arXiv: arXiv:1311.4115

BELIEF PROPAGATION, ROBUST RECONSTRUCTION AND OPTIMAL RECOVERY OF BLOCK MODELS

By Elchanan Mossel[*,‡,¶], Joe Neeman[*,‖] and Allan Sly[†,‡,§]

U.C. Berkeley[‡], Australian National University[§], University of Pennsylvania[¶], and U.T. Austin[‖]

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. Ours is the only algorithm proven to achieve the optimal fraction of 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 more than 30 years ago [13] 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 [27], belief propagation [10],

[*]Supported by NSF grant DMS-1106999 and DOD ONR grants N000141110140 and N00014-14-1-0823 and grant 328025 from the Simons Foundation

[†]Supported by an Alfred Sloan Fellowship and NSF grant DMS-1208338.

MSC 2010 subject classifications: Primary 05C80, 60J20; secondary 91D30

Keywords and phrases: stochastic block model, unsupervised learning, belief propagation, robust reconstruction

spectral methods [21], modularity maximization [2], and a number of combinatorial methods [7, 9, 11, 15].

Early work on the stochastic block model mainly focused on fairly dense graphs: Dyer and Frieze [11]; Snijders and Nowicki [27]; and Condon and Karp [9] 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. In a substantial improvement, McSherry [21] gave a spectral algorithm that succeeds when a and b are logarithmic in n; this had been anticipated previously by Boppana [5], but his proof was incomplete. McSherry's parameter range was later equalled by Bickel and Chen [2] using an algorithm based on modularity maximization.

We also note that related but different problems of planted coloring were studied in Blum and Spencer [4] in the dense case, and Alon and Kahale [1] in the sparse case.

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. [18] and Strogatz [28] 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. Until recently, there was only one algorithm – due to [8], 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 recently emerged, supported by simulations and deep but non-rigorous physical intuition. We are referring specifically to work of Decelle et al. [10], 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 the authors [24], while sharp upper bounds were given independently by Massoulié [20] and by the authors [25].

1.2. *Our results: optimal reconstruction.* Given that 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 [24], 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 main result may be stated informally as follows:[1]

THEOREM 1.1. *Let $p_G(a,b)$ be the highest asymptotic accuracy that any algorithm can achieve in reconstructing communities of the block model with parameters a and b. We provide an algorithm that achieves accuracy of $p_G(a,b)$ with probability tending to 1 as $n \to \infty$, provided that $(a-b)^2/(a+b)$ is sufficiently large.*

To put Theorem 1.1 into the context of earlier work [20, 24, 25] by the authors and Massoulié, those works showed that $p_G(a,b) > 1/2$ if and only if $(a-b)^2 > 2(a+b)$; in the case that $p_G(a,b) > 1/2$, they also provided algorithms whose accuracy was bounded away from $1/2$. However, those algorithms were not guaranteed (and are not expected) to have *optimal* accuracy, only *non-trivial* accuracy. In other words, previous results have shown that for every value of a, b such that $(a-b)^2 > 2(a+b)$ there exists an algorithm that recovers (with high probability) a fraction $q(a,b) > 1/2$ of the nodes correctly. Our results provide an algorithm that (when $(a-b)^2 > C(a+b)$ for a large constant C) recovers the optimal fraction of nodes $p_G(a,b)$ in the sense that it is information theoretically impossible for any other algorithms to recover a bigger fraction.

Our new algorithm, which is based on belief propagation, is essentially an algorithm for locally improving an initial guess at the communities. In our current analysis, the initial guess is provided by a previous algorithm of the

[1] An extended abstract stating the results of the current paper [23] appeared in the proceedings of COLT 2014 (where it won the best paper award).

authors [25], which we use as a black box. We should mention that standard belief propagation with random uniform initial messages and without our modifications and also without a good initial guess, is also conjectured to have optimal accuracy [10]. However, at the moment, we don't know of any approach to analyze the vanilla version of BP for this problem.

As a major part of our analysis, we prove a result about broadcast processes on trees that 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.

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., [17]). This is despite the recent appearance of two new spectral algorithms, due to [17] and [20], which were specifically designed for clustering sparse block models. The algorithm of [17] 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, any method based on linear updates will be suboptimal.

1.3. *Dramatis personae.* Before defining everything carefully, we briefly introduce the three main objects and their relationships.

- The *block model detection problem* is the problem of detecting communities in a sparse stochastic block model.
- In the *tree reconstruction problem*, there is a two-color branching process in which every node has some children of its own color and some children of the other color. We observe the family tree of this process and also all of the colors in some generation; the goal is to guess the color of the original node.
- The *robust tree reconstruction problem* is like the tree reconstruction problem, except that instead of observing exactly the colors in some

generation, our observations contain some noise.

The two tree problems are related to the block model problem because a neighborhood in the stochastic block model model looks like a random tree from one of the tree problems. This connection was proved in [24], who also showed that tree reconstruction is "easier" than the block model detection (in a sense that we will make precise later). The current work has two main steps: we show that block model detection is "easier" than robust tree reconstruction, and we show that – for a certain range of parameters – robust tree reconstruction is exactly as hard as tree reconstruction.

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 2.1 (Stochastic 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 [10], the relationship between $(a-b)^2$ and $(a+b)$ turns out to be of critical importance for the reconstructability of the block model:

THEOREM 2.2 (Threshold for non-trivial detection [20, 24, 25]). *For the block model with parameters a and b it holds that*

- *If $(a-b)^2 < 2(a+b)$ 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.* Our study of optimal reconstruction accuracy is based on the local structure of $\mathcal{G}(n, a/n, b/n)$, which requires the notion of 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 2.3 (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, independently for every $v \in L_1(u)$ 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 answered definitively by Evans et al. [12], after many other contributions including [3, 16]. The complete statement of the theorem requires the notion of *branching number*, which we would prefer not to define here (see [12]). For our purposes it suffices to know that a d-ary 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 2.4 (Tree reconstruction threshold [12]). *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.

2.3. *Robust reconstruction on trees.* Janson and Mossel [14] considered a version of the tree broadcast process that has extra noise at the leaves:

DEFINITION 2.5 (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 [14] showed that the range of parameters for which σ_ρ may be non-trivially reconstructed from σ_{L_k} is the same as the range for which σ_ρ may be non-trivially 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 2.11): 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.

2.4. *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 [10] and made rigorous in [24]. 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 conjectured 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 [22]. 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 [10, 24] for some conjectures.

2.5. *Reconstruction probabilities on trees and graphs.* Note that Theorem 2.4 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

$$\mathbb{E}\left|\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - \frac{1}{2}\right| = \frac{1}{2}\mathbb{E}[|\mathbb{E}[\sigma_\rho \mid \sigma_{L_k(\rho)}]|].$$

It is now natural to define:

DEFINITION 2.6 (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 on the tree with parameter η and define*

$$p_T(a,b) = \frac{1}{2} + \lim_{k\to\infty} \mathbb{E}\left|\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - \frac{1}{2}\right|. \tag{2.1}$$

In words, $p_T(a,b)$ is the probability of correctly inferring σ_ρ given the "labels at infinity."

Note that by Theorem 2.4, $p_T(a,b) > 1/2$ if and only if $(a-b)^2 > 2(a+b)$.

We remark that the limit in Definition 2.6 always exists because the right-hand side is non-increasing in k. To see this, it helps to write $p_T(a,b)$ in a different way: let μ_k^+ be the distribution of $\sigma_{L_k(\rho)}$ given $\sigma_\rho = +$ and let μ_k^- be the distribution of $\sigma_{L_k(\rho)}$ given $\sigma_\rho = -$. Then

$$\mathbb{E}\left|\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - \frac{1}{2}\right| = \frac{1}{2} d_{TV}(\mu_k^+, \mu_k^-),$$

where d_{TV} denotes the total variation distance. Next, note that since labels at levels $k' > k$ are independent of σ_ρ given $\sigma_{L_k(\rho)}$,

$$\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) = \Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}, \sigma_{L_{k+1}(\rho)}, \sigma_{L_{k+2}(\rho)}, \dots).$$

Hence, if we set ν_k^+ to be the distribution of $\{\sigma_{L_{k'}}(\rho) : k' \geq k\}$ and similarly for ν_k^-, we have

$$\mathbb{E}\left|\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - \frac{1}{2}\right| = \frac{1}{2} d_{TV}(\nu_k^+, \nu_k^-).$$

Now the right hand side is clearly non-increasing in k, because ν_{k+1}^+ can be obtained from ν_k by marginalization.

One of the main results of [24] is that the graph reconstruction problem is at least as hard as 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 2.7 (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*

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

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

$$p_{G,n,\epsilon}(a,b) = \sup_f \sup \{p : \Pr(\mathrm{acc}(f,G,\sigma) \geq p) \geq \epsilon\}.$$

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

$$p_G(a,b) = \lim_{\epsilon \to 0} \limsup_{n \to \infty} p_{G,n,\epsilon}(a,b),$$

where the limit exists because $p_{G,n,\epsilon}(a,b)$ is monotonic in ϵ.

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 [24] implies that graph reconstruction is always at most as accurate as tree reconstruction:

THEOREM 2.8 (Graph detection is harder than tree reconstruction [24]).

$$p_G(a,b) \leq p_T(a,b).$$

We remark that Theorem 2.8 is not stated explicitly in [24]; because the authors were only interested in the case $(a-b)^2 \leq 2(a+b)$, the claimed result was that $(a-b)^2 \leq 2(a+b)$ implies $p_G(a,b) = \frac{1}{2}$. However, a cursory examination of the proof of [24, Theorem 1] reveals that the claim was proven in two stages: first, they prove via a coupling argument that $p_G(a,b) \leq p_T(a,b)$ and then they apply Theorem 2.4 to show that $(a-b)^2 \leq 2(a+b)$ implies $p_T(a,b) = \frac{1}{2}$.

2.6. *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 2.9. *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 $p_G(a,b) - \epsilon$.

We will assume for simplicity that our algorithm is given the parameters a and b. This is a minor assumption because a and b can be estimated from the data to arbitrary accuracy [24, Theorem 3].

A key ingredient of Theorem 2.9's 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" on trees.

DEFINITION 2.10 (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* $\mathrm{Pois}((a+b)/2)$. *We define the robust reconstruction accuracy as:*

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

Our main technical result is that when $a - b$ is large enough then in fact the extra noise does not have any effect on the reconstruction probability.

THEOREM 2.11. *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 the robust reconstruction accuracy is independent of δ for any parameters, and also for more general trees; however, our proof does not naturally extend to cover these cases.

2.7. *Algorithmic amplification and robust reconstruction.* The second main ingredient in Theorem 2.9 connects the community detection problem to the robust tree reconstruction problem: we show that given a suitable algorithm for providing a better-than-random 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 2.12. *For all a and b, $p_G(a,b) \geq \widetilde{p}_T(a,b)$. 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 $\widetilde{p}_T(a,b) - \epsilon$.*

Combining Theorem 2.12 with Theorems 2.8 and 2.11 proves Theorem 2.9. We remark that Theorem 2.12 easily extends to other versions of the block model (i.e., models with more clusters or unbalanced classes); however, Theorem 2.11 does not. In particular, Theorem 2.9 may not hold for general block models. In fact, one fascinating conjecture of [10] says that for general block models, computational hardness enters the picture (whereas it does not play any role in our current work).

2.8. *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 $0 \leq \delta < \frac{1}{2}$, and so we can guess the label of u from them using robust tree reconstruction. (In the previous sentence, we are implicitly claiming that the errors made by the black-box algorithm are independent of the neighborhood of u. This is because the edges in the neighborhood of u are independent of the edges in the rest of the graph, a fact that we will justify more carefully later.) 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 easier to analyze.

Finally, a word about notation. Throughout this article, we will use the letters C and c to denote positive constants whose value may change from line to line. We will also write statements like "for all $k \geq K(\theta, \delta)$ ..." as

abbreviations for statements like "for every θ and δ there exists K such that for all $k \geq K$ ..."

3. Robust reconstruction on regular trees. Our main effort is devoted to proving Theorem 2.11. Since the proof is quite involved, we begin with a somewhat easier case of regular trees which already contains the main ideas of the proof. The adaptation to the case of Poisson random trees will be carried in Section 4.

First, we need to define the reconstruction and robust reconstruction probabilities for regular trees. Their definitions are analogous to Definitions 2.6 and 2.10.

DEFINITION 3.1. *Let σ be distributed according to the broadcast process with parameter η on an infinite d-ary tree. Let τ be distributed according to the noisy broadcast process with parameters η and δ on the same tree. We define*

$$p_{reg}(d,\eta) = \frac{1}{2} + \lim_{k\to\infty} \mathbb{E}\left|\Pr(\sigma_\rho = + \mid \sigma_{L_k(\rho)}) - \frac{1}{2}\right|$$
$$\widetilde{p}_{reg}(d,\eta) = \frac{1}{2} + \liminf_{\delta\to 1/2} \liminf_{k\to\infty} \mathbb{E}\left|\Pr(\sigma_\rho = + \mid \tau_{L_k(\rho)}) - \frac{1}{2}\right|.$$

THEOREM 3.2. *Consider the broadcast process on the infinite d-ary tree where if $u \in L_1(v)$ then $\Pr(\sigma_u = \sigma_v) = \frac{1}{2}(1+\theta)$ (equivalently $\mathbb{E}[\sigma_u\sigma_v] = \theta$). There exists a constant C such that if $d\theta^2 > C$ then*

$$\widetilde{p}_{reg}(d,\eta) = p_{reg}(d,\eta),$$

3.1. *Magnetization.* Define

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

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 [6]) shows that $(1+\mathbb{E}|X_{\rho,k}|)/2$ is the probability of estimating σ_ρ correctly given $\sigma_{L_k(\rho)}$.

We may also define the noisy magnetization Y:

$$Y_{u,k} = \Pr(\sigma_u = + \mid \tau_{L_k(u)}) - \Pr(\sigma_u = - \mid \tau_{L_k(u)}) \tag{3.1}$$
$$y_k = \mathbb{E}(Y_{u,k} \mid \sigma_u = +).$$

As above, $(1+\mathbb{E}|Y_{\rho,k}|)/2$ is the probability of estimating σ_ρ correctly given $\tau_{L_k(\rho)}$. In particular, the analogue of Theorem 2.11 for d-ary trees may be written as follows:

THEOREM 3.3. *There exists a constant C such that if $\theta^2 d > C$ and $\delta < \frac{1}{2}$ then*

$$\lim_{k\to\infty} \mathbb{E}|X_{\rho,k}| = \lim_{k\to\infty} \mathbb{E}|Y_{\rho,k}|.$$

Our main method for proving Theorem 3.3 (and also Theorem 2.11) is by studying certain recursions. Indeed, Bayes' rule implies the following recurrence for X (see, eg., [26]):

$$X_{u,k} = \frac{\prod_{i\in\mathcal{C}(u)}(1+\theta X_{ui,k-1}) - \prod_{i\in\mathcal{C}(u)}(1-\theta X_{ui,k-1})}{\prod_{i\in\mathcal{C}(u)}(1+\theta X_{ui,k-1}) + \prod_{i\in\mathcal{C}(u)}(1-\theta X_{ui,k-1})}. \tag{3.2}$$

The same reasoning that gives (3.2) also shows that (3.2) also holds when every instance of X is replaced by Y. Since our entire analysis is based on the recurrence (3.2), the only meaningful (for us) difference between X and Y is that their initial conditions are different: $X_{u,0} = \pm 1$ while $Y_{u,0} = \pm(1-2\delta)$. In fact, we will see later that Theorem 3.3 also holds for some more general estimators Y satisfying (3.2).

