Statistica Sinica **27** (2017), 1639-1659
doi:https://doi.org/10.5705/ss.202015.0279

GENERIC SAMPLE SPLITTING FOR REFINED COMMUNITY RECOVERY IN DEGREE CORRECTED STOCHASTIC BLOCK MODELS

Jing Lei and Lingxue Zhu

Carnegie Mellon University

Abstract: We study the problem of community recovery in stochastic block models and degree corrected block models. We show that a simple sample splitting trick can refine almost any approximately correct community recovery method to achieve exactly correct community recovery when the expected node degrees are of order $\log n$ or higher. Our results simplify and extend some of the previous work on exact community recovery using sample splitting, and provide better theoretical guarantees for degree corrected stochastic block models.

Key words and phrases: Block models, clustering, community detection, network data, sample splitting.

1. Introduction

Stochastic block models (Holland, Laskey and Leinhardt (1983)) are a popular tool in modeling the co-occurrence of pairwise interactions between individuals in a population of interest. In recent years, the stochastic block model and its variants, such as the degree corrected block model (Karrer and Newman (2011)), have been the focus of much research effort in statistics and machine learning, with wide applications in social networks (Faust and Wasserman (1992)), biological networks and information networks (see, e.g., Kemp et al. (2006); Bickel and Chen (2009)).

Consider a network data set on n nodes recorded in the form of an n by n adjacency matrix A, where an entry $A_{ij} = 1$ if an interaction is observed between nodes i and j, and $A_{ij} = 0$ otherwise. The stochastic block model assumes that the nodes are partitioned into K disjoint communities, and that given the community partition, A_{ij} is an independent Bernoulli random variable whose parameter only depends on the community membership of i and j. Such a model naturally captures the community structure commonly observed in network data. It is also possible to incorporate node specific connectivity parameters, and the

resulting degree corrected block model allows the network data to have arbitrary degree distribution.

A key inference problem in the stochastic block model and its variants is to recover the hidden communities from an observed adjacency matrix. Various methods have been developed in the last decade, including modularity based methods (Newman and Girvan (2004)), likelihood methods (Daudin, Picard and Robin (2008); Bickel and Chen (2009); Zhao, Levina and Zhu (2012); Celisse, Daudin and Pierre (2012)), convex optimization (Chen, Sanghavi and Xu (2012); Le, Levina and Vershynin (2014); Abbe, Bandeira and Hall (2014)), spectral methods (McSherry (2001); Coja-Oghlan (2010); Rohe, Chatterjee and Yu (2011); Jin (2012); Chaudhuri, Chung and Tsiatas (2012); Fishkind et al. (2013); Massoulie (2013); Lei and Rinaldo (2015); Vu (2014)), and others (Decelle et al. (2011); Mossel, Neeman and Sly (2013b); Anandkumar et al. (2014)). These methods are proved to be successful under different assumptions with different types of performance guarantee. A related problem, in a setting where the membership is assumed to be generated at random, is to estimate the membership probability and the community-wise edge probability (see Bickel et al. (2013) for example). In this paper we focus on the community recovery problem.

There are three levels of accuracy for community recovery methods usually considered. The first is *proportional recovery*, which means that an algorithm can correctly recover the community memberships for a subset of nodes whose proportion is bounded away from one but better than random guessing. Such a proportional recovery is usually of interest in very sparse networks, where the node degrees are of constant order and do not grow as the number of nodes increases. In this difficult regime, theory and efficient algorithms are available only for simple special cases and have been studied in Coja-Oghlan (2010); Decelle et al. (2011); Mossel, Neeman and Sly (2012, 2013a,b); Massoulie (2013); Krzakala et al. (2013). The second level is *approximate recovery*, where the proportion of correctly clustered nodes tends to one as the number of nodes grows. Approximate recovery can only be achieved when the expected node degrees diverge to infinity as the number of nodes grows. In this regime, more general models and more practical algorithms have been studied, such as Rohe, Chatterjee and Yu (2011); Jin (2012); Lei and Rinaldo (2015); Amini et al. (2012).

We focus on the third level of recovery accuracy, *exact recovery*. With exact recovery, an algorithm can correctly recover the community memberships for *all* nodes with high probability. Understanding the problem of exact recovery provides useful insight to both the model and the algorithms. There has been

much research effort on exact community recovery for stochastic block models in recent years. Existing methods and results vary in terms of model generality and theoretical sharpness, as we now summarize.

