JMLR: Workshop and Conference Proceedings vol 35:1–15, 2014

Belief Propagation, Robust Reconstruction and Optimal Recovery of Block Models

Elchanan Mossel MOSSEL@STAT.BERKELEY.EDU
Department of Statistics and Department of Computer Science, U.C. Berkeley

Joe Neeman JOENEEMAN@GMAIL.COM
Department of Electrical and Computer Engineering and Department of Math, U.T. Austin

Allan Sly SLY@STAT.BERKELEY.EDU
Department of Statistics, U.C. Berkeley

Abstract

We consider the problem of reconstructing sparse symmetric block models with two blocks and connection probabilities a/n and b/n for inter- and intra-block edge probabilities respectively. It was recently shown that one can do better than a random guess if and only if $(a-b)^2 > 2(a+b)$. Using a variant of Belief Propagation, we give a reconstruction algorithm that is *optimal* in the sense that if $(a-b)^2 > C(a+b)$ for some constant C then our algorithm maximizes the fraction of the nodes labelled correctly. Along the way we prove some results of independent interest regarding *robust reconstruction* for the Ising model on regular and Poisson trees.

1. Introduction

1.1. Sparse Stochastic Block Models

Stochastic block models were introduced almost 30 years ago by Holland et al. (1983) in order to study the problem of community detection in random graphs. In these models, the nodes in a graph are divided into two or more communities, and then the edges of the graph are drawn independently at random, with probabilities depending on which communities the edge lies between. In its simplest incarnation – which we will study here – the model has n vertices divided into two classes of approximately equal size, and two parameters: a/n is the probability that each within-class edge will appear, and b/n is the probability that each between-class edge will appear. Since their introduction, a large body of literature has been written about stochastic block models, and a multitude of efficient algorithms have been developed for the problem of inferring the underlying communities from the graph structure. To name a few, we now have algorithms based on maximum-likelihood methods (Snijders and Nowicki (1997)), belief propagation (Decelle et al. (2011)), spectral methods (McSherry (2001)), modularity maximization (Bickel and Chen (2009)), and a number of combinatorial methods (Bui et al. (1987); Dyer and Frieze (1989); Jerrum and Sorkin (1998); Condon and Karp (2001)).

Early work on the stochastic block model mainly focused on fairly dense graphs: Dyer and Frieze (1989); Snijders and Nowicki (1997); and Condon and Karp (2001) all gave algorithms that will correctly recover the exact communities in a graph from the stochastic block model, but only when a and b are polynomial in n. McSherry (2001) broke this polynomial barrier by giving a spectral algorithm which succeeds when a and b are logarithmic in n; this was later equalled by Bickel and Chen (2009) using an algorithm based on modularity maximization.

The $O(\log n)$ barrier is important because if the average degree of a block model is logarithmic or larger, it is possible to exactly recover the communities with high probability as $n \to \infty$. On the other hand, if the average degree is less than logarithmic then some fairly straightforward probabilistic arguments show that it is not possible to completely recover the communities. When the average degree is constant, as it will be in this work, then one cannot get more than a constant fraction of the labels correct.

Despite these apparent difficulties, there are important practical reasons for considering block models with constant average degree. Indeed, many real networks are very sparse. For example, Leskovec et al. (2008) and Strogatz (2001) collected and studied a vast collection of large network datasets, many of which had millions of nodes, but most of which had an average degree of no more than 20; for instance, the LinkedIn network studied by Leskovec et al. had approximately seven million nodes, but only 30 million edges. Moreover, the very fact that sparse block models are impossible to infer exactly may be taken as an argument for studying them: in real networks one does not expect to recover the communities with perfect accuracy, and so it makes sense to study models in which this is not possible either.

Although sparse graphs are immensely important, there is not yet much known about very sparse stochastic block models. In particular, there is a gap between what is known for block models with a constant average degree and those with an average degree that grows with the size of the graph. In the latter case, it is often possible – by one of the methods mentioned above – to exactly identify the communities with high probability. On the other hand, simple probabilistic arguments show that complete recovery of the communities is not possible when the average degree is constant. Until very recently, there was only one algorithm – due to Coja-Oghlan (2010), and based on spectral methods – which was guaranteed to do anything at all in the constant-degree regime, in the sense that it produced communities which have a better-than-50% overlap with the true communities.

Despite the lack of rigorous results, a beautiful conjectural picture has emerged in the last few years, supported by simulations and deep but non-rigorous physical intuition. We are referring specifically to work of Decelle et al. (2011), who conjectured the existence of a threshold, below which is it not possible to find the communities better than by guessing randomly. In the case of two communities of equal size, they pinpointed the location of the conjectured threshold. This threshold has since been rigorously confirmed; a sharp lower bound on its location was given by Mossel et al. (2013), while sharp upper bounds were given independently by Massoulié (2014); Mossel et al. (2014a).

1.2. Our results: optimal reconstruction

Given that even above the threshold, it is not possible to completely recover the communities in a sparse block model, it is natural to ask how accurately one may recover them. In Mossel et al. (2013), we gave an upper bound on the recovery accuracy; here, we will show that that bound is tight – at least, when the signal to noise ratio is sufficiently high – by giving an algorithm which performs as well as the upper bound.

