices is used to classify the vertices in the withheld subset. The idea is to gain independence: the outcome of BP based on the reduced set of vertices is independent of the data corresponding to edges between the withheld vertices and the reduced set of vertices. The full description of the process is given in Algorithm 3.2.

Algorithm 3.2. (*BP plus clean-up for exact recovery.*) The algorithm comprises four steps.

(i) Input: $n \in \mathbb{N}$, $K > 0$, $p > q > 0$, adjacency matrix $A \in \{0,1\}^{n \times n}$, $t_f \in \mathbb{N}$, and $\delta \in (0,1)$ with $1/\delta, n\delta \in \mathbb{N}$.

(ii) Partition: partition $[n]$ into $1/\delta$ subsets S_k of size $n\delta$, uniformly at random.

(iii) Approximate recovery: for each $k = 1, \dots, 1/\delta$, let A_k denote the restriction of A to the rows and columns with index in $[n] \setminus S_k$, run Algorithm 3.1 (BP for weak recovery) with input $(n(1-\delta), \lceil K(1-\delta) \rceil, p, q, A_k, t_f)$, and let $\hat{C}_k$ denote the output.

(iv) Clean-up: for each $k = 1, \dots, 1/\delta$, compute $r_i = \sum_{j \in \hat{C}_k} A_{ij}$ for all $i \in S_k$ and return $\tilde{C}$, the set of K indices in $[n]$ with the largest values of r_i.

Theorem 3.2. *Suppose that Assumption 3.1 holds with* $\lambda > 1/\mathrm{e}$ *and* $(np)^{\log^* \nu} = n^{o(1)}$. *Consider the planted dense subgraph model (Definition 1.1) with* $|C^*| \equiv K$. *Select* $\delta > 0$ *so small that* $(1-\delta)\lambda \mathrm{e} > 1$. *Let* $t_f = \bar{t}_0 + \log^*(\nu) + 2$, *where* $\bar{t}_0$ *is a constant depending only on* $\lambda(1-\delta)$. *Also, suppose that* p *is bounded away from* 1 *and the following condition is satisfied:*

$$\liminf_{n \to \infty} \frac{K d(\tau^* \| q)}{\log n} > 1, \tag{3.5}$$

where

$$\tau^* = \frac{\log((1-q)/(1-p)) + (1/K)\log(n/K)}{\log(p(1-q)/q(1-p))} \tag{3.6}$$

and $d(p\|q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ *denotes the Kullback–Leibler divergence between Bernoulli distributions with mean* p *and* q. *Let* $\tilde{C}$ *be produced by Algorithm 3.2. Then* $\mathbb{P}\{\tilde{C} = C^*\} \to 1$ *as* $n \to \infty$. *The running time is* $O(|E(G)| \log^* n)$.

Note that condition (3.5) was shown in [16] to be the necessary (if '>' is replaced by '$\geq$') and sufficient condition for the success of the clean-up procedure in upgrading weak recovery to exact recovery.

Next, we provide a lower bound on the error probability achievable by any local algorithm for estimating the label σ_u of a given vertex u. Let $p_{\mathrm{err}} = \pi_0 p_{\mathrm{err},0} + \pi_1 p_{\mathrm{err},1}$ for prior probabilities $\pi_0 = (n-K)/n$ and $\pi_1 = K/n$, where $p_{\mathrm{err},0} = \mathbb{P}\{\hat{\sigma}_u = 1 \mid \sigma_u = 0\}$ and $p_{\mathrm{err},1} = \mathbb{P}\{\hat{\sigma}_u = 0 \mid \sigma_u = 1\}$.

Theorem 3.3. (Converse for local algorithms.) *Suppose that Assumption 3.1 holds with* $0 < \lambda \leq 1/\mathrm{e}$. *Let* $t_f \in \mathbb{N}$ *depend on* n *such that* $(2+np)^{t_f} = n^{o(1)}$. *Consider the planted dense subgraph model (Definition 1.1) with* C^* *random and uniformly distributed over all subsets of* $[n]$ *such that* $|C^*| \equiv K$. *Then, for any estimator* $\hat{C}$ *such that for each vertex* u *in* G, σ_u *is estimated based on* G *in a neighborhood of radius* t_f *from* u. *Then*

$$\mathbb{E}[|\hat{C} \triangle C^*|] \geq \frac{K(n-K)}{n} \mathrm{e}^{-\lambda \mathrm{e}/4} - n^{o(1)} \tag{3.7}$$

and

$$p_{\mathrm{err},0} + p_{\mathrm{err},1} \geq \tfrac{1}{2}\mathrm{e}^{-1/4} - n^{-1+o(1)}. \tag{3.8}$$

Furthermore, $\liminf_{n\to\infty} np_{\mathrm{err}}/K \geq 1$ *or, equivalently,*

$$\liminf_{n \to \infty} \frac{\mathbb{E}[|\hat{C} \triangle C^*|]}{K} \geq 1. \tag{3.9}$$

The assumption $(2+np)^{t_f} = n^{o(1)}$ is needed to ensure that the neighborhood of radius t_f from any given vertex u is a tree with high probability.

Note that an estimator is said to achieve weak recovery (see [30]) if

$$\lim_{n\to\infty} p_{\mathrm{err},0} + p_{\mathrm{err},1} = 0.$$

From condition (3.8) we see that weak recovery in this sense is not possible. If C^* is uniformly distributed over $\{C \subset [n]\colon |C| = K\}$, among all estimators that disregard the graph, the one that minimizes the mean number of classification errors is $\hat{C} \equiv \varnothing$ (declaring no community), which achieves $\mathbb{E}[|\hat{C}\triangle C^*|]/K = 1$ or, equivalently, $p_{\mathrm{err}} = K/n$. From condition (3.9) we see that in the asymptotic regime $\nu \to \infty$ with $\lambda < 1/\mathrm{e}$, improving upon random guessing is impossible.

3.2. Upper and lower bounds for linear message passing

Results are given in this section to show that a particular spectral method—linear message passing—achieves weak recovery if and only if $\lambda > 1$. Spectral algorithms are used to estimate the communities based on the principal eigenvectors of the adjacency matrix; see, e.g. [2], [27], and [37] and the references therein. Under the single community model, $\mathbb{E}[A] = (p - q)(\sigma\sigma^\top - \mathrm{diag}\{\sigma\}) + q(\boldsymbol{J} - \boldsymbol{I})$, where $\mathrm{diag}\{\sigma\}$ denotes the diagonal matrix with the diagonal entries given by σ; $\boldsymbol{I}$ denotes the identity matrix, and $\boldsymbol{J}$ denotes the matrix of all 1s. By the Davis–Kahan $\sin\theta$ theorem (see [8]), the principal eigenvector of $A - q(\boldsymbol{J} - \boldsymbol{I})$ is almost parallel to σ provided that the spectral norm $\|A - \mathbb{E}[A]\|$ is much smaller than $K(p-q)$; thus, one can estimate C^* by thresholding the principal eigenvector entry-wise. Therefore, if we apply the spectral method, a natural matrix to start with is $A - q(\boldsymbol{J} - \boldsymbol{I})$ or $A - q\boldsymbol{J}$. Finding the principal eigenvector of $A - q\boldsymbol{J}$ according to the power method is carried out by starting with some vector and repeatedly multiplying by $A - q\boldsymbol{J}$ sufficiently many times. We will consider the scaled matrix $(A - q\boldsymbol{J})/\sqrt{m}$, where $m = (n-K)q$. Of course, the scaling does not change the eigenvectors. This suggests the following linear message-passing update equation:

$$\theta_i^{t+1} = -\frac{q}{\sqrt{m}} \sum_{\ell\in[n]} \theta_\ell^t + \frac{1}{\sqrt{m}} \sum_{\ell\in\partial i} \theta_\ell^t. \tag{3.10}$$

The first sum is over all vertices in the graph and does not depend on i. An idea is to appeal to the law of large numbers and replace the first sum by its expectation. Also, in the sparse graph regime $np = o(\log n)$, there exist vertices of high degrees $\omega(np)$, and the spectrum of A is very sensitive to high-degree vertices; see, e.g. [15, Appendix A] for a proof. To deal with this issue, as proposed in [6] and [25], we associate the messages in (3.10) with directed edges and prevent the message transmitted from j to i from being immediately reflected back as a term in the next message from i to j, resulting in the following linear message-passing algorithm:

$$\theta_{i\to j}^{t+1} = -\frac{q((n-K)A_t + KB_t)}{\sqrt{m}} + \frac{1}{\sqrt{m}} \sum_{\ell\in\partial i\setminus\{j\}} \theta_{\ell\to i}^t \tag{3.11}$$

with initial values $\theta_{\ell\to i}^0 = 1$, where $A_t \approx \mathbb{E}[\theta_{\ell\to i}^t \mid \sigma_\ell = 0]$ and $B_t \approx \mathbb{E}[\theta_{\ell\to i}^t \mid \sigma_\ell = 1]$. Note that when computing $\theta_{i\to j}^{t+1}$, the contribution of $\theta_{j\to i}^t$ is subtracted out. Since we focus on the regime $np = n^{o(1)}$, the graph is locally tree-like with high probability. In the Poisson random tree limit of the neighborhood of a vertex, the expectations $\mathbb{E}[\theta_{\ell\to i}^t \mid \sigma_\ell = 0]$ and $\mathbb{E}[\theta_{\ell\to i}^t \mid \sigma_\ell = 1]$ can be calculated exactly, and as a result we take $A_0 = 1$, $A_t = 0$ for $t \geq 1$, and $B_t = \lambda^{t/2}$ for $t \geq 0$.

The update equation (3.11) can be expressed in terms of the *nonbacktracking matrix* associated with graph G. It is the matrix $\boldsymbol{B} \in \{0,1\}^{2m\times 2m}$ with $B_{ef} = \mathbf{1}_{\{e_2=f_1\}}\mathbf{1}_{\{e_1\neq f_2\}}$, where

$e = (e_1, e_2)$, $f = (f_1, f_2)$ are directed edges, and $\mathbf{1}_{\{A\}}$ is the indicator function on the event A. Let $\Theta^t \in \mathbb{R}^{2m}$ denote the messages on directed edges with $\Theta^t_e = \theta^t_{e_1 \to e_2}$. Then (3.11) in matrix form reads

$$\Theta^{t+1} = -\frac{q((n-K)A_t + KB_t)}{\sqrt{m}}\mathbf{1} + \frac{1}{\sqrt{m}}\boldsymbol{B}^\top \Theta^t.$$

As demonstrated by Borndenave *et al.* [6], the spectral properties of the nonbacktracking matrix closely match those of the original adjacency matrix. It is therefore reasonable to take the linear update equation (3.11) as a form of spectral method for the community recovery problem. Finally, to estimate C^*, we define the belief at vertex u as

$$\theta^{t+1}_u = -\frac{q((n-K)A_t + KB_t)}{\sqrt{m}} + \frac{1}{\sqrt{m}}\sum_{i \in \partial u} \theta^t_{i \to u}, \tag{3.12}$$

and select the vertices u such that θ^t_u exceeds a certain threshold. The full description of the algorithm is given in Algorithm 3.3.

Algorithm 3.3. (*Spectral algorithm for weak recovery.*) The algorithm comprises six steps.

(i) Input: n, $K \in \mathbb{N}$, $p > q > 0$, and adjacency matrix $A \in \{0, 1\}^{n \times n}$.

(ii) Set

$$\lambda = \frac{K^2(p-q)^2}{(n-K)q} \quad \text{and} \quad T = \left\lceil 2\alpha \frac{\log((n-K)/K)}{\log \lambda} \right\rceil,$$

where $\alpha = \frac{1}{4}$ (in fact any $\alpha < 1$ works).

(iii) Initialize: set $\theta^0_{i \to j} = 1$ for all $i \in [n]$ and $j \in \partial i$.