Many authors in probability, machine learning, and theoretical computer science focus on the special case of stochastic block model where there are two equal-sized communities, with edge probability p within community and q between communities. Under this model, sharp threshold for exact recovery can be established, using convex optimization (Abbe, Bandeira and Hall (2014); Hajek, Wu and Xu (2014); Bandeira (2015)), and spectral methods (Mossel, Neeman and Sly (2014)). The method studied in Mossel, Neeman and Sly (2014) uses a sample splitting technique, called the "replica trick", which is similar to the sample splitting considered here. Roughly speaking, these results imply that exact recovery is possible only when the average node degree is of order $\log n$ or higher. Yun and Proutiere (2014) extended the spectral method to the case of more than two communities, with edge probabilities p within community and q between communities.

Exact recovery for general stochastic block models was first studied by McSherry (2001), where a combinatorial projection spectral method was combined with sample splitting. Recently Vu (2014) modified and improved this method using singular value decomposition combined with multiple sample splittings. Bickel and Chen (2009) proved exact recovery for a profile likelihood estimator, for models with node degrees growing faster than $\log n$. It is computationally demanding to maximize the profile likelihood, and commonly used heuristic algorithms are not guaranteed to converge to the global maximum. Abbe and Sandon (2015) provided sharp thresholds for exact community recovery in general stochastic block models, where the recovery is achieved using a spectral method.

There are fewer exact recovery results for degree corrected block models. Chaudhuri, Chung and Tsiatas (2012) extended the method of McSherry (2001) to a special case of degree corrected block models under stronger assumptions on the growth rate of node degrees. Zhao, Levina and Zhu (2012) extended the method of Bickel and Chen (2009) to general degree corrected block models, requiring the node degree to grow faster than $\log n$.

A common idea used in many of the aforementioned works is sample splitting (McSherry (2001); Chaudhuri, Chung and Tsiatas (2012); Vu (2014); Mossel, Neeman and Sly (2014)). In this paper we further study the sample splitting approach for exact community recovery in stochastic block models and degree

corrected block models. We provide further insights to exact recovery methods using sample splitting. Given a preliminary algorithm that achieves approximate recovery, we prove that sample splitting can refine the result to exactly recover the communities with high probability when the expected node degrees are at least $C \log n$ for some constant C. Compared with previous methods using sample splitting, our method is more general and can be combined with many natural and simpler initial community recovery algorithms. In particular, we show that exact recovery can be achieved by combining sample splitting with the simple spectral clustering method, which applies k-means to the top eigenvectors of the adjacency matrix. This insight extends beyond the stochastic block model. We give the first exact recovery method for general degree corrected block models with expected node degrees of order $\log n$.

2. Background

In a stochastic block model, the nodes of a network are partitioned into K disjoint communities. Let $g_i \in \{1, \dots, K\}$ be the community label of node i. The observed data is an $n \times n$ symmetric binary random matrix A with independent upper-diagonal entries A_{ij} $(1 \le i < j \le n)$: $P(A_{ij} = 1) = 1 - P(A_{ij} = 0) = B_{g_i g_j}$, where $B \in [0,1]^{K \times K}$ is a symmetric matrix representing the community-wise edge probabilities. For convenience we assume $A_{ii} = 0$ for all i. Throughout this paper we assume that K, the number of communities, is known.

The community recovery problem concerns estimating the membership vector $g = (g_1, \dots, g_n)$, up to a label permutation. It is well-known that the difficulty of community recovery depends on (i) the sample size n, (ii) the number of communities K, (iii) the differences between the rows of B, and (iv) the magnitude of the entries in B, which controls the overall density of edges in the observed network. In most theoretical studies of community recovery, it is common to consider the large sample behavior as n grows to infinity, while other model parameters, such as K and B, change as functions of n. For simplicity, we focus on the network edge density and fix other parameters as constants. Our analysis can be used to investigate dependence on other model parameters, as discussed in Section 6. To this end, we assume that

$$B = \alpha_n B_0 \,, \tag{2.1}$$

where B_0 is a $K \times K$ non-negative symmetric constant matrix with maximum entry 1. In Abbe, Bandeira and Hall (2014) it is shown that, for a special class of stochastic block models where $K = 2$ and $B_0(1,1) = B_0(2,2) = 1$, exact

Algorithm 1: Cross Clustering (CrossClust)

Input: adjacency matrix A; subset of nodes $\mathcal{V}_1$; subset of nodes $\mathcal{V}_2$; membership vector $\widehat{g}^{(1)}$ on $\mathcal{V}_1$.

Require subroutine: distance based clustering algorithm $\mathcal{D}$.