3.2. *The simple majority method.* Our first step in proving Theorem 3.3 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 (3.2), 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 [16]; however, we include most of the details for the sake of completeness.

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. We will show this by calculating the first two moments of $S_{u,k}$ and $\tilde{S}_{u,k}$; we write $\mathbb{E}^+$ and Var^+ for the conditional expectation and conditional variance given $\sigma_\rho = +$. The first moments are trivial, and we omit the proof:

LEMMA 3.4.

$$\mathbb{E}^+ S_{\rho,k} = \theta^k d^k$$
$$\mathbb{E}^+ \tilde{S}_{\rho,k} = (1-2\delta)\theta^k d^k.$$

The second moment calculation uses the recursive structure of the tree. The argument is not new, but we include it for completeness.

LEMMA 3.5.

$$\operatorname{Var}^+ S_{\rho,k} = 4\eta(1-\eta)d^k \frac{(\theta^2 d)^k - 1}{\theta^2 d - 1}$$
$$\operatorname{Var}^+ \tilde{S}_{\rho,k} = 4d^k\delta(1-\delta) + 4(1-2\delta)^2\eta(1-\eta)d^k \frac{(\theta^2 d)^k - 1}{\theta^2 d - 1}.$$

PROOF. We decompose the variance of S_k by conditioning on the first level of the tree:

$$\operatorname{Var}^+ S_{\rho,k} = \mathbb{E}\operatorname{Var}^+(S_{\rho,k} \mid \sigma_1, \ldots, \sigma_d) + \operatorname{Var}^+ \mathbb{E}(S_{\rho,k} \mid \sigma_1, \ldots, \sigma_d). \tag{3.3}$$

Now, $S_{\rho,k} = \sum_{u \in L_1} S_{u,k-1}$, and $S_{u,k-1}$ are i.i.d. under $\Pr^+$. Thus, the first term of (3.3) decomposes into a sum of variances:

$$\mathbb{E}\operatorname{Var}^+(S_{\rho,k} \mid \sigma_1, \ldots, \sigma_d) = \sum_{u \in L_1} \mathbb{E}\operatorname{Var}^+(S_{u,k-1} \mid \sigma_u) = d\operatorname{Var}^+(S_{\rho,k-1}).$$

For the second term of (3.3), note that (by Lemma 3.4), $\mathbb{E}(S_{u,k-1} \mid \sigma_u)$ is $(\theta d)^{k-1}$ with probability $1-\eta$ and $-(\theta d)^{k-1}$ otherwise. Since $\mathbb{E}(S_{u,k-1} \mid \sigma_u)$ are independent as u varies, we have

$$\operatorname{Var}^+ \mathbb{E}(S_{\rho,k} \mid \sigma_1, \ldots, \sigma_d) = 4d\eta(1-\eta)(\theta d)^{2k-2}.$$

Plugging this back into (3.3), we get the recursion

$$\operatorname{Var}^+ S_{\rho,k} = d\operatorname{Var}^+ S_{\rho,k-1} + 4d\eta(1-\eta)(\theta d)^{2k-2}.$$

Since $\operatorname{Var}^+ S_{\rho,0} = 0$, we solve this recursion to obtain

$$\begin{aligned}\operatorname{Var}^+ S_{\rho,k} &= d\sum_{\ell=1}^{k} 4\eta(1-\eta)(\theta d)^{2\ell-2} d^{k-\ell} \\ &= 4\eta(1-\eta)d^k \sum_{\ell=0}^{k-1} (\theta^2 d)^\ell \\ &= 4\eta(1-\eta)d^k \frac{(\theta^2 d)^k - 1}{\theta^2 d - 1}.\end{aligned} \tag{3.4}$$

To compute $\mathrm{Var}^+ \tilde{S}_{\rho,k}$, we condition on $S_{\rho,k}$: conditioned on $S_{\rho,k}$, $\tilde{S}_{\rho,k}$ is a sum of d^k i.i.d. terms, of which $(d^k + S_{\rho,k})/2$ have mean $1-2\delta$, $(d^k - S_{\rho,k})/2$ have mean $2\delta - 1$, and all have variance $4\delta(1-\delta)$. Hence, $\mathbb{E}(\tilde{S}_k \mid S_k) = (1-2\delta)S_k$ and $\mathrm{Var}(\tilde{S}_k \mid S_k) = 4d^k\delta(1-\delta)$. By the decomposition of variance,

$$\begin{aligned}\mathrm{Var}^+(\tilde{S}_k) &= \mathbb{E}^+(4d^k\delta(1-\delta)) + \mathrm{Var}^+((1-2\delta)S_k) \\ &= 4d^k\delta(1-\delta) + 4(1-2\delta)^2\eta(1-\eta)d^k\frac{(\theta^2 d)^k - 1}{\theta^2 d - 1},\end{aligned}$$

where the last equality follows from (3.4) and the fact that $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$. □

Taking $k \to \infty$ in Lemmas 3.4 and 3.5, we see that if $\theta^2 d > 1$ then

$$\left.\begin{matrix}\frac{\mathrm{Var}^+ S_k}{(\mathbb{E}^+ S_k)^2} \\ \frac{\mathrm{Var}^+ \tilde{S}_k}{(\mathbb{E}^+ \tilde{S}_k)^2}\end{matrix}\right\} \overset{k\to\infty}{\to} \frac{4\eta(1-\eta)}{\theta^2 d}.$$

By Chebyshev's inequality,

$$\liminf_{k\to\infty} \Pr^+(S_k > 0) \geq 1 - \frac{4\eta(1-\eta)}{\theta^2 d}.$$

In other words, the estimators $\mathrm{sgn}(S_k)$ and $\mathrm{sgn}(\tilde{S}_k)$ succeed with probability at least $1 - \frac{4\eta(1-\eta)}{\theta^2 d^2}$ as $k \to \infty$. Now, $\mathrm{sgn}(Y_{\rho,k})$ is the optimal estimator of σ_ρ given τ_{L_k}, and its success probability is exactly $(1 + \mathbb{E}|Y_{\rho,k}|)/2$. Hence this quantity must be larger than the success probability of $\mathrm{sgn}(\tilde{S}_k)$ (and similarly for X and $\mathrm{sgn}(S_k)$). Putting this together, we arrive at the following estimates: if $\theta^2 d > 1$ and $k \geq K(\delta)$ then

$$\mathbb{E}|X_{\rho,k}| \geq 1 - \frac{10\eta(1-\eta)}{\theta^2 d} \tag{3.5}$$

$$\mathbb{E}|Y_{\rho,k}| \geq 1 - \frac{10\eta(1-\eta)}{\theta^2 d}. \tag{3.6}$$

Now, given that $\sigma_\rho = +$, the optimal estimator makes a mistake whenever $X_{\rho,k} < 0$; hence, $\Pr^+(X_{\rho,k} < 0) \leq (1 - \mathbb{E}|X_{\rho,k}|)/2$. Since $X_{u,k} \geq -1$, this implies

$$\mathbb{E}^+ X_{\rho,k} \geq \mathbb{E}^+|X_{\rho,k}| - 2\Pr^+(X_{\rho,k} < 0) \geq 1 - \frac{C\eta(1-\eta)}{\theta^2 d}.$$

We will use this fact repeatedly, so let us summarize in a lemma:

LEMMA 3.6. *There is a constant C such that if $\theta^2 d > 1$ and $k \geq K(\delta)$ then*

$$\mathbb{E}^+ X_{\rho,k} \geq 1 - \frac{C\eta(1-\eta)}{\theta^2 d}$$
$$\mathbb{E}^+ Y_{\rho,k} \geq 1 - \frac{C\eta(1-\eta)}{\theta^2 d}.$$

By Markov's inequality, we find that $X_{u,k}$ is large with high probability:

LEMMA 3.7. *There is a constant C such that for all $k \geq K(\delta)$ and all $t > 0$*

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

As we will see, Lemma 3.6 and the recursion (3.2) are really the only properties of Y that we will use. Hence, from now on $Y_{u,k}$ need not be defined by (3.1). Rather, we will make the following assumptions on $Y_{u,k}$:

ASSUMPTION 3.1. *There is a $K = K(\delta)$ such that for all $k \geq K$, the following hold:*

1. $Y_{u,k+1} = \dfrac{\prod_{i\in\mathcal{C}(u)}(1+\theta Y_{ui,k}) - \prod_{i\in\mathcal{C}(u)}(1-\theta Y_{ui,k})}{\prod_{i\in\mathcal{C}(u)}(1+\theta Y_{ui,k}) + \prod_{i\in\mathcal{C}(u)}(1-\theta Y_{ui,k})}$
2. *The distribution of $Y_{u,k}$ given $\sigma_u = +$ is equal to the distribution of $-Y_{u,k}$ given $\sigma_u = -$.*
3. $\mathbb{E}^+ Y_{\rho,k} \geq 1 - \dfrac{C\eta(1-\eta)}{\theta^2 d}$ *for some constant C.*

We will prove Theorem 3.3 under Assumption 3.1. Note that part 2 above immediately implies

$$\mathbb{E}(Y_{ui,k} \mid \sigma_u = +) = \theta\mathbb{E}(Y_{ui,k} \mid \sigma_{ui} = +).$$

Also, part 3 implies that Lemma 3.7 holds for Y.

3.3. *The recursion for small θ.* Our proof of Theorem 3.3 proceeds in two cases, with two different analyses. In the first case, we suppose that θ is small (i.e., smaller than a fixed, small constant). In this case, we proceed by Taylor-expanding the recursion (3.2) in θ. For the rest of this section, we will assume that X and Y satisfy parts 1 and 2 of Assumption 3.1, and that

$x_k, y_k \geq 5/6$ for $k \geq K(\delta)$. This restriction will allow us to reuse most of the argument in the Galton-Watson case (where part 3 of Assumption 3.1 fails to hold, but we nevertheless have $x_k, y_k \geq 5/6$).

PROPOSITION 3.8. *There are absolute constants C and $\theta^* > 0$ such that if $d\theta^2 \geq C$ and $\theta \leq \theta^*$ then for all $k \geq K(\theta, d, \delta)$,*

$$\mathbb{E}(X_{\rho,k+1} - Y_{\rho,k+1})^2 \leq \frac{1}{2}\mathbb{E}(X_{\rho,k} - Y_{\rho,k})^2.$$

Note that Proposition 3.8 immediately implies that if $d\theta^2 \geq C$ and $\theta \leq \theta^*$ then $\mathbb{E}(X_{\rho,k} - Y_{\rho,k})^2 \to 0$ as $k \to \infty$, which implies Theorem 3.3 in the case that $\theta \leq \theta^*$.

In proving Proposition 3.8, the first step is to replace the right hand side of (3.2) with something easier to work with; in particular, we would like to have something without X in the denominator. For this, we note that

$$\frac{a-b}{a+b} = \frac{1 - b/a}{1 + b/a} = \frac{2}{1 + b/a} - 1.$$

Hence, if $a = \prod_i (1 + \theta X_{ui,k})$, $b = \prod_i (1 - \theta X_{ui,k})$, and a' and b' are the same quantities with Y replacing X, then

$$|X_{u,k+1} - Y_{u,k+1}| = \left|\frac{a-b}{a+b} - \frac{a'-b'}{a'+b'}\right| = 2\left|\frac{1}{1+b/a} - \frac{1}{1+b'/a'}\right|. \tag{3.7}$$

Using Taylor's theorem, the right hand side can be bounded in terms of $|(b/a)^p - (b'/a')^p|$ for some $0 < p < 1$ of our choice:

LEMMA 3.9. *For any $0 < p < 1$ and any $x, y \geq 0$,*

$$\left|\frac{1}{1+x} - \frac{1}{1+y}\right| \leq \frac{1}{p}|x^p - y^p|$$

PROOF. Let $f(x) = \frac{1}{1+x}$ and $g(x) = x^p$. By the fundamental theorem of calculus, the proof would follow from the inequality $|f'(x)| \leq p^{-1} g'(x)$. Now, $|f'(x)| = \frac{1}{(1+x)^2}$ and $g'(x) = px^{p-1}$. When $x \geq 1$, we have $|f'(x)| \leq x^{-2} \leq x^{p-1}$, while if $x \leq 1$ then $|f'(x)| \leq 1 \leq x^{p-1}$. □

As an immediate consequence of Lemma 3.9 (for $p = 1/4$) and (3.7),

$$|X_{u,k+1} - Y_{u,k+1}| \leq 8\left|\left(\prod_i \frac{1 - \theta X_{ui,k}}{1 + \theta X_{ui,k}}\right)^{1/4} - \left(\prod_i \frac{1 - \theta Y_{ui,k}}{1 + \theta Y_{ui,k}}\right)^{1/4}\right|. \tag{3.8}$$

Next, we present a general bound on the second moment of differences of products. Of course, we have in mind the example $A_i = (\frac{1-\theta X_{ui,k}}{1+\theta X_{ui,k}})^{1/4}$ and similarly for B_i and Y_i.

LEMMA 3.10. *Let $(A_1, B_1), \ldots, (A_d, B_d)$ be i.i.d. copies of (A, B). Then*

$$\mathbb{E}\left(\prod_{i=1}^{d} A_i - \prod_{i=1}^{d} B_i\right)^2 \le dm^{d-1}(\mathbb{E}A^2 - \mathbb{E}B^2)^2 + 2dm^{d-1}\mathbb{E}(A-B)^2,$$

where $m = \max\{\mathbb{E}A^2, \mathbb{E}B^2\}$.

