Global and Local Information in Clustering Labeled Block Models*

Varun Kanade, Elchanan Mossel, and Tselil Schramm

University of California, Berkeley
vkanade@eecs.berkeley.edu, mossel@stat.berkeley.edu, tschramm@cs.berkeley.edu

Abstract

The stochastic block model is a classical cluster-exhibiting random graph model that has been widely studied in statistics, physics and computer science. In its simplest form, the model is a random graph with two equal-sized clusters, with intra-cluster edge probability p, and inter-cluster edge probability q. We focus on the sparse case, *i. e.*, $p, q = O(1/n)$, which is practically more relevant and also mathematically more challenging. A conjecture of Decelle, Krzakala, Moore and Zdeborová, based on ideas from statistical physics, predicted a specific threshold for clustering. The negative direction of the conjecture was proved by Mossel, Neeman and Sly (2012), and more recently the positive direction was proven independently by Massoulié and Mossel, Neeman, and Sly.

In many real network clustering problems, nodes contain information as well. We study the interplay between node and network information in clustering by studying a *labeled* block model, where in addition to the edge information, the true cluster labels of a small fraction of the nodes are revealed. In the case of two clusters, we show that below the threshold, a small amount of node information does not affect recovery. On the other hand, we show that for any small amount of information efficient local clustering is achievable as long as the number of clusters is sufficiently large (as a function of the amount of revealed information).

1998 ACM Subject Classification G.3 Probability and Statistics

Keywords and phrases stochastic block models, information flow on trees

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2014.779

1 Introduction

The stochastic block model is one of the most popular models for networks with clusters. The model has been extensively studied in statistics [14, 27, 4], computer science (where it is called the planted partition problem) [10, 15, 7, 19] and theoretical statistical physics [8, 29, 9].

The simplest block model has k clusters of equal size, and is generated as follows. Starting with n nodes, each node v is randomly assigned a label σ_v from the set $\{1, \ldots, k\}$. For each pair of nodes, (u, v), if their labels are identical an edge is added between them with probability p, otherwise an edge is added with probability q. Often the case when $p > q$ is considered, and the question of interest is understanding how large $p - q$ must be for correct clusters recovery to be possible. In the recovery problem the input consists of the unlabeled graph and the desired output is a partition of the graph.

* Varun Kanade is supported by a Simons Postdoctoral Fellowship, Elchanan Mossel acknowledges the support of the NSF (grants DMS 1106999 and CCF 1320105) and ONR (DOD ONR grant N000141110140), and Tselil Schramm is supported by a Berkeley Chancellor's Fellowship and the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1106400.

17th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'14) / 18th Int'l Workshop on Randomization and Computation (RANDOM'14).
Editors: Klaus Jansen, José Rolim, Nikhil Devanur, and Cristopher Moore; pp. 779–792

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Real world networks are typically sparse. Thus, an interesting setting in the block model is when p and q are in $O(1/n)$. Here, it is more convenient to parametrize the problem by setting $p = a/n$ and $q = b/n$, where a, b are constants. In the sparse setting, exact recovery is impossible as the resulting graph will have isolated nodes. Moreover, it is easy to see that even nodes with constant degree cannot be classified accurately given all other nodes in the graph. Thus the goal is to find a partition that has non-trivial correlation with the original clusters (up to permutation of cluster labels). This has sometimes been referred to as the *cluster detection* problem (see e.g. [8]); throughout the paper we refer to it as the *cluster recovery* problem (though note that the goal is not to recover every cluster with probability 1).

General results of Coja-Oghlan [6] imply that it is possible to identify a partition that is correlated with the true hidden partition when $(a-b)^2 \geq Ck^4(a+(k-1)b)$. A beautiful physics paper by Decelle *et al.* [8] conjectured that the recovery problem is feasible for the case of two clusters when $(a-b)^2 > 2(a+b)$ and impossible when $(a-b)^2 < 2(a+b)$. The non-reconstructability in the case where $(a-b)^2 < 2(a+b)$ was proved by Mossel, Neeman and Sly [22], and more recently the same authors [24] and Massoulié [18] independently showed that recovery is possible when $(a-b)^2 > 2(a+b)$.

1.1 The Labeled Stochastic Block Model

The aforementioned results along with previous results for denser block models provide a detailed picture of recovery in the stochastic block model. However, the model they consider is idealized and does not capture many aspects of real network problems. One such aspect is that in many realistic settings, node label information is available for some of the nodes. For example, in social networks, the group label of some individuals (nodes) is known. In metabolic networks, the function of some of the nodes may be known. Indeed, there has been much recent work in the machine learning and applied networks communities on combining node and network information (see for example [5, 2, 3]). There are several ways in which node and edge information can be incorporated; in real applications nodes and edges contain rich information which is noisy, but correlated with the node's "true" label and with the "similarity" of pairs of nodes.

In this paper, we study a simple model which incorporates both node and edge information which we call the *labeled* stochastic block model. This model has been considered previously in the physics literature [8, 28, 1]. In addition to having the unlabeled graph as an input, a *small* random fraction of the nodes' labels are also provided as input to the clustering algorithm.

1.2 The Big Effect of a Small Number of Node Labels

It is easy to see that even a vanishing fraction of node labels can play a major role in the cluster recovery problem. For example, consider the denser case where the clusters $C_1, \ldots, C_k$ can be identified accurately [19]. Here, it is impossible to distinguish between a clustering $C_1, \ldots, C_k$ where the nodes in cluster C_i have label i and the same clustering where the nodes in cluster i have label $\pi(i)$ for any permutation π of the labels. However, note that for any $p > 0$, given a p-fraction of the node labels, it is possible to identify the permutation π correctly with high probability. It is natural to ask if the same result holds in the sparse case, and it is not hard to see that a similar statement can be made (see Proposition 14).