1. For each $v \in \mathcal{V}_2$, let $\widehat{h}_v = (\widehat{h}_{v,1}, \ldots, \widehat{h}_{v,K})$ with
$$\widehat{h}_{v,k} = \frac{\sum_{v' \in \mathcal{V}_1, \widehat{g}^{(1)}_{v'} = k} A_{v,v'}}{\#\{v' : v' \in \mathcal{V}_1, \widehat{g}^{(1)}_{v'} = k\}}.$$
2. Output $\widehat{g}^{(2)} = \mathcal{D}(\{\widehat{h}_v : v \in \mathcal{V}_2\}, K)$.

recovery is possible if and only if $\alpha_n > C \log n/n$ for a constant C depending on the off-diagonal entry $B_0(1,2)$. For general stochastic block models, McSherry (2001) studied a spectral method with exact recovery when $\alpha_n \geq (\log n)^{3.5}/n$. Vu (2014) improved this result to $\alpha_n \geq C \log n/n$ for a sufficiently large C. Other methods, such as convex optimization (Chen, Sanghavi and Xu (2012)) and profile likelihood (Bickel and Chen (2009)), require stronger conditions on α_n.

We argue that a single sample splitting can lead to exact recovery when combined with a wide range of community recovery methods with approximate recovery. The key idea is that once we have a roughly correct community membership for a subset of the nodes, it can be used to produce exact recovery for the remaining nodes. This is described in detail in Algorithm 1 (CrossClust).

The intuition behind Algorithm 1 is natural. Suppose we have approximately correct memberships for nodes in $\mathcal{V}_1$. Then we can use this membership to estimate the community-wise edge probability for each node in $\mathcal{V}_2$. For $v \in \mathcal{V}_2$ we must have $\widehat{h}_{v,k} \approx B_{g_v,k}, \quad k = 1, \ldots, K$. Therefore, if two nodes v and v' are in the same community, their corresponding $\widehat{h}$ vectors are close to the same row of B. This gives us a good embedding of these nodes in a K dimensional Euclidean space with a nearly perfect clustering structure. If such an embedding is good enough, for example, the distance between any two cluster centers is at least four times larger than the distance between any point to its center; then, as pointed out in Vu (2014), some classical distance-based clustering algorithms such as minimum spanning tree can perfectly recover the communities.

Our main algorithm for stochastic block models based on sample splitting is described in a more general form in Algorithm 2 (V-Clust), where we further extend the sample splitting method to a full V-fold cross clustering method. In V-fold cross clustering, the nodes are divided into V disjoint subsets (folds), and for the jth fold, we apply the preliminary community recovery algorithm

Algorithm 2: V-fold Community Recovery (V-Clust)

Input: adjacency matrix A; number of communities K; number of folds V

Require subroutine: CrossClust; Merge; initial community recovery algorithm $\mathcal{S}$.

1. Randomly split the nodes into V equal sized subsets, $\mathcal{V}^{(1)}, \cdots, \mathcal{V}^{(V)}$.
2. For $j = 1, \cdots, V$:
   (a) $\widehat{g}^{(-j)} = \mathcal{S}(A^{(-j)}, K)$, where $A^{(-j)}$ is the induced adjacency matrix over $\mathcal{V}^{(-j)} = \cup_{j' \neq j} \mathcal{V}^{(j')}$.
   (b) $\widehat{g}^{(j)} = \mathsf{CrossClust}(A, \mathcal{V}^{(-j)}, \mathcal{V}^{(j)}, \widehat{g}^{(-j)})$.
3. Output $\widehat{g} = \mathsf{Merge}(A, (\mathcal{V}^{(j)} : 1 \le j \le V), (\widehat{g}^{(j)} : 1 \le j \le V))$.

Algorithm 3: Merge V-fold Labels (Merge)

Input: adjacency matrix A; the nodes and estimated community membership labels in V folds $\{\mathcal{V}^{(j)}, \widehat{g}^{(j)}\}_{j=1,\ldots,V}$.