(iv) Run $T-1$ iterations of message passing as in (3.11) to compute $\theta^{T-1}_{i \to j}$ for all $i \in [n]$ and $j \in \partial i$.

(v) Run one more iteration of message passing to compute θ^T_i for all $i \in [n]$ as per (3.12).

(vi) Return $\hat{C}$, the set of K indices in $[n]$ with largest values of θ^T_i.

Theorem 3.4. *Suppose that Assumption 3.1 holds with* $\lambda > 1$ *and* $(np)^{\log(n/K)} = n^{o(1)}$. *Consider the planted dense subgraph model (Definition 1.1) with*

$$\mathbb{P}\{||C^*| - K| \geq \sqrt{3K \log n}\} \leq n^{-1/2+o(1)}.$$

Let $\hat{C}$ *be the estimator produced by Algorithm 3.3. Then* $\mathbb{E}[|C^* \triangle \hat{C}|] = o(K)$.

One can upgrade the weak recovery result of linear message passing to exact recovery under condition $\lambda > 1$ and condition (3.5) in a similar manner as described in Algorithm 3.2 and the proof of Theorem 3.2.

The following theorem is the converse and we show that if $\lambda \leq 1$ then improving the estimate beyond the random guessing by linear message passing is not possible.

Theorem 3.5. (Converse for the linear-message passing algorithm.) *Suppose that Assumption 3.1 holds with* $0 < \lambda \leq 1$ *and consider the planted dense subgraph model (Definition 1.1) with* C^* *random and uniformly distributed over all subsets of* $[n]$ *such that* $|C^*| \equiv K$. *Assume that* $t \in \mathbb{N}$, *with* t *possibly depending on* n *such that* $(np)^t = n^{o(1)}$ *and* $t = O(\log((n-K)/K))$. *Let* $(\theta^t_u : u \in [n])$ *be computed using the message passing updates (3.11) and (3.12) and*

let $\hat{C} = \{u : \theta_u^t \geq \gamma\}$ *for some threshold* γ*, which may also depend on* n*. Equivalently,* σ_u *is estimated for each* u *by* $\hat{\sigma}_u = \mathbf{1}_{\{\theta_u^t \geq \gamma\}}$*. Then* $\liminf_{n\to\infty} p_{\mathrm{err}} n/K \geq 1$.

The proofs of Theorems 3.4 and 3.5 are similar to the counterparts for BP and can be found in the arXiv version of this paper [17].

4. Inference problem on a random tree by BP

In the regime we consider, the graph is locally tree-like with mean degree converging to ∞. We begin by deriving the exact BP algorithm for an infinite tree network, and then deduce performance results when using that algorithm on the original graph.

The related inference problem on a Galton–Watson tree with a Poisson number of offspring is defined as follows. Fix a vertex u and let T_u denote the infinite Galton–Watson undirected tree rooted at vertex u. The neighbors of vertex u are considered to be the children of vertex u, and u is the parent of those children. The other neighbors of each child are the children of the child, and so on. For vertex i in T_u, let T_i^t denote the subtree of T_u of height t rooted at vertex i, induced by the set of vertices consisting of vertex i and its descendants for t generations. Let $\tau_i \in \{0, 1\}$ denote the label of vertex i in T_u. Assume that $\tau_u \sim \mathrm{Bernoulli}(K/n)$. For any vertex $i \in T_u$, let L_i denote the number of its children j with $\tau_j = 1$, and M_i denote the number of its children j with $\tau_j = 0$. Suppose that $L_i \sim \mathrm{Poisson}(Kp)$ if $\tau_i = 1$, $L_i \sim \mathrm{Poisson}(Kq)$ if $\tau_i = 0$, and $M_i \sim \mathrm{Poisson}((n-K)q)$ for either value of τ_i.

We are interested in estimating the label of root u given an observation of the tree T_u^t. Note that the labels of vertices in T_u^t are not observed. The probability of error for an estimator $\hat{\tau}_u(T_u^t)$ is defined by

$$p_{\mathrm{err}}^t := \frac{K}{n}\mathbb{P}(\hat{\tau}_u = 0 \mid \tau_u = 1) + \frac{n-K}{n}\mathbb{P}(\hat{\tau}_u = 1 \mid \tau_u = 0).$$

The estimator that minimizes p_{err}^t is the maximum *a posteriori* probability (MAP) estimator, which can be expressed either in terms of the log-belief ratio or log-likelihood ratio:

$$\hat{\tau}_{\mathrm{MAP}} = \mathbf{1}_{\{\xi_u^t \geq 0\}} = \mathbf{1}_{\{\Lambda_u^t \geq \nu\}}, \tag{4.1}$$

where

$$\xi_u^t := \log \frac{\mathbb{P}\{\tau_u = 1 \mid T_u^t\}}{\mathbb{P}\{\tau_u = 0 \mid T_u^t\}}, \qquad \Lambda_u^t := \log \frac{\mathbb{P}\{T_u^t \mid \tau_u = 1\}}{\mathbb{P}\{T_u^t \mid \tau_u = 0\}}, \qquad \nu = \log \frac{n-K}{K}.$$

From Bayes' formula, $\xi_u^t = \Lambda_u^t - \nu$ and, by definition, $\Lambda_u^0 = 0$. By a standard result in the theory of binary hypothesis testing (due to [24], stated without proof in [35], proved in the special case $\pi_0 = \pi_1 = \frac{1}{2}$ in [22], and the same proof easily extends to the general case), the probability of error for the MAP decision rule is bounded by

$$\pi_1\pi_0\rho_{\mathrm{B}}^2 \leq p_{\mathrm{err}}^t \leq \sqrt{\pi_1\pi_0}\rho_{\mathrm{B}}, \tag{4.2}$$

where the Bhattacharyya coefficient (or Hellinger integral) ρ_{B} is defined by

$$\rho_{\mathrm{B}} = \mathbb{E}[\mathrm{e}^{\Lambda_u^t/2} \mid \tau_u = 0],$$

and π_1 and π_0 are the prior probabilities on the hypotheses.

We comment briefly on the parameters of the model. The distribution of the tree T_u is determined by the three parameters $\lambda = K^2(p-q)^2/(n-K)q$, ν, and the ratio p/q. Indeed, vertex u has label $\tau_u = 1$ with probability $K/n = 1/(1+\mathrm{e}^{\nu})$, and the mean number of children of a vertex i is given by

$$\mathbb{E}[L_i \mid \tau_i = 1] = Kp = \frac{\lambda(p/q)\mathrm{e}^{\nu}}{(p/q-1)^2}, \qquad \mathbb{E}[L_i \mid \tau_i = 0] = Kq = \frac{\lambda \mathrm{e}^{\nu}}{(p/q-1)^2}, \tag{4.3}$$
$$\mathbb{E}[M_i] = (n-K)q = \frac{\lambda \mathrm{e}^{2\nu}}{(p/q-1)^2}.$$

The parameter λ can be interpreted as a signal-to-noise ratio in case $K \ll n$ and $p/q = O(1)$, since $\operatorname{var} M_i \gg \operatorname{var} L_i$ and

$$\lambda = \frac{(\mathbb{E}[M_i + L_i \mid \tau_i = 1] - \mathbb{E}[M_i + L_i \mid \tau_i = 0])^2}{\operatorname{var} M_i}.$$

In this section, the parameters are allowed to vary with n as long as $\lambda > 0$ and $p/q > 1$, although the focus is on the asymptotic regime: λ fixed, $p/q = O(1)$, and $\nu \to \infty$. This entails that the mean number of children given in (4.3) converges to ∞. Montanari [30] considered the case of ν fixed with $p/q \to 1$, which also leads to the mean vertex degrees converging to ∞.

Remark 4.1. It turns out that $\lambda = 1$ coincides with the Kesten–Stigum threshold [23]. To see this, let $O = (O_{ab})$ denote the 2×2 matrix with O_{ab} equal to the expected number of children of type b given a parent of type a for $a, b \in \{0, 1\}$. Then

$$O = \begin{bmatrix} (n-K)q & Kq \\ (n-K)q & Kp \end{bmatrix}.$$

Let $\lambda_+ \geq \lambda_-$ denote the two largest eigenvalues of M. The Kesten–Stigum threshold [23] is defined to be $\lambda_-^2/\lambda_+ = 1$. A direct calculation yields

$$\lambda_\pm = \frac{1}{2}\left(nq + K(p-q) \pm |nq - K(p-q)|\sqrt{1 + \frac{4K^2(p-q)q}{(nq-K(p-q))^2}}\right).$$

Since $K(p-q) = o(nq)$ and $K = o(n)$, it follows that $\lambda_+ = (1+o(1))nq$ and $\lambda_- = (1+o(1))K(p-q)$. Hence,

$$\lambda = (1+o(1))\frac{\lambda_-^2}{\lambda_+}.$$

Thus, $\lambda = 1$ is asymptotically equivalent to the Kesten–Stigum threshold $\lambda_-^2/\lambda_+ = 1$.

It is well known that the likelihoods can be computed via a BP algorithm. Let ∂i denote the set of children of vertex i in T_u and $\pi(i)$ denote the parent of i. For every vertex $i \in T_u$ other than u, define

$$\Lambda^t_{i \to \pi(i)} := \log \frac{\mathbb{P}\{T_i^t \mid \tau_i = 1\}}{\mathbb{P}\{T_i^t \mid \tau_i = 0\}}.$$

In the following lemma we provide a recursive formula to compute Λ_u^t; no approximations are needed.

Lemma 4.1. *For $t \geq 0$,*

$$\Lambda_u^{t+1} = -K(p-q) + \sum_{\ell \in \partial u} \log\left(\frac{e^{\Lambda^t_{\ell\to u} - \nu}(p/q)+1}{e^{\Lambda^t_{\ell\to u}-\nu}+1}\right),$$

$$\Lambda_{i\to\pi(i)}^{t+1} = -K(p-q) + \sum_{\ell \in \partial i} \log\left(\frac{e^{\Lambda^t_{\ell\to i} - \nu}(p/q)+1}{e^{\Lambda^t_{\ell\to i}-\nu}+1}\right) \quad \textit{for all } i \neq u,$$

$$\Lambda_{i\to\pi(i)}^{0} = 0 \quad \textit{for all } i \neq u.$$

Proof. The last equation follows by definition. We prove the first equation; the second one follows similarly. A key point is to use the independent splitting property of the Poisson distribution to obtain an equivalent description of the number of children with each label for any vertex in the tree. Instead of separately generating the number of children of with each label, we can first generate the total number of children and then independently and randomly label each child. Specifically, for every vertex i in T_u, let N_i denote the total number of its children. Let $d_1 = Kp + (n-K)q$ and $d_2 = Kq + (n-K)q = nq$. If $\tau_i = 1$ then $N_i \sim \text{Poisson}(d_1)$, and for each child $j \in \partial i$, independently of everything else, $\tau_j = 1$ with probability Kp/d_1 and $\tau_j = 0$ with probability $(n-K)q/d_1$. If $\tau_i = 0$ then $N_i \sim \text{Poisson}(d_0)$, and for each child $j \in \partial i$, independently of everything else, $\tau_j = 1$ with probability K/n and $\tau_j = 0$ with probability $(n-K)/n$. With this view, the observation of the total number of children N_u of vertex u gives some information on the label of u, and then the conditionally independent messages from those children yield additional information. To be precise, we have