The above observation shows that even a small amount of node information can overcome the problems of symmetry in the stochastic block model. Another problem of symmetry

present in the unlabeled model is that there is no *local algorithm* that can identify clusters better than random guessing. Informally, a local algorithm determines the label of a node based solely on an $o(\log n)$ neighborhood of that node, including possibly uniform independent random variables attached to each node of the graph (see A.2 for a formal definition and [17, 13] for examples). The proof that a local algorithm cannot detect better than random guessing in this case is folklore, and we include it (in the full version [16]) for completeness. This limitation in detection may be compared to the problem of finding independent sets, where local algorithms can have non-trivial power (while still being less powerful than global algorithms) [12]. It is therefore natural to ask:

▶ Question 1. Does a vanishing fraction of labeled nodes allow *local algorithms* to detect clusters? If so, when?

An even a more direct question relates to the statistical power of revealing some of the node labels. While it is clear that revealing a large fraction of the node labels allows non-trivial recovery, it is far from clear what the effect is when this fraction is vanishingly small. On the one hand, we might expect by continuity that revealing a vanishing fraction of the node labels will be identical in the limit to revealing no labels. On the other hand, we might imagine how a small fraction of the node labels could be used as seeds for recovery algorithms. We thus ask:

▶ Question 2. Does revealing a vanishing fraction of the node labels change the detectability threshold? Does it change the fraction of correctly labeled nodes?

The latter question was considered in recent work in statistical physics [30, 28, 1].

1.3 Our Results

To set the stage for our contributions, we begin with some observations regarding the utility of local information. More formal versions of these propositions are provided in Appendix A. The proofs of these propositions are straightforward (provided in the full version [16]), but they are useful for establishing context of how information about (a small fraction of) node labels may help. The first is that even a vanishingly small proportion of node labels aids in breaking the symmetry and assigning labels to the cluster assignments.

▶ **Proposition 1** (Informal version). *Given a clustering algorithm which outputs clusters correlated with the true clustering, a small fraction of revealed node labels is sufficient to output a labeling which is correlated with the true labeling.*

In the absence of any node information, it is an easy folklore result that any local algorithm cannot recover clusters. However, we show that in the case of two clusters, when a small fraction of node labels are revealed, a local algorithm is able to recover the clusters optimally. This latter result is a direct corollary of a robust reconstruction result on trees of [23].

▶ **Proposition 2** (Informal version). *In the unlabeled stochastic block model, no local algorithm can find a clustering correlated with the true clustering.*

▶ **Proposition 3** (Informal version). *In an instance of the labeled stochastic block model, when $k = 2$, if $(a-b)^2 > C(a+b)$ for some large constant C, then there is a local algorithm which given a vanishing fraction of labeled nodes, reconstruct the label of all nodes with the same accuracy as the optimal (non-local) algorithm for the unlabeled problem.*

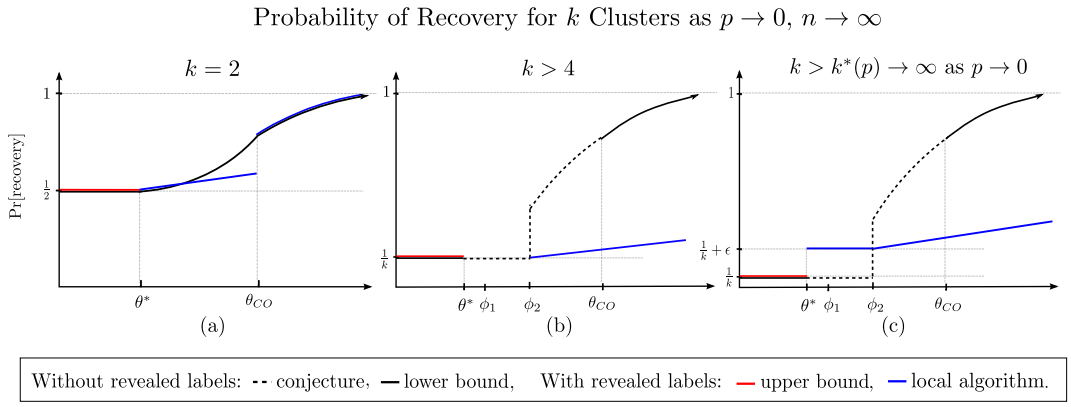

Figure 1 Previous work (black) and our contributions (colored). The x-axes represent the second eigenvalue of the corresponding broadcast process on the coupled Galton-Watson Tree when the average degree is fixed–in simpler terms, this is an increasing function of the ratio $\frac{a-b}{a}$. In all three cases, θ^* is the reconstruction threshold corresponding to the root reconstruction problem on trees, and θ_{CO} is the threshold of [6]. In the two-cluster case (Subfigure (a)), θ^* corresponds exactly to the Kesten-Stigum bound of $(a-b)^2 < 2(a+b)$ [8, 22]. For the case of larger k, $\theta^* < (a-b)^2/k(a+(k-1)b)$ (see Subfigures (b), (c) and Proposition 10). We prove analogously that recovery is not possible below θ^* in the labeled model as $p \to 0$ for all k (Theorem 11). In the two-cluster case, recent results of [24] and [18] show that recovery is possible in the range (θ^*, θ_{CO}); above θ_{CO}, a combination of the results of [6] and [23] give optimal recovery in the standard model for $k = 2$; we observe that in the labeled model for $k = 2$, one can reconstruct better than randomly in the range (θ^*, θ_{CO}) and optimally above θ_{CO} using *local* algorithms (see Propositions 19 and 18). The results of [6] also give non-trivial recovery guarantees above θ_{CO} for all k. In the k-cluster case (Figures (b), (c)), the picture is more complicated: ϕ_1 and ϕ_2 are conjectured brute-force and efficient solvability thresholds respectively, both conjectured by [8]—above ϕ_1 recovery is possible via brute-force enumeration, and above ϕ_2 an efficient algorithm for recovery exists. Above ϕ_2, Proposition 19 shows that recovery is possible for k clusters via a *local* algorithm. In Subfigure (c), for any b, p, if $k > k^*(p)$ and $(a-b)/k > 1$, as in Theorem 8, we give an efficient *local* recovery algorithm that correctly labels $\frac{1}{k} + \epsilon$ of the nodes, even below the conjectured efficient recovery threshold ϕ_2.