PROOF. Let $\epsilon = \mathbb{E}(A_i - B_i)^2$, so that $\mathbb{E}A_iB_i = \frac{1}{2}(\mathbb{E}A_i^2 + \mathbb{E}B_i^2 - \epsilon)$. Then

$$\begin{aligned}\mathbb{E}\left(\prod_{i=1}^{d} A_i - \prod_{i=1}^{d} B_i\right)^2 &= \mathbb{E}\prod_{i=1}^{d} A_i^2 + \mathbb{E}\prod_{i=1}^{d} B_i^2 - 2\mathbb{E}\prod_{i=1}^{d} A_iB_i \\ &= (\mathbb{E}A^2)^d + (\mathbb{E}B^2)^d - 2\prod_{i=1}^{d}\frac{\mathbb{E}A_i^2 + \mathbb{E}B_i^2 - \epsilon}{2} \\ &= (\mathbb{E}A^2)^d + (\mathbb{E}B^2)^d - 2\left(\frac{\mathbb{E}A^2 + \mathbb{E}B^2 - \epsilon}{2}\right)^d. \qquad (3.9)\end{aligned}$$

By a second-order Taylor expansion, any twice differentiable f satisfies $f(x) + f(y) \le 2f((x+y)/2) + \frac{1}{4}(x-y)^2 \max_z f''(z)$, where the maximum ranges over z between x and y. Applying this for $f(x) = x^d$ yields

$$(\mathbb{E}A^2)^d + (\mathbb{E}B^2)^d \le d^2m^{d-2}(\mathbb{E}A^2 - \mathbb{E}B^2)^2 + 2\left(\frac{\mathbb{E}A^2 + \mathbb{E}B^2}{2}\right)^d.$$

Hence,

$$\begin{aligned}(3.9) &\le d^2m^{d-2}(\mathbb{E}A^2 - \mathbb{E}B^2)^2 + 2\left(\frac{\mathbb{E}A^2 + \mathbb{E}B^2}{2}\right)^d - 2\left(\frac{\mathbb{E}A^2 + \mathbb{E}B^2 - \epsilon}{2}\right)^d \\ &\le d^2m^{d-2}(\mathbb{E}A^2 - \mathbb{E}B^2)^2 + 2dm^{d-1}\epsilon,\end{aligned}$$

where the second inequality follows from a first-order Taylor expansion of the function $f(x) = x^d$ around $x = (\mathbb{E}A^2 + \mathbb{E}B^2)/2$. □

As we said before, we will apply Lemma 3.10 with $A_i = \left(\frac{1-\theta X_{ui,k}}{1+\theta X_{ui,k}}\right)^{1/4}$ and $B_i = \left(\frac{1-\theta Y_{ui,k}}{1+\theta Y_{ui,k}}\right)^{1/4}$. To make the lemma useful, we will need to bound $\mathbb{E}A_i^2$, $\mathbb{E}B_i^2$, and their difference. First, we will bound $\mathbb{E}A_i^2$ and $\mathbb{E}B_i^2$. In other words, we will bound

$$\mathbb{E}\sqrt{\frac{1-\theta X_{ui,k}}{1+\theta X_{ui,k}}}$$

and the same expression with Y instead of X.

LEMMA 3.11. *There is a constant $\theta^* > 0$ such that if $x_k, y_k \geq 5/6$ then*

$$\mathbb{E}(A_i^2 \mid \sigma_u = +) \leq 1 - \frac{\theta^2 x_k}{4}$$
$$\mathbb{E}(B_i^2 \mid \sigma_u = +) \leq 1 - \frac{\theta^2 y_k}{4}.$$

PROOF. First, note that for sufficiently small x,

$$(1+x)(1-x+\frac{5}{8}x^2)^2 = (1+x)(1-2x+\frac{18}{8}x^2+O(x^3)) = 1-x+\frac{1}{4}x^2+O(x^3) \geq 1-x,$$

which may be rearranged to read

$$\sqrt{\frac{1-x}{1+x}} \leq 1 - x + \frac{5}{8}x^2.$$

Now, if θ^* is sufficiently small then we may apply this with $x = \theta X_{ui,k}$, obtaining

$$\mathbb{E}(A_i^2 \mid \sigma_u = +) \leq 1 - \mathbb{E}(\theta X_{ui,k} \mid \sigma_u = +) + \frac{5}{8}\mathbb{E}(\theta^2 X_{ui,k}^2 \mid \sigma_u = +).$$

Recalling the assumption that $x_k \geq 5/6$, we have

$$1 - \mathbb{E}(\theta X_{ui,k} \mid \sigma_u = +) + \frac{5}{8}\mathbb{E}(\theta^2 X_{ui,k}^2 \mid \sigma_u = +) \leq 1 - \theta^2 x_k + \frac{3\theta^2}{4}x_k = 1 - \frac{\theta^2}{4}x_k.$$

The same argument applies to B_i, but using Y_i instead of X_i. □

3.4. *The $\mathbb{E}A^2 - \mathbb{E}B^2$ term.* In this section, we will bound the $|\mathbb{E}A^2 - \mathbb{E}B^2|$ term in Lemma 3.10, bearing in mind that the bound has to be at most of order θ^4 in order for $d^2(\mathbb{E}A^2 - \mathbb{E}B^2)^2$ to be a function of $d\theta^2$. Note that the distribution of A_i conditioned on $\sigma_v = +$ is equal to the distribution of $1/A_i$ conditioned on $\sigma_v = -$. Hence,

$$\begin{aligned}\mathbb{E}(A_i^2 \mid \sigma_u = +) &= (1-\eta)\mathbb{E}(A_i^2 \mid \sigma_{ui} = +) + \eta\mathbb{E}(A_i^2 \mid \sigma_{ui} = -) \\ &= \mathbb{E}\left((1-\eta)A_i^2 + \eta A_i^{-2} \mid \sigma_{ui} = +\right). \qquad (3.10)\end{aligned}$$

Now,

$$\begin{aligned}(1-\eta)A_i^2 + \eta A_i^{-2} &= (1-\eta)\left(\frac{1-\theta X_{ui,k}}{1+\theta X_{ui,k}}\right)^{1/2} + \eta\left(\frac{1+\theta X_{ui,k}}{1-\theta X_{ui,k}}\right)^{1/2} \\ &= \frac{(1-\eta)(1-\theta X_{ui,k}) + \eta(1+\theta X_{ui,k})}{\sqrt{(1+\theta X_{ui,k})(1-\theta X_{ui,k})}} \\ &= \frac{1-\theta^2 X_{ui,k}}{\sqrt{1-\theta^2 X_{ui,k}^2}} \qquad (3.11)\end{aligned}$$

(recalling in the last line that $\theta = 1 - 2\eta$).

LEMMA 3.12. *There is a $\theta^* > 0$ such that if $\theta < \theta^*$ then*

$$\left| \frac{d}{dx} \frac{1-\theta^2 x}{\sqrt{1-\theta^2 x^2}} \right| \le 3\theta^2$$

for all $x \in [-1,1]$.

PROOF. By a direct computation,

$$\frac{d}{dx} \frac{1-\theta^2 x}{\sqrt{1-\theta^2 x^2}} = \frac{\theta^2 x(1-\theta^2 x^2)^{-1/2}(1-\theta^2 x) - \theta^2\sqrt{1-\theta^2 x^2}}{1-\theta^2 x^2}.$$

Since $|x| \le 1$, we have

$$\left| \frac{d}{dx} \frac{1-\theta^2 x}{\sqrt{1-\theta^2 x^2}} \right| \le \frac{\theta^2(1-\theta^2)^{-1/2}(1+\theta^2) + \theta^2}{1-\theta^2} = \theta^2 \frac{(1-\theta^2)^{-1/2}(1+\theta^2)+1}{1-\theta^2}.$$

The result follows because $1-\theta^2$ and $1+\theta^2$ can be made arbitrarily close to 1 by taking θ^* small enough. □

Now we apply (3.11) with Lemma 3.12 to obtain the promised bound on $\mathbb{E}A_i^2 - \mathbb{E}B_i^2$.

LEMMA 3.13. *There is a $\theta^* > 0$ such that for all $\theta < \theta^*$,*

$$\mathbb{E}(A_i^2 - B_i^2 \mid \sigma_u = +) \le 3\theta^2 \sqrt{\mathbb{E}((X_{ui,k} - Y_{ui,k})^2 \mid \sigma_u = +)}.$$

PROOF. By (3.10) and (3.11) (and analogously with A replaced by B), we have

$$\mathbb{E}(A_i^2 - B_i^2 \mid \sigma_u = +) = \mathbb{E}\left(\frac{1-\theta^2 X_{ui,k}}{\sqrt{1-\theta^2 X_{ui,k}^2}} - \frac{1-\theta^2 Y_{ui,k}}{\sqrt{1-\theta^2 Y_{ui,k}^2}} \Big| \sigma_{ui} = + \right).$$

For a general function f we have $\mathbb{E}|f(X)-f(Y)| \le \mathbb{E}|X-Y| \max_x \left|\frac{df}{dx}\right|$. Applying this fact with the function $f(x) = \frac{1-\theta^2 x}{\sqrt{1-\theta^2 x^2}}$ and the bound of Lemma 3.12,

$$\begin{aligned}\mathbb{E}(A_i^2 - B_i^2 \mid \sigma_u = +) &\le 3\theta^2 \mathbb{E}(|X_{ui,k} - Y_{ui,k}| \mid \sigma_{ui} = +) \\ &\le 3\theta^2 \sqrt{\mathbb{E}((X_{ui,k} - Y_{ui,k})^2 \mid \sigma_{ui} = +)}.\end{aligned}$$

Finally, note that

$$\mathbb{E}((X_{ui,k} - Y_{ui,k})^2 \mid \sigma_{ui} = +) = \mathbb{E}((X_{ui,k} - Y_{ui,k})^2 \mid \sigma_u = +).$$ □

3.5. *Combining the estimates to complete the proof.* Next, we combine Lemma 3.10 with the estimates provided in Lemmas 3.11 and 3.13.

LEMMA 3.14. *There is some constant $\theta^* > 0$ such that the following holds. Suppose that X and Y satisfy parts 1 and 2 of Assumption 3.1 and that $x_k, y_k \geq 5/6$ for $k \geq K(\delta)$. If u has $d \geq 4$ children and $\theta \leq \theta^*$ then for $k \geq K(\delta)$,*

$$\frac{\mathbb{E}\left((X_{u,k+1} - Y_{u,k+1})^2 \mid \sigma_u = +\right)}{\mathbb{E}\left((X_{u1,k} - Y_{u1,k})^2 \mid \sigma_{u1} = +\right)} \leq C(d^2\theta^4 + d\theta^2)e^{-\frac{\theta^2 d}{5}},$$

for a universal constant C.

PROOF. Taking the square of (3.8) and taking the expectation on both sides, we have

$$\mathbb{E}\big((X_{u,k+1} - Y_{u,k+1})^2 \mid \sigma_u = +\big) \leq 64\mathbb{E}\left(\left(\prod_{i=1}^{d} A_i - \prod_{i=1}^{d} B_i\right)^2 \;\middle|\; \sigma_u = +\right).$$

Conditioned on σ_u, the pairs (A_i, B_i) are i.i.d. and so Lemma 3.10 implies that

$$\begin{aligned}(3.12)\quad &\mathbb{E}\left((X_{u,k+1} - Y_{u,k+1})^2 \mid \sigma_u = +\right)\\ &\qquad \leq 64d^2 m^{d-2}(a-b)^2 + 128dm^{d-1}\mathbb{E}((A_i - B_i)^2 \mid \sigma_u = +),\end{aligned}$$

where

$$\begin{aligned} a &= \mathbb{E}(A_i^2 \mid \sigma_u = +)\\ b &= \mathbb{E}(B_i^2 \mid \sigma_u = +)\\ m &= \max\{a, b\}.\end{aligned}$$

Now, if θ^* is suffiently small then the function $x \mapsto (\frac{1-\theta x}{1+\theta x})^{1/4}$ has derivative at most θ for $x \in [-1, 1]$. Hence,

$$\begin{aligned}\mathbb{E}((A_i - B_i)^2 \mid \sigma_u = +) &\leq \theta^2\mathbb{E}((X_{u1,k} - Y_{u1,k})^2 \mid \sigma_u = +)\\ &= \theta^2\mathbb{E}((X_{u1,k} - Y_{u1,k})^2 \mid \sigma_{u1}) \qquad (3.13)\end{aligned}$$

provided that θ^* is sufficiently small. Define

$$z = \mathbb{E}((X_{u1,k} - Y_{u1,k})^2 \mid \sigma_{u1}) = \mathbb{E}((X_{u1,k} - Y_{u1,k})^2 \mid \sigma_{u1} = +).$$

By Lemma 3.11 and the assumption that $x_k, y_k \geq 5/6$, if θ^* is sufficiently small then $m \leq 1 - \theta^2/5 \leq \exp(-\theta^2/5)$. Moreover, Lemma 3.13 implies that $(a-b)^2 \leq 9\theta^4 z$. Plugging these and (3.13) back into (3.12), we have

$$\mathbb{E}\left((X_{u,k+1} - Y_{u,k+1})^2 \mid \sigma_u = +\right) \leq 64\left(9d^2\theta^4 e^{-\frac{\theta^2(d-2)}{5}} + 2d\theta^2 e^{-\frac{\theta^2(d-1)}{5}}\right) z,$$

which proves the claim. □

PROOF OF PROPOSITION 3.8. If $\theta^2 d$ is sufficiently large then Lemma 3.6 implies that $x_k, y_k \geq 5/6$ for $k \geq K(\delta)$; hence, the conditions of Lemma 3.14 are satisfied. Finally, if $d\theta^2$ is large enough then the right hand side in Lemma 3.14 is at most $\frac{1}{2}$. □

3.6. *The recursion for large θ.* To handle the case in which θ is not small, we require a different argument. In this case, we study the derivatives of the recurrence, obtaining the following result:

PROPOSITION 3.15. *For any $0 < \theta^* < 1$, there is some $d^* = d^*(\theta^*)$ such that for all $\theta \geq \theta^*$, $d \geq d^*$, and $k \geq K(\theta, d, \delta)$,*

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

Combined with Proposition 3.8, this proves Theorem 3.3. Indeed, to complete the choices of parameters we first take θ^* to be the universal constant in Proposition 3.8. Then let $d^* = d^*(\theta^*)$ be given by Proposition 3.15 (note that d^* is also a universal constant). Finally, choose C to be the maximum of d^* and the C from Proposition 3.8. Now, if $\theta^2 d \geq C$ then either $\theta \leq \theta^*$ in which case Proposition 3.8 applies, or $\theta \geq \theta^*$ in which case $\theta \leq 1$ implies that $d \geq C \geq d^*$ and so Proposition 3.15 applies. In either case, we deduce Theorem 3.3.

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

$$g(x) = \frac{\prod_{i=1}^d (1+\theta x_i) - \prod_{i=1}^d (1-\theta x_i)}{\prod_{i=1}^d (1+\theta x_i) + \prod_{i=1}^d (1-\theta x_i)}. \tag{3.14}$$

Then the recurrence (3.2) 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 $X_{u,k+1} = g(X_{L_1(u),k})$.

Define $g_1(x) = \prod_{i=1}^d (1+\theta x_i)$ and $g_2(x) = \prod_{i=1}^d (1-\theta x_i)$ so that g can be written as $g = \frac{g_1 - g_2}{g_1 + g_2}$. Since $\frac{\partial g_1}{\partial x_i} = \theta \frac{g_1}{1+\theta x_i}$ and $\frac{\partial g_2}{\partial x_i} = -\theta \frac{g_2}{1-\theta x_i}$, we have

$$\begin{aligned}\frac{\partial g}{\partial x_i} &= \frac{\partial}{\partial x_i}\frac{g_1 - g_2}{g_1 + g_2} \\ &= 2\frac{g_2 \frac{\partial g_1}{\partial x_i} - g_1 \frac{\partial g_2}{\partial x_i}}{(g_1+g_2)^2} \\ &= 4\theta \frac{g_1 g_2}{(g_1+g_2)^2(1-\theta^2 x_i^2)}. \end{aligned} \tag{3.15}$$

If $|x_i| \le 1$ then g_1 and g_2 are both positive, so $\frac{g_1 g_2}{(g_1+g_2)^2} \le \frac{g_1 g_2}{g_1^2} = \frac{g_2}{g_1}$; of course, we also have the symmetric bound $\frac{g_1 g_2}{(g_1+g_2)^2} \le \frac{g_1}{g_2}$. Define

$$\begin{aligned} h_i^+(x) &= 4\frac{g_2}{(1-\theta^2 x_i^2) g_1} = \frac{4}{(1+\theta x_i)^2} \prod_{j \ne i} \frac{1-\theta x_j}{1+\theta x_j} \\ h_i^-(x) &= 4\frac{g_1}{(1-\theta^2 x_i^2) g_2} = \frac{4}{(1-\theta x_i)^2} \prod_{j \ne i} \frac{1+\theta x_j}{1-\theta x_j} \\ h_i(x) &= \min\{h_i^+(x), h_i^-(x)\}. \end{aligned}$$

By (3.15) and since $|\theta| \le 1$,

$$\left|\frac{\partial g}{\partial x_i}\right| \le h_i(x). \tag{3.16}$$

The point is that if $\sigma_u = +$ then for most $v \in L_1(u)$, $X_{v,k}$ will be close to 1 and so $h_i^+(X_{L_1(u),k})$ will be small. On the other hand, if $\sigma_u = -$ then for most $v \in L_1(u)$, $X_{v,k}$ will be close to -1 and so $h_i^-(X_{L_1(u),k})$ will be small.