Our algorithm, which is based on belief propagation (BP), is essentially an algorithm for locally improving an initial guess at the communities. In our current analysis, we assume that we are given a black-box algorithm for providing this initial guess; both Mossel et al. (2014a) and Massoulié (2014) provide algorithms meeting our requirements. We should mention that standard BP with random uniform initial messages, without our modifications, and also without a good initial guess, is also conjectured by Decelle et al. (2011) to have optimal accuracy. However, we do not know of any approach to analyze the vanilla version of BP for this problem. Indeed, the performance of global BP on graphs is a major open question (see, e.g., Ihler et al. (2005)) and even our local analysis is highly non-trivial.

We should point out that spectral algorithms – which, due to their efficiency, are very popular algorithms for this model – empirically do not perform as well as BP on very sparse graphs (see, e.g., Krzakala et al. (2013)). This is despite the recent appearance of two new spectral algorithms, due to Krzakala et al. (2013) and Massoulié (2014), which were specifically designed for clustering sparse block models. The algorithm of Krzakala et al. (2013) is particularly relevant here, because it was derived by linearizing belief propagation; empirically, it performs well all the way to the impossibility threshold, although not quite as well as BP. Intuitively, the linear aspects of spectral algorithms (i.e., the fact that they can be implemented – via the power method – using local linear updates) explain why they cannot achieve optimal performance. Indeed, since the optimal local updates – those given by BP – are non-linear, then any method based on linear updates will be suboptimal.

As a major part of our analysis, we prove a result about broadcast processes on trees, which may be of independent interest. Specifically, we prove that if the signal-to-noise ratio of the broadcast process is sufficiently high, then adding extra noise at the leaves of a large tree does not hurt our ability to guess the label of the root given the labels of the leaves. In other words, we show that for a certain model on trees, belief propagation initialized with arbitrarily noisy messages converges to the optimal solution as the height of the tree tends to infinity. We prove our result for regular trees and Galton-Watson trees with Poisson offspring, but we conjecture that it also holds for general trees, and even if the signal-to-noise ratio is low.

2. Definitions and main results

2.1. The block model

In this article, we restrict the stochastic block model to the case of two classes with roughly equal size.

Definition 1 (Block Model) *The* block model *on n nodes is constructed by first labelling each node $+$ or $-$ with equal probability independently. Then each edge is included in the graph independently, with probability a/n if its endpoints have the same label and b/n otherwise. Here a and b are two positive parameters. We write $\mathcal{G}(n, a/n, b/n)$ for this distribution of (labelled) graphs.*

For us, a and b will be fixed, while n tends to infinity. More generally one may consider the case where a and b may be allowed to grow with n.

As conjectured by Decelle et al. (2011), the relationship between $(a-b)^2$ and $(a+b)$ turns out to be of critical importance for the reconstructability of the block model:

Theorem 2 (Mossel et al. (2013, 2014a); Massoulié (2014)) *For the block models with parameters a and b it holds that*

- *If $(a-b)^2 \leq 2(a+b)$ then the node labels cannot be inferred from the unlabelled graph with better than 50% accuracy (which could also be done just by random guessing).*
- *if $(a-b)^2 > 2(a+b)$ then it is possible to infer the labels with better than 50% accuracy.*

2.2. Broadcasting on Trees

The proof in Mossel et al. (2013) will be important to us here, so we will introduce one of its main ingredients, the *broadcast process on a tree.*

Consider an infinite, rooted tree. We will identify such a tree T with a subset of $\mathbb{N}^*$, the set of finite strings of natural numbers, with the property that if $v \in T$ then any prefix of v is also in T. In this way, the root of the tree is naturally identified with the empty string, which we will denote by ρ. We will write uv for the concatenation of the strings u and v, and $L_k(u)$ for the kth-level descendents of u; that is, $L_k(u) = \{uv \in T : |v| = k\}$. Also, we will write $\mathcal{C}(u) \subset \mathbb{N}$ for the indices of u's children relative to itself. That is, $i \in \mathcal{C}(u)$ if and only if $ui \in L_1(u)$.

Definition 3 (Broadcast process on a tree) *Given a parameter $\eta \neq 1/2$ in $[0,1]$ and a tree T, the* broadcast process on T *is a two-state Markov process $\{\sigma_u : u \in T\}$ defined as follows: let σ_ρ be + or − with probability $\frac{1}{2}$. Then, for each u such that σ_u is defined and for each $v \in L_1(u)$, let $\sigma_v = \sigma_u$ with probability $1-\eta$ and $\sigma_v = -\sigma_\rho$ otherwise.*

This broadcast process has been extensively studied, where the major question is whether the labels of vertices far from the root of the tree give any information on the label of the root. For general trees, this question was solved definitively by Evans et al. (2000), after many other contributions including those by Kesten and Stigum (1966); Bleher et al. (1995). The complete statement of the theorem requires the notion of *branching number*, which we would prefer not to define here (see Evans et al. (2000)). For our purposes it suffices to know that a $(d+1)$-regular tree has branching number d and that a Poisson branching process tree with mean $d > 1$ has branching number d (almost surely, and conditioned on non-extinction).

Theorem 4 (Tree reconstruction threshold, Evans et al. (2000)) *Let $\theta = 1 - 2\eta$ and d be the branching number of T. Then $\mathbb{E}[\sigma_\rho \mid \sigma_u : u \in L_k(\rho)] \to 0$ in probability as $k \to \infty$ if and only if $d\theta^2 \leq 1$*

The theorem implies in particular that if $d\theta^2 > 1$ then for every k there is an algorithm which guesses σ_ρ given $\sigma_{L_k(\rho)}$, and which succeeds with probability bounded away from 1/2. If $d\theta^2 \leq 1$ there is no such algorithm.