We also observe that results on census reconstruction [25] imply that above the Kesten-Stigum bound a vanishingly small fraction of revealed nodes suffices for the cluster recovery problem.

▶ **Proposition 4** (Informal version). *For any fixed k, above the robust reconstruction threshold (i.e. when $(a-b)^2 > k(a+(k-1)b)$), when the fraction of revealed node labels is vanishingly small, the cluster recovery problem is solvable.*

In this context, one might expect that labels could allow clustering in the labeled model in regimes which cannot be effectively clustered in the unlabeled model. The case of two clusters is the case we understand the best. Here, utilizing results for the reconstruction problem on trees and of [22], we answer Question 2 in the *negative* (Theorem 5) and at the same time answer Question 1 positively (Propositions 18 and 19). The complete picture for the case of two clusters is presented in Figure 1(a).

For any fixed $k > 2$, the picture is much more complicated. In this case, we observe that below the tree reconstruction threshold (this corresponds to θ^* in Figure 1(b)), a vanishing fraction of node labels do not assist in the cluster recovery problem (see Theorem 5 in Section 4).

▶ **Theorem 5** (Informal version). *For any fixed k, below the associated tree reconstruction threshold (to be defined later), when the fraction of revealed node labels is vanishingly small, the cluster recovery problem is not solvable. In particular, when $k = 2$, the threshold is the Kesten-Stigum bound of $(a-b)^2 < 2(a+b)$; for $k \geq 2$, if $a - b < k$ then recovery is impossible.*

Our main interest is in the case when the number of clusters is very large. Here, we consider the setting when the fraction of revealed nodes $p \to 0$, and simultaneously the number of clusters $k = k(p) \to \infty$. In this setting, we show that revealing node labels has a dramatic effect on the threshold for cluster recovery. We show that a local algorithm successfully solves the cluster recovery problem even below the conjectured algorithmic threshold in the unlabeled case, $(a-b)^2 = k(a+(k-1)b)$. As the number of clusters $k \to \infty$, our algorithm works all the way down to the tree reconstruction threshold of $(a-b)/k > 1$. Moreover, it is impossible to recover (locally or globally) with a vanishing fraction of labeled nodes if $(a-b)/k < 1$. Both results follow from the corresponding results on trees.

▶ **Theorem 6** (Informal version). *For every $p > 0$, if k is large enough as a function of p, and $a - b > (1+\delta)k$, then the label of a random node can be recovered with probability at least $\frac{1}{k} + \epsilon$, where ϵ depends on δ but is independent of p.*

We give a more formal statement of Theorem 8 in Section 3.

Recent work in statistical physics [30] argues that for every *fixed number of clusters* k, a vanishing fraction of labels does not provide any advantage in the detection probability over having no labels at all. We note that in our results, the order of limits is exchanged as the number of clusters k needed for our results to hold, depends on the fraction of nodes revealed. Thus, there is no contradiction between the results (see also [1, 28]). Figure 1(c) provides a detailed picture of the case in which the number of clusters is very large (in the setting of Theorem 6).

Open Problems

In the case of two clusters, we conjecture that whenever any fraction of node labels are revealed, there is a *local algorithm* that recovers the clusters optimally. This would follow from a related conjecture regarding information flow on trees stated below. We report some simulations suggesting the veracity of the conjecture in Appendix B.

▶ **Conjecture 7** (Informal version). *Let T be an infinite tree with root ρ. The tree is labeled from the set $\{\pm 1\}$ as follows. First, the root is assigned a label from $\{\pm 1\}$ at random. Along each edge the label is propagated with probability $1-\eta$ and flipped with probability η. Let (T, τ) denote the resulting labeled tree. Add each node independently to a set R with probability p. Finally for any r, let ∂T_r denote the set of leaves at depth r. Then, for any value of $p > 0$ and $\eta < 1/2$,*

$$\lim_{r\to\infty} \mathbb{E}\left| \Pr[\tau_\rho = 1 \mid \tau_R] - \Pr[\tau_\rho = 1 \mid \tau_R, \tau_{\partial T_r}] \right| = 0$$

In addition to Conjecture 7, several interesting questions remain, particularly in the regime where k is large. When k is large, is it possible to use global and local information together to obtain better recovery guarantees? Which algorithmic tools might allow one to use global and local information simultaneously?

Another open problem relates to different types noise models. The assumption in the current paper is that each label is revealed accurately with a vanishing probability. But one may consider other types of noise. In particular, we may assume for example that for each

node independently we are given the correct label with small probability δ and otherwise a uniformly chosen label. Is it true that the same results hold for this noise model as for the noise model considered here? For most of the results presented here, it is easy to see that the answer is yes. However, for one of our main results, Theorem 6, the proof *does not* extend to the latter noise model. It is an interesting open problem to determine the effect of the noisy information in this setup.

2 Model

2.1 Stochastic Block Model

The stochastic block model is a generative model for modular random networks, defined by the following set of parameters: the number of clusters k, the expected fraction of nodes in each cluster i, $\langle f_i \rangle_{i=1}^{k}$, and a $k \times k$ symmetric affinity matrix $P_{i,j}$ indicating the edge probability between nodes of type i and j. A random network G on n nodes is generated as follows:

1. First, each node v is assigned a label $\sigma_v \in \{1, \ldots, k\}$, s.t. $\Pr[\sigma_v = i] = f_i$.
2. For every pair of nodes u, v, an edge is added between them with probability P_{σ_u, σ_v}, independently for each pair.