Note that h_i^+ is convex on $[-1,1]^d$ because it is the tensor product of nonnegative, convex functions. Hence for any $x, y \in [-1,1]^d$ and any $0 < \lambda < 1$,

$$\left|\frac{\partial g}{\partial x_i}(\lambda x + (1-\lambda)y)\right| \le h_i^+(\lambda x + (1-\lambda)y) \le \max\{h_i^+(x), h_i^+(y)\}.$$

Then the mean value theorem implies that

$$|g(x) - g(y)| \le \sum_i |x_i - y_i| \max\{h_i^+(x), h_i^+(y)\}.$$

Applied for $x = X_{L_1(u),k} = (X_{u1,k}, \dots, X_{ud,k})$ and $y = Y_{L_1(u),k} = (Y_{u1,k}, \dots, Y_{ud,k})$, this yields

$$|X_{u,k+1} - Y_{u,k+1}| \le \sum_i |X_{ui,k} - Y_{ui,k}| \max\{h_i^+(X_{L_1(u),k}), h_i^+(Y_{L_1(u),k})\}. \tag{3.17}$$

Note that the two terms on the right hand side of (3.17) are dependent on one another. Hence, it will be convenient to bound $h_i^+(X_{L_1(u),k})$ by something that doesn't depend on X_{ui}. To that end, note that for $|x_i| \le 1$, we have $1+\theta x_i \ge 1-\theta = 2\eta$, and so

$$h_i^+(x) = \frac{4}{(1+\theta x_i)^2}\prod_{j\neq i}\frac{1-\theta x_j}{1+\theta x_j} \le \frac{1}{\eta^2}\prod_{j\neq i}\frac{1-\theta x_j}{1+\theta x_j} =: m_i(x). \tag{3.18}$$

Since $m_i(x)$ doesn't depend on x_i, it follows that $m_i(X_{L_1(u),k})$ is independent of $X_{ui,k}$ given σ_u (and similarly with Y instead of X). Hence, (3.17) implies that

$$\begin{aligned}&\mathbb{E}\left(\sqrt{|X_{u,k+1}-Y_{u,k+1}|}\Big|\sigma_u=+\right)\\ &\le \sum_i \mathbb{E}\left(\sqrt{|X_{ui,k}-Y_{ui,k}|}\Big|\sigma_u=+\right)\mathbb{E}\left(\sqrt{\max\{m_i(X_{L_1(u),k}),m_i(Y_{L_1(u),k})\}}\Big|\sigma_u=+\right).\end{aligned} \tag{3.19}$$

To prove Proposition 3.15, 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. The following lemma will help bounding these terms.

LEMMA 3.16. *For any $0<\theta^*<1$, there is some $d^*=d^*(\theta^*)$ and some $\lambda=\lambda(\theta^*)<1$ such that for all $\theta\ge\theta^*$, $d\ge d^*$ and $k\ge K(\theta,d,\delta)$,*

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

The proof of Lemma 3.16 is straightforward but tedious, and we postpone it until the appendix. Instead, we will now prove Proposition 3.15.

PROOF OF PROPOSITION 3.15. By Lemma 3.16, and the definition (3.18) of m_i, it follows that

$$\mathbb{E}(\sqrt{m_i(X_{ui,k})}\mid\sigma_u=+)\le \eta^{-1}\min\{\lambda,\eta^{1/4}\}^{d-1}\le\min\{\lambda,\eta^{1/4}\}^{d-5}\le\lambda^{d-5}. \tag{3.20}$$

In particular, if $d^*(\theta^*)$ is sufficiently large then $d\lambda^{d-5}\le 1/4$ for all $d\ge d^*$. The same argument applies with Y replacing X, and hence

$$\mathbb{E}\left(\sqrt{\max\{m_i(X_{L_1(u),k}),m_i(Y_{L_1(u),k})\}}\Big|\sigma_u=+\right)\le\frac{1}{2d}. \tag{3.21}$$

By (3.19), we have

$$\mathbb{E}\left(\sqrt{|X_{u,k+1} - Y_{u,k+1}|} \;\middle|\; \sigma_u = +\right) \le \frac{1}{2}\mathbb{E}\left(\sqrt{|X_{u,k} - Y_{u,k}|} \;\middle|\; \sigma_u = +\right),$$

and so we have proved Proposition 3.15. □

4. Reconstruction accuracy on Galton-Watson trees. In this section, we will adapt the proof of the d-ary case (Theorem 3.3) to the Galton-Watson case (Theorem 2.11). Let $T \subset \mathbb{N}^*$ be a Galton-Watson tree with offspring distribution Pois(d). Recall that such a tree may be constructed by taking, for each $u \in \mathbb{N}^*$, an independent Pois(d) random variable D_u. Then define $T \subset \mathbb{N}^*$ recursively by starting with $\varnothing \in T$ and then taking $ui \in T$ for $i \in \mathbb{N}$ if $u \in T$ and $i \le D_u$.

As in Section 3, we let $\{\sigma_u : u \in T\}$ be distributed as the two-state broadcast process on T with parameter η, and let $\{\tau_u : u \in T\}$ be the noisy version, with parameter δ. We recall the magnetization

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

Note that unlike in Section 3, $X_{u,k}$ now depends on both the randomness of the tree and the randomness of σ. Hence, x_k now averages over both the randomness of the tree and the randomness of σ.

We recall that X satisfies the recursion (3.2). As in Section 3, we will let $\{Y_{u,k}\}$ be any collection of random variables which satisfies the same recursion (for large enough k), and for which $Y_{u,k}$ is a good estimator of σ_u given $\sigma_{L_k(u)}$.

ASSUMPTION 4.1. *There is a $K = K(\delta)$ and a constant C such that for all $k \ge K$, the following hold:*

1. $Y_{u,k+1} = \dfrac{\prod_{i\in\mathcal{C}(u)}(1+\theta Y_{ui,k}) - \prod_{i\in\mathcal{C}(u)}(1-\theta Y_{ui,k})}{\prod_{i\in\mathcal{C}(u)}(1+\theta Y_{ui,k}) + \prod_{i\in\mathcal{C}(u)}(1-\theta Y_{ui,k})}$.
2. *The distribution of $Y_{u,k}$ given $\sigma_u = +$ is equal to the distribution of $-Y_{u,k}$ given $\sigma_u = -$.*
3. *With probability at least $1 - e^{-cd}$ over T,*

$$\mathbb{E}(Y_{u,k} \mid \sigma_u = +, T) \ge 1 - \frac{C\eta}{\theta^2 d}.$$

Note that Assumption 4.1 is the same as Assumption 3.1 except for part 3. Indeed, the change in part 3 between Assumption 3.1 and Assumption 4.1

points to the main change, and biggest challenge, in extending our previous argument to Galton-Watson trees: unlike for a regular tree, there is always some chance that a Galton-Watson tree will go extinct, or that it will be thinner and more spindly than expected. In this case, we will not be able to reconstruct the broadcast process as well as we might want, even as $\eta \to 0$.

In any case, in order to prove Theorem 2.11 it suffices to prove that Y satisfies part 3 of Assumption 4.1 as well as the following theorem:

THEOREM 4.1. *Under Assumption 4.1, there is a universal constant C such that if $\theta^2 d \geq C$ then $\lim_{k\to\infty} \mathbb{E}|X_{\rho,k}| = \lim_{k\to\infty} \mathbb{E}|Y_{\rho,k}|$.*

Recall that $p_T(a,b)$ is equal to $\lim_{k\to\infty}(1 + \mathbb{E}|X_{\rho,k}|)/2$ in the case $d = (a+b)/2$ and $\eta = b/(a+b)$, and that $\widetilde{p}_T(a,b)$ is equal to $\lim_{k\to\infty}(1+\mathbb{E}|Y_{\rho,k}|)/2$ in the same case. In particular, Theorem 4.1 immediately implies Theorem 2.11.

4.1. *Large expected magnetization.* The first step towards extending Theorem 3.3 to the Galton-Watson case is to show that the magnetization of each node tends to be large.

PROPOSITION 4.2. *There is a universal constant $c > 0$ such that for all $k \geq K(\theta, d, \delta)$,*

$$\Pr\left(\mathbb{E}(X_{\rho,k} \mid \sigma_\rho = +, T) \geq 1 - \frac{16\eta}{\theta^2 d}\right) \geq 1 - e^{-cd}.$$

and similarly for $Y_{\rho,k}$. Hence, $x_k, y_k \geq 1 - \frac{8\eta}{\theta^2 d} - 2e^{-cd}$.

Note that the proposition implies that Y satisfies part 3 of Assumption 4.1.

In the regular case, the proof of Lemma 3.6 was based on the fact that a simple majority vote at the leaves estimates the root well. Here, we will follow Evans et al. [12] by using a weighted majority vote. For this, we will need to use the terminology of electrical networks, in particular the notion of effective conductance and effective resistance. An introduction to these concepts may be found in [19]; the essential properties that we will need are that conductances add over parallel paths, while resistances add over consecutive paths.

Put a resistance of $(1 - \theta^2)\theta^{-2k}$ on each edge e in T whose child is in generation k (where ρ is generation zero). We write $\mathscr{C}_{\text{eff}}(k)$ for the effective conductance between ρ and level k and $\mathscr{R}_{\text{eff}}(k)$ for $1/\mathscr{C}_{\text{eff}}(k)$. Also, attach an additional "noisy" node to each node at level k, with resistance $4\delta(1 - \delta)(1 - 2\delta)^{-2}\theta^{-2k}$; then let $\mathscr{C}'_{\text{eff}}(k)$ be the effective conductance between the

root and these nodes and let $\mathscr{R}'_{\text{eff}}(k) = 1/\mathscr{C}'_{\text{eff}}(k)$. Note that $\mathscr{C}_{\text{eff}}(k)$ and $\mathscr{C}'_{\text{eff}}(k)$ are random quantities which depend on the Galton-Watson tree. The importance of $\mathscr{C}_{\text{eff}}$ and $\mathscr{C}'_{\text{eff}}$ for estimating σ_ρ was shown by [12] (Lemma 5.1):

THEOREM 4.3. *There exist weights $w(u)$ such that if $R_k = \sum_{v\in L_k(\rho)} w(v)\sigma_v$ and $S_k = (1-2\delta)^{-1}\sum_{v\in L_k(\rho)} w(v)\tau_v$ then*

$$\begin{aligned}\mathbb{E}(R_k \mid \sigma_\rho) &= \sigma_\rho\\ \mathbb{E}(S_k \mid \sigma_\rho) &= \sigma_\rho\\ \operatorname{Var}(R_k \mid \sigma_\rho) &= \mathscr{R}_{\mathit{eff}}(k)\\ \operatorname{Var}(S_k \mid \sigma_\rho) &= \mathscr{R}'_{\mathit{eff}}(k).\end{aligned}$$

We mention that $w(v)$ in Theorem 4.3 is proportional to the unit current flow from ρ to v; for our work, however, we only need to know that it exists and that it can be easily computed.

Consider the estimators $\operatorname{sgn}(R_k)$ and $\operatorname{sgn}(S_k)$ for σ_ρ. By Chebyshev's inequality,

$$\Pr(S_k \le 0 \mid \sigma_\rho = +) \le \operatorname{Var}(S_k) = \mathscr{R}_{\text{eff}}(k) = \frac{1}{\mathscr{C}_{\text{eff}}(k)}$$

and similarly $\Pr(R_k \le 0 \mid \sigma_\rho = +) \le 1/\mathscr{C}'_{\text{eff}}(k)$. In particular, if we can show that $\mathscr{C}_{\text{eff}}(k)$ and $\mathscr{C}'_{\text{eff}}(k)$ are large, we will have shown that $\operatorname{sgn}(S_k)$ and $\operatorname{sgn}(R_k)$ are good estimators of σ_ρ. Since $\operatorname{sgn}(X_{k,\rho})$ and $\operatorname{sgn}(Y_{k,\rho})$ are the optimal estimators of σ_ρ given, respectively, $\sigma_{L_k(\rho)}$ and $\tau_{L_k(\rho)}$, this will prove that x_k and y_k are large. Note that this is exactly the same method that we used to show that x_k and y_k are large in the d-regular case; the difference here is that we need to consider a weighted linear estimator instead of an unweighted one.

LEMMA 4.4. *There is a universal constant $c > 0$ such that for all $k \ge K(\theta, d, \delta)$,*

$$\begin{aligned}\Pr\left(\mathscr{C}_{\mathit{eff}}(k) \ge \frac{\theta^2 d}{16\eta}\right) &\ge e^{-cd}\\ \Pr\left(\mathscr{C}'_{\mathit{eff}}(k) \ge \frac{\theta^2 d}{16\eta}\right) &\ge e^{-cd}.\end{aligned}$$

PROOF. The proof is by a recursive argument. Note that $\mathscr{C}_{\text{eff}}(0) = \infty$ and $\mathscr{C}'_{\text{eff}}(0) = (4\delta(1-\delta))^{-1}(1-2\delta)^{-2} > 0$. We will write the rest of the proof only for $\mathscr{C}_{\text{eff}}$, but the same argument holds with $\mathscr{C}'_{\text{eff}}$ replacing $\mathscr{C}_{\text{eff}}$ everywhere.

Let $\alpha_{k-1} = \min\{(4\eta)^{-1}, M\}$ where M is the largest median of $\mathscr{C}_{\text{eff}}(k-1)$ (in the case of $\mathscr{C}_{\text{eff}}(0)$, M is any positive value). Now fix k and let $Z_1, Z_2, \ldots$ be independent copies of $\mathscr{C}_{\text{eff}}(k-1)$. Then $\Pr(Z_i \geq \alpha_{k-1}) \geq 1/2$ for all i.