$$\begin{aligned}
\Lambda_u^{t+1} &= \log \frac{\mathbb{P}\{T_u^{t+1} \mid \tau_u = 1\}}{\mathbb{P}\{T_u^{t+1} \mid \tau_u = 0\}} \\
&\overset{(a)}{=} \log \frac{\mathbb{P}\{N_u \mid \tau_u = 1\}}{\mathbb{P}\{N_u \mid \tau_u = 0\}} + \sum_{i\in\partial u} \log \frac{\mathbb{P}\{T_i^t \mid \tau_u = 1\}}{\mathbb{P}\{T_i^t \mid \tau_u = 0\}} \\
&\overset{(b)}{=} -K(p-q) + N_u \log \frac{d_1}{d_0} + \sum_{i\in\partial u} \log \frac{\sum_{x\in\{0,1\}} \mathbb{P}\{\tau_i = x \mid \tau_u = 1\}\mathbb{P}\{T_i^t \mid \tau_i = x\}}{\sum_{\tau_i\in\{0,1\}} \mathbb{P}\{\tau_i = x \mid \tau_u = 0\}\mathbb{P}\{T_i^t \mid \tau_i = x\}} \\
&\overset{(c)}{=} -K(p-q) + \sum_{i\in\partial u} \log \frac{Kp\mathbb{P}\{T_i^t \mid \tau_i = 1\} + (n-K)q\mathbb{P}\{T_i^t \mid \tau_i = 0\}}{Kq\mathbb{P}\{T_i^t \mid \tau_i = 1\} + (n-K)q\mathbb{P}\{T_i^t \mid \tau_i = 0\}} \\
&\overset{(d)}{=} -K(p-q) + \sum_{i\in\partial u} \log \frac{e^{\Lambda^t_{i\to u} - \nu}(p/q)+1}{e^{\Lambda^t_{i\to u}-\nu}+1},
\end{aligned}$$

where (a) holds since N_u and T_i^t for $i \in \partial u$ are independent conditional on τ_u; (b) follows since $N_u \sim \text{Poisson}(d_1)$ if $\tau_u = 1$ and $N_u \sim \text{Poisson}(d_0)$ if $\tau_u = 0$, and T_i^t is independent of τ_u conditional on τ_i; (c) follows from the fact $\tau_i \sim \text{Bernoulli}(Kp/d_1)$ given $\tau_u = 1$, and $\tau_i \sim \text{Bernoulli}(Kq/d_0)$ given $\tau_u = 0$; (d) follows from the definition of $\Lambda^t_{i\to u}$. □

Note that Λ_u^t is a function of T_u^t alone; and it is statistically correlated with the vertex labels. Also, since the construction of a subtree T_i^t and its vertex labels are the same as the construction of T_u^t and its vertex labels, the conditional distribution of T_i^t given τ_i is the same as the conditional distribution of T_u^t given τ_u. Therefore, for any $i \in \partial u$, the conditional distribution of $\Lambda^t_{i\to u}$ given τ_i is the same as the conditional distribution of Λ_u^t given τ_u. For $i = 0$ or 1, let Z_i^t denote a random variable that has the same distribution as Λ_u^t given $\tau_u = i$.

The above update rules can be viewed as an infinite-dimensional recursion that determines the probability distribution of Z_0^{t+1} in terms of that of Z_0^t.

The remainder of this section is devoted to the analysis of BP on the Poisson tree model, and is organized into two main parts. In Section 4.1 we provide expressions for exponential moments of the log-likelihood messages, which are applied in Section 4.2 to yield an upper bound, in Lemma 4.8 on the error probability for the problem of classifying the root vertex of the tree. That bound, together with a standard coupling result between the Poisson tree and the local neighborhood of G (stated in Appendix C), is enough to establish weak recovery for the BP algorithm run on graph G, given in Theorem 3.1. In Section 4.3 we focus on lower bounds on the probability of correct classification. Those bounds, together with the coupling lemmas, are used to establish the converse results for local algorithms.

4.1. Exponential moments of log-likelihood messages for the Poisson tree

In the following lemma we provide formulas for some exponential moments of Z_0^t and Z_1^t, based on Lemma 4.1. Although the formulas are not recursions, they are close enough to permit useful analysis.

Lemma 4.2. *For $t \geq 0$ and any integer $h \geq 2$,*

$$\begin{aligned}\mathbb{E}[\mathrm{e}^{hZ_0^{t+1}}] &= \mathbb{E}[\mathrm{e}^{(h-1)Z_1^{t+1}}] \\ &= \exp\left(K(p-q)\sum_{j=2}^{h}\binom{h}{j}\left(\frac{\lambda}{K(p-q)}\right)^{j-1}\mathbb{E}\left[\left(\frac{\mathrm{e}^{Z_1^t}}{1+\mathrm{e}^{Z_1^t-\nu}}\right)^{j-1}\right]\right). \qquad (4.4)\end{aligned}$$

Proof. We first illustrate the proof for $h = 2$. By the definition of Λ_u^t and a change of measure, we have $\mathbb{E}[g(\Lambda_u^t) \mid \tau_u = 0] = \mathbb{E}[g(\Lambda_u^t)\mathrm{e}^{-\Lambda_u^t} \mid \tau_u = 1]$, where g is any measurable function such that the expectations above are well defined. It follows that

$$\mathbb{E}[g(Z_0^t)] = \mathbb{E}[g(Z_1^t)\mathrm{e}^{-Z_1^t}]. \qquad (4.5)$$

Substituting in for $g(z) = \mathrm{e}^z$ and $g(z) = \mathrm{e}^{2z}$, we have $\mathbb{E}[\mathrm{e}^{Z_0^t}] = 1$ and $\mathbb{E}[\mathrm{e}^{2Z_0^t}] = \mathbb{E}[\mathrm{e}^{Z_1^t}]$. Moreover,

$$\mathrm{e}^{\nu}\mathbb{E}[g(Z_0^t)] + \mathbb{E}[g(Z_1^t)] = \mathbb{E}[g(Z_1^t)(\mathrm{e}^{-Z_1^t+\nu}+1)].$$

Substituting $g(z) = (1+\mathrm{e}^{-z+\nu})^{-1}$ and $g(z) = (1+\mathrm{e}^{-z+\nu})^{-2}$ into the last displayed equation, we have

$$\mathrm{e}^{\nu}\mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_0^t+\nu}}\right] + \mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_1^t+\nu}}\right] = 1, \qquad (4.6)$$

$$\mathrm{e}^{\nu}\mathbb{E}\left[\frac{1}{(1+\mathrm{e}^{-Z_0^t+\nu})^2}\right] + \mathbb{E}\left[\frac{1}{(1+\mathrm{e}^{-Z_1^t+\nu})^2}\right] = \mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_1^t+\nu}}\right]. \qquad (4.7)$$

In view of Lemma 4.1, by defining $f(x) = (x(p/q)+1)/(x+1)$, we have

$$\mathrm{e}^{2\Lambda_u^{t+1}} = \mathrm{e}^{-2K(p-q)}\prod_{\ell\in\partial u} f^2(\mathrm{e}^{\Lambda_{\ell\to u}^t-\nu}).$$

Since the distribution of $\Lambda_{\ell\to u}^t$ conditional on $\tau_u = 0$ and $\tau_u = 1$ is the same as the distribution of Z_0^t and Z_1^t, respectively, it follows that

$$\mathbb{E}[\mathrm{e}^{2Z_0^{t+1}}] = \mathrm{e}^{-2K(p-q)}\mathbb{E}[(\mathbb{E}[f^2(\mathrm{e}^{Z_1^{t+1}-\nu})])^{L_u}]\mathbb{E}[(\mathbb{E}[f^2(\mathrm{e}^{Z_0^{t+1}-\nu})])^{M_u}].$$

Using the fact that $\mathbb{E}[c^X] = \mathrm{e}^{\lambda(c-1)}$ for $X \sim \text{Poisson}(\lambda)$ and $c > 0$, we have

$$\mathbb{E}[\mathrm{e}^{2Z_0^{t+1}}] = \exp(-2K(p-q) + Kq(\mathbb{E}[f^2(\mathrm{e}^{Z_1^{t+1}-\nu})] - 1) + (n-K)q(\mathbb{E}[f^2(\mathrm{e}^{Z_0^{t+1}-\nu})] - 1)).$$

Note that

$$f^2(x) = \left(1 + \frac{p/q - 1}{1 + x^{-1}}\right)^2 = 1 + \frac{2(p/q-1)}{1+x^{-1}} + \frac{(p/q-1)^2}{(1+x^{-1})^2}.$$

It follows that

$$\begin{aligned}
&Kq(\mathbb{E}[f^2(\mathrm{e}^{Z_1^{t+1}-\nu})] - 1) + (n-K)q(\mathbb{E}[f^2(\mathrm{e}^{Z_0^{t+1}-\nu})] - 1) \\
&\quad = 2Kq\left(\frac{p}{q} - 1\right)\left(\mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_1^t+\nu}}\right] + \mathrm{e}^{\nu}\mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_0^t+\nu}}\right]\right) \\
&\qquad + Kq\left(\frac{p}{q} - 1\right)^2\left(\mathbb{E}\left[\frac{1}{(1+\mathrm{e}^{-Z_1^t+\nu})^2}\right] + \mathrm{e}^{\nu}\mathbb{E}\left[\frac{1}{(1+\mathrm{e}^{-Z_0^t+\nu})^2}\right]\right) \\
&\quad \overset{\text{(a)}}{=} 2K(p-q) + Kq\left(\frac{p}{q} - 1\right)^2\mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_1^t+\nu}}\right] \\
&\quad = 2K(p-q) + \lambda\mathbb{E}\left[\frac{\mathrm{e}^{Z_1^t}}{1+\mathrm{e}^{Z_1^t-\nu}}\right],
\end{aligned}$$

where (a) follows by applying (4.6) and (4.7). Combining the above proves (4.4) with $h = 2$. For general $h \geq 2$, we expand $f^h(x) = (1+(p/q-1)/(1+x^{-1}))^h$ using binomial coefficients as already illustrated for $h = 2$. □

Using the notation

$$a_t = \mathbb{E}[\mathrm{e}^{Z_1^t}], \qquad b_t = \mathbb{E}\left[\frac{\mathrm{e}^{Z_1^t}}{1+\mathrm{e}^{Z_1^t-\nu}}\right], \tag{4.8}$$

(4.4) with $h = 2$ becomes

$$a_{t+1} = \mathrm{e}^{\lambda b_t}. \tag{4.9}$$

In the following lemma we provide upper bounds on some exponential moments in terms of b_t.

Lemma 4.3. *Let* $C := \lambda(2+p/q)$ *and* $C' := \lambda(3+2(p/q)+(p/q)^2)$. *Then* $\mathbb{E}[\exp(2Z_1^{t+1})] \leq \exp(Cb_t)$ *and* $\mathbb{E}[\exp(3Z_1^{t+1})] \leq \exp(C'b_t)$. *More generally, for any integer* $h \geq 2$,

$$\mathbb{E}[\exp(hZ_0^{t+1})] = \mathbb{E}[\exp((h-1)Z_1^{t+1})] \leq \exp\left(\lambda b_t \sum_{j=2}^{h}\binom{h}{j}\left(\frac{p}{q} - 1\right)^{j-2}\right). \tag{4.10}$$

Proof. Note that $\mathrm{e}^z/(1+\mathrm{e}^{z-\nu}) \leq \mathrm{e}^{\nu}$ for all z. Therefore, for any $j \geq 2$,

$$\left(\frac{\mathrm{e}^z}{1+\mathrm{e}^{z-\nu}}\right)^{j-1} \leq \mathrm{e}^{(j-2)\nu}\left(\frac{\mathrm{e}^z}{1+\mathrm{e}^{z-\nu}}\right).$$

Applying this inequality to (4.4) yields (4.10). □

4.2. Upper bound on the classification error via exponential moments

Note that $b_t \approx a_t$ if $\nu \gg 0$, in which case (4.9) is approximately a recursion for $\{b_t\}$. In the following two lemmas we use this intuition to show that if $\lambda > 1/\mathrm{e}$ and ν is large enough, the b_t eventually grow large. In turn, that fact will be used to show that the Bhattacharyya coefficient mentioned in (4.2), which can be expressed as $\rho_{\mathrm{B}} = \mathbb{E}[\mathrm{e}^{Z_0^t/2}] = \mathbb{E}[\mathrm{e}^{-Z_1^t/2}]$, becomes small, culminating in Lemma 4.8, yielding an upper bound on the classification error for the root vertex.