In this work, we are mainly interested in the sparse case, *i. e.*, when the average degree of the graph is constant. We focus on the setting where edge probabilities only depend on whether the labels of the endpoint are same or different. Thus, $P_{ii} = a/n$ for $1 \leq i \leq k$ and $P_{ij} = b/n$ for $i \neq j$, for constants $a > b$.[1] Also, we focus on the case where $f_i = 1/k$ for each i, *i. e.*, each cluster is roughly of the same size. The model is denoted by $\mathcal{G}(n, k, a, b)$, and $(G, \sigma) \sim \mathcal{G}(n, k, a, b)$ denotes an instance of a graph generated according to the model, where σ are the cluster labels of the nodes.

Labeled Block Model: The labeled block model has an additional parameter p, which is the probability with which the true cluster label of any given node is revealed. Thus, if $(G, \sigma) \sim \mathcal{G}(n, k, a, b)$ is an instance of the block model, $R \subseteq [n]$ is chosen by placing each node of G in R independently with probability p. We denote this by $(G, \sigma, R) \sim \mathcal{G}(n, k, a, b, p)$. The clustering algorithm has access to the edges of G and the cluster labels σ_R of nodes in R, *i. e.*, (G, R, σ_R).

We also introduce the following notation for convenience. For any two nodes $u, v \in G$, let $d(u, v)$ denote the distance between u and v. We let $G_r(v) = \{u \in G \mid d(u, v) \leq r\}$ denote the neighborhood of radius r around v; at times we will use G_r when v is clear from context. Let $\partial G_r(v) = \{u \in G \mid d(u, v) = r\}$ denote the boundary of $G_r(v)$.

Cluster Recovery: The *cluster recovery* problem is the problem of recovering the cluster label of nodes in the stochastic block model or labeled stochastic block model with better-than-random probability. Note that correct recovery of all nodes is not the aim, nor is it possible due to the sparsity of the graph. This problem has also been called the *cluster detection* problem and the *cluster reconstruction* problem; for consistency we will use the term recovery throughout the paper when referring to graphs, and use reconstruction when referring to broadcast processes on trees.

[1] This is the so-called assortative model.

▶ **Algorithm 1.**
Input: $(G, R) \sim \mathcal{G}(n, k, a, b, p)$, *radius* r, *max-degree* D, *revealed cluster labels* σ_R

For each node $v \notin R$
1. *Let* $G_r(v)$ *denote the (tree-like) neighborhood of* v *up to distance* r
2. *From* $G_r(v)$ *delete every subtree rooted at a node with degree larger than* D
3. *Let* L *denote the set of labels* $l \in \Sigma$ *for which there exist* $x, y \in R$ *such that* $\sigma_x = \sigma_y = l$, $d(x, v) = d(y, v) = r$, *and* v *is* x *and* y*'s first common ancestor*
4. *Assign a random label from* L *to node* v

2.2 Information Flow on Trees

We use some results regarding information flow on trees. For a detailed survey on this topic, the reader is referred to [21].

Let T be an infinite rooted tree, with the root note denoted by ρ. A Galton-Watson tree is obtained by starting with a root node, ρ, and recursively adding offspring drawn from some distribution D with mean d. In particular, we will often be interested in the case when D is Poisson(d). For any node $v \in T$, let $d(v, \rho)$ denote the distance of v from the root. Throughout the paper, we denote $T_r = \{v \in T \mid d(v, \rho) \leq r\}$ as the subtree of T up to depth r, and $\partial T_r = \{v \in T \mid d(v, \rho) = r\}$ as the boundary at depth r.

Broadcast Process: Let T be an infinite rooted tree with root ρ. Each node in the tree is assigned a label from some finite alphabet $\Sigma = \{1, \ldots, k\}$. The root is labeled by choosing a label $\tau_\rho \in \Sigma$ uniformly at random. For any edge (u, v), with $d(u, \rho) < d(v, \rho)$, τ_v is conditionally independent given τ_u, and is chosen as follows: $\tau_v = \tau_u$ with probability $1-(k-1)\eta$, and $\tau_v \in \Sigma \setminus \{\tau_u\}$ randomly otherwise, where $\eta < 1/k$ is the broadcast parameter. We denote this process by $\mathcal{T}(T, k, \eta)$ and an instance generated according to this process by $(T, \tau) \sim \mathcal{T}(T, k, \eta)$. As in the block model, we can consider the process when the label of each node is revealed with probability p, *i. e.*, $R \subseteq T$ is obtained by adding each $v \in T$ to R independently with probability p. We denote this process by $(T, \tau, R) \sim \mathcal{T}(T, k, \eta, p)$. The *reconstruction problem* is to identify the label of the root, ρ given the labeled nodes up to some depth r. Thus, the algorithm has access to (T_r, R_r, τ_{R_r}), where R_r denotes $T_r \cap R$.

Percolation Process: Let T be an infinite rooted tree with root ρ. For percolation parameter λ, each edge $e \in T$ is deleted independently with probability λ. Let $C(\rho)$ denote the component of T containing the root after percolation.

3 Recovery in the Many Clusters Regime

We show that when the number of clusters is very large, even a very small fraction of revealed node labels allow for cluster recovery, and even in some regimes below the conjectured algorithmic threshold in the standard model. More formally, if p is the probability that the label of a node is revealed, and if the number of clusters is at least $k^* = k(p)$, then even as $p \to 0$, the algorithm performs better than random assignment. The algorithm (Algorithm 1) is simple and *local*—it considers a neighborhood around each node and uses the revealed node information in the neighborhood to make its prediction.