Now, the first k levels of a Galton-Watson tree consist of a root with $\text{Pois}(d)$ independent subtrees of $k-1$ levels each. For each child i, the conductance between i and $L_{k-1}(i)$ is distributed like $\theta^2 Z_i$ (the factor θ^2 arises because at each level of the tree the conductances are multiplied by an extra factor of θ^2). Since the edge between ρ and i has conductance $\theta^2(1-\theta^2)^{-1}$, the conductance between ρ and $L_{k-1}(i)$ is distributed like

$$\frac{1}{\theta^{-2} Z_i^{-1} + \theta^{-2}(1-\theta^2)} = \frac{\theta^2 Z_i}{(1-\theta^2)Z_i + 1}.$$

Summing over the children of ρ, we see that $\mathscr{C}_{\text{eff}}(k)$ has the same distribution as

$$\sum_{i=1}^{\text{Pois}(d)} \frac{\theta^2 Z_i}{(1-\theta^2)Z_i + 1} \geq \theta^2 \sum_{i=1}^{\text{Pois}(d)} \frac{Z_i}{4\eta Z_i + 1}.$$

Recall that $\Pr(Z_i \geq \alpha_{k-1}) \geq 1/2$ and $\alpha_{k-1} \leq (4\eta)^{-1}$. Hence, $\alpha_{k-1}/(4\eta\alpha_{k-1} + 1) \geq \alpha_{k-1}/2$, and so

$$\begin{aligned}
\mathscr{C}_{\text{eff}}(k) &\geq \theta^2 \sum_{i=1}^{\text{Pois}(d)} 1_{\{Z_i \geq \alpha_{k-1}\}} \frac{\alpha_{k-1}}{4\eta\alpha_{k-1} + 1} \\
&\geq \frac{\theta^2}{2} \sum_{i=1}^{\text{Pois}(d)} 1_{\{Z_i \geq \alpha_{k-1}\}} \alpha_{k-1} \\
&\geq \frac{\theta^2 \alpha_{k-1}}{2} \text{Pois}(d/2).
\end{aligned}$$

Now, there is a universal constant $c > 0$ such that $\Pr(\text{Pois}(d/2) \leq d/4) \leq e^{-cd}$; hence

$$\Pr(\mathscr{C}_{\text{eff}}(k) \leq \theta^2 d\alpha_{k-1}/4) \leq e^{-cd}. \tag{4.1}$$

In particular, if d is sufficiently large then $e^{-cd} < 1/2$ and hence every median of $\mathscr{C}_{\text{eff}}(k)$ is larger than $\theta^2 d\alpha_{k-1}/4$. In particular, $\alpha_k \geq \min\{(4\eta)^{-1}, \theta^2 d\alpha_{k-1}/4\}$. Hence, if $\theta^2 d > 4$ and k is sufficiently large then $\alpha_k \geq (4\eta)^{-1}$. Applying this to (4.1) completes the proof for $\mathscr{C}_{\text{eff}}(k)$, and an identical argument applies to $\mathscr{C}'_{\text{eff}}(k)$. □

Now Proposition 4.2 follows directly from Theorem 4.3 and Lemma 4.4.

4.2. *The small-θ case.* The proof of Proposition 3.8 extends fairly easily to the Galton-Watson case. The weakening of Lemma 3.6 to Proposition 4.2 makes hardly any difference because the proof of Proposition 3.8 only needed $x_k \geq 1/2$.

PROPOSITION 4.5. *Consider the broadcast process on a Poisson Galton-Watson tree. Then there are absolute constants C and $\theta^* > 0$ such that if $d\theta^2 \geq C$ and $\theta \leq \theta^*$ then for all $k \geq K(\theta, d, \delta)$,*

$$\mathbb{E}(X_{\rho,k+1} - Y_{\rho,k+1})^2 \leq \frac{1}{2}\mathbb{E}(X_{\rho,k} - Y_{\rho,k})^2.$$

PROOF. Let D be the number of children of u, so that $D \sim \text{Pois}(d)$. If $\theta^2 d$ is sufficiently large then Proposition 4.2 implies that $x_k, y_k \geq 5/6$ and so applying Lemma 3.14 conditioned on D yields

$$\mathbb{E}\big((X_{u,k+1} - Y_{u,k+1})^2 \mid D, \sigma_u = +\big) \leq C(D^2\theta^4 + D\theta^2)e^{-\frac{\theta^2 D}{5}} z \quad \leq C' e^{-\frac{\theta^2 D}{10}} z$$

where $z = \mathbb{E}((X_{u1,k} - Y_{u1,k})^2 \mid \sigma_{u1} = +)$. Now we integrate out D. Since $D \sim \text{Pois}(d)$, its moment generating function is $\mathbb{E}e^{tD} = e^{d(e^t-1)}$. Setting $t = -\theta^2/10$, we have $e^t \leq 1 + t/2$ for all $\theta \in [0, 1]$; hence,

$$\mathbb{E}e^{tD} \leq e^{td/2} = e^{-\frac{\theta^2 d}{20}}.$$

That is,

$$\mathbb{E}\big((X_{u,k+1} - Y_{u,k+1})^2 \mid \sigma_u = +\big) \leq Cz\mathbb{E}e^{-\frac{\theta^2 D}{10}} \leq Cze^{-\frac{\theta^2 d}{20}}.$$

In particular, the right hand side is smaller than $z/2$ if $\theta^2 d$ is sufficiently large. □

4.3. *The large-θ case.* We now give an analogue of Proposition 3.15 in the Galton-Watson case.

PROPOSITION 4.6. *For any $0 < \theta^* < 1$, there is some $d^* = d^*(\theta^*)$ such that for the broadcast process on the Poisson mean d tree it holds that for all $\theta \geq \theta^*$, $d \geq d^*$, and $k \geq K(\theta, d, \delta)$,*

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

This completes the proof of Theorem 4.1 (by the same argument that followed Proposition 3.15).

4.3.1. *The case where one child has large error.* Our eventual goal is to prove Proposition 3.15 by a similar analysis of the partial derivatives of g that led to the proof of Proposition 3.15. In this section, however, we will deal with one case where the derivatives of g cannot be controlled well. First, we introduce a parameter $\epsilon = \epsilon(d) > 0$ that will be specified later. Next, fix a vertex u and let Ω be the event that all children i of u satisfy $|X_{ui,k} - Y_{ui,k}| \le \epsilon$. On Ω, we will analyze derivatives of g; off Ω we have the following lemma (recalling that D is the number of children of u):

LEMMA 4.7. *For any $0 < \theta^* < 1$, there exist $c, C > 0$ such that if $\eta < c$, $\theta \in [\theta^*, 1)$, and $\theta^2 d > C$ then for any $\epsilon > 0$ and $k \ge K(\theta, d, \delta)$*

$$\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|}1_{\Omega^c} \mid D) \le \frac{C}{\sqrt{\epsilon}} D e^{-cD} \mathbb{E}\sqrt{|X_{ui,k} - Y_{ui,k}|}1_{\{|X_{ui,k} - Y_{ui,k}| > \epsilon\}}.$$

PROOF. First, we condition on D; we may then write $1_{\Omega^c} \le \sum_{i=1}^{D} 1_{\{|X_{ui,k} - Y_{ui,k}| > \epsilon\}}$. Hence,

$$\begin{aligned}\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|}1_{\Omega^c} \mid D) &\le \mathbb{E}\left(\sum_{i=1}^{D} \sqrt{|X_{u,k+1} - Y_{u,k+1}|}1_{\{|X_{ui,k} - Y_{ui,k}| > \epsilon\}} \;\middle|\; D\right) \\ &= D\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|}1_{\{|X_{ui,k} - Y_{ui,k}| > \epsilon\}} \mid D),\end{aligned}$$

where the equality follows because all the terms in the sum have the same distribution. Now we will condition on $X_{ui,k}$ and $Y_{ui,k}$, and we will show that on the event $\{|X_{ui,k} Y_{ui,k}| \ge \epsilon\}$ we have

$$D\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|} \mid D, X_{ui,k}, Y_{ui,k}) \le CDe^{-cD}. \tag{4.2}$$

After bounding $1 \le \epsilon^{-1/2}\sqrt{|X_{ui,k} - Y_{ui,k}|}$ on the event $\{|X_{ui,k} Y_{ui,k}| \ge \epsilon\}$ and then integrating out $X_{ui,k}$ and $Y_{ui,k}$, the proof will be complete.

Now we prove (4.2). Condition on σ_u, and suppose without loss of generality that $\sigma_u = +$. If $\theta^2 d$ is sufficiently large then Proposition 4.2 implies that (conditioned on $\sigma_u = +$) every child $j \ne i$ of u independently satisfies

$$\Pr(X_{uj,k} \ge 1 - \eta \mid \sigma_u = +) \ge 7/8.$$

If we condition also on D, Hoeffding's inequality implies that there is a constant $c > 0$ such that with probability at least e^{-cD^2}, at least $3/4$ of u's children j satisfy $X_{uj,k} \ge 1 - \eta$. The remaining children (which possibly include i) satisfy $X_{uj,k} \ge -1$, and so on this event

$$A := \prod_{j=1}^{D} \frac{1 - \theta X_{uj,k}}{1 + \theta X_{uj,k}} \le \left(\frac{1 - \theta(1-\eta)}{1 + \theta(1-\eta)}\right)^{3D/4} \left(\frac{1+\theta}{1-\theta}\right)^{D/4} \le (3\eta)^{3D/4} \eta^{-D/4}.$$

Now, $X_{u,k+1} = \frac{1-A}{1+A} \geq 1 - 2A$, and so we conclude that

$$\Pr(X_{u,k+1} \geq 1 - 2 \cdot 3^{3D/4}\eta^{D/2} \mid X_{ui,k}, Y_{ui,k}, \sigma_u = +, D) \geq 1 - e^{-cD^2}.$$

The previous argument applies equally well with X replaced by Y; hence the union bound implies

$$\Pr(|X_{u,k+1} - Y_{u,k+1}| \geq 4 \cdot 3^{3D/4}\eta^{D/2} \mid X_{ui,k}, Y_{ui,k}, \sigma_u = +, D) \geq 1 - 2e^{-cD^2}.$$

On the other hand, we always have the bound $|X_{u,k+1} - Y_{u,k+1}| \leq 2$, and so

$$\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|} \mid X_{ui,k}, Y_{ui,k}, \sigma_u = +, D) \leq 2 \cdot 3^{3D/8}\eta^{D/4} + 2\sqrt{2}e^{-cD^2}.$$

Now, if $\eta < c$ for c sufficiently small, the right hand side is bounded by Ce^{-cD}. This proves (4.2) in the case that $\sigma_u = +$. To complete the proof, we apply the symmetric argument conditioned on $\sigma_u = -$. □

4.3.2. *An analogue of Lemma 3.16.* The proof of Proposition 4.6 proceeds by analysing the derivatives of the recurrence (3.14). Recalling that these derivatives involve a large product, an important ingredient in the analysis is a bound on the expectation of each term. The following lemma is analogous to Lemma 3.16 in the regular case; an important difference is that Lemma 4.8 does not improve as $\eta \to 0$. In fact, as we remarked after Assumption 4.1, we cannot expect such behavior because of the possibility of extinction.

LEMMA 4.8. *For any $0 < \theta^* < 1$, there are some $\lambda = \lambda(\theta^*) < 1$ and $d^* = d^*(\theta^*)$ such that for all $\theta \geq \theta^*$, $d \geq d^*$ and $k \geq K(\theta, d, \delta)$,*

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

The same holds with Y replacing X.

We postpone the details of Taylor expansion and approximation to the appendix, but we will include here one of the main ingredients of Lemma 4.8's proof. The point is that in the Galton-Watson case (unlike the d-ary case) if d is fixed and $\eta \to 0$ then we cannot expect $X_{\rho,k}$ to be large (i.e. close to 1) with probability converging to 1. It turns out to be enough, however, to show that $X_{\rho,k}$ is *non-negative* with probability converging to 1.

LEMMA 4.9. *There is a constant C such that if $\theta^2 d \geq C$ then for any $k \geq K(\theta, d, \delta)$,*

$$\Pr(X_{u,k} < 0 \mid \sigma_u = +) \leq \eta, \tag{4.3}$$

and similarly for Y.

PROOF. We will give the argument for X only (the argument for Y is identical). First, note that if $\eta \geq 1/12$ then (4.3) follows directly from Proposition 4.2 if $d^\star$ is sufficiently large. Hence, we may assume that $\eta < 1/12$. Let $p_k = \Pr^+(X_{\rho,k} < 0)$. Then by Proposition 4.2, if C is sufficiently large then $p_k \leq 1/12$ for $k \geq K(\delta)$.

Let Z_- be the number of children i of the root with $X_{i,k} < 0$ and Z_+ be the number with $X_{i,k} \geq 1 - \eta$. Consider the quantity

$$Z := \prod_{i=1}^{D} \frac{1 - \theta X_{ui,k}}{1 + \theta X_{ui,k}},$$

and note that $X_{u,k} < 0$ if and only if $Z > 1$. Now, Z is increasing in each $X_{ui,k}$, and Z only increases if we drop some terms i with $X_{ui,k} \geq 0$. Hence,

$$Z \leq \left(\frac{1 - \theta(1-\eta)}{1 + \theta(1-\eta)}\right)^{Z_+} \left(\frac{1+\theta}{1-\theta}\right)^{Z_-} \leq (3\eta)^{Z_+} \eta^{-Z_-}. \tag{4.4}$$

Now, by the definition of p_k,

$$\Pr^+(X_{1,k} < 0) \leq \Pr(X_{1,k} < 0 \mid \sigma_1 = +) + \Pr(\sigma_1 = - \mid \sigma_\rho = +) = p_k + \eta. \tag{4.5}$$

Conditioned on σ_ρ and D, $Z_+ - Z_-$ is a sum of i.i.d. variables with values $1, -1$, and 0. Moreover, Proposition 4.2 with d sufficiently large implies that the probability of $X_{i,k} \geq 1 - \eta$ is at least $5/6$, while (4.5) implies that the probability of $X_{i,k} < 0$ is at most $p_k + \eta \leq 1/6$. Hence, Hoeffding's inequality implies that

$$\Pr^+(Z_+ - Z_- \leq D/3 + 1 \mid D) \leq Ce^{-cD^2},$$

for universal constants $c, C > 0$. Note also that if $Z_- = 0$ then $Z \geq 1$ and that in order to have $Z_- > 0$, there must be some i with $X_{i,k} < 0$. Note also that if $Z_+ - Z_- \geq D/3$ then $Z \leq 3^D \eta^{D/3} \leq (3/4)^{D/3} < 1$. Thus, applying a union bound, Hoeffding's inequality, and (4.5),

$$\begin{aligned}
\Pr^+(Z > 1 \mid D) &\leq \Pr^+(Z_+ - Z_- \leq D/3, Z_- > 0 \mid D) \\
&\leq D \Pr^+(Z_+ - Z_- \leq D/3, X_{1,k} < 0 \mid D) \\
&= D \Pr^+(Z_+ - Z_- \leq D/3 \mid D, X_{1,k} < 0) \Pr^+(X_{1,k} < 0 \mid D) \\
&\leq CDe^{-cD^2}(\eta + p_k).
\end{aligned} \tag{4.6}$$

Now, if d is large enough (which can be enforced by taking C large) then $\mathbb{E}De^{-cD^2} \le \frac{1}{4}$, which implies that

$$p_{k+1} = \Pr^+(X_{\rho,k+1} < 0) = \Pr^+(Z > 1) \le \frac{\eta + p_k}{4} \le \max\{\eta/2, p_k/2\}.$$

Recursing with k, we see that $\lim_{k\to\infty} \Pr^+(X_{\rho,k} < 0) \le \eta/2$, which implies that $\Pr^+(X_{\rho,k} < 0) \le \eta$ for sufficiently large k. □

4.3.3. *Analysis of the derivatives of g.* Our goal in this section is the following lemma, for which we recall that Ω is the event that all children i of u satisfy $|X_{ui,k} - Y_{ui,k}| \le \epsilon$. Let Ω_i be the event that $|X_{ui,k} - Y_{ui,k}| \le \epsilon$.