Lemma 4.4. *Let* $C := \lambda(2 + p/q)$. *Then*

$$b_{t+1} \geq \mathrm{e}^{\lambda b_t}(1 - \mathrm{e}^{-\nu/2}) \quad \text{if } b_t \leq \frac{\nu}{2(C-\lambda)}. \tag{4.11}$$

Proof. Note that $C - \lambda > 0$. If $b_t \leq \nu/2(C-\lambda)$, we have

$$b_{t+1} \overset{\text{(a)}}{\geq} a_{t+1} - \mathbb{E}[\mathrm{e}^{-\nu+2Z_1^{t+1}}] \overset{\text{(b)}}{\geq} \mathrm{e}^{\lambda b_t} - \mathrm{e}^{-\nu+Cb_t} = \mathrm{e}^{\lambda b_t}(1 - \mathrm{e}^{-\nu+(C-\lambda)b_t}) \overset{\text{(c)}}{\geq} \mathrm{e}^{\lambda b_t}(1 - \mathrm{e}^{-\nu/2}),$$

where (a) follows by the definitions in (4.8) and the fact $1/(1+x) \geq 1-x$ for $x \geq 0$; (b) follows from Lemma 4.3; (c) follows from the condition $b_t \leq \nu/2(C-\lambda)$. □

Lemma 4.5. *The variables* a_t *and* b_t *are nondecreasing in* t *and* $\mathbb{E}[\mathrm{e}^{Z_0^t/2}]$ *is nonincreasing in* t *over all* $t \geq 0$. *More generally,* $\mathbb{E}[\Upsilon(\mathrm{e}^{Z_0^t})]$ *is nondecreasing (nonincreasing) in* t *for any convex (concave) function* Υ *with domain* $(0, \infty)$.

Proof. Note that, in view of (4.5), $\mathbb{E}[\Upsilon(\mathrm{e}^{Z_0^t})]$ becomes a_t for the convex function $\Upsilon(x) = x^2$, b_t for the convex function $\Upsilon(x) = x^2/(1 + x\mathrm{e}^{-\nu})$, and $\mathbb{E}[\mathrm{e}^{Z_0^t/2}]$ for the concave function $\Upsilon(x) = \sqrt{x}$. It thus suffices to prove the last statement of the lemma.

It is well known that for a nonsingular binary hypothesis testing problem with a growing amount of information indexed by some parameter s (that is, an increasing family of σ-algebras as usual in martingale theory), the likelihood ratio $\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$ is a martingale under measure $\mathbb{Q}$. Therefore, the likelihood ratios $\{\mathrm{e}^{\Lambda_u^t} : t \geq 0\}$ (where Λ_s denotes the log-likelihood ratio) at the root vertex u for the infinite tree, conditioned on $\tau_u = 0$, form a martingale. Thus, the random variables $\{\mathrm{e}^{Z_0^t} : t \geq 0\}$ can be constructed on a single probability space to be a martingale. The lemma therefore follows from Jensen's inequality. □

Recall that $\log^*(\nu)$ denotes the number of times the logarithm function must be iteratively applied to ν to obtain a result less than or equal to 1.

Lemma 4.6. *Suppose that* $\lambda > 1/\mathrm{e}$. *There are constants* $\bar{t}_0$ *and* $\nu_o > 0$ *depending only on* λ *such that*

$$b_{\bar{t}_0 + \log^*(\nu) + 2} \geq \exp\left(\frac{\lambda\nu}{2(C-\lambda)}\right)\left(1 - \exp\left(-\frac{\nu}{2}\right)\right),$$

where $C = \lambda(p/q + 2)$, *whenever* $\nu \geq \nu_o$ *and* $\nu \geq 2(C-\lambda)$.

Proof. Given λ with $\lambda > 1/\mathrm{e}$, select the following constants, depending only on λ:

- D and ν_0 so large that $\lambda\mathrm{e}^{\lambda D}(1 - \mathrm{e}^{-\nu_o/2}) > 1$ and $\lambda\mathrm{e}(1 - \mathrm{e}^{-\nu_o/2}) \geq \sqrt{\lambda\mathrm{e}}$;
- $w_0 > 0$ so large that $w_0\lambda\mathrm{e}^{\lambda D}(1 - \mathrm{e}^{-\nu_o/2}) - \lambda D \geq w_0$;
- a positive integer $\bar{t}_0$ so large that $\lambda((\lambda\mathrm{e})^{\bar{t}_0/2-1} - D) \geq w_0$.

Throughout the remainder of the proof we assume without further comment that $\nu \geq \nu_o$ and $\nu \geq 2(C-\lambda)$. The latter condition and the fact $b_0 = 1/(1+\mathrm{e}^{-\nu})$ ensure that $b_0 < \nu/2(C-\lambda)$. Let $t^* = \max\{t \geq 0\colon b_t < \nu/2(C-\lambda)\}$ and let $\bar{t}_1 = \log^*(\nu)$. The first step of the proof is to show that $t^* \leq \bar{t}_0 + \bar{t}_1$. For that purpose we will show that the b_t increase at least geometrically to reach a certain large constant (specifically, so that (4.12) below holds), and then they increase as fast as a sequence produced by iterated exponentiation.

Since $b_0 \geq 0$, it follows from (4.11) and the choice of ν_0 that $b_1 \geq (1-\mathrm{e}^{-\nu_o/2}) \geq (\lambda \mathrm{e})^{-1/2}$. Note that $\mathrm{e}^u \geq \mathrm{e}u$ for all $u > 0$, since e^u/u is minimized at $u = 1$. Thus, $\mathrm{e}^{\lambda b_t} \geq \lambda \mathrm{e} b_t$, which combined with the choice of ν_0 and (4.11) shows that if $b_t \leq \nu/2(C-\lambda)$ then $b_{t+1} \geq \sqrt{\lambda \mathrm{e}}\, b_t$. It follows that $b_t \geq (\lambda \mathrm{e})^{t/2-1}$ for $1 \leq t \leq t^*+1$.

If $b_{\bar{t}_0-1} \geq \nu/2(C-\lambda)$ then $t^* \leq \bar{t}_0 - 2$ and the claim $t^* \leq \bar{t}_0 + \bar{t}_1$ is proved (that is, the geometric growth phase alone was enough), so to cover the other possibility, suppose that $b_{\bar{t}_0-1} < \nu/2(C-\lambda)$. Then $\bar{t}_0 \leq t^*+1$ and, therefore, $b_{\bar{t}_0} \geq (\lambda \mathrm{e})^{\bar{t}_0/2-1}$. Let $t_0 = \min\{t\colon b_t \geq (\lambda \mathrm{e})^{\bar{t}_0/2-1}\}$. It follows that $t_0 \leq \bar{t}_0$ and, by the choice of $\bar{t}_0$ and the definition of t_0,

$$\lambda(b_{t_0} - D) \geq w_0. \tag{4.12}$$

Define the sequence $(w_t\colon t \geq 0)$ beginning with w_0 already chosen, and satisfying the recursion $w_{t+1} = \mathrm{e}^{w_t}$. It follows by induction that

$$\lambda(b_{t_0+t} - D) \geq w_t \quad \text{for } t \geq 0,\ t_0 + t \leq t^*+1. \tag{4.13}$$

Indeed, the base case is (4.12), and if (4.13) holds for some t with $t_0 + t \leq t^*$, then $b_{t_0+t} \geq w_t/\lambda + D$, so that

$$\lambda(b_{t_0+t+1} - D) \geq \lambda(\mathrm{e}^{\lambda b_{t_0+t}}(1-\mathrm{e}^{-\nu/2}) - D) \geq w_{t+1}\lambda \mathrm{e}^{\lambda D}(1-\mathrm{e}^{-\nu/2}) - \lambda D \geq w_{t+1},$$

where the last inequality follows from the choice of w_0 and the fact $w_{t+1} \geq w_0$. The proof of (4.13) by induction is complete.

Let $\bar{t}_1 = \log^*(\nu)$. Since $w_1 \geq 1$, it follows that $w_{\bar{t}_1+1} \geq \nu$ (verify by applying the log function $\bar{t}_1$ times to each side). Therefore, $w_{\bar{t}_1+1} \geq \lambda\nu/2(C-\lambda) - \lambda D$, where we use the fact that $C - \lambda \geq 2\lambda$. If $t_0 + \bar{t}_1 < t^*$, from (4.13) with $t = t_0 + \bar{t}_1 + 1$ we would have

$$b_{t_0+\bar{t}_1+1} \geq \frac{w_{\bar{t}+1}}{\lambda} + D \geq \frac{\nu}{2(C-\lambda)},$$

which would imply $t^* \leq t_0 + \bar{t}_1$, which would be a contradiction. Therefore, $t^* \leq t_0 + \bar{t}_1 \leq \bar{t}_0 + \bar{t}_1$, as required.

Since t^* is the last iteration index t such that $b_t < \nu/2(C-\lambda)$, either $b_{t^*+1} = \nu/2(C-\lambda)$, and we say the threshold $\nu/2(C-\lambda)$ is exactly reached at iteration t^*+1, or $b_{t^*+1} > \nu/2(C-\lambda)$, in which case we say there was overshoot at iteration t^*+1. First, consider the case that the threshold is exactly reached at iteration t^*+1. Then, $b_{t^*+1} = \nu/2(C-\lambda)$, and (4.11) can be applied with $t = t^*+1$, yielding

$$b_{t^*+2} \geq \exp(\lambda b_{t^*+1})\left(1 - \exp\left(-\frac{\nu}{2}\right)\right) = \exp\left(\frac{\lambda\nu}{2(C-\lambda)}\right)\left(1 - \exp\left(-\frac{\nu}{2}\right)\right).$$

Since $t^*+2 \leq \bar{t}_0 + \bar{t}_1 + 2 = \bar{t}_0 + \log^*(\nu) + 2$, it follows from Lemma 4.5 that $b_{\bar{t}_0+\log^*(\nu)+2} \geq b_{t^*+2}$, which completes the proof of the lemma in the case when the threshold is exactly reached at iteration t^*+1.

To complete the proof, we explain how the information available for estimation can be reduced through a *thinning* method, leading to a reduction in the value of b_{t^*+1}, so that we can assume without loss of generality that the threshold is always exactly reached at iteration t^*+1. Let ϕ be a parameter with $0 \le \phi \le 1$. As before, we will be considering a total of t^*+2 iterations, so consider a random tree with labels $(\mathcal{T}_u^{t^*+2}, \tau_{\mathcal{T}_u^{t^*+2}})$, with root vertex u, and maximum depth t^*+2. For the original model, each vertex of depth t^*+1 or less with label 0 or 1 has a Poisson number of children with labels 0 and 1, respectively, with means specified in the construction. For the thinning method, for each $\ell \in \partial u$ and each child i of $\partial \ell$ (that is, for each grandchild of u), we generate a random variable $U_{\ell,i}$ that is uniformly distributed on the interval $[0,1]$. Then we retain i if $U_{\ell,i} \le \phi$, and we delete i, and all its decedents, if $U_{\ell,i} > \phi$. That is, the grandchildren of the root vertex u are each deleted with probability $1-\phi$. It is equivalent to reducing p and q to ϕp and ϕq, respectively, for that one generation. Consider the calculation of the likelihood ratio at the root vertex for the thinned tree. The log-likelihood ratio messages begin at the leaf vertices at depth t^*+2.