▶ **Theorem 8.** *Let $b > 1$ be fixed, let $a = b + (1+\delta)k$ for some $\delta > 0$, let $p > 0$ be fixed. Then, there exists an $\epsilon = \epsilon(b,\delta)$ and $k^* = k^*(b,\delta,p)$, such that for every $k \geq k^*$, if $(G, R, \sigma_R) \sim \mathcal{G}(n,k,a,b,p)$, Algorithm 1 labels any random node of G correctly with probability at least ϵ. In particular, there exists settings where $(a-b)^2 < k(a+(k-1)b)$ and recovery is still possible.*

We give a proof of Theorem 8 in the full version [16]; here we give a high-level idea of the proof. First, we utilize a coupling between local neighborhoods in $\mathcal{G}(n,k,a,b)$ and a broadcast process on a rooted Galton-Watson tree with offspring distribution Poisson($\frac{a+(k-1)b}{k}$). Fix $v \in [n]$ and let $(G,\sigma) \sim \mathcal{G}(n,k,a,b)$. For large values of n, and when r is not too large (though increasing as a function of n), $G_r(v)$ looks like a tree. The degree distribution of any node in G is Binomial($n, \frac{a+(k-1)b}{kn}$) $\approx$ Poisson($\frac{a+(k-1)b}{k}$). If $\eta = \frac{b}{a+(k-1)b}$, the distribution (G_r, σ_{G_r}) resembles the distribution (T_r, τ_r), where $(T,\tau) \sim \mathcal{T}(T,k,\eta)$ corresponds to the broadcast process on a Galton-Watson tree process T with offspring distribution Poisson($\frac{a+(k-1)b}{k}$). This coupling was formally proved in [22].

▶ **Lemma 9** ([22]). *Let $r < r(n) = \frac{1}{10\log(2(a+(k-1)b))}\log(n)$. There exists a coupling between (G,σ) and (T,τ) such that $(G_r,\sigma_{G_r}) = (T_r,\tau_{T_r})$ a.a.s.*

In [20] it is shown that for larger alphabet sizes, $d(1-k\eta)^2 \geq 1$ is not the threshold for reconstruction for regular trees. As our results show, this is also the case for Galton-Watson trees. In order to understand the intuition behind Algorithm 1, it is useful to consider an *infinite color* broadcast process on a tree. Let $\tilde{\eta} \ll 1$ be a small broadcast parameter. Suppose the root ρ is given some color, which is propagated away from the root as follows. With $(1-\tilde{\eta})$ probability the neighboring node gets the same color, with $\tilde{\eta}$ probability the neighboring node gets a completely new color. The color of each node is revealed with probability p. Consider the following event: there are two nodes in the tree with the same color, for which the root ρ is the first common ancestor. If such an event occurs, this color *must* also be the color of the root. We show that this infinite-color picture is more or less accurate when k is large enough.

4 Upper Bounds Below the Threshold

In this section, we consider the setting where there are a fixed number of clusters and the fraction of revealed node labels is vanishingly small. We show that below a certain threshold that arises from the reconstruction problem on trees, in the limit as $p \to 0$, cluster recovery is not possible. We first note that a threshold exists for the tree problem.

▶ **Proposition 10.** *Let T be a Galton-Watson tree with average degree $d > 1$. Let $(T,\tau) \sim \mathcal{T}(T,k,\eta)$ be the labels obtained by the broadcast process with parameter η. There there exists a predicate, $\pi_k(d,\eta)$, monotonically decreasing in η and monotonically increasing in d, such that if $\pi_k(d,\eta)$ is false, then for each $i \in [k]$,*

$$\lim_{r\to\infty} \Pr\left[\tau_\rho = i \mid \tau_{\partial T_r}\right] \to \frac{1}{k}, \quad a.a.s.$$

For the case of $k = 2$, the exact form of π_2 is known, $\pi_2(d,\eta) = \mathbb{1}[d(1-2\eta)^2 > 1]$, which follows from [11]. In [26], the exact threshold is given for $k = 3$, and bounds on the thresholds are given for $k \geq 5$. For $k \geq 4$, the exact form π_k is not known, but it holds that if $(1-k\eta)d < 1$, $\pi_k(d,\eta)$ is false. (This was proved for the case of regular trees in [20]; the proof for Galton-Watson trees is essentially identical). For all k, a reconstructability

threshold in η, d provably exists in the limit as $n \to \infty$; the proof of Proposition 10 relies on the monotonicity of π_k in η and d, and the existence of points where reconstruction is feasible and also points where it is impossible.

The threshold from Proposition 10 can be translated to an equivalent threshold $\theta_k(a, b)$ in the stochastic block model. We show that even in the labeled stochastic block model (where each node's label is revealed with probability p), if p is small and θ_k is false then it is impossible to recover node labels with better accuracy than random guessing. Specifically, we study the setting where k is fixed, θ_k is false, and $p \to 0$. We first prove this for the general k-cluster case, then give an alternative proof for the case of two clusters (which results in a more explicit dependence on p).

▶ **Theorem 11.** *Fix $v \in [n]$, and let $(G, R, \sigma) \sim \mathcal{G}(n, k, a, b, p)$, for $a + (k-1)b > k$. Then if the predicate $\theta_k(a, b) = \pi_k(\frac{a+(k-1)b}{k}, \frac{b}{a+(k-1)b})$ is not satisfied, then for all $i \in \Sigma = [k]$,*

$$\lim_{p \to 0} \lim_{n \to \infty} \Pr[\sigma_v = i | G, R, \sigma_R] = \frac{1}{k}, \quad a.a.s.$$

The above result says that as the amount of revealed node information goes to zero, recovering a clustering that is correlated with the true clustering is not possible if θ_k is false. The proof of Theorem 11 is given in the full version [16], but we give a high-level overview of the proof here.

We again utilize a coupling between local neighborhoods in $\mathcal{G}(n, k, a, b)$ and a broadcast process on a rooted Galton-Watson tree. As in Section 3, let T be a Galton-Watson tree with offspring distribution Poisson($\frac{a+(k-1)b}{k}$) and broadcast parameter $\eta = \frac{b}{a+(k-1)b}$. We fix $v \in [n]$ and let $(G, \sigma) \sim \mathcal{G}(n, k, a, b)$. The distribution $(G_r(v), \sigma_{G_r(v)})$ resembles the distribution (T_r, τ_r).