LEMMA 4.10. *For any $0 < \theta^* < 1$, there are constants $c, C > 0$ such that for all $0 < \epsilon < 1/4$, all $d \ge d^*(\theta^*)$, and for any $k \ge K(\theta, d, \delta)$,*

$$\mathbb{E}(1_\Omega \sqrt{|X_{u,k+1} - Y_{u,k+1}|} \mid D) \le CD(\epsilon^{-1}e^{-cD} + \sqrt{\epsilon})\mathbb{E}1_{\Omega_i}\sqrt{|X_{ui,k} - Y_{ui,k}|}.$$

We begin with an slightly improved version of (3.17): since $|X_{u,k+1} - Y_{u,k+1}| \le 2$, we can trivially improve (3.17) to

$$|X_{u,k+1} - Y_{u,k+1}| \le \sum_{i=1}^{D} \min\{2, |X_{ui,k} - Y_{ui,k}| \max\{h_i(X_{L_1(u),k}), h_i(Y_{L_1(u),k})\}\}. \tag{4.7}$$

Note that $1_\Omega \le 1_{\Omega_i}$ for any i (recall that $\Omega_i = \{|X_{ui,k} - Y_{ui,k}| \le \epsilon\}$), and so

$$|X_{u,k+1} - Y_{u,k+1}|1_\Omega \le \sum_{i=1}^{D} 1_{\Omega_i} \min\{2, |X_{ui,k} - Y_{ui,k}| \max\{h_i(X_{L_1(u),k}), h_i(Y_{L_1(u),k})\}\}.$$

Now, the terms on the right hand side have identical distributions; hence, taking conditional expectations gives

$$\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|}1_\Omega \mid D) \le D\mathbb{E}(1_{\Omega_i} \min\{2, \sqrt{|X_{ui,k} - Y_{ui,k}| \max\{h_i(X_{L_1(u),k}), h_i(Y_{L_1(u),k})\}\}} \mid D)$$

Defining

$$Z_X = \min\{1, \sqrt{|X_{ui,k} - Y_{ui,k}| h_i(X_{L_1(u),k})}\}$$

and similarly for Z_Y, we see that to prove Lemma 4.10 it suffices to show that

$$\mathbb{E}(1_{\Omega_i} Z_X \mid D) \le C(\epsilon^{-1} e^{-cD} + \sqrt{\epsilon})\mathbb{E}1_{\Omega_i}\sqrt{|X_{ui,k} - Y_{ui,k}|},$$

and similarly for Z_Y. We will show this by conditioning on $X_{ui,k}$ and $Y_{ui,k}$; that is, we will show the stronger statement that on the event Ω_i,

$$\mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}) \le C(\epsilon^{-1} e^{-cD} + \sqrt{\epsilon})\sqrt{|X_{ui,k} - Y_{ui,k}|} \tag{4.8}$$

(and similarly for Z_Y).

We split the analysis of Z_X and Z_Y into two cases. The first case is the easy case: if η is bounded away from zero or $|X_{ui,k}|$ and $|Y_{ui,k}|$ are bounded away from 1 then the denominator in h_i is bounded above:

LEMMA 4.11. *For any $0 < \theta^* < 1$, there are constants $c, C > 0$ such that for all $\epsilon \ge 0$, all $d \ge d^*(\theta^*)$, and for any $k \ge K(\theta, d, \delta)$, if $\max\{|X_{ui,k}|, |Y_{ui,k}|\} \le 1 - \epsilon$ then*

$$\mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}) \le \frac{C\lambda^{D-1}}{\max\{\sqrt{\eta}, \epsilon\}}\sqrt{|X_{ui,k} - Y_{ui,k}|},$$

and similarly for Z_Y.

PROOF. By the definition of h_i, and because $|X_{ui,k}| \le 1 - \epsilon$,

$$h_i(X_{ui,k}) \le \frac{4}{\max\{\eta, \epsilon^2\}} \min\left\{\prod_{j\ne i} \frac{1 - \theta X_{uj,k}}{1 + \theta X_{uj,k}}, \prod_{j\ne i} \frac{1 + \theta X_{uj,k}}{1 - \theta X_{uj,k}}\right\}.$$

Conditioning on $\sigma_u = +$ and considering the first term in the minimum, Lemma 4.8 implies that

$$\mathbb{E}\Big(\sqrt{|X_{ui,k} - Y_{ui,k}| h_i(X_{L_1(u),k})} \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = +\Big) \le \frac{2\lambda^{D-1}}{\max\{\sqrt{\eta}, \epsilon\}}\sqrt{|X_{ui,k} - Y_{ui,k}|}.$$

By symmetry, the same bound holds if we condition on $\sigma_u = -$. Recalling that $Z_X \le \sqrt{|X_{ui,k} - Y_{ui,k}| h_i(X_{L_1(u),k})}$, this completes the proof for Z_X. The exact same argument applies to Z_Y also. □

If $X_{ui,k}$ and $Y_{ui,k}$ are allowed to be arbitrarily close to 1 and η is allowed to be arbitrarily close to zero, then the argument is somewhat more tricky. The basic idea is that if $X_{ui,k}$ is close to 1 then σ_u is very likely to be +,

in which case the denominator in h_i^+ is at least 1 and so h_i^+ is small. Bad things happen if $\sigma_u = -$ because then we need to consider h_i^-, which has a small denominator. However, this event is very unlikely conditioned on $X_{ui,k}$ being close to 1, and so its contribution can be controlled.

LEMMA 4.12. *For any $0 < \theta^* < 1$, there are constants $c, C > 0$ such that for all $0 < \epsilon < 1/4$, all $d \geq d^*(\theta^*)$, and for any $k \geq K(\theta, d, \delta)$, if $|X_{ui,k} - Y_{ui,k}| \leq \epsilon$ and $\max\{|X_{ui,k}|, |Y_{ui,k}|\} \geq 1 - \epsilon$ then*

$$\mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}) \leq C(\lambda^{D-1} + \sqrt{\epsilon})\sqrt{|X_{ui,k} - Y_{ui,k}|},$$

and similarly for Z_Y.

Before proving Lemma 4.12, note that together with Lemma 4.11 it proves (4.8) and hence Lemma 4.10.

PROOF. Fix $\theta^* \in (0,1)$ and take $\lambda < 1$ satisfying Lemma 4.8. Since $\epsilon \leq 1/4$, it follows that $X_{ui,k}$ and $Y_{ui,k}$ have the same sign. Without loss of generality, they are both positive; hence, if $A = (1 - \min\{X_{ui,k}, Y_{ui,k}\})/2$ and $B = (1 - \max\{X_{ui,k}, Y_{ui,k}\})/2$ then $0 \leq B \leq A \leq \epsilon$. Note that $|X_{ui,k} - Y_{ui,k}| = 2|A - B|$. Now,

$$\Pr(\sigma_{ui} = + \mid X_{ui,k}, Y_{ui,k}) = \frac{1 + X_{ui,k}}{2} \geq 1 - A,$$

and so

$$\Pr(\sigma_u = + \mid X_{ui,k}, Y_{ui,k}) \geq 1 - A - \eta.$$

Since $X_{ui,k}$ is positive,

$$h_i^+(X_{L_1(u),k}) = \frac{4}{(1 + \theta X_{ui,k})^2} \prod_{j \neq i} \frac{1 - \theta X_{uj,k}}{1 + \theta X_{uj,k}} \leq 4 \prod_{j \neq i} \frac{1 - \theta X_{uj,k}}{1 + \theta X_{uj,k}}$$

and similarly for Y. By Lemma 4.8, if d^* is sufficiently large then

$$\begin{aligned} (4.9) \qquad & \mathbb{E}\left(\sqrt{|X_{ui,k} - Y_{ui,k}| h_i^+(X_{L_1(u),k})} \;\middle|\; D, X_{ui,k}, Y_{ui,k}, \sigma_u = +\right) \\ & \leq 4\mathbb{E}\left(\sqrt{|X_{ui,k} - Y_{ui,k}| \prod_{j \neq i} \frac{1 - \theta X_{uj,k}}{1 + \theta X_{uj,k}}} \;\middle|\; D, X_{ui,k}, Y_{ui,k}, \sigma_u = +\right) \\ & \leq 4\lambda^{D-1}\sqrt{|X_{ui,k} - Y_{ui,k}|}, \end{aligned}$$

since the $X_{uj,k}$ are independent conditioned on σ_u. On the other hand, since $Z_X \geq 0$ we have

$$\begin{aligned}\mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}) &\leq \mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = +) \\ &\quad + \Pr(\sigma_u = - \mid X_{ui,k}, Y_{ui,k})\mathbb{E}(Z \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = -) \\ &\leq \mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = +) \\ &\quad + (A + \eta)\mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = -). \qquad (4.10)\end{aligned}$$

By (4.9), the first term of (4.10) is bounded by $4\lambda^{D-1}\sqrt{|X_{ui,k} - Y_{ui,k}|}$.

Next, we consider the second term of (4.10); we will consider the coefficients A and η separately. Now, $Z_X \leq \sqrt{|X_{ui,k} - Y_{ui,k}| h_i^-(X_{L_1(u),k})}$ and

$$h_i^-(X_{L_1(u),k}) = \frac{4}{(1 - \theta X_{ui,k})^2} \prod_{j \neq i} \frac{1 + \theta X_{uj,k}}{1 - \theta X_{uj,k}} \leq \frac{1}{\max\{\eta, B\}^2} \prod_{j \neq i} \frac{1 + \theta X_{uj,k}}{1 - \theta X_{uj,k}}.$$

Then Lemma 4.8 implies that for d^* sufficiently large,

$$\begin{aligned}\mathbb{E}(\sqrt{h_i^-(X_{L_1(u),k})} \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = -) &\leq \frac{1}{\max\{\eta, B\}} \prod_{j \neq i} \mathbb{E}\left(\sqrt{\frac{1 + \theta X_{uj,k}}{1 - \theta X_{uj,k}}} \;\middle|\; D, \sigma_u = -\right) \\ &\leq \frac{\lambda^{D-1}}{\max\{\eta, B\}}. \qquad (4.11)\end{aligned}$$

In particular, we have

$$\begin{aligned}\eta\mathbb{E}(Z \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = -) &\leq \eta\sqrt{|X_{ui,k} - Y_{ui,k}|}\mathbb{E}(\sqrt{h_i^-(X_{L_1(u),k})} \mid D, X_{ui,k}, Y_{ui,k}, \sigma_i) \\ &\leq \lambda^{D-1}\sqrt{|X_{ui,k} - Y_{ui,k}|}, \qquad (4.12)\end{aligned}$$

which handles the term in (4.10) involving η.

Next, we consider the term involving A. If $A \leq 2B$ then we may use (4.11) for the bound

$$\mathbb{E}(\sqrt{h_i^-(X_{L_1(u),k})} \mid D, X_{ui,k}, Y_{ui,k}, \sigma_u = -) \leq \frac{\lambda^{D-1}}{B} \leq \frac{2\lambda^{D-1}}{A}. \qquad (4.13)$$

Alternatively, if $A \geq 2B$ then $|X_{ui,k} - Y_{ui,k}| = 2|A - B| \geq A$; since $Z \leq 1$, we have

$$A\mathbb{E}(Z \mid X_{ui,k}, Y_{ui,k}, \sigma_u = -) \leq A \leq \sqrt{A|X_{ui,k} - Y_{ui,k}|} \leq \sqrt{\epsilon|X_{ui,k} - Y_{ui,k}|}.$$

Combining this with (4.13), we have

$$A\mathbb{E}(Z \mid X_{ui,k}, Y_{ui,k}, \sigma_u = -) \leq \max\{2\lambda^{D-1}, \sqrt{\epsilon}\}\sqrt{|X_{ui,k} - Y_{ui,k}|}$$

in either case. Combining this with (4.12) and going back to (4.10), we have

$$\mathbb{E}(Z \mid D, X_{ui,k}, Y_{ui,k}) \le (C\lambda^{D-1} + \sqrt{\epsilon})\sqrt{|X_{ui,k} - Y_{ui,k}|},$$

which completes the proof. □

4.3.4. *Putting it together.* Finally, we put together the various cases and prove Proposition 4.6. First, fix θ^* and put $\epsilon = d^{-4}$. The easy case is when $\eta \ge c$, where c is the constant from Lemma 4.7. In this case, Lemma 4.11 with $\epsilon = 0$ implies that

$$\mathbb{E}(Z_X \mid D, X_{ui,k}, Y_{ui,k}) \le Ce^{-cD}\sqrt{|X_{ui,k} - Y_{ui,k}|}$$

and similarly for Z_Y. Taking the expectation over $X_{ui,k}$ and applying (4.7) implies that

$$\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|} \mid D) \le CDe^{-cD}\mathbb{E}\sqrt{|X_{ui,k} - Y_{ui,k}|}. \tag{4.14}$$

Now consider the case where $\eta \le c$. By Lemma 4.7 (recalling that $\epsilon = d^{-4}$), we have

$$\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|}1_{\Omega^c} \mid D) \le Cd^2De^{-cD}\mathbb{E}1_{\Omega_i^c}\sqrt{|X_{ui,k} - Y_{ui,k}|}.$$

By Lemma 4.10, we have

$$\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|}1_{\Omega} \mid D) \le C(d^4De^{-cD} + d^{-2}D)\mathbb{E}1_{\Omega_i}\sqrt{|X_{ui,k} - Y_{ui,k}|}.$$

Putting these together, we have

$$\mathbb{E}(\sqrt{|X_{u,k+1} - Y_{u,k+1}|} \mid D) \le C(d^4De^{-cD} + d^{-2}D)\mathbb{E}\sqrt{|X_{ui,k} - Y_{ui,k}|}. \tag{4.15}$$

Noting that the right hand side of (4.15) is larger than the right hand side of (4.14), we see that (4.15) holds without extra conditions on η. Finally, we integrate out D in (4.15). Since $D \sim \mathrm{Pois}(d)$, we have $\mathbb{E}D = d$ and $\mathbb{E}De^{-cD} \le e^{-c'd}$ for some constant c' depending on c. In particular, if d is sufficiently large (depending on C and c, which depend in turn on θ^*) then

$$C\mathbb{E}(d^4De^{-cD} + d^{-2}D) \le \frac{1}{2},$$

which proves Proposition 4.6.

5. From trees to graphs. In this section, we will give our reconstruction algorithm and prove that it performs optimally. It will be convenient for us to work with block models on fixed vertex sets instead of random ones; therefore, let $\mathcal{G}(V^+, V^-, p, q)$ denote the random graph on the vertices $V^+ \cup V^-$ where pairs of vertices within V^+ or V^- are connected with probability p and pairs of vertices spanning V^+ and V^- are included with probability q. Note that if V^- and V^+ are chosen to be a uniformly random partition of $[n]$ then $\mathcal{G}(V^+, V^-, \frac{a}{n}, \frac{b}{n})$ is simply $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$.