For any vertex $\ell \ne u$, let $\Lambda_{\ell \to \pi(\ell),\phi}$ denote the log-likelihood message passed from vertex ℓ to its parent, $\pi(\ell)$. Also, let $\Lambda_{u,\phi}$ denote the log-likelihood computed at the root vertex. For brevity, we remove the superscript t on the log-likelihood ratios, though t on the message $\Lambda_{\ell \to \pi(\ell),\phi}$ would be t^*+2 minus the depth of ℓ. The messages of the form $\Lambda_{\ell \to \pi(\ell),\phi}$ do not actually depend on ϕ unless $\ell \in \partial u$. For a vertex $\ell \in \partial u$, the message $\Lambda_{\ell \to u,\phi}$ has the nearly the same representation as in Lemma 4.1; namely,

$$\Lambda_{\ell \to u,\phi} = -\phi K(p-q) + \sum_{i \in \partial \ell:\, U_{\ell,i} \le \phi} \log\left(\frac{e^{\Lambda_{i \to \ell,\phi} - \nu}(p/q)+1}{e^{\Lambda_{i \to \ell,\phi}-\nu}+1} \right). \tag{4.14}$$

The representation of $\Lambda_{u,\phi}$ is the same as the representation of Λ_u^{t+1} in Lemma 4.1, except with $\Lambda_{\ell \to u}^t$ replaced in both places on the right-hand side by $\Lambda_{\ell \to u,\phi}$.

Let $Z_{0,\phi}^t$ and $Z_{1,\phi}^t$ denote random variables for analyzing the message-passing algorithm for this depth t^*+2 tree. Their laws are as follows. For $0 \le t \le t^*+1$, $\mathcal{L}(Z_{0,\phi}^t)$ is the law of $\Lambda_{\ell \to \pi(\ell),\phi}$ given $\tau_\ell = 0$ for a vertex ℓ of depth t^*+2-t. And $\mathcal{L}(Z_{0,\phi}^{t^*+2})$ is the law of $\Lambda_{u,\phi}$ given $\tau_u = 0$. Note that $Z_{0,\phi}^0 \equiv 0$. The laws $\mathcal{L}(Z_{1,\phi}^t)$ are determined similarly, conditioning on the labels of the vertices to be 1. For t fixed, $\mathcal{L}(Z_{0,\phi}^t)$ and $\mathcal{L}(Z_{1,\phi}^t)$ each determine the other because they represent distributions of the log-likelihood for a binary hypothesis testing problem.

The message-passing equations for the log-likelihood ratios translate into recursions for the laws $\mathcal{L}(Z_{0,\phi}^t)$ and $\mathcal{L}(Z_{1,\phi}^t)$. We have not focused directly on the full recursions of the laws, but rather looked at equations for exponential moments. The basic recursions under consideration for $\mathcal{L}(Z_{0,\phi}^t)$ are exactly as before for $0 \le t \le t^*-1$ and for $t = t^*+1$. For $t = t^*$, the thinning needs to be taken into account, resulting, for example, in the following updates for $t = t^*$:

$$\mathbb{E}[e^{Z_1^{t^*+1}}] = \mathbb{E}[e^{2Z_0^{t^*+1}}] = \exp\left(\lambda\phi \mathbb{E}\left[\frac{e^{Z_1^{t^*}}}{1+e^{Z_1^{t^*}-\nu}} \right] \right),$$

$$\mathbb{E}[e^{2Z_1^{t^*+1}}] = \exp\left(3\lambda\phi \mathbb{E}\left[\frac{e^{Z_1^{t^*}}}{1+e^{Z_1^{t^*}-\nu}} \right] + \frac{\lambda^2\phi}{K(p-q)} \mathbb{E}\left[\left(\frac{e^{Z_1^{t^*}}}{1+e^{Z_1^{t^*}-\nu}} \right)^2 \right] \right).$$

Let

$$a_{t,\phi} = \mathbb{E}[e^{Z_{1,\phi}^t}], \qquad b_{t,\phi} = \mathbb{E}\left[\frac{e^{Z_{1,\phi}^t}}{1+e^{Z_{1,\phi}^t-\nu}} \right] \quad \text{for } 0 \le t \le t^*+2.$$

Note that $a_{t,\phi}$ and $b_{t,\phi}$ do not depend on ϕ for $0 \le t \le t^*$. We have

$$a_{t+1,\phi} = \begin{cases} \exp(\lambda b_{t,\phi}), & t \ne t^*, \\ \exp(\lambda \phi b_{t,\phi}), & t = t^*. \end{cases} \tag{4.15}$$

We will not need (4.15) for $t = t^*$ but we will use it for $t = t^* + 1$.

On the one hand, if $\phi = 0$ then $\Lambda_{\ell \to u,\phi} \equiv 0$ for all $\ell \in \partial u$, so that

$$Z_{0,\phi=0}^{t^*+1} = Z_{1,\phi=0}^{t^*+1} \equiv 0 \quad \text{and} \quad b_{t^*+1,\,\phi=0} = \frac{1}{1+\mathrm{e}^{-\nu}} = \frac{n-K}{n} \le 1 < \frac{\nu}{2(C-\lambda)}.$$

On the other hand, by the definition of t^*, we know that $b_{t^*+1,\,\phi=1} \ge \nu/2(C-\lambda)$. We will show that there exists a value of $\phi \in [0,1]$ so that $b_{t^*+1,\phi} = \nu/2(C-\lambda)$. To do this we next prove that $b_{t^*+1,\phi}$ is a continuous and, in fact, nondecreasing function of ϕ using a variation of the proof of Lemma 4.5. Let ℓ denote a fixed neighbor of the root node u. Note that $\exp(\Lambda_{\ell \to u,\phi})$ is the likelihood ratio for detection of τ_ℓ based on the thinned subtree of depth t^*+1 with root ℓ. As ϕ increases from 0 to 1, the amount of thinning decreases, so larger values of ϕ correspond to larger amounts of information. Therefore, conditioned on $\tau_u = 0$, $(\exp(\Lambda_{\ell,\phi}) : 0 \le \phi \le 1)$ is a martingale. Moreover, the independent splitting property of Poisson random variables implies that, given $\tau_\ell = 0$, the random process

$$\phi \mapsto |\{i \in \partial\ell : U_{\ell,i} \le \phi\}|$$

is a Poisson process with intensity nq and, therefore, the sum in (4.14), as a function of ϕ over the interval $[0,1]$, is a compound Poisson process. Compound Poisson processes, just like Poisson processes, are almost surely continuous at any fixed value of ϕ and, therefore, the random process $\phi \mapsto \Lambda_{\ell \to u,\phi}$ is continuous in distribution. Therefore, the random variables $\exp(Z_{0,\phi}^{t^*+1})$ can be constructed on a single probability space for $0 \le \phi \le 1$ to form a martingale which is continuous in distribution. Since $b_{t^*+1,\phi}$ is the expectation of a bounded, continuous, convex function of $\exp(Z_{0,\phi}^{t^*+1})$, it follows that $b_{t^*+1,\phi}$ is continuous and nondecreasing in ϕ. Therefore, we can conclude that there exists a value of ϕ so that $b_{t^*+1,\phi} = \nu/2(C-\lambda)$, as claimed.

Since there is no overshoot, we obtain, as before (by using (4.15) for $t = t^* + 1$ to modify Lemma 4.4 to handle (b_{t+1}, b_t) replaced by $(b_{t^*+2,\phi}, b_{t^*+1,\phi})$),

$$b_{t^*+2,\phi} \ge \exp(\lambda b_{t^*+1,\phi})(1-\mathrm{e}^{-\nu/2}) = \exp\left(\frac{\lambda\nu}{2(C-\lambda)}\right)(1-\mathrm{e}^{-\nu/2}).$$

The same martingale argument used in the previous paragraph can be used to show that $b_{t^*+2,\phi}$ is nondecreasing in ϕ and, in particular, $b_{t^*+2} = b_{t^*+2,1} \ge b_{t^*+2,\phi}$ for $0 \le \phi \le 1$. Hence, by Lemma 4.5 and the fact $t^* + 2 \le \bar{t}_0 + \log^*(\nu) + 2$, we have $b_{\bar{t}_0+\log^*(\nu)+2} \ge b_{t^*+2} \ge b_{t^*+2,\phi}$, completing the proof of the lemma. □

Lemma 4.7. *Let* $B = (p/q)^{3/2}$. *Then*

$$\exp\left(-\frac{\lambda}{8}b_t\right) \le \mathbb{E}[\exp\left(\frac{Z_0^{t+1}}{2}\right)] \le \exp\left(-\frac{\lambda}{8B}b_t\right).$$

Proof. We prove the upper bound first. In view of Lemma 4.1, by defining

$$f(x) = \frac{x(p/q)+1}{x+1},$$

we obtain

$$\exp\bigl(\tfrac{1}{2}\Lambda_u^{t+1}\bigr) = \exp\bigl(-\tfrac{1}{2}K(p-q)\bigr)\prod_{\ell\in\partial u} f^{1/2}(\exp(\Lambda_{\ell\to u}^t - \nu)).$$

Thus,

$$\begin{aligned}&\mathbb{E}\bigl[\exp\bigl(\tfrac{1}{2}Z_0^{t+1}\bigr)\bigr]\\ &\quad= \exp\bigl(-\tfrac{1}{2}K(p-q)\bigr)\mathbb{E}[(\mathbb{E}[f^{1/2}(\exp(Z_1^t-\nu))])^{L_u}]\mathbb{E}[(\mathbb{E}[f^{1/2}(\exp(Z_0^t-\nu))])^{M_u}].\end{aligned}$$

Using the fact that $\mathbb{E}[c^X] = \mathrm{e}^{\lambda(c-1)}$ for $X \sim \mathrm{Poisson}(\lambda)$ and $c > 0$, we have

$$\begin{aligned}\mathbb{E}\bigl[\exp\bigl(\tfrac{1}{2}Z_0^{t+1}\bigr)\bigr] = \exp\bigl(&-\tfrac{1}{2}K(p-q) + Kq(\mathbb{E}[f^{1/2}(\exp(Z_1^t-\nu))]-1)\\ &+ (n-K)q(\mathbb{E}[f^{1/2}(\exp(Z_0^t-\nu))]-1)\bigr)\end{aligned} \tag{4.16}$$

By the intermediate value form of Taylor's theorem, for any $x \geq 0$, there exists y with $1 \leq y \leq x$ such that $\sqrt{1+x} = 1 + x/2 - x^2/8(1+y)^{3/2}$. Therefore,

$$\sqrt{1+x} \leq 1 + \frac{x}{2} - \frac{x^2}{8(1+A)^{3/2}} \quad \text{for all } 0 \leq x \leq A. \tag{4.17}$$

Letting $A \coloneqq p/q - 1$ and noting that $B = (1+A)^{3/2}$, we have

$$\left(\frac{\mathrm{e}^{z-\nu}(p/q)+1}{1+\mathrm{e}^{z-\nu}}\right)^{1/2} = \left(1 + \frac{p/q-1}{1+\mathrm{e}^{-z+\nu}}\right)^{1/2} \leq 1 + \frac{1}{2}\frac{(p/q-1)}{(1+\mathrm{e}^{-z+\nu})} - \frac{1}{8B}\frac{(p/q-1)^2}{(1+\mathrm{e}^{-z+\nu})^2}.$$

It follows that

$$\begin{aligned}&Kq(\mathbb{E}[f^{1/2}(\mathrm{e}^{Z_1^t-\nu})]-1) + (n-K)q(\mathbb{E}[f^{1/2}(\mathrm{e}^{Z_0^t-\nu})]-1)\\ &\quad\leq \frac{1}{2}Kq\left(\frac{p}{q}-1\right)\left(\mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_1^t+\nu}}\right] + \mathrm{e}^{\nu}\mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_0^t+\nu}}\right]\right)\\ &\qquad - \frac{1}{8B}Kq\left(\frac{p}{q}-1\right)^2\left(\mathbb{E}\left[\frac{1}{(1+\mathrm{e}^{-Z_1^t+\nu})^2}\right] + \mathrm{e}^{\nu}\mathbb{E}\left[\frac{1}{(1+\mathrm{e}^{-Z_0^t+\nu})^2}\right]\right)\\ &\quad= \frac{K(p-q)}{2} - \frac{1}{8B}Kq\left(\frac{p}{q}-1\right)^2\mathbb{E}\left[\frac{1}{1+\mathrm{e}^{-Z_1^t+\nu}}\right]\\ &\quad= \frac{K(p-q)}{2} - \frac{\lambda}{8B}\underbrace{\mathbb{E}\left[\frac{\mathrm{e}^{Z_1^t}}{1+\mathrm{e}^{Z_1^t-\nu}}\right]}_{b_t},\end{aligned}$$

where the first equality follows from (4.6) and (4.7), and the last equality holds as a result of $Kq(p/q-1)^2\mathrm{e}^{\nu} = \lambda$. Combining the last displayed equation with (4.16) yields the desired upper bound.