We also use a result of [22] which states that conditioned on $\sigma_{\partial G_r}$, information from further nodes is not helpful in clustering.

▶ **Lemma 12** ([22]). *Fix $v \in [n]$, and let $(G, R, \sigma) \sim \mathcal{G}(n, k, a, b, p)$, with $a + (k-1)b > k$. For $r \leq \frac{1}{10 \log(2(a+(k-1)b))} \log n$, let $C = \{u \in G \mid d(u, v) > r\}$, $B = \partial G_r$, and $A = \{u \in G \mid d(u, v) \leq r\}$. Then*

$$\Pr[\sigma_A \mid \sigma_B, \sigma_C, G] = (1 + o(1)) \Pr[\sigma_A \mid \sigma_B, G].$$

In [22], the lemmas above are stated for the case when $k = 2$; however, the same proofs apply for any value of k. Armed with Lemmas 9, 12 and Proposition 10, we can prove Theorem 11 by choosing p small enough that there is no label information in the local neighborhood of any vertex with high probability, then showing that the global graph information is not helpful in recovering the labels.

In the special case of $k = 2$ clusters, it is possible to prove the same result using a slightly different technique. Here, we get a more explicit convergence rate in terms of p. Note that the RHS in the statement of Theorem 13 cannot be smaller than p, since with probability p the node of the label itself is revealed.

▶ **Theorem 13.** *Fix $v \in [n]$, and let $(G, R, \sigma) \sim \mathcal{G}(n, 2, a, b, p)$, for $a + b > 2$. Then if $(a-b)^2 < 2(a+b)$, then*

$$\lim_{n \to \infty} \mathbb{E} \left| \Pr[\sigma_v = 1 \mid G, R, \sigma_R] - \frac{1}{2} \right| \leq \frac{1}{2} \sqrt{\frac{p}{1 - \frac{(a-b)^2}{2(a+b)}}}$$

We give a proof of this better dependence in the full version [16].

Acknowledgments. E. M. thanks Cris Moore, Joe Neeman, Allan Sly and Lenka Zdeborová for many interesting discussions related to the block model. We would like to thank the authors of [30] for discussion of their work at its early stages. The authors would like to thank the Simons Institute for the Theory of Computing where much of the work reported here was carried out.

References

1 Armen E. Allahverdyan, Greg Ver Steeg, and Aram Galstyan. Community detection with and without prior information. *Europhysics Letters*, 90:18002, 2010.

2 Sugato Basu, Arindam Banerjee, and Raymond J Mooney. Semi-supervised clustering by seeding. In *ICML*, volume 2, pages 27–34, 2002.

3 Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 59–68. ACM, 2004.

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

5 Olivier Chapelle, Jason Weston, and Bernhard Schoelkopf. Cluster kernels for semi-supervised learning. In *NIPS*, pages 585–592, 2002.

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

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

8 Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011.

9 Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Lett.*, 107:065701, 2011.

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

11 William Evans, Claire Kenyon, Yuval Peres, and Leonard J. Schulman. Broadcasting on trees and the Ising model. *The Annals of Applied Probability*, 10(2):410–433, 2000.

12 David Gamarnik and Madhu Sudan. Limits of local algorithms over sparse random graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 369–376. ACM, 2014.

13 Hamed Hatami, László Lovász, and Balázs Szegedy. Limits of local-global convergent graph sequences. *arXiv preprint arXiv:1205.4356*, 2012.

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

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

16 Varun Kanade, Elchanan Mossel, and Tselil Schramm. Global and local information in clustering labeled block models. Available at `http://arxiv.org/abs/1404.6325`, 2014.

17 Russell Lyons and Fedor Nazarov. Perfect matchings as iid factors on non-amenable groups. *European Journal of Combinatorics*, 32(7):1115–1125, 2011.

18 Laurent Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the Symposium on the Theory of Computation (STOC)*, 2014.

19 Frank McSherry. Spectral partitioning of random graphs. In *Proceedings of IEEE Conference on the Foundations of Computer Science (FOCS)*, pages 529–537, 2001.

20 Elchanan Mossel. Reconstruction on trees: Beating the second eigenvalue. *The Annals of Applied Probability*, 11(1):285–300, 2001.

21 Elchanan Mossel. Survey: Information flow on trees. Available at `arxiv.org/abs/math/0406446`, 2004.

22 Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. Preprint avaiable at `arxiv.org/abs/1202.1499`, 2012.

23 Elchanan Mossel, Joe Neeman, and Allan Sly. Belief propogation, robust reconstruction, and optimal recovery of block models. Preprint available at `arxiv.org/abs/1309.1380`, 2013.

24 Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. Preprint available at `http://arxiv.org/abs/1311.4115`, 2013.

25 Elchanan Mossel and Yuval Peres. Information flow on trees. *Ann. Appl. Probab.*, 13(3):817–1230, 2003.

26 A. Sly. Reconstruction of symmetric potts models. In *Proceedings of the 41st ACM Symposium on Theory of Computing*, pages 581–590, 2009.

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 Greg Ver Steeg, Cristopher Moore, Aram Galstyan, and Armen E. Allahverdyan. Phase transitions in community detection: A solvable toy model. Available at `http://www.santafe.edu/media/workingpapers/13-12-039.pdf`, 2013.

29 Pan Zhang, Florent Krzakala, Jörg Reichardt, and Lenka Zdeborová. Comparitive study for inference of hidden classes in stochastic block models. *Journal of Statistical Mechanics: Theory and Experiment*, 2012.

30 Pan Zhang, Cristopher Moore, and Lenka Zdeborová. Phase transitions in semisupervised clustering of sparse networks. Available at `http://arxiv.org/abs/1404.7789`, 2014.