Janson and Mossel (2004) considered a version of the tree broadcast process that has extra noise at the leaves:

Definition 5 (Noisy broadcast process on a tree) *Given a broadcast process σ on a tree T and a parameter $\delta \in [0, 1/2)$, the* noisy broadcast process on T *is the process $\{\tau_u : u \in T\}$ defined by independently taking $\tau_u = -\sigma_u$ with probability δ and $\tau_u = \sigma_u$ otherwise.*

We observe that the noise present in σ and the noise present in τ have qualitatively different roles, since the noise present in σ propagates down the tree while the noise present in τ does not. Janson and Mossel (2004) showed that the range of parameters for which σ_ρ may be reconstructed from σ_{L_k} is the same as the range for which σ_ρ may be reconstructed from τ_{L_k}. In other words, additional noise at the leaves has no effect on whether the root's signal propagates arbitrarily far. One of our main results is a quantitative version of this statement (Theorem 11): we show that for a certain range of parameters, the presence of noise at the leaves does not even affect the accuracy with which the root can be reconstructed.

2.3. The block model and broadcasting on trees

The connection between the community reconstruction problem on a graph and the root reconstruction problem on a tree was first pointed out in Decelle et al. (2011) and made rigorous in Mossel et al. (2013). The basic idea is the following:

- A neighborhood in G looks like a Galton-Watson tree with offspring distribution $\mathrm{Pois}((a+b)/2)$ (which almost surely has branching number $d = (a+b)/2$).
- The labels on the neighborhood look as though they came from a broadcast process with parameter $\eta = \frac{b}{a+b}$.
- With these parameters, $\theta^2 d = \frac{(a-b)^2}{2(a+b)}$, and so the threshold for community reconstruction is the same as the proven threshold for tree reconstruction.

This local approximation can be formalized as convergence locally on average, a type of local weak convergence defined in Montanari et al. (2012). We should mention that in the case of more than two communities (i.e. in the case that the broadcast process has more than two states) then the picture becomes rather more complicated, and much less is known, see Decelle et al. (2011); Mossel et al. (2013) for some conjectures.

2.4. Reconstruction probabilities on trees and graphs

Note that Theorem 4 only answers the question of whether one can achieve asymptotic reconstruction accuracy better than 1/2. Here, we will be interested in more detailed information about the actual accuracy of reconstruction, both on trees and on graphs.

Note that in the tree reconstruction problem, the optimal estimator of σ_ρ given $\sigma_{L_k(\rho)}$ is easy to write down: it is simply the sign of $X_{\rho,k} := 2\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - 1$. Compared to the trivial procedure of guessing σ_ρ completely at random, this estimator has an expected gain of $\mathbb{E}\left|\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - \frac{1}{2}\right|$. It is therefore natural to define:

Definition 6 (Tree reconstruction accuracy) *Let T be an infinite Galton-Watson tree with $\mathrm{Pois}((a+b)/2)$ offspring distribution, and $\eta = \frac{b}{a+b}$. Consider the broadcast process σ on the tree with parameters a, b and define:*

$$p_T(a,b) = \frac{1}{2} + \lim_{k\to\infty} \mathbb{E}\left|\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - \frac{1}{2}\right| \tag{1}$$

to be the probability of correctly inferring σ_ρ given the "labels at infinity."

We remark that the limit always exists because the right-hand side is non-increasing in k. Moreover, Theorem 4 implies that $p_T(a,b) > 1/2$ if and only if $(a-b)^2 > 2(a+b)$.

One of the main results of Mossel et al. (2013) is that the graph reconstruction problem is harder than the tree reconstruction problem in the sense that for any community-detection algorithm, the asymptotic accuracy of that algorithm is bounded by $p_T(a,b)$.

Definition 7 (Graph reconstruction accuracy) *Let (G,σ) be a labelled graph on n nodes. If f is a function that takes a graph and returns a labelling of it, we write*

$$\operatorname{acc}(f,G,\sigma) = \frac{1}{2} + \left|\frac{1}{n}\sum_v 1((f(G))_v = \sigma_v) - \frac{1}{2}\right|$$

for the accuracy of f in recovering the labels σ. For $\epsilon > 0$, let

$$p_{G,n,\epsilon}(a,b) = \sup_f \sup \left\{p : \Pr(\operatorname{acc}(f,G,\sigma) \ge p) \ge \epsilon\right\}.$$

where the first supremum ranges over all functions f, and the probability is taken over $(G,\sigma) \sim \mathcal{G}(n,a/n,b/n)$. Let $p_G(a,b) = \sup_{\epsilon>0} \limsup_{n\to\infty} p_{G,n,\epsilon}(a,b)$.

One should think of $p_G(a,b)$ as the optimal fraction of nodes that can be reconstructed correctly by any algorithm (not necessarily efficient) that only gets to observe an unlabelled graph. More precisely, for any algorithm and any $p > p_G(a,b)$, the algorithm's probability of achieving accuracy p or higher converges to zero as n grows. Note that the symmetry between the $+$ and $-$ is reflected in the definition of acc (for example, in the appearance of the constant $1/2$), and also that acc is defined to be large if f gets most labels *incorrect* (because there is no way for an algorithm to break the symmetry between $+$ and $-$).