The proof for the lower bound is similar. Instead of (4.17), we use $\sqrt{1+x} \geq 1 + x/2 - \frac{1}{8}x^2$ for all $x \geq 0$, and the lower bound readily follows by the same argument as above. □

Lemma 4.8. (Upper bound on the classification error for the random tree model.) *Consider the random tree model with parameters λ, ν, and p/q. Let λ be fixed with $\lambda > 1/\mathrm{e}$. There are constants $\bar{t}_0$ and ν_o depending only on λ such that if $\nu \geq \nu_o$ and $\nu \geq 2(C-\lambda)$, then*

after $\bar{t}_0 + \log^*(\nu) + 2$ *iterations of the BP algorithm, the average error probability for the MAP estimator* $\hat{\tau}_u$ *of* τ_u *satisfies*

$$p_{\mathrm{err}}^t \le \left(\frac{K(n-K)}{n^2}\right)^{1/2} \exp\left(-\frac{\lambda}{8B} e^{\nu\lambda/2(C-\lambda)}(1-e^{-\nu/2})\right), \tag{4.18}$$

where $B = (p/q)^{3/2}$ *and* $C = \lambda(p/q+2)$. *In particular, if* $p/q = O(1)$, *and* r *is any positive constant, then if* ν *is sufficiently large,*

$$p_{\mathrm{err}}^t \le \frac{K e^{-r\nu}}{n} = \frac{K}{n}\left(\frac{K}{n-K}\right)^r. \tag{4.19}$$

Proof. We use the Bhattacharyya upper bound in (4.2) with $\pi_1 = K/n$ and $\pi_0 = (n-K)/n$, and the fact that $\rho = \mathbb{E}[e^{Z_0^t/2}]$. Substituting in the lower bound on $b_{\bar{t}_0+\log^*(\nu)+2}$ from Lemma 4.6 into the upper bound on $\mathbb{E}[e^{Z_0^t/2}]$ from Lemma 4.7 yields (4.18). If $p/q = O(1)$ and $r > 0$ then, for large enough ν,

$$\frac{\lambda}{8B} e^{\nu\lambda/2(C-\lambda)}(1-e^{-\nu/2}) \ge \nu\left(r+\frac{1}{2}\right),$$

which, together with (4.18), implies (4.19). □

4.3. Lower bounds on the classification error for the Poisson tree

The bounds in this section will be combined with the coupling lemmas of Appendix C to yield converse results for recovering a community by local algorithms.

Lemma 4.9. (Lower bounds for the Poisson tree model.) *Fix* λ *with* $0 < \lambda \le 1/e$. *For any estimator* $\hat{\tau}_u$ *of* τ_u *based on the observation of the tree up to any depth* t, *the average error probability satisfies*

$$p_{\mathrm{err}}^t \ge \frac{K(n-K)}{n^2} e^{-\lambda e/4}, \tag{4.20}$$

and the sum of type I and type II error probabilities satisfies

$$p_{\mathrm{err},0}^t + p_{\mathrm{err},1}^t \ge \tfrac{1}{2} e^{-\lambda e/4}. \tag{4.21}$$

Furthermore, if $p/q = O(1)$ *and* $\nu \to \infty$ *then*

$$\liminf_{n\to\infty} \frac{n}{K} p_{\mathrm{err}}^t \ge 1. \tag{4.22}$$

Proof. From Lemma 4.7 we see that the Bhattacharyya coefficient $\rho_{\mathrm{B}} = \mathbb{E}[e^{Z_0^{t+1}/2}]$ satisfies $\rho_{\mathrm{B}} \ge e^{-\lambda b_t/8}$. Note that $b_{t+1} \le a_{t+1} = e^{\lambda b_t}$ for $t \ge 0$ and $b_0 = 1/(1+e^{-\nu})$. It follows from induction and the assumption $\lambda e \le 1$ that $b_t \le e$ for all $t \ge 0$. Therefore, $\rho_{\mathrm{B}} \ge e^{-\lambda e/8}$. Applying the Bhattacharyya lower bound on p_{err}^t in (4.2) (which holds for any estimator) with $(\pi_0, \pi_1) = ((n-K)/n, K/n)$ yields (4.20) and with $(\pi_0, \pi_1) = (\frac{1}{2}, \frac{1}{2})$ yields (4.21), respectively.

It remains to prove (4.22), so suppose that $p/q = O(1)$ and $\nu \to \infty$. It suffices to prove (4.22) for the MAP estimator, $\hat{\tau}_u = \mathbf{1}_{\{\Lambda_u^t \ge \nu\}}$, since the MAP estimator minimizes the average error probability.

In the arXiv version of this paper [17] we show that the distribution of Λ_u^{t+1} is well approximated by the $\mathcal{N}(\frac{1}{2}\lambda b_t, \lambda b_t)$ distribution if $\sigma_u = 1$ and by the $\mathcal{N}(-\frac{1}{2}\lambda b_t, \lambda b_t)$ distribution if $\sigma_u = 0$. It follows that as $n \to \infty$, the type I and type II error probabilities satisfy

$$p_{\mathrm{e},1}^t - Q\left(\frac{\lambda b_{t-1}/2 - \nu}{\sqrt{\lambda b_{t-1}}}\right) \to 0 \quad \text{and} \quad p_{\mathrm{e},0}^t - Q\left(\frac{\lambda b_{t-1}/2 + \nu}{\sqrt{\lambda b_{t-1}}}\right) \to 0,$$

where Q is the complementary cumulative distribution function of the standard normal distribution. Recall that $b_t \le \mathrm{e}$ for all $t \ge 0$. Also, b_t is bounded away from 0, since $b_t \ge b_0 = 1/(1+\mathrm{e}^{-\nu})$. Since $\nu \to \infty$, we have $p_{\mathrm{e},1}^t \to 1$. By definition, $(n/K)p_{\mathrm{err}}^t \ge p_{\mathrm{err},1}^t$ and, consequently, $\liminf_{n\to\infty}(n/K)p_{\mathrm{err}}^t \ge 1$. □

5. Proofs of the main results of BP

Proof of Theorem 3.1. The proof basically consists of combining Lemma 4.8 and the Lemma C.1. Lemma 4.8 holds by the assumptions $K^2(p-q)^2/(n-K)q \equiv \lambda$ for a constant λ with $\lambda > 1/\mathrm{e}$, $\nu \to \infty$, and $p/q = O(1)$. From Lemma 4.8 we can also determine the given expression for t_f. In turn, the assumptions $(np)^{\log^* \nu} = n^{o(1)}$ and $\mathrm{e}^{\log^* \nu} \le \nu = n^{o(1)}$ ensure that $(2+np)^{t_f} = n^{o(1)}$, so that Lemma C.1 holds.

A subtle point is that the performance bound of Lemma 4.8 is for the MAP rule (4.1) for detecting the label of the root vertex. The same rule could be implemented at each vertex of the graph G which has a locally tree-like neighborhood of radius $t_0 + \log^*(\nu) + 2$ by using the estimator $\hat{C}_o = \{i : R_i^{t_f} \ge \nu\}$. We first bound the performance for $\hat{C}_o$ and then do the same for $\hat{C}$ produced by Algorithm 3.1. (We could have taken $\hat{C}_o$ to be the output of Algorithm 3.1, but returning a constant size estimator leads to simpler analysis of the algorithm for exact recovery.)

The average probability of misclassification of any given vertex u in G by $\hat{C}_o$ (for prior distribution $(K/n, (n-K)/n)$) is less than or equal to the sum of the two terms. The first term is $n^{-1+o(1)}$ in the $|C^*| \equiv K$ case or $n^{-1/2+o(1)}$ in the other case (due to failure of the tree coupling of radius t_f neighborhood; see Lemma C.1). The second term is $(K/n)\mathrm{e}^{-\nu r}$ (bound on average error probability for the detection problem associated with a single vertex u in the tree model; see Lemma 4.8.) Multiplying by n bounds the expected total number of misclassification errors, $\mathbb{E}[|C^* \triangle \hat{C}_o|]$; dividing by K gives the bounds stated in the lemma with $\hat{C}$ replaced by $\hat{C}_o$ and the factor 2 dropped in the bounds.

The set $\hat{C}_o$ is defined by a threshold condition whereas $\hat{C}$ similarly corresponds to using a data dependent threshold and tie breaking rule to arrive at $|\hat{C}| \equiv K$. Therefore, with probability 1, either $\hat{C}_o \subset \hat{C}$ or $\hat{C} \subset \hat{C}_o$. Together with the fact that $|\hat{C}| \equiv K$, we have

$$|C^* \triangle \hat{C}| \le |C^* \triangle \hat{C}_o| + |\hat{C}_o \triangle \hat{C}| = |C^* \triangle \hat{C}_o| + ||\hat{C}_o| - K|$$

and, furthermore,

$$||\hat{C}_o| - K| \le ||\hat{C}_o| - |C^*|| + ||C^*| - K| \le |C^* \triangle \hat{C}_o| + ||C^*| - K|.$$

So

$$|C^* \triangle \hat{C}| \le 2|C^* \triangle \hat{C}_o| + ||C^*| - K|.$$

If $|C^*| \equiv K$ then $|C^* \triangle \hat{C}| \le 2|C^* \triangle \hat{C}_o|$ and (3.3) follows from what was proved for $\hat{C}_o$. In the other case, $\mathbb{E}[||C^*| - K|] \le n^{1/2+o(1)}$ and (3.4) follows from what was proved for $\hat{C}_o$.

As for the computational complexity guarantee, note that in each BP iteration, each vertex i needs to transmit the outgoing message $R_{i\to j}^{t+1}$ to its neighbor j according to (3.1). To do so, vertex i can first compute R_i^{t+1} and then subtract neighbor j's contribution from it to obtain the desired message $R_{i\to j}^{t+1}$. In this way, each vertex i needs $O(|\partial i|)$ basic operations and the total time complexity of one BP iteration is $O(|E(G)|)$, where $|E(G)|$ is the total number of edges. Since $\nu \le n$, at most $O(\log^* n)$ iterations are needed and, hence, the algorithm terminates in $O(|E(G)|\log^* n)$ time. □

Proof of Theorem 3.2. The theorem follows from the fact that the BP algorithm achieves weak recovery, even if the cardinality $|C^*|$ is random and is only known to satisfy $\mathbb{P}\{||C^*|-K| \ge \sqrt{3K\log n}\} \le n^{-1/2+o(1)}$ and the results of [16]. We include the proof for completeness. Let $C_k^* = C^* \cap ([n]\setminus S_k)$ for $1 \le k \le 1/\delta$. As explained in Remark C.2, C_k^* is obtained by sampling the vertices in $[n]$ without replacement and, thus, the distribution of C_k^* is hypergeometric with $\mathbb{E}[|C_k^*|] = K(1-\delta)$. A result of Hoeffding [18] implies that the Chernoff bounds for the $\mathrm{binom}(n(1-\delta), K/n)$ distribution also hold for $|C_k^*|$, so (C.2) with $np = K(1-\delta)$ and $\varepsilon = \sqrt{3\log n/[K(1-\delta)]}$ imply that

$$\mathbb{P}\{||C_k^*| - K(1-\delta)| \ge \sqrt{3K(1-\delta)\log n}\} \le 2n^{-1} \le n^{-1/2+o(1)}.$$

Hence, it follows from Theorem 3.1 and the condition $\lambda > 1/\mathrm{e}$ that

$$\mathbb{P}\left\{|\hat{C}_k \Delta C_k^*| \le \delta K \text{ for } 1 \le k \le \frac{1}{\delta}\right\} \to 1 \quad \text{as } n \to \infty,$$

where $\hat{C}_k$ is the output of the BP algorithm in step (iii) of Algorithm 3.2. Applying [16, Theorem 3] together with assumption (3.5), it follows that $\mathbb{P}\{\tilde{C} = C^*\} \to 1$ as $n \to \infty$. □

Proof of Theorem 3.3. The average error probability p_{err} for classifying the label of a vertex in the graph G is greater than or equal to the lower bound (4.20) on the average error probability for the tree model, minus the upper bound $n^{-1+o(1)}$ on the coupling error provided by Lemma C.1. Multiplying the lower bound on the average error probability per vertex by n yields (3.7). Similarly, $p_{\mathrm{err},0}$ and $p_{\mathrm{err},1}$ for the community recovery problem can be approximated by the respective conditional error probabilities for the random tree model by the last part of the coupling lemma (Lemma C.1) so (3.8) follows from (4.21).