A When Little Information Helps

Here, we give formal statements of the simple observations described in Section 1 which illustrate the power and limitations of revealed labels in the stochastic block model.

A.1 Revealed Labels and Cluster Labeling

▶ **Proposition 14.** *Let $C : [n] \to [k]$ be the output of some clustering algorithm with the guarantee that there exists a permutation $\pi : [k] \to [k]$ such that*

$$\frac{1}{n}\sum_i \mathbb{1}[\pi(C(i)) = \sigma_i] \geq \frac{1}{k} + \epsilon,$$

Then for $p \geq \frac{1}{n}\frac{512k}{\epsilon^3}\log\frac{4k}{\delta}$, if a p-fraction of node labels are revealed, we can find a function $g : [k] \to [k]$ such that

$$\frac{1}{n}\sum_i \mathbb{1}[g(C(i)) = \sigma_i] \geq \frac{1}{k} + \frac{\epsilon}{2}$$

with probability at least $1 - \delta$.

The proof follows easily from the following lemma, which is a simple application of the Chernoff-Hoeffding bound.

▶ **Lemma 15.** *Let D be a probability distribution over $[k]$, and let $S \sim D^m$ be a sample. When $m \geq \frac{64}{\epsilon^2}\log(\frac{4k}{\delta})$, for $i = \text{plurality}(S)$ (ties may be broken arbitrarily), with probability at least $1 - (\delta/2)$,*

$$|D_i - \max_j D_j| \leq \frac{\epsilon}{4},$$

where D_j is the probability of j under D.

The complete proofs are given in the full version [16].

A.2 Limitations of Local Algorithms in the Unlabeled Model

Now, we discuss the impact of revealed labels in the context of local algorithms. We use the definition of local algorithms as in [12]. (The reader is referred to their paper and references therein for more background on local algorithms.)

▶ **Definition 16.** Let G be a graph with node set V, and for each $v \in V$, let $X_v \in [0,1]$ uniformly at random. An *r-local algorithm* on G is one in which the value of each node $v \in V$ is decided by a function $f_v(G_r(v), X_r(v))$, where $X_r(v)$ is the set of samples from D associated with $G_r(v)$.

The proposition below formalizes the intuitive statement that no r-local algorithm can accurately reconstruct clusters in the unlabeled stochastic block model for $r = o(\log n)$. The proof is provided in the full version [16].

▶ **Proposition 17.** *In the unlabeled stochastic block model, let A be a local algorithm with node functions $\{f_v\} : G_r(v) \to \Sigma$, where here $G_r(v)$ denotes the structural information and random variables on the neighborhood of radius $r = o(\log n)$ around v. Then for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \Pr_{G,X}[\max_{\pi} \frac{1}{n}\sum_{v} 1(f_v(G_r(v)) = \pi(\sigma_v)) \geq \frac{1}{k} + \epsilon] = 0,$$

where the maximum is taken over all possible permutations of the labels.

A.3 Optimal Local Reconstruction in the Labeled Model when $k = 2$

Before giving a formal statement of Proposition 18, we need to introduce some notation related to broadcast processes on trees. Let $(T, \tau) \sim \mathcal{T}(T, 2, \eta)$, where T is a Galton-Watson tree with offspring distribution Poisson(d). Let

$$\mathsf{T}^*(d,\eta) = \lim_{r\to\infty} \mathbb{E}\left|\Pr[\tau_\rho = 1 \mid \tau_{\partial T_r}] - \frac{1}{2}\right|$$

It follows from the work of Evans *et al.* that $\mathsf{T}^*(d,\eta) > 0$ if and only if $d(1-2\eta)^2 > 1$ [11].

Mossel *et al.* [23] looked at the robust reconstruction problem on trees. Let $(T,\tau) \sim \mathcal{T}(T,2,\eta)$ be as defined above. For some parameter $\delta \in [0, 1/2)$, let $\tilde{\tau}_u$ be the random variable, such that $\tilde{\tau}_u = \tau_u$ with probability $1-\delta$, and $\tilde{\tau}_u = 1 - \tau_u$ with probability δ. In [23], the authors consider the question of reconstruction of the root label given the noisy labels, $\tilde{\tau}_{\partial T_r}$, in the limit as $r \to \infty$. They showed that if

$$\widetilde{\mathsf{T}}^*(d,\eta) = \lim_{r\to\infty} \mathbb{E}\left|\Pr[\tau_\rho = 1 \mid \tilde{\tau}_{\partial T_r}] - \frac{1}{2}\right|,$$

then for any $\delta \in [0, 1/2)$, whenever $d(1-2\eta)^2 \geq C$ for a sufficiently large constant C, $\widetilde{\mathsf{T}}^*(d,\eta) = \mathsf{T}^*(d,\eta)$.

▶ **Proposition 18.** *Let $(G, R, \sigma_R) \sim \mathcal{G}(n, 2, a, b, p)$, with $a + b > 2$. Then, there exists a large constant C, such that if $(a-b)^2 > C(a+b)$, there is a local algorithm A such that if $A(v)$ denotes the label output by the algorithm, for a random node v,*

$$\lim_{p\to 0}\lim_{n\to\infty} \Pr[A(v) = \sigma_v] = \frac{1}{2} + \mathsf{T}^*(\tfrac{a+b}{2}, \tfrac{b}{a+b})$$

A.4 Better-than-random Reconstruction with Local Algorithms in the Labeled Model for Any k

Given an instance of the stochastic block model $(G, \sigma, R) \sim \mathcal{G}(n, k, a, b, p)$ and the corresponding Galton-Watson tree and broadcast process $(T, \tau, R) \sim \mathcal{T}(T, k, \eta, p)$, we now prove that if $d\lambda^2 = (a-b)^2/(k(a+(k-1)b)) > 1$, the plurality of labels at distance ℓ from a node v provides a robust way to recover a v's label for every information p. The argument is based on the reconstruction argument for the label of a root in a broadcast process on trees, and the fact that the application of the second moment method in this argument is robust to noise in the leaf labels. This was implicit in [25] and more explicit in [23]. Interestingly, the proof will show that in the case of Poisson Galton-Watson tree, a simple plurality style rule is sufficient for reconstruction.