An immediate corollary of the analysis of Mossel et al. (2013) implies that graph reconstruction is always less accurate than tree reconstruction:

Theorem 8 (Mossel et al. (2013)) $p_G(a,b) \le p_T(a,b)$.

Note that Theorems 4 and 8 imply the first part of Theorem 2. We remark that Theorem 8 is not stated explicitly in Mossel et al. (2013); because the authors were only interested in the case $(a-b)^2 \le 2(a+b)$, the claimed result was that $(a-b)^2 \le 2(a+b)$ implies $p_G(a,b) = \frac{1}{2}$. However, a cursory examination of the proof of (Mossel et al., 2013, Theorem 1) reveals that the claim was proven in two stages: first, they prove via a coupling argument that $p_G(a,b) \le p_T(a,b)$ and then they apply Theorem 4 to show that $(a-b)^2 \le 2(a+b)$ implies $p_T(a,b) = \frac{1}{2}$.

2.5. Our results

In this paper, we consider the high signal-to-noise case, namely the case that $(a-b)^2$ is significantly larger than $2(a+b)$. In this regime, we give an algorithm (Algorithm 1) which achieves an accuracy of $p_T(a,b)$.

Theorem 9 *There exists a constant C such that if $(a-b)^2 \geq C(a+b)$ then*

$$p_G(a,b) = p_T(a,b),$$

Moreover, there is a polynomial time algorithm such that for all such a,b and every $\epsilon > 0$, with probability tending to one as $n \to \infty$, the algorithm reconstructs the labels with accuracy $p_G(a,b) - \epsilon$.

A key ingredient of the proof is a procedure for amplifying a clustering that is a slightly better than a random guess to obtain optimal clustering. In order to discuss this procedure, we define the problem of "robust reconstruction" problem on trees.

Definition 10 (Robust tree reconstruction accuracy) *Consider the noisy tree broadcast process with parameters $\eta = \frac{a}{a+b}$ and $\delta \in [0,1/2)$ on a Galton-Watson tree with offspring distribution* Pois$((a+b)/2)$*. We define the robust reconstruction accuracy as:*

$$\widetilde{p}_T(a,b) = \frac{1}{2} + \lim_{\delta \to 1/2} \lim_{k \to \infty} \mathbb{E} \left| \Pr(\sigma_\rho = + \mid \tau_{L_k(\rho)}) - \frac{1}{2} \right|$$

In our main technical result we show that when $a-b$ is large enough then in fact the extra noise does not have any effect on the reconstruction accuracy.

Theorem 11 *There exists a constant C such that if $(a-b)^2 \geq C(a+b)$ then $\widetilde{p}_T(a,b) = p_T(a,b)$.*

We conjecture that $p_T = \widetilde{p}_T$ for any parameters, and also for more general trees; however, our proof does not naturally extend to cover these cases.

2.6. Algorithmic amplification and robust reconstruction

Our second main result connects the community detection problem to the robust tree reconstruction problem: we show that given a suitable algorithm for providing an initial guess at the communities, the community detection problem is easier than the robust reconstruction problem, in the sense that one can achieve an accuracy of $\widetilde{p}_T(a,b)$.

Theorem 12 *Consider an algorithm for reconstructing the block models which satisfies that with high probability it labels $\frac{1}{2} + \delta$ of the nodes accurately. Then the algorithm can be used in a black box manner to provide an algorithm whose reconstruction accuracy (with high probability) is $\widetilde{p}_T(a,b)$.*

Combining Theorem 12 with Theorem 11 proves that our algorithm obtains accuracy p_T provided that $(a-b)^2 \geq C(a+b)$. By Theorem 8 this accuracy is optimal, thereby justifying the claim that our algorithm is optimal. We remark that Theorem 12 easily extends to other versions of the block model (i.e., models with more clusters or unbalanced classes); however, Theorem 11 does not. In particular, Theorem 9 does not hold for general block models. In fact, one fascinating conjecture of Decelle et al. (2011) says that for general block models, computational hardness enters the picture (whereas it does not play any role in our current work).

2.7. Algorithm Outline

Before getting into the technical details, let us give an outline of our algorithm: for every node u, we remove a neighborhood (whose radius r is slowly increasing with n) of u from the graph G. We then run a black-box community-detection algorithm on what remains of G. This is guaranteed to produce some communities which are correlated with the true ones, but they may not be optimally accurate. Then we return the neighborhood of u to G, and we consider the inferred communities on the boundary of that neighborhood. Now, the neighborhood of u is like a tree, and the true labels on its boundary are distributed like $\sigma_{L_r(u)}$. The inferred labels on the boundary are hence distributed like $\tau_{L_r(u)}$ for some $0 \le \delta < 1/2$, and so we can guess the label of u from them using robust tree reconstruction. Since robust tree reconstruction succeeds with probability p_T regardless of δ, our algorithm attains this optimal accuracy even if the black-box algorithm does not.

To see the connection between our algorithm and belief propagation, note that finding the optimal estimator for the tree reconstruction problem requires computing $\Pr(\sigma_u \mid \tau_{L_r(u)})$. On a tree, the standard algorithm for solving this is exactly belief propagation. In other words, our algorithm consists of multiple local applications of belief propagation. Although we believe that a single global run of belief propagation would attain the same performance, these local instances are rather more feasible to analyze.