By Lemma 4.9, assuming that $p/q = O(1)$ and $\nu \to \infty$, $\liminf_{n\to\infty}(n/K)\tilde{p}_{\mathrm{err}}^t \ge 1$, where $\tilde{p}_{\mathrm{err}}^t$ is the average error probability for any estimator for the corresponding random tree network. By the Lemma C.1, $|\tilde{p}_{\mathrm{err}}^t - p_{\mathrm{err}}^t| \le n^{-1+o(1)}$. From the assumption that $n/K = n^{o(1)}$, $|(n/K)\tilde{p}_{\mathrm{err}}^t - (n/K)p_{\mathrm{err}}^t| \le n^{-1+o(1)}$. The conclusion $\liminf_{n\to\infty}(n/K)p_{\mathrm{err}} \ge 1$ follows from the triangle inequality. □

Appendix A. Degree thresholding when $K \asymp n$

A simple algorithm for recovering C^* is degree thresholding. Specifically, let d_i denote the degree of vertex i. Then d_i is distributed as the sum of two independent random variables, with distributions $\mathrm{binom}(K-1, p)$ and $\mathrm{binom}(n-K, q)$, respectively, if $i \in C^*$, while $d_i \sim \mathrm{binom}(n-1, q)$ if $i \notin C^*$. The mean-degree difference between these two distributions is $(K-1)(p-q)$, and the degree variance is $O(nq)$. By assuming that p/q is bounded, it follows from the Bernstein's inequality that $|d_i - \mathbb{E}[d_i]| \ge \frac{1}{2}(K-1)(p-q)$ with probability at most $\mathrm{e}^{-\Omega((K-1)^2(p-q)^2/(nq))}$. Let $\hat{C}$ be the set of vertices with degrees

larger than $nq + \frac{1}{2}(K-1)(p-q)$ and, thus, $\mathbb{E}[|\hat{C}\triangle C^*|] = ne^{-\Omega((K-1)^2(p-q)^2/(nq))}$. Hence, if $(K-1)^2(p-q)^2/(nq) = \omega(\log(n/K))$ then $\mathbb{E}[|\hat{C}\triangle C^*|] = o(K)$, that is, weak recovery is achieved. In the regime $K \asymp n - K \asymp n$ and p is bounded away from 1, the necessary and sufficient condition for the existence of estimators providing weak recovery is $K^2(p-q)^2/(nq) \to \infty$ as shown in [16]. Thus, degree thresholding provides weak recovery in this regime whenever it is information-theoretically possible. Under the additional condition (3.5), an algorithm attaining exact recovery can be built using degree thresholding for weak recovery followed by a linear-time voting procedure, as in Algorithm 3.2; see [16, Theorem 3] and its proof. In the regime $(n/K)\log(n/K) = o(\log n)$ or, equivalently, $K = \omega(n \log\log n/\log n)$, the information-theoretic necessary condition for exact recovery given by (B.3) and (B.4) imply that $K^2(p-q)^2/(nq) = \omega(\log(n/K))$ and, hence, in this regime the degree thresholding attains exact recovery whenever it is information-theoretically possible.

Appendix B. Comparison with information-theoretic limits

As noted in the introduction, in the $K = \Theta(n)$ regime, degree thresholding achieves weak recovery and, if a voting procedure is also used, exact recovery whenever it is information-theoretically possible. In this section we compare the recovery thresholds by BP to the information-theoretic thresholds established in [16], in the regime of

$$K = o(n), \qquad np = n^{o(1)}, \qquad \frac{p}{q} = O(1), \tag{B.1}$$

which is the main focus of this paper.

The information-theoretic threshold for weak recovery was established in [16, Corollary 1], which, in the regime of (B.1), reduces to the following. If

$$\liminf_{n\to\infty} \frac{Kd(p\|q)}{2\log(n/K)} > 1, \tag{B.2}$$

then weak recovery is possible. On the other hand, if weak recovery is possible then

$$\liminf_{n\to\infty} \frac{Kd(p\|q)}{2\log(n/K)} \geq 1. \tag{B.3}$$

To compare with BP, we rephrase the above sharp threshold in terms of the signal-to-noise ratio λ defined in (1.1). Note that $d(p\|q) = (p\log(p/q) + q - p)(1 + o(1))$ provided that $p/q = O(1)$ and $p \to 0$. Therefore, the information-theoretic weak recovery threshold is given by

$$\lambda > \left(C\left(\frac{p}{q}\right) + \varepsilon\right)\frac{K}{n}\log\frac{n}{K} \quad \text{for any } \varepsilon > 0,$$

where $C(\alpha) \coloneqq 2(\alpha-1)^2/(1-\alpha+\alpha\log\alpha)$. In other words, in principle weak recovery only demands a vanishing signal-to-noise ratio $\lambda = \Theta((K/n)\log(n/K))$, while, in contrast, BP requires $\lambda > 1/\mathrm{e}$ to achieve weak recovery. No polynomial-time algorithm is known to succeed for $\lambda \leq 1/\mathrm{e}$, suggesting that computational complexity constraints might incur a severe penalty on the statistical optimality in the sublinear regime of $K = o(n)$.

Next we turn to exact recovery. The information-theoretic optimal threshold was established by Hajek *et al.* [16, Corollary 3]. In the regime of interest (B.1), exact recovery is possible

via the maximum likelihood estimator provided that (B.2) and (3.5) hold. Conversely, if exact recovery is possible then (B.3) and

$$\liminf_{n\to\infty} \frac{Kd(\tau^*\|q)}{\log n} \geq 1 \tag{B.4}$$

must hold. Note that the information-theoretic sufficient condition for exact recovery has two parts: one is the information-theoretic sufficient condition (B.2) for weak recovery; the other is the sufficient condition (3.5) for the success of the linear-time voting procedure. Similarly, recall that the sufficient condition for exact recovery by BP also has two parts: one is the sufficient condition $\lambda > 1/\mathrm{e}$ for weak recovery, and the other is again (3.5).

Clearly, the information-theoretic sufficient conditions for exact recovery and $\lambda > 1/\mathrm{e}$, which is needed for weak recovery by local algorithms, are both at least as strong as the information-theoretic necessary conditions (B.3) for weak recovery. It is thus of interest to compare them by assuming that (B.3) holds. If p/q is bounded, p is bounded away from 1, and (B.3) holds, then $d(\tau^*\|q) \asymp d(p\|q) \asymp (p-q)^2/q$ as shown by Hajek *et al.* [16]. So under those conditions on p, q, and (B.3), and if K/n is bounded away from 1,

$$\frac{Kd(\tau^*\|q)}{\log n} \asymp \frac{K(p-q)^2}{q\log n} \asymp \left(\frac{n}{K\log n}\right)\lambda.$$

Hence, the information-theoretic sufficient condition for exact recovery (3.5) demands a signal-to-noise ratio

$$\lambda = \Theta\left(\frac{K\log n}{n}\right). \tag{B.5}$$

Therefore, on the one hand, if $K = \omega(n/\log n)$ then condition (3.5) is stronger than $\lambda > 1/\mathrm{e}$ and, thus, condition (3.5) alone is sufficient for local algorithms to attain exact recovery. On the other hand, if $K = o(n/\log n)$ then $\lambda > 1/\mathrm{e}$ is stronger than condition (B.4) and, thus, for local algorithms to achieve exact recovery it requires $\lambda > 1/\mathrm{e}$, which far exceeds the information-theoretic optimal level (B.5). The critical value of K for this crossover is $K = \Theta(n/\log n)$. To determine the precise crossover point, we solve for K^* which satisfies

$$\frac{Kd(\tau^*\|q)}{\log n} = 1, \tag{B.6}$$

$$\lambda = \frac{K^2(p-q)^2}{nq} = \frac{1}{\mathrm{e}}. \tag{B.7}$$

Let $c = p/q = O(1)$. From (B.7), it follows that

$$q = \frac{n}{K^2(c-1)^2\mathrm{e}}. \tag{B.8}$$

Substituting (B.8) into the definition of τ^* in (3.6), we obtain

$$\tau^* = (1+o(1))q\frac{c-1}{\log c}.$$

It follows that

$$d(\tau^*\|q) = (1+o(1))q\left(1 - \frac{c-1}{\log c}\log\frac{\mathrm{e}\log c}{c-1}\right).$$

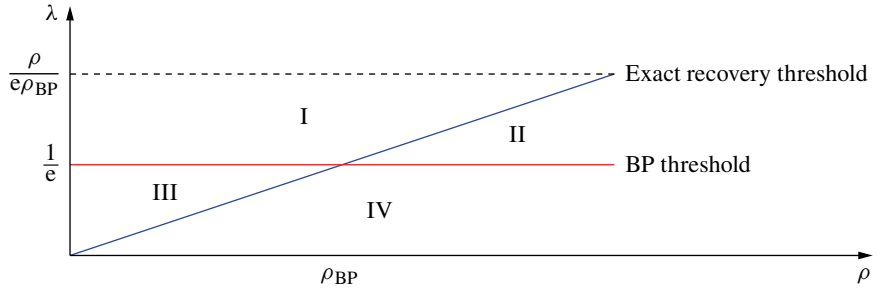

FIGURE 1: Phase diagram with $K = \rho n/\log n$ and $p/q = c$ for fixed constants $c \geq 1$, ρ, and λ as $n \to \infty$. In region I, exact recovery is provided by the BP algorithm plus voting procedure. In region II, weak recovery is provided by the BP algorithm, but exact recovery is not information-theoretically possible. In region III exact recovery is information-theoretically possible, but no polynomial-time algorithm is known for even weak recovery. In region IV, with $\lambda > 0$ and $\rho > 0$, weak recovery, but not exact recovery, is information-theoretically possible and no polynomial-time algorithm is known for weak recovery.

Combining the last displayed equation with (B.6) and (B.8) yields the crossover point K^* given by

$$K^* = \frac{n}{\log n}(\rho_{\mathrm{BP}}(c) + o(1)),$$

where

$$\rho_{\mathrm{BP}}(c) = \frac{1}{\mathrm{e}(c-1)^2}\left(1 - \frac{c-1}{\log c}\log\frac{\mathrm{e}\log c}{c-1}\right).$$

In Figure 1 we present the phase diagram with $K = \rho n/\log n$ for a fixed constant ρ. The line $\{(\rho, \lambda)\colon \lambda = 1/\mathrm{e}\}$ corresponds to the weak recovery, while the line $\{(\rho, \lambda)\colon \lambda = \rho/(\mathrm{e}\rho_{\mathrm{BP}})\}$ corresponds to the information-theoretic exact recovery threshold. Therefore, BP plus voting (Algorithm 3.2) achieves optimal exact recovery whenever the former line lies below the latter or, equivalently, $\rho > \rho_{\mathrm{BP}}(c)$.