▶ **Proposition 19.** *Let $(G, \sigma, R) \sim \mathcal{G}(n, k, a, b, p)$, with $a + (k-1)b > k$. Then, there exists a constant $\epsilon = \epsilon(a, b, k, p)$, such that if $(a-b)^2 > k(a+(k-1)b)$, there is a local algorithm A such that if $A(v)$ denotes the label output by the algorithm, for a random node v,*

$$\Pr[A(v) = \sigma_v] \geq \frac{1}{k} + \epsilon.$$

The result also holds for the noisy-label model.

The proof follows more or less directly from previous results [25], but we also provide it in the full version [16] for completeness.

B Conjecture

B.1 The Uselessness of Global Information

In the case of two clusters, we conjecture that whenever any node label information is present, a local algorithm is already able to recover the clusters optimally. The algorithm is the following: Fix some radius r, for each $v \in G$, look at the neighborhood $G_r(v)$, let $R_r \subseteq G_r(v)$ denote the revealed nodes in the neighborhood. As long as $r \leq c\log(n)$ for a sufficiently small constant c, the neighborhood is a tree with high probability. Then $\Pr[\sigma_v = 1 \mid R_r, \sigma_{R_r}]$ can be computed exactly by belief propagation. We conjecture that this is optimal. This would follow from a related conjecture regarding the broadcast process on trees and an application of Lemma 9.

▶ **Conjecture 20.** *Let T be infinite tree with root ρ. Let $(T, \tau, R) \sim \mathcal{T}(T, 2, \eta, p)$ (see Section 2). Then for any $p > 0$ and $\eta < 1/2$,*

$$\lim_{r\to\infty} \mathbb{E}\big|\Pr[\tau_{\rho=1} \mid \tau_R] - \Pr[\tau_{\rho=1} \mid \tau_R, \tau_{\partial T_r}]\big| = 0.$$

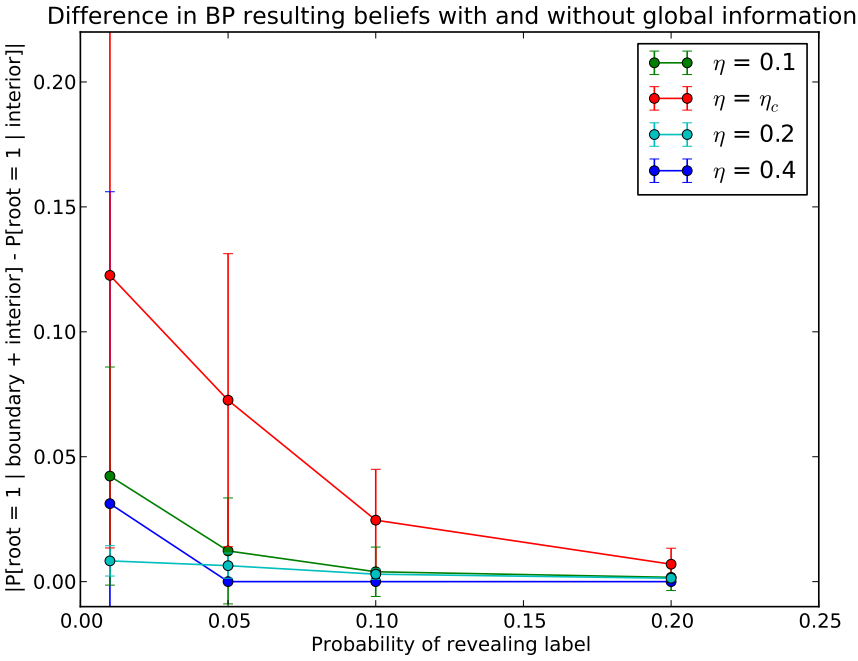

Figure 2 The average distance $|p_{R,L} - p_R|$ is shown for $\eta = 0.1, \eta_c, 0.3, 0.4$ and $p = 0.01, 0.05, 0.1, 0.2$.

B.2 Simulation

To test this conjecture, we ran the Belief Propagation algorithm on 3-regular trees of depth 10, in which labels were assigned to nodes according to broadcast processes starting at the root. Let L denote the set of leaves at level 10. Each node in the interior was revealed independently with probability p, to get the set R. We considered $p \in \{0.01, 0.05, 0.10, 0.20\}$. We also tried various settings of the broadcast parameter, η. We chose $\eta \in \{0.1, \eta_c, 0.3, 0.4\}$, where $\eta_c = \frac{1}{2}\left(1 - \frac{1}{\sqrt{3}}\right)$ is the threshold value for the setting considered.

The labeling process was always initiated with the root having label 1. Thus, we were interested in the posterior probability of the root being labeled 1 in various cases. We computed this posterior probability in three cases: (i) using only the labels at the leaves, denoted by p_L (ii) using only the interior nodes, denoted p_R, and (iii) using both the leaves and the interior nodes, denoted by $p_{L,R}$.

In the first case, only global information is used—*i. e.*, the set of labels at the boundary is the maximum possible information that can be inferred using the global properties of the graph. Thus, in some sense this is an upper bound on the utility of global information. In the second case, only local information in the form revealed nodes in the neighborhood is used. Finally, in the the third case, both local and global information is used.

Our conjecture suggests that as $r \to \infty$, $|p_{R,L} - p_R| \to 0$. Figure 2 shows our results. Each plot corresponds to a fixed value of η, and displays the average distance $|p_{R,L} - p_R|$ for different values of p. We ran the simulation multiple times for each setting of p and η and the standard deviation is marked on the plot.