3. Robust Reconstruction on Regular Trees

The main technical effort of this paper is the proof of Theorem 11. Since the proof is quite involved we will only give an outline, only in the case of d-regular trees (instead of the Galton-Watson case claimed in Theorem 11), and without the sharp dependence of d on θ. For the complete proof, see the full version of this paper (Mossel et al. (2014b)).

Theorem 13 *Consider the broadcast process on the infinite $\frac{a+b}{2} = d$-ary tree with parameter $\eta = \frac{a}{a+b}$. Set $\theta = 1 - 2\eta$. For any $0 < \theta^* < 1$, there is some $d^* = d^*(\theta^*)$ such that if $\theta \ge \theta^*$ and $d \ge d^*$ then $\widetilde{p}_T(a,b) = p_T(a,b)$.*

We remark that the statement of Theorem 13 is not precise, because we have only defined $p_T(a,b)$ for Galton-Watson trees. A more precise statement will follow shortly.

3.1. Magnetization

Define

$$\begin{aligned} X_{u,k} &= \Pr(\sigma_u = + \mid \sigma_{L_k(u)}) - \Pr(\sigma_u = - \mid \sigma_{L_k(u)}) \\ x_k &= \mathbb{E}(X_{u,k} \mid \sigma_u = +). \end{aligned}$$

Here, we say that $X_{u,k}$ is the *magnetization* of u given $\sigma_{L_k(u)}$. Note that by the homogeneity of the tree, the definition of x_k is independent of u. A simple application of Bayes' rule (see Lemma 1 of Borgs et al. (2006)) shows that $(1 + x_k)/2$ is the probability of estimating σ_ρ correctly given $\sigma_{L_k(\rho)}$.

We may also define the noisy magnetization Y:

$$\begin{aligned} Y_{u,k} &= \Pr(\sigma_u = + \mid \tau_{L_k(u)}) - \Pr(\sigma_u = - \mid \tau_{L_k(u)}) \\ y_k &= \mathbb{E}(Y_{u,k} \mid \sigma_u = +). \end{aligned} \tag{2}$$

As above, $(1 + y_k)/2$ is the probability of estimating σ_ρ correctly given $\tau_{L_k(\rho)}$. In particular, Theorem 13 may be re-stated (more precisely) as follows:

Theorem 14 *Under the assumptions of Theorem 13,* $\lim_{k\to\infty} x_k = \lim_{k\to\infty} y_k$.

We will prove Theorem 14 by studying certain recursions. Indeed, Bayes' rule implies the following recurrence for X (see, eg., Sly (2011)):

$$X_{u,k} = \frac{\prod_{i\in\mathcal{C}(u)}(1 + \theta X_{ui,k-1}) - \prod_{i\in\mathcal{C}(u)}(1 - \theta X_{ui,k-1})}{\prod_{i\in\mathcal{C}(u)}(1 + \theta X_{ui,k-1}) + \prod_{i\in\mathcal{C}(u)}(1 - \theta X_{ui,k-1})}. \tag{3}$$

3.2. The simple majority method

Our first step in proving Theorem 14 is to show that when $\theta^2 d$ is large, then both the exact reconstruction and the noisy reconstruction do quite well. While it is possible to do so by studying the recursion (3), such an analysis is actually quite delicate. Instead, we will show this by studying a completely different estimator: the one which is equal to the most common label among $\sigma_{L_k(\rho)}$. This estimator is easy to analyze, and it performs quite well; since the estimator based on the sign of $X_{\rho,k}$ is optimal, it performs even better. The study of the simple majority estimator is quite old, having essentially appeared in the paper of Kesten and Stigum (1966) that introduced the tree broadcast model. Therefore, we omit the proofs of what follows.

Suppose $d\theta^2 > 1$. Define $S_{u,k} = \sum_{v\in L_k(u)} \sigma_v$ and set $\tilde{S}_{u,k} = \sum_{v\in L_k(u)} \tau_v$. We will attempt to estimate σ_ρ by $\operatorname{sgn}(S_{\rho,k})$ or $\operatorname{sgn}(\tilde{S}_{\rho,k})$; when $\theta^2 d$ is large enough, these estimators turn out to perform quite well. This may be shown by calculating the first two moments of $S_{u,k}$ and $\tilde{S}_{u,k}$.

Lemma 15 *If* $\theta^2 d > 1$ *then* $\frac{\operatorname{Var}^+ S_k}{(\mathbb{E}^+ S_k)^2} \overset{k\to\infty}{\to} \frac{4\eta(1-\eta)}{\theta^2 d}$ *and* $\frac{\operatorname{Var}^+ \tilde{S}_k}{(\mathbb{E}^+ \tilde{S}_k)^2} \overset{k\to\infty}{\to} \frac{4\eta(1-\eta)}{\theta^2 d}$.

By Chebyshev's inequality, the estimators $\operatorname{sgn}(S_k)$ and $\operatorname{sgn}(\tilde{S}_k)$ succeed with probability at least $1 - \frac{4\eta(1-\eta)}{\theta^2 d^2}$ as $k \to \infty$. Now, $\operatorname{sgn}(Y_{\rho,k})$ is the optimal estimator of σ_ρ given τ_{L_k}, and its success probability is exactly $(1 + y_k)/2$. Hence $(1 + y_k)/2$ must be larger than the success probability of $\operatorname{sgn}(\tilde{S}_k)$ (and similarly for x_k and $\operatorname{sgn}(S_k)$). Putting this all together, we see that x_k and y_k are both at least $1 - \frac{10\eta(1-\eta)}{\theta^2 d}$ for large k. Finally, we apply Markov's inequality to show that $X_{u,k}$ and $Y_{u,k}$ are large with high probability (conditioned on $\sigma_u = +$).