Let `BBPartition` denote the algorithm of [25], which satisfies the following guarantee, where V^i denotes $\{v \in V(G) : \sigma_v = i\}$:

THEOREM 5.1. *Suppose that* $G \sim \mathcal{G}(V^+, V^-, \frac{a}{n}, \frac{b}{n})$, *where* $|V^+| + |V^-| = n + o(n)$, $|V^+| - |V^-| = O(\sqrt{n})$ *and* $(a-b)^2 > 2(a+b)$. *There exists some* $0 \le \delta < \frac{1}{2}$ *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* $|W^+ \Delta V^i| \le \delta n$ *for some* $i \in \{+, -\}$.

Moreover, `BBPartition` *runs in time* $O(n^{1+o(1)})$.

REMARK 5.2. *We should point out that [25] only claims Theorem 5.1 when* V^+ *and* V^- *are uniformly random partitions of* $[n]$; *however, one easily deduce the result for almost-balanced partitions from the result for uniformly random partitions: choose* $\epsilon > 0$ *so that* $\frac{(a-b)^2}{2(a+b)} > \frac{1}{1-\epsilon}$. *Given a graph* G *from* $\mathcal{G}(V^+, V^-, \frac{a}{n}, \frac{b}{n})$, *let* H *be the graph obtained by deleting all but* $\lceil (1-\epsilon)n \rceil$ *vertices at random from* G. *If* (W^+, W^-) *is the partition of* H *according to its vertex labels then one can check that the sizes of* W^+ *and* W^- *are contiguous with the sizes of a uniformly random partition of* $\lceil (1-\epsilon)n \rceil$. *Hence, the distribution of* H *is contiguous with* $\mathcal{G}(\lceil (1-\epsilon)n \rceil, \frac{a}{n}, \frac{b}{n})$. *The results of [25] then imply that the labels of* H *can be recovered adequately (i.e., as claimed in Theorem 5.1); by randomly labelling the vertices of* G *that were deleted, we recover Theorem 5.1 as stated.*

Note that by symmetry, Theorem 5.1 also implies that $|W^- \Delta V^j| \le \delta n$ for $j \ne i \in \{+, -\}$. 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 $\delta(G)$ be the (random) fraction of vertices that are mis-labelled.

For $v \in G$ and $R \in \mathbb{N}$, define $B(v, R) = \{u \in G : d(u, v) \le R\}$ 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 τ on its leaves, we consider the following estimator of v's label: first, take K large enough so that Proposition 4.2 holds for $k = K$. For

$u \in S(v, R-K)$, define $Y_{u,K}(\tau)$ as the sign of $S'_k(\tau)$, where S'_k is given as in the proof of Proposition 4.2. That is, $Y_{u,K}(\tau)$ is the sign of a weighted sum of the labelling τ on $S(v,R)$. For $k > K$ and $u \in B(v, R-k)$, define $Y_{u,k}(\tau)$ recursively by $Y_{u,k} = g(Y_{L_1(u),k-1})$, where g is given by (3.14). Then Y satisfies Assumption 4.1.

We remark that the reason for taking this two-stage definition of Y is because we don't necessarily know how much noise there is on the leaves (i.e., δ), and so we cannot define Y by (3.1). Defining Y as we have done avoids the need to know δ, while still satisfying the required assumptions.

Before presenting the algorithm, we will mention one issue that we glossed over in our earlier sketch: since we will run the black-box algorithm several times, and since the labels + and − are symmetric, we need some way to break the symmetry between the various runs of the algorithm. We do this by holding out a single vertex of high degree (that we call u_*) and breaking symmetry according to the sign of most of its neighbors.

Algorithm 1 Optimal graph reconstruction algorithm

$R \leftarrow \lfloor \frac{1}{20(a+b)} \log n \rfloor$
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 $V \setminus U$
$W_*^+, W_*^- \leftarrow \varnothing$
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 if
    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
for $v \in U$ **do**
    Assign v to W_*^+ or W_*^- uniformly at random
end for

REMARK 5.3. *Our analysis of Algorithm 1 will assume that we can compute with arbitrary precision numbers in constant time. However, Propositions 4.5 and 4.6 can also be used to analyze an implementation of Algorithm 1 with finite-precision arithmetic. Indeed, the only part of Algorithm 1 where continuous quantities appear is in the computation of $Y_{v,R}$, and the main question in the computation of $Y_{v,R}$ is whether the numerical errors accumulate as we repeatedly apply the recursion $g(x)$ defined in (3.14).*

Consider the following finite-precision implementation of the recursion: first, compute $\hat{Y}_{ui,k}$ to the desired precision for all children i of u. Then compute $g(\hat{Y}_{u,L_1(k)})$ to arbitrary precision, and finally define $\hat{Y}_{u,k}$ to be $g(\hat{Y}_{u,L_1(k)})$ truncated to the desired precision. Let us see what Proposition 4.5 has to say about this procedure (Proposition 4.6 has similar consequences for the other range of parameters): if X denotes the true magnetizations and the rounding error is bounded by ϵ then

$$\begin{aligned}\mathbb{E}(X_{u,k+1} - \hat{Y}_{u,k+1})^2 &\le \mathbb{E}(X_{u,k+1} - g(\hat{Y}_{L_1(u),k}) + \epsilon)^2 \\ &\le O(\epsilon) + \mathbb{E}(X_{u,k+1} - g(\hat{Y}_{L_1(u),k}))^2 \\ &\le O(\epsilon) + \frac{1}{2}\mathbb{E}(X_{u,k} - \hat{Y}_{u,k})^2,\end{aligned}$$

which implies that the asymptotic accuracy of our finite-precision scheme is within $O(\sqrt{\epsilon})$ of optimal.

As presented, our algorithm is not particular efficient (although it does run in polynomial time) because we need to re-run `BBPartition` for almost every vertex in V. However, one can modify Algorithm 1 to run in $O(n^{1+o(1)})$ time by processing $o(n)$ vertices in each iteration (a similar idea is used in [25]). 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 5.4. *Algorithm 1 produces a partition $W_*^+ \cup W_*^- = V(G)$ such that a.a.s. $|W_*^+ \Delta V^i| \le (1 + o(1))n(1 - p_T(a,b))$ for some $i \in \{+,-\}$.*

Theorem 2.8 implies that for any algorithm, $|W_*^+ \Delta V^i| \ge (1 - o(1))n(1 - p_T(a,b))$ a.a.s. Hence, it is enough to show that $\mathbb{E}|W_*^+ \Delta V^i| \le (1 + o(1))n(1 - p_T(a,b))$. Since Algorithm 1 treats every node equally, it is enough to show that there is some i such that for every $v \in V^i$,

$$\Pr(v \in W_*^+) \to p_T(a,b). \tag{5.1}$$

Moreover, since $\Pr(v \in U) \to 0$, it is enough to show (5.1) for every $v \in V^i \setminus U$.

The proof of (5.1) will take the remainder of this section. First, we will deal with a technicality: in line 6, we are applying `BBPartition` to the subgraph of G induced by $V \setminus B(v, R-1) \setminus U$; call this graph G_v. We need to justify the fact that G_v satisfies the requirements of Theorem 5.1. Now, if $W^+ = V^+ \setminus B(v, R-1) \setminus U$ and $W^- = V^- \setminus B(v, R-1) \setminus U$ then $G_v \sim \mathcal{G}(W^+, W^-, \frac{a}{n}, \frac{b}{n})$. Since

$$|W^+| + |W^-| = n - |B(v, R-1)| - \lfloor \sqrt{n} \rfloor$$

and

$$\big||W^+|-|W^-|\big| \le \big||V^+|-|V^-|\big| + |B(v,R-1)| + \lfloor\sqrt{n}\rfloor \le O(\sqrt{n}) + |B(v,R-1)|,$$

we see that the hypothesis of Theorem 5.1 is satisfied as long as $|B(v,R-1)| = O(\sqrt{n})$. This is indeed the case; Lemma 4.4 of [24] shows that $|B(v,R)| = O(n^{1/8})$ for the value of R that we have chosen:

LEMMA 5.5. *$|B(v,R)| = O(n^{1/8})$ a.a.s.*

We conclude, therefore, that Theorem 5.1 applies in line 6 of Algorithm 1:

LEMMA 5.6. *There is some $0 \le \delta < \frac{1}{2}$ such that for any $v \in V \setminus U$, there a.a.s. exists some $i \in \{+,-\}$ such that $|W_v^+ \Delta V^i| \le \delta n$, with W_v^+ defined as in line 6.*

5.1. *Aligning the calls to* `BBPartition`. Next, let us discuss in more detail the purpose of u_* and line 8. Recall 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`.

We will break the symmetry between + and − by assuming, from now on, that u_* is labelled +. Next, let us note some properties of u_*:

LEMMA 5.7. *In line 3, there a.a.s. exists at least one $u \in U$ with more than $\sqrt{\log n}$ neighbors in $V \setminus U$; hence, u_* is well-defined. Moreover, there is some $\eta > 0$ such that a.a.s. at least a $(1+\eta)/2$-fraction of u_*'s neighbors in $V \setminus U$ either are labelled + (if $a > b$) or − (if $a < b$). Finally, for any $v \in V \setminus U$, u_* a.a.s. has no neighbors in $B(v, R-1)$.*

PROOF. For the first claim, note that every $u \in U$ independently has more than $\mathrm{Binom}(\lceil n(1-\epsilon/2)\rceil, \frac{\min\{a,b\}}{n})$ neighbors in $V \setminus U$, and the maximum of $\sqrt{n}$ such variables is of order $\Theta(\log n/\log\log n) \gg \sqrt{\log n}$.

For the second claim, let d be the number of neighbors that u_* has in $V \setminus U$ and note that $d = O(\log n)$ a.a.s., because the maximum degree of any vertex in G is $O(\log n)$. Conditioned on d, the number of u_*'s +-labelled neighbors in $V \setminus U$ is dominated by $\mathrm{Binom}(d, \frac{a}{a+b} \cdot \frac{|V^+|-d}{|V^-|})$; this is because the neighborhood of u_* may be generated by sequentially choosing d neighbors without replacement from $V \setminus U$, where a +-labelled neighbor is chosen with probability $\frac{a}{a+b}$ times the fraction of +-labelled vertices remaining. Since $|V^+| = n/2 \pm O(n^{1/2})$ and $d = o(n)$, we see that u_* a.a.s. has at least $d(\frac{a}{a+b} -$

$o(1)$) +-labelled neighbors. If $a > b$ then this verifies the second claim; if $a < b$ then we repeat the argument with + replaced by −.

For the final claim, note that if u_* has a neighbor in $B(v, R-1)$ then $u_* \in B(v,R)$. But (by Lemma 5.5) $|B(v,R)| = O(n^{1/8})$ a.a.s., and so with probability tending to 1, $B(v,R)$ does not intersect U at all; in particular, it does not contains u_*. □

From now on, suppose without loss of generality that $\sigma_{u^*} = +$. Thanks to the previous paragraph and Theorem 5.1, we see that the relabelling in lines 8 and 10 correctly aligns W_v^+ with V^+:

LEMMA 5.8. *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 8 or line 10.*

PROOF. Assume for now that $a > b$. Just for the duration of this proof, let W_v^+ and W_v^- denote the partition as defined in line 6 of Algorithm 1, while $\tilde{W}_v^+$ and $\tilde{W}_v^-$ denote the partition defined by line 8 or line 10.

Recall from Lemma 5.7 that u_* has at least $\sqrt{\log n}$ neighbors in $V \setminus B(v,R-1) \setminus U$, of which at least a $(1+\eta)/2$-fraction are labelled +; let $d \ge \sqrt{\log n}$ be the number of neighbors that u_* has in $V \setminus B(v,R-1) \setminus U$, and let $p \ge (1+\eta)/2$ be the fraction that are actually labelled +. Note that the labelling W_v^+, W_v^- produced in line 6 is independent of the set of u_*'s neighbors in $V \setminus B(v,R-1) \setminus U$, because W_v^+ and W_v^- depend only on edges within $V \setminus B(v,R-1) \setminus U$ and these are independent of the edges adjoining u_*. That is, conditioned on d, p, W_v^+ and W_v^-, the neighbors of u_* can be generated by taking u_*'s +-labelled neighbors to be a uniformly random set of pd +-labelled vertices and then taking u_*'s −-labelled neighbors to be a uniformly random set of $(1-p)d$ −-labelled vertices. Hence, if N_{ij} (for $i,j \in \{+,-\}$) is the number of u_*'s neighbors in $V^i \cap W_v^j$ then conditioned on d, p, and W_v^+, N_{++} is distributed as $\mathrm{HyperGeom}(dp, |W_v^+ \cap V^+|, |V^+|)$ and N_{-+} is distributed as $\mathrm{HyperGeom}(d(1-p), |W_v^+ \cap V^-|, |V^-|)$. Since $d = o(|V^+|) = o(|V^-|)$ and $d \to \infty$ a.a.s., we have

$$N_{++} \ge (1-o(1))dp\frac{|W_v^+ \cap V^+|}{|V^+|} = (1-o(1))\frac{2dp|W_v^+ \cap V^+|}{n}$$
$$N_{-+} \ge (1-o(1))d(1-p)\frac{|W_v^+ \cap V^-|}{|V^-|} = (1-o(1))\frac{2d(1-p)|W_v^+ \cap V^-|}{n}.$$

Adding these together, we have

$$N_{++} + N_{-+} = (1-o(1))\frac{d}{n}(\alpha + \beta + (2p-1)(\alpha - \beta)) \tag{5.2}$$

where $\alpha = |W_v^+ \cap V^+|$ and $\beta = |W_v^+ \cap V^-|$.

Now, Lemma 5.6 admits two cases: if $i = +$ then

$$\delta n \geq |W_v^+ \Delta V^+| = |W_v^+ \cap V^-| + |W_v^- \cap V^+| = |W_v^+ \cap V^-| + \frac{n}{2} + o(n) - |W_v^+ \cap V^+|,$$

and we conclude that $\alpha - \beta \geq (\frac{1}{2} - \delta - o(1)))n$. A similar argument when $i = -$ in Lemma 5.6 shows that in that case $\alpha - \beta \leq -(\frac{1}{2} - \delta - o(1))n$. In either case, $\alpha + \beta = (1 + o(1))n/2$.

If $i = +$ in Lemma 5.6 then since $p - 1/2 \geq \eta/2$, (5.2) implies

$$N_{++} + N_{-+} = (1 - o(1))d\left(\frac{1}{2} + \frac{(\frac{1}{2} - \delta)\eta}{2}\right)$$

a.a.s. Since $N_{++} + N_{-+} + N_{+-} + N_{--} = d$, we have in particular $N_{++} + N_{-+} > N_{+-} + N_{--}$ a.a.s., and so u_* has most of its neighbors in W_v^+. Hence, $\tilde{W}_v^+ = W_v^+$ and so Lemma 5.6 with $i = +$ implies the the conclusion of Lemma 5.8 holds. On the other hand, if $i = -$ in Lemma 5.6 then $\alpha - \beta < -(\frac{1}{2} - \delta)n$; by (5.2), $N_{+-} + N_{--} > N_{++} + N_{-+}$. Then u_* has most of its neighbors in W_v^- and so $\tilde{W}_v^+ = W_v^-$. By Lemma 5.6 with $i = -$, the conclusion of Lemma 5.8 holds.