Appendix C. Coupling lemma

Consider a sequence of planted dense subgraph models $G = (E, V)$ as described in the introduction. For each $i \in V$, σ_i denotes the indicator of $i \in C^*$. For $u \in V$, let G_u^t denote the subgraph of G induced by the vertices whose distance from u is at most t. Recall from Section 4 that T_u^t is defined similarly for the random tree graph, and τ_i denotes the label of a vertex i in the tree graph. In the following lemma we show there is a coupling such that $(G_u^{t_f}, \sigma_{G_u^{t_f}}) = (T_u^{t_f}, \tau_{T_u^{t_f}})$ with probability converging to 1, where t_f is growing slowly with n. A version of the lemma for fixed t, assuming $p, q = \Theta(1/n)$, was proved by Mossel *et al.* [33, Proposition 4.2], and the argument used there extends to the proof of this version; see [17] for the proof.

Lemma C.1. (Coupling lemma.) *Let $d = np$. Suppose that p, q, K, and t_f depend on n such that t_f is positive integer-valued, and $(2+d)^{t_f} = n^{o(1)}$. Consider an instance of the planted dense subgraph model. Suppose that C^* is random and all $\binom{n}{|C^*|}$ choices of C^* are equally likely to give its cardinality $|C^*|$. (If this is not true, this lemma still applies to the random graph obtained by randomly, uniformly permuting the vertices of G.) If the planted dense subgraph model (Definition 1.1) is such that $|C^*| \equiv K$ then, for any fixed $u \in [n]$, there exists a coupling*

between (G, σ) *and* (T_u, τ_{T_u}) *such that*

$$\mathbb{P}\big\{(G_u^{t_f}, \sigma_{G_u^{t_f}}) = (T_u^{t_f}, \tau_{T_u^{t_f}})\big\} \geq 1 - n^{-1+o(1)}.$$

If the planted dense subgraph model is such that $|C^*| \sim \text{binom}(n, K/n)$ *then, for any fixed* $u \in [n]$*, there exists a coupling between* (G, σ) *and* (T_u, τ_{T_u}) *such that*

$$\mathbb{P}\big\{(G_u^{t_f}, \sigma_{G_u^{t_f}}) = (T_u^{t_f}, \tau_{T_u^{t_f}})\big\} \geq 1 - n^{-1/2+o(1)}. \tag{C.1}$$

If the planted dense subgraph model is such that $K \geq 3 \log n$ *and* $|C^*|$ *is random such that* $\mathbb{P}\{||C^*| - K| \geq \sqrt{3K \log n}\} \leq n^{-1/2+o(1)}$*, then there exists a coupling between* (G, σ) *and* (T_u, τ_{T_u}) *such that (C.1) holds.*

Furthermore, the bounds stated remain true if the label σ_u *of the vertex* u *in the planted community graph, and the label* τ_u *of the root vertex in the tree graph, are both conditioned to be* 0 *or are both conditioned to be* 1.

Remark C.1. The condition $(2 + d)^{t_f} = n^{o(1)}$ of Lemma C.1 is satisfied, for example, if $t_f = O(\log^* n)$ and $d \leq n^{o(1/\log^* n)}$, or if $t_f = O(\log \log n)$ and $d = O((\log n)^s)$ for some constant $s > 0$. In particular, the condition is satisfied if $t_f = O(\log^* n)$ and $d = O((\log n)^s)$ for some constant $s > 0$.

Remark C.2. The part of Lemma C.1 involving $||C^*| - K| \geq \sqrt{3K \log n}$ is included to handle the case that $|C^*|$ has a certain hypergeometric distribution. In particular, if we begin with the planted dense subgraph model (Definition 1.1) with n vertices and a planted dense community with $|C^*| \equiv K$, for a clean-up procedure we will use for exact recovery (See Algorithm 3.2), we need to withhold a small fraction δ of vertices and run the BP algorithm on the subgraph induced by the set of $n(1 - \delta)$ retained vertices. Let C^{**} denote the intersection of C^* with the set of $n(1 - \delta)$ retained vertices. Then $|C^{**}|$ is obtained by sampling the vertices of the original graph without replacement. Thus, the distribution of $|C^{**}|$ is hypergeometric, and $\mathbb{E}[|C^{**}|] = K(1 - \delta)$. Therefore, by a result of Hoeffding [18], the distribution of $|C^{**}|$ is convex order dominated by the distribution that would result by sampling with replacement, namely, by $\text{binom}(n(1 - \delta), K/n)$. That is, for any convex function Ψ,

$$\mathbb{E}[\Psi(|C^{**}|)] \leq \mathbb{E}\left[\Psi\left(\text{binom}\left(n(1-\delta), \frac{K}{n}\right)\right)\right].$$

Therefore, Chernoff bounds for $\text{binom}(n(1 - \delta), K/n))$ also hold for $|C^{**}|$. We use the following Chernoff bounds for binomial distributions [29, Theorems 4.4 and 4.5]. For $X \sim \text{binom}(n, p)$,

$$\begin{aligned} \mathbb{P}\{X \geq (1+\varepsilon)np\} &\leq \mathrm{e}^{-\varepsilon^2 np/3} \\ \mathbb{P}\{X \leq (1-\varepsilon)np\} &\leq \mathrm{e}^{-\varepsilon^2 np/2} \end{aligned} \quad \text{for all } 0 \leq \varepsilon \leq 1. \tag{C.2}$$

Thus, if $K(1 - \delta) \geq 3 \log n$ then (C.2) with $\varepsilon = \sqrt{3 \log n/[K(1 - \delta)]}$ imply that

$$\mathbb{P}\{||C^{**}| - K(1-\delta)| \geq \sqrt{3K(1-\delta)\log n}\} \leq n^{-1}.$$

Thus, Lemma C.1 can be applied with K replaced by $K(1 - \delta)$.

Acknowledgements

The authors are grateful to the anonymous referees for their many useful suggestions. This work was supported in part by the National Science Foundation (grant numbers CCF 14-09106, OIS 18-23145, IIS-14-47879, and CCF-15-27105) and a CAREER award (grant number CCF-1651588).

References

[1] Abbe, E. and Sandon, C. (2016). Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic BP, and the information-computation gap. Preprint. Available at https://arxiv.org/abs/1512.09080.

[2] Alon, N., Krivelevich, M. and Sudakov, B. (1998). Finding a large hidden clique in a random graph. *Random Structures Algorithms* **13,** 457–466.

[3] Ames, B. P. and Vavasis, S. A. (2011). Nuclear norm minimization for the planted clique and biclique problems. *Math. Program.* **129,** 69–89.

[4] Arias-Castro, E. and Verzelen, N. (2014). Community detection in dense random networks. *Ann. Statist.* **42,** 940–969.

[5] Banks, J., Moore, C., Neeman, J. and Netrapalli, P. (2016). Information-theoretic thresholds for community detection in sparse networks. In *Proceedings of the 29th Conference on Learning Theory*, pp. 383–416.

[6] Bordenave, C., Lelarge, M. and Massoulié, L. (2015). Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, IEEE, New York, pp. 1347–1357.

[7] Chen, Y. and Xu, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *J. Mach. Learn. Res.* **17,** 27.

[8] Davis, C. and Kahan, W. M. (1970). The rotation of eigenvectors by a perturbation. III. *SIAM J. Numer. Anal.* **7,** 1–46.

[9] Decelle, A., Krzakala, F., Moore, C. and Zdeborová, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84,** 066106.

[10] Dekel, Y., Gurel-Gurevich, O. and Peres, Y. (2014). Finding hidden cliques in linear time with high probability. *Combin. Prob. Comput.* **23,** 29–49.

[11] Deshpande, Y. and Montanari, A. (2015). Finding hidden cliques of size $\sqrt{N/e}$ in nearly linear time. *Found. Comput. Math.* **15,** 1069–1128.

[12] Deshpande, Y. and Montanari, A. (2015). Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. In *Proceedings of the 28th Conference on Learning Theory*, pp. 523–562.

[13] Feige, U. and Ron, D. (2010). Finding hidden cliques in linear time. In *Proceedings of 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms*, pp. 189–204.

[14] Hajek, B., Wu, Y. and Xu, J. (2015). Computational lower bounds for community detection on random graphs. In *Proceedings of the 28th Conference on Learning Theory*, pp. 899–928.

[15] Hajek, B., Wu, Y. and Xu, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inf. Theory* **62,** 2788–2797.

[16] Hajek, B., Wu, Y. and Xu, J. (2017). Information limits for recovering a hidden community. *IEEE Trans. Inf. Theory* **63,** 4729–4745.

[17] Hajek, B., Wu, Y. and Xu, J. (2018). Recovering a hidden community beyond the Kesten-Stigum threshold in $O(|E|\log^*|V|)$ time. Preprint. Available at https://arxiv.org/abs/1510.02786.

[18] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58,** 13–30.

[19] Holland, P. W., Laskey, K. B. and Leinhardt, S. (1983). Stochastic blockmodels: first steps. *Social Networks* **5,** 109–137.

[20] Hopkins, S. B., Kothari, P. K. and Potechin, A. (2015). SoS and planted clique: tight analysis of MPW moments at all degrees and an optimal lower bound at degree four. Preprint. Available at https://arxiv.org/abs/1507.05230.

[21] Jerrum, M. (1992). Large cliques elude the Metropolis process. *Random Structures Algorithms* **3,** 347–359.

[22] Kailath, T. (1967). The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans. Commun. Tech.* **15,** 52–60.

[23] Kesten, H. and Stigum, B. P. (1966). Additional limit theorems for indecomposable multidimensional Galton–Watson processes. *Ann. Math. Statist.* **37,** 1463–1481.

[24] Kobayashi, H. and Thomas, J. B. (1967). Distance measures and related criteria. In *Proceedings of Fifth Allerton Conference Circuit and System Theory* (Monticello, IL), pp. 491–500.

[25] Krzakala, F. *et al.* (2013). Spectral redemption in clustering sparse networks. *Proc. Nat. Acad. Sci. USA* **110,** 20935–20940.
[26] Massoulié, L. (2014). Community detection thresholds and the weak Ramanujan property. In *STOC'14—Proceedings of the 2014 ACM Symposium on Theory of Computing*, ACM, New York, pp. 694–703.
[27] McSherry, F. (2001). Spectral partitioning of random graphs. In *42nd IEEE Symposium on Foundations of Computer Science*, IEEE, Los Alamitos, CA, pp. 529–537.
[28] Meka, R., Potechin, A. and Wigderson, A. (2015). Sum-of-squares lower bounds for planted clique. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, ACM, New York, pp. 87–96.
[29] Mitzenmacher, M. and Upfal, E. (2005). *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.
[30] Montanari, A. (2015). Finding one community in a sparse graph. *J. Statist. Phys.* **161,** 273–299.
[31] Mossel, E. (2004). Survey: information flow on trees. In *Graphs, Morphisms and Statistical Physics*, American Mathematical Society, Providence, RI, pp. 155–170.
[32] Mossel, E., Neeman, J. and Sly, A. (2015). Consistency thresholds for the planted bisection model. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, ACM, New York, pp. 69–75.
[33] Mossel, E., Neeman, J. and Sly, A. (2015). Reconstruction and estimation in the planted partition model. *Prob. Theory Relat. Fields* **162,** 431–461.
[34] Mossel, E., Neeman, J. and Sly, A. (2017). A proof of the block model threshold conjecture. *Combinatorica* **37,** 44 pp.
[35] Poor, H. V. (1994). *An Introduction to Signal Detection and Estimation*, 2nd edn. Springer, New York.
[36] Raghavendra, P. and Schramm, T. (2016). Tight lower bounds for planted clique in the degree-4 SOS program. Preprint. Available at https://arxiv.org/abs/1507.05136.
[37] Yun, S.-Y. and Proutiere, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms. Preprint. Available at https://arxiv.org/abs/1412.7335.