Lemma 16 *There is a constant C such that for all $k \ge K(\delta)$ and all $t > 0$*

$$\begin{aligned} \Pr\left(X_{u,k} \ge 1 - t\frac{\eta}{\theta^2 d} \;\middle|\; \sigma_u = +\right) \ge 1 - Ct^{-1} \\ \Pr\left(Y_{u,k} \ge 1 - t\frac{\eta}{\theta^2 d} \;\middle|\; \sigma_u = +\right) \ge 1 - Ct^{-1}. \end{aligned}$$

As we will see, Lemma 16 and the recursion (3) are really the only properties of Y that we will use. This fact is actually important for the full proof of Theorem 9, where we will take a slightly different definition of Y. See the full version for details.

3.3. Analysis of the magnetization recurrence

In this section, we study the recurrence (3) and derive the following recurrence for the distance between X and Y:

Proposition 17 *For any $0 < \theta^* < 1$, there is some $d^* = d^*(\theta^*)$ such that for all $\theta \geq \theta^*$, $d \geq d^*$, and $k \geq K(\theta, d, \delta)$,*

$$\mathbb{E}\sqrt{|X_{\rho,k+1} - Y_{\rho,k+1}|} \leq \frac{1}{2}\mathbb{E}\sqrt{|X_{\rho,k} - Y_{\rho,k}|}.$$

By sending $k \to \infty$, Proposition 17 immediately implies Theorem 13.

Let $g : \mathbb{R}^d \to \mathbb{R}$ denote the function

$$g(x) = \frac{\prod_{i=1}^d (1 + \theta x_i) - \prod_{i=1}^d (1 - \theta x_i)}{\prod_{i=1}^d (1 + \theta x_i) + \prod_{i=1}^d (1 - \theta x_i)}. \tag{4}$$

Then the recurrence (3) may be written as $X_{u,k+1} = g(X_{u1,k}, \ldots, X_{ud,k})$. We will also abbreviate $(X_{u1,k}, \ldots, X_{ud,k})$ by $X_{L_1(u),k}$, so that we may write $X_{u,k+1} = g(X_{L_1(u),k})$. After some calculations (involving differentiating g and applying the mean value theorem), we have

$$|g(x) - g(y)| \leq \sum_{i=1}^d |x_i - y_i| \max\{m_i(x), m_i(y)\}, \tag{5}$$

where

$$m_i(x) := \frac{1}{\eta^2} \prod_{j \neq i} \frac{1 - \theta x_j}{1 + \theta x_j}$$

The point is that if $\sigma_u = +$ (which we may assume, by symmetry) then for most $v \in L_1(u)$, $X_{v,k}$ will be close to 1 and so $m_i(X_{L_1(u),k})$ will be small; applying the same logic to Y, we will deduce from (5) that $|X_{u,k+1} - Y_{u,k+1}| = |g(X_{L_1(u),k}) - g(Y_{L_1(u),k})|$ is typically much smaller than $|X_{ui,k} - Y_{ui,k}|$. One difficulty in this analysis is that there is some low probability of having many $X_{v,k}$ close to -1. If θ is large, then $m_i(X_{L_1(u),k})$ will be very large on this event: so large that this low-probability event may have an effect after taking expectations. One solution is to take the square root, which reduces the impact of this bad event:

$$\sqrt{|g(x) - g(y)|} \leq \sum_{i=1}^d \sqrt{|x_i - y_i|} \max\{\sqrt{m_i(x)}, \sqrt{m_i(y)}\}, \tag{6}$$

Note that (conditioned on $\sigma_u = +$) $X_{ui,k}$ is independent of $m_i(X_{L_1(u),k})$ because $m_i(x)$ does not depend on x_i. Taking expectations of (6), and using the fact that $X_{u,k+1} = g(X_{L_1(u),k})$, we obtain

$$\begin{aligned}
&\mathbb{E}\left(\sqrt{|X_{u,k+1} - Y_{u,k+1}|}\middle|\sigma_u = +\right) \\
&\leq \sum_i \mathbb{E}\left(\sqrt{|X_{ui,k} - Y_{ui,k}|}\middle|\sigma_u = +\right) \mathbb{E}\left(\sqrt{\max\{m_i(X_{L_1(u),k}), m_i(Y_{L_1(u),k})\}}\middle|\sigma_u = +\right).
\end{aligned} \tag{7}$$

To prove Proposition 17, it therefore suffices to show that $\mathbb{E}(\sqrt{m_i(X_{L_1(u),k})} \mid \sigma_u = +)$ and $\mathbb{E}(\sqrt{m_i(Y_{L_1(u),k})} \mid \sigma_u = +)$ are both small. Since $m_i(X_{L_1(u),k})$ is a product of independent (when conditioned on σ_u) terms, it is enough to show that each of these terms has small expectation. This is done using Lemma 16:

Lemma 18 *For any $0 < \theta^* < 1$, there is some $d^* = d^*(\theta^*)$ and some $\lambda = \lambda(\theta^*) < 1$ such that for all $\theta \geq \theta^*$, $d \geq d^*$ and $k \geq K(\theta, d, \delta)$,*

$$\mathbb{E}\left(\sqrt{\frac{1-\theta X_{ui,k}}{1+\theta X_{ui,k}}} \;\middle|\; \sigma_u = +\right) \leq \min\{\lambda, 4\eta^{1/4}\}.$$

Finally, applying Lemma 18 to (7) completes the proof of Proposition 17.