Finally, we mention the case $a < b$: essentially the same argument holds except that instead of $p \geq (1+\eta)/2$ we have $p \leq (1-\eta)/2$. Then $i = +$ implies that u_* has most of its neighbors in W_v^-, while $i = -$ implies that u_* has most of its neighbors in W_v^+. □

5.2. *Calculating v's label.* To complete the proof of (5.1) (and hence Theorem 5.4), we need to discuss the coupling between graphs and trees. We will invoke a lemma from [24] which says that a neighborhood in G can be coupled with a multi-type branching process of the sort that we considered in Section 4. Indeed, let T be the Galton-Watson tree of Section 4 (with $d = (a+b)/2$) and let σ' be a labelling on it, given by running the two-state broadcast process with parameter $\eta = b/(a+b)$. We write T_R for $T \cap \mathbb{N}^R$; that is, the part of T which has depth at most R.

LEMMA 5.9. *For any fixed $v \in G$, there is a coupling between (G, σ) and (T, σ') such that $(B(v, R), \sigma_{B(v,R)}) = (T_R, \sigma'_{T_R})$ a.a.s.*

Armed with Lemma 5.9, we will consider a slightly different method of generating G, which is nevertheless equivalent to the original model in the sense that the new method and the old method may be coupled a.a.s. In the new construction, we begin by assigning labels to $V(G)$ uniformly at random. Beginning with a fixed vertex v, we construct $B(v, R-1)$ by drawing

a Galton-Watson tree of depth $R-1$ rooted at v, with labels distributed according to the broadcast process. On the vertices that remain (i.e., those that were not used in $B(v,R-1)$), we construct a graph G' according to the stochastic block model with parameters a/n and b/n. Finally, we join $B(v,R-1)$ to the rest of the graph: for every vertex $u \in S(v,R-1)$, we draw $\mathrm{Pois}(a/(a+b))$ vertices at random from G' with label σ_u and $\mathrm{Pois}(b/(a+b))$ vertices from G' with label $-\sigma_u$; we connect all these vertices to u. It follows from Lemma 5.9 that this construction is equivalent to the original construction. It also follows from Lemma 5.5 that $|G'| \geq n - O(n^{1/8})$ a.a.s.

The advantage of the construction above is that it becomes obvious that the edges of $G' = G \setminus B(v,R-1) \setminus U$ are independent of both $B(v,R-1)$ and the edges joining $B(v,R-1)$ to G'. Since W_v^+ and W_v^- are both functions of G' only, it follows that $B(v,R-1)$ and its edges to G' are also independent of W_v^+ and W_v^-. Using this observation, we can improve Lemma 5.9 to include the noisy labels. In particular, we claim that the labelling ξ produced in line 12 of Algorithm 1 has the same distribution as the noisy labelling τ of the noisy broadcast process.

In view of Lemma 5.9, it suffices to condition on σ, $B(v,R-1)$ and G', and to show that the conditional distribution of ξ is essentially the same as the conditional distribution of τ given T and σ' in the noisy broadcast process. Since the edges joining $B(v,R-1)$ to G' are independent of W_v^+ and W_v^-, for any $u \in S(v,R-1)$ with $\sigma_u = +$ we have

$$\begin{aligned}
\#\{w \sim u : w \in G', \sigma_w = +, \xi_w = +\} &\sim \mathrm{Binom}\left(|V^+ \cap W_v^+|, \frac{a}{n}\right)\\
\#\{w \sim u : w \in G', \sigma_w = +, \xi_w = -\} &\sim \mathrm{Binom}\left(|V^+ \cap W_v^-|, \frac{a}{n}\right)\\
\#\{w \sim u : w \in G', \sigma_w = -, \xi_w = -\} &\sim \mathrm{Binom}\left(|V^- \cap W_v^-|, \frac{b}{n}\right)\\
\#\{w \sim u : w \in G', \sigma_w = -, \xi_w = +\} &\sim \mathrm{Binom}\left(|V^- \cap W_v^+|, \frac{b}{n}\right).
\end{aligned}$$

Moreover, the random variables above are independent as u ranges over $S(v,R-1)$. Now, if we define $\delta = \frac{1}{n}|V^+ \Delta W_v^+|$ then $\mathrm{Binom}(|V^+ \cap W_v^+|, a/n)$ and $\mathrm{Pois}(a(1-\delta)/2)$ are at total variation distance at most $O(n^{-1/2})$; here, we are using the fact that $|V^+ \cap W_v^+| = (1-\delta)n/2 \pm O(n^{1/2})$, which follows because V^+, V^- are an equipartition of $V(G)$ and W_v^+, W_v^- are an equipartition of $V(G')$, which contains all but at most $O(\sqrt{n})$ vertices of G. Similarly, we

have

$$\#\{w \sim u : w \in G', \sigma_w = +, \xi_w = +\} \overset{d}{\approx} \mathrm{Pois}(a(1-\delta)/2)$$
$$\#\{w \sim u : w \in G', \sigma_w = +, \xi_w = -\} \overset{d}{\approx} \mathrm{Pois}(a\delta/2)$$
$$\#\{w \sim u : w \in G', \sigma_w = -, \xi_w = -\} \overset{d}{\approx} \mathrm{Pois}(b(1-\delta)/2)$$
$$\#\{w \sim u : w \in G', \sigma_w = -, \xi_w = +\} \overset{d}{\approx} \mathrm{Pois}(b\delta/2)$$

where "$\overset{d}{\approx}$" means that the distributions are at total variation distance at most $O(n^{-1/2})$. Note that the distributions on the right hand side are exactly the distributions of the noisy labels τ under the noisy broadcast process. By a similar argument for $\sigma_u = -$, and a union bound over the $O(n^{1/8})$ choices for u, we see that the joint distribution of $B(v,R)$ and $\{\xi_u : u \in S(v,R)\}$ a.a.s. the same as the joint distribution of T_R and $\{\tau_u : u \in \partial T_R\}$. Hence, by Theorem 4.1,

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

By line 13 of Algorithm 1, this completes the proof of (5.1).

Acknowledgement. The authors thank Jiaming Xu for his careful reading of the manuscript and his helpful comments and corrections.

APPENDIX A: BOUNDS ON $\mathbb{E}\sqrt{\frac{1-\theta X}{1+\theta X}}$

Because of the form of the recursion (3.14), at various points in our analysis we require bounds on quantities of the form $\mathbb{E}\sqrt{\frac{1-\theta X}{1+\theta X}}$, under various assumptions on X. These estimates are elementary but tedious to check, and so we have collected them here.

Proof of Lemma 3.16. By Lemma 3.7, we have

$$\Pr(X_{ui,k} \geq 1 - \eta\alpha t \mid \sigma_u = +) \geq \Pr(X_{ui,k} \geq 1 - \eta\alpha t \mid \sigma_{ui} = +) - \eta \geq 1 - t^{-1} - \eta,$$

where $\alpha = C/(\theta^2 d)$ can be taken arbitrarily small if we require $\theta^2 d$ to be large.

Fix some $\epsilon = \epsilon(\theta^*) > 0$ to be determined later. Take $t = \epsilon^{-1}\eta^{-3/4}$ so that

$$\Pr(X \geq 1 - \frac{\alpha\eta^{1/4}}{\epsilon}) \geq 1 - \epsilon\eta^{3/4} - \eta.$$

Now, suppose that α is small enough so that $\alpha\epsilon^{-1} \leq \epsilon$. Then

$$\Pr(X \geq 1 - \epsilon\eta^{1/4}) \geq 1 - \epsilon\eta^{3/4} - \eta. \tag{A.1}$$

Now consider the function

$$f(x) := \sqrt{\frac{1-\theta x}{1+\theta x}}.$$

Note that $f(x)$ is decreasing in x, and hence

$$\mathbb{E} f(X) \le f(s)\Pr(X \ge s) + f(-1)\Pr(X \le s).$$

for any random variable X supported on $[-1,1]$ and for any $s \in [-1,1]$. Applying this for $s = 1-\epsilon\eta^{1/4}$, we have (by (A.1))

$$\mathbb{E} f(X) \le f(1-\epsilon\eta^{1/4})(1-\epsilon\eta^{3/4}-\eta) + f(-1)(\epsilon\eta^{3/4}+\eta). \tag{A.2}$$

We will now check that if $\eta \le \frac{1-\theta^*}{2} < 1/2$ then each term on the right hand side of (A.2) can be made strictly smaller than $1/2$, and also smaller than $2\eta^{1/4}$, by taking $\epsilon = \epsilon(\theta^*)$ small enough. This will complete the proof of the Lemma.

We consider the term involving $f(-1)$ first:

$$f(-1)(\epsilon\eta^{3/4}+\eta) = \epsilon\eta^{1/4}\sqrt{1-\eta} + \sqrt{\eta(1-\eta)}. \tag{A.3}$$

On the interval $\eta \in [0, \frac{1-\theta^*}{2}]$, $\sqrt{\eta(1-\eta)}$ is bounded away from $1/2$, and $\eta^{1/4}\sqrt{1-\eta}$ is bounded above. Hence, (A.3) is bounded away from $1/2$ as long as $\epsilon(\theta^*)$ is small enough. On the other hand, (A.3) is also bounded by $2\eta^{1/4}$ as long as $\epsilon \le 1$.

Next, we consider the $f(1-\epsilon\eta^{1/4})$ term of (A.2). Note that $\theta(1-\epsilon\eta^{1/4}) \ge 1-2\eta-\epsilon\eta^{1/4}$ and so

$$f(1-\epsilon\eta^{1/4}) \le \sqrt{\frac{2\eta+\epsilon\eta^{1/4}}{2-(2\eta+\epsilon\eta^{1/4})}} \le \sqrt{\frac{\eta}{1-\eta}} + C\epsilon\eta^{1/4},$$

where the second inequality follows from applying a first-order Taylor expansion to the function $\sqrt{x/(1-x)}$ near $x=\eta$. Here, C is a universal constant because the assumptions $\eta \le 1/2$ and $\epsilon \le 1$ ensure that the derivative of $\sqrt{x/(1-x)}$ is universally bounded on the interval of interest. Thus,

$$\begin{aligned} f(1-\epsilon\eta^{1/4})(1-\epsilon\eta^{1/4}-\eta) &\le f(1-\epsilon\eta^{1/4})(1-\eta) \\ &\le \sqrt{\eta(1-\eta)} + C\epsilon\eta^{1/4}(1-\eta). \end{aligned} \tag{A.4}$$

As before, on the interval $\eta \in [0, \frac{1-\theta^*}{2}]$, $\sqrt{\eta(1-\eta)}$ is bounded away from $1/2$, and $\eta^{1/4}(1-\eta)$ is bounded above. Hence, (A.4) is bounded away from $1/2$ as long as $\epsilon(\theta^*)$ is small enough. On the other hand, (A.4) is also smaller than $2\eta^{1/4}$ as long as ϵ is small enough compared to C. □

PROOF OF LEMMA 4.8. Fix some $\epsilon = \epsilon(\theta^*) > 0$ to be determined. If $\theta^2 d$ is sufficiently large compared to ϵ, Proposition 4.2 implies that

$$\Pr(X_{ui,k} \geq 1-\epsilon \mid \sigma_u = +) \geq 1 - \epsilon - \Pr(\sigma_{ui} = - \mid \sigma_u = +) \geq 1 - \epsilon - \eta.$$

Now, if f is any decreasing function then

$$\begin{aligned} \mathbb{E} f(X) \leq & f(1-\epsilon) \Pr(X \geq 1-\epsilon) \\ & + f(0) \Pr(0 \leq X < 1-\epsilon) \\ & + f(-1) \Pr(X < 0). \end{aligned} \tag{A.5}$$

We will apply this with $f(x) = \sqrt{\frac{1-\theta x}{1+\theta x}}$; note that $f(0) = 1$ and $f(-1) = \sqrt{(1-\eta)/\eta}$, where $\eta = \frac{1-\theta}{2}$.

Now, we consider two regimes. If $\sqrt{\eta} \geq \theta^*/10$, we bound

$$\begin{aligned} \mathbb{E}(f(X_{ui,k}) \mid \sigma_u = +) \leq & \Pr(X_{ui,k} \geq 1-\epsilon \mid \sigma_u = +) f(1-\epsilon) \\ & + \Pr(X_{ui,k} < 1-\epsilon \mid \sigma_u = +) f(-1) \\ \leq & (1-\epsilon-\eta) f(1-\epsilon) + \frac{\epsilon+\eta}{\sqrt{\eta}} \\ \leq & (1-\eta) f(1-\epsilon) + \sqrt{\eta(1-\eta)} + \frac{10\epsilon}{\theta^*}. \end{aligned} \tag{A.6}$$

Now, $f(1-\epsilon) = \frac{\eta}{1-\eta} + O(\epsilon)$, and so

$$\mathbb{E}(f(X_{ui,k}) \mid \sigma_u = +) \leq 2\sqrt{\eta(1-\eta)} + O(\epsilon),$$

where the constants in $O(\epsilon)$ depend on θ^*. Since $2\sqrt{\eta(1-\eta)}$ is bounded away from 1 while η is bounded away from 1/2, it follows that for small enough ϵ (depending on θ^*), $\mathbb{E}(f(X_{ui,k}) \mid \sigma_u = +)$ is bounded away from 1.

On the other hand, if $\sqrt{\eta} \leq \theta^*/10$ then we use (A.5) and the fact (from Lemma 4.9) that $\Pr(X_{ui,k} < 0 \mid \sigma_u = +) \leq 2\eta$ to bound

$$\begin{aligned} \mathbb{E} f(X) &\leq (1-\epsilon) f(1-\epsilon) + \epsilon f(0) + 2\eta f(-1) \\ &\leq f(1-\epsilon) + \epsilon + 2\sqrt{\eta}. \end{aligned}$$

Now, if $\epsilon \leq \frac{1}{2}$ then $f(1-\epsilon) \leq \sqrt{1-\theta^*/2} \leq 1 - \theta^*/4$, so

$$\mathbb{E} f(X) \leq 1 - \theta^*/4 + \epsilon + 2\sqrt{\eta} \leq 1 - \frac{\theta^*}{20} + \epsilon,$$

which is bounded away from 1 if ϵ is small enough. □

REFERENCES

[1] Noga Alon and Nabil Kahale. A spectral technique for coloring random 3-colorable graphs. *SIAM Journal on Computing*, 26(6):1733–1748, 1997.

[2] 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.

[3] 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.

[4] Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995.

[5] R.B. Boppana. Eigenvalues and graph bisection: An average-case analysis. In *28th Annual Symposium on Foundations of Computer Science*, pages 280–285. IEEE, 1987.

[6] 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.

[7] 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.

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

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

[10] 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.

[11] 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.

[12] 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.

[13] P.W. Holland, K.B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109 – 137, 1983.

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

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

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

[17] 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.

[18] 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.

[19] R. Lyons. Probability on trees and networks. 2013.

[20] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

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

[22] 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.

[23] E. Mossel, J. Neeman, and A. Sly. Belief propagation, robust reconstruction, and optimal recovery of block models (extended abstract). *JMLR Workshop and Conference Proceedings (COLT proceedings)*, 35:1–35, 2014. Winner of best paper award at COLT 2014.

[24] E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *Probability Theory and Related Fields*, 2014. (to appear).

[25] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. (submitted to Combinatorica), 2014.

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

[27] 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.

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