4. From trees to graphs

In this section, we will give our reconstruction algorithm and prove that it performs optimally. We begin by letting `BBPartition` denote the algorithm of Mossel et al. (2014a), which satisfies the following guarantee, where V^i denotes $\{v \in V(G) : \sigma_v = i\}$:

Theorem 19 *Suppose that $G \sim \mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$. There exists some $0 \leq \delta < \frac{1}{2}$ such that as $n \to \infty$, `BBPartition` a.a.s. produces a partition $W^+ \cup W^- = V(G)$ such that $|W^+| = |W^-| + o(n) = \frac{n}{2} + o(n)$ and $|W^+ \Delta V^i| \leq \delta n$ for some $i \in \{+, -\}$.*

Remark 20 *Theorem 19 does not appear in Mossel et al. (2014a) exactly as we have quoted it. In the full version of this paper, we remark on how the result we have quoted (in particular, the condition $|W^+|, |W^-| = n/2 \pm o(n)$) follows from Theorem 2.1 of Mossel et al. (2014a).*

Note that by symmetry, Theorem 19 also implies that $|W^- \Delta V^j| \leq \delta n$ for $j \neq i \in \{+, -\}$. In other words, `BBPartition` recovers the correct partition up to a relabelling of the classes and an error bounded away from $\frac{1}{2}$. Note that $|W^+ \Delta V^i| = |W^- \Delta V^j|$. Let δ be the (random) fraction of vertices that are mis-labelled.

For $v \in G$ and $R \in \mathbb{N}$, define $B(v, R) = \{u \in G : d(u, v) \leq R\}$ and $S(v, R) = \{u \in G : d(u, v) = R\}$. If $B(v, R)$ is a tree (which it is a.a.s.), and τ is a labelling on $S(v, R)$, then we define $Y_{v,R}(\tau)$ as follows: set $Y_{u,0}(\tau) = \tau_u(1 - 2\delta)$ for $u \in S(v, R)$ and then for $u \in S(v, R - k)$ with $k \geq 1$, define $Y_{u,k}(\tau)$ recursively by $Y_{u,k} = g(Y_{L_1(u),k-1})$.

We remark that this definition requires knowing the exact δ that is guaranteed by Theorem 19. In the full version, we show how to get by without knowing δ.

Algorithm 1: Optimal graph reconstruction algorithm

```
$R \leftarrow \lfloor \frac{1}{10(2(a+b))} \log n \rfloor$ ;
Take $U \subset V$ to be a random subset of size $\lfloor \sqrt{n} \rfloor$ ;
Let $u_* \in U$ be a random vertex in $U$ with at least $\sqrt{\log n}$ neighbors in $V \setminus U$ ;
$W_*^+, W_*^- \leftarrow \emptyset$ ;
for $v \in V \setminus U$ do
    $W_v^+, W_v^- \leftarrow$ BBPartition($G \setminus B(v, R-1) \setminus U$) ;
    if $a > b$ then
        relabel $W_v^+, W_v^-$ so that $u_*$ has more neighbors in $W_v^+$ than $W_v^-$
    else
        relabel $W_v^+, W_v^-$ so that $u_*$ has more neighbors in $W_v^-$ than $W_v^+$
    end
    Define $\xi \in \{+,-\}^{S(v,R)}$ by $\xi_u = i$ if $u \in W_v^i$ ;
    Add $v$ to $W_*^{\mathrm{sgn}(Y_{v,R}(\xi))}$ ;
end
for $v \in U$ do
    Assign $v$ to $W_*^+$ or $W_*^-$ uniformly at random ;
end
```

As presented, our algorithm is not particular efficient (although it does run in polynomial tiem) because we need to re-run `BBPartition` for almost every vertex in V. However, one can modify Algorithm 1 to run in almost-linear time by processing $o(n)$ vertices in each iteration (a similar idea is used in Mossel et al. (2014a)). Since vanilla belief propagation is much more efficient than Algorithm 1 and reconstructs (in practice) just as well, we have chosen not to present the faster version of Algorithm 1.

Theorem 21 *Algorithm 1 produces a partition $W_*^+ \cup W_*^- = V(G)$ such that a.a.s. $|W_*^+ \Delta V^i| \leq (1+o(1))n(1-p_T(a,b))$ for some $i \in \{+,-\}$.*

Note that Theorem 8 shows that for any algorithm, $|W_*^+ \Delta V^i| \geq (1-o(1))n(1-p_T(a,b))$ a.a.s. Hence, it is enough to show that $\mathbb{E}|W_*^+ \Delta V^i| \leq (1+o(1))n(1-p_T(a,b))$. Since Algorithm 1 treats every node equally, it is enough to show that there is some i such that for every $v \in V^i$,

$$\Pr(v \in W_*^+) \to p_T(a,b) \text{ a.s.} \tag{8}$$

Moreover, since $\Pr(v \in U) \to 0$, it is enough to show (8) for every $v \in V^i \setminus U$.

For the remainder of the section, we will sketch the proof of (8). First, we will deal with a technicality: in line 1, we are applying `BBPartition` to the subgraph of G induced by $V \setminus B(v, R-1) \setminus U$; this induced subgraph is not exactly distributed according to a block model. However, since $B(v, R-1)$ and U are both small sets it is possible to show that `BBPartition` nevertheless produces a labelling satisfying the claim in Theorem 19.

Next, let us discuss the purpose of u_* and line 1. Note that Algorithm 1 relies on multiple applications of `BBPartition`, each of which is only guaranteed to give a good labelling up to swapping $+$ and $-$. In order to get a consistent labelling at the end, we need to "align" these multiple applications of BBPartition. This is the purpose of u_*: since u_* has high degree and BBPartition labels most vertices correctly, it follows from the law of large numbers that with high probability, BBPartition gives most of u_*'s neighbors the same label as u_*. (For this to hold, we need independence between the label of u_* and the output of BBPartition; this is why we remove u_* from the graph before running BBPartition.)

From now on, suppose without loss of generality that $\sigma_{u^*} = +$. Thanks to the previous paragraph and Theorem 19, we see that the relabelling in lines 1 and 1 correctly aligns W_v^+ with V^+:

Lemma 22 *There is some $0 \le \delta < \frac{1}{2}$ such that for any $v \in V \setminus U$, $|W_v^+ \Delta V^+| \le \delta n$ a.a.s., with W_v^+ defined as in line 1 or line 1.*

To complete the proof of (8) (and hence Theorem 21), we need to discuss the coupling between graphs and trees. We will invoke a lemma from Mossel et al. (2013) which says that a neighborhood in G can be coupled with a multi-type branching process. Indeed, let T be a Galton-Watson tree with offspring distribution $\mathrm{Pois}((a+b)/2)$ and let σ' be a labelling on it, given by running the tree broadcast process with parameter $\eta = b/(a+b)$. We write T_R for $T \cup \mathbb{N}^R$; that is, the part of T which has depth at most R.

Lemma 23 *For any fixed $v \in G$, there is a coupling between (G, σ) and (T, σ') such that*

$$(B(v,R), \sigma_{B(v,R)}) = (T_R, \sigma'_{T_R}) \text{ a.a.s.}$$

Let us therefore examine the labelling $\{\xi_u : u \in S(v,R)\}$ produced in line 1 of Algorithm 1. Since ξ is independent of the edges from $B(v, R-1)$ to G', it follows that for every neighbor $w \in G'$ of $u \in B(v, R-1)$, we may generate (independently of the other neighbors) ξ_w by flipping σ_w with probability $(1 - o(1))\delta$. Hence, we see that ξ can be coupled a.a.s. with τ', where τ'_w is defined by flipping the label of σ'_w (independently for each w) with probability δ. In other words, the joint distribution of $B(v,R)$ and $\{\xi_u : u \in S(v,R)\}$ is a.a.s. the same as the joint distribution of T_R and $\{\tau'_u : u \in \partial T_R\}$. Hence, by Theorem 11,

$$\lim_{n\to\infty} \Pr(Y_{v,R}(\xi) = \sigma_v) = p_T(a,b).$$

By line 1 of Algorithm 1, this completes the proof of (8).

References

P.J. Bickel and A. Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Statist. Phys.*, 79(1-2):473–482, 1995.

C. Borgs, J. Chayes, E. Mossel, and S. Roch. The Kesten-Stigum reconstruction bound is tight for roughly symmetric binary channels. In *Proceedings of IEEE FOCS 2006*, pages 518–530, 2006.

T.N. Bui, S. Chaudhuri, F.T. Leighton, and M. Sipser. Graph bisection algorithms with good average case behavior. *Combinatorica*, 7(2):171–191, 1987.

A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

A. Condon and R.M. Karp. Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140, 2001.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physics Review E*, 84:066106, Dec 2011.

M.E. Dyer and A.M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

W. S. Evans, C. Kenyon, Yuval Y. Peres, and L. J. Schulman. Broadcasting on trees and the Ising model. *Ann. Appl. Probab.*, 10(2):410–433, 2000.

P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137, 1983. ISSN 0378-8733.

Alexander T Ihler, John Iii, and Alan S Willsky. Loopy belief propagation: Convergence and effects of message errors. In *Journal of Machine Learning Research*, pages 905–936, 2005.

S. Janson and E. Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Ann. Probab.*, 32:2630–2649, 2004.

M. Jerrum and G.B. Sorkin. The Metropolis algorithm for graph bisection. *Discrete Applied Mathematics*, 82(1-3):155–175, 1998.

H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.*, 37:1463–1481, 1966.

Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborov, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

J. Leskovec, K.J. Lang, A. Dasgupta, and M.W. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceeding of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

L. Massoulié. Community detection thresholds and the weak Ramanujan property. arXiv:1311:3085, 2014.

F. McSherry. Spectral partitioning of random graphs. In *42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.

Andrea Montanari, Elchanan Mossel, and Allan Sly. The weak limit of Ising models on locally tree-like graphs. *Probability Theory and Related Fields*, 152:31–51, 2012.

E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. arXiv:1202.4124, 2013.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. arXiv:1311.4115, 2014a.

Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propagation, robust reconstruction, and optimal recovery of block models. arXiv:1309.1380, 2014b.

A. Sly. Reconstruction for the Potts model. *Annals of Probability*, 39(4):1365–1406, 2011.

T.A.B. Snijders and K. Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, 1997.

S.H. Strogatz. Exploring complex networks. *Nature*, 410(6825):268–276, 2001.