1. For each $j = 1, \cdots, V$, calculate $\widehat{B}^{(j)} = (\widehat{B}^{(j)}_{kl})_{k,l=1}^{K}$ by
$$\widehat{B}^{(j)}(k, \cdot) = \frac{\sum_{e \in \mathcal{V}^{(j)},\ \widehat{g}_e^{(j)} = k} \widehat{h}_e}{\#\{e : e \in \mathcal{V}^{(j)},\ \widehat{g}_e^{(j)} = k\}},$$
where $\widehat{h}_e = (\widehat{h}_{e,1}, \cdots, \widehat{h}_{e,K})$ is calculated by
$$\widehat{h}_{e,l} = \frac{\sum_{e' \in \mathcal{V}^{(1)},\ \widehat{g}_{e'}^{(1)} = l,\ e \neq e} A_{e,e'}}{\#\{e' : e' \in \mathcal{V}^{(1)},\ \widehat{g}_{e'}^{(1)} = l,\ e \neq e'\}}.$$
2. For each $j = 2, \cdots, V$:
$$\widehat{\sigma}_j = \arg\min_{\sigma} \sum_{1 \le k \le K} \left\| \widehat{B}^{(j)}(\sigma(k), \cdot) - \widehat{B}^{(1)}(k, \cdot) \right\|^2,$$
where the minimum is taken over all permutations over $\{1, \cdots, K\}$.
3. Output $\widehat{g} = (\widehat{g}^{(1)}, \widehat{\sigma}_2(\widehat{g}^{(2)}), \cdots, \widehat{\sigma}_V(\widehat{g}^{(V)}))$.

on the remaining $V - 1$ folds, and use Algorithm 1 (CrossClust) to obtain exact community recovery on the jth fold. Finally, the exact community recovery on all V folds are combined to obtain a single community recovery for the whole set of nodes. The detailed algorithms are described in Algorithm 2 (V-Clust) and Algorithm 3 (Merge) for stochastic block models. The special case of $V = 2$ corresponds to a half-half split. In practice, V-fold cross clustering with more than two folds may lead to better performance due to a larger sample size used in the preliminary step.

The initial community recovery algorithm $\mathcal{S}$ can be chosen by the user. As shown in Section 3, it only needs to satisfy some mild accuracy requirements that can be achieved by such popular and practical methods as spectral methods and

likelihood-based methods.

The computational cost of our method depends on the initial clustering method chosen by the user. Here we assume that simple spectral clustering is used. For the special case of $V = 2$, the cost of our method is less than that of McSherry (2001), because our method applies clustering to two $n \times K$ matrices, while the method of McSherry (2001) needs to cluster the columns of an $n \times n$ matrix. The cost of our method is also lower than that of Vu (2014), because both methods use spectral clustering but the latter may require more than one split.

Remark 1. We consider the particular form of Algorithm 3 (Merge) for its generality in our theoretical development. In practice, after Step 2(b) in Algorithm 2 (V-Clust), the estimated community in each fold is usually more accurate than the initial estimate. One can actually use many other simpler heuristic methods instead of Algorithm 3. If we know that the within-community edge probability is higher than between-community edge probability, then we can merge the subclusters in all folds by maximizing the total number of within-community edges among all label permutations.

3. Main Results for Stochastic Block Models

Terminology and notation Here the term "with high probability" means "with probability at least $1 - O(n^{-1})$". The rate $O(n^{-1})$ is chosen for convenience and can be changed to $O(n^{-r})$ for any fixed $r > 0$. For two community membership vectors g and $\widehat{g}$, we say $\widehat{g}$ makes m recovery errors as an estimate of g, where m is the smallest integer such that there exists a label permutation on $\widehat{g}$ under which $\widehat{g}$ and g disagree at exactly m entries. We write $\widehat{g} = g$ if $\widehat{g}$ makes zero error. For a membership vector g on $\{1, \ldots, n\}$, and a subset $\mathcal{V} \subseteq \{1, \ldots, n\}$, $g^{(\mathcal{V})}$ denotes the membership vector obtained by confining g on $\mathcal{V}$. We use $g^{(j)}$ in place of $g^{(\mathcal{V}_j)}$ ($j = 1, 2$, in the notation of Algorithm 1) for simplicity. We use $\|\cdot\|$ to denote the ℓ_2 norm of vectors in Euclidean spaces. For any matrix A, we refer to its (i, j)-th element as $A(i, j)$, and its i-th row as $A(i, \cdot)$. Sometimes $A_{i,j}$ is used in place of $A(i, j)$ for brevity.

Our analysis does keep track of the network sparsity and the number of communities. To facilitate the presentation, we assume that the smallest community size is proportional to n/K. The proofs of all theoretical results in this paper are provided in the Supplementary Material (Lei and Zhu, 2016).

Definition 1 (Proper membership)**.** *Given a subset $\mathcal{V} \subseteq \{1, \ldots, n\}$, a membership vector g on $\mathcal{V}$, and a positive constant $\pi_0 \in (0, 1]$, we say $g^{(\mathcal{V})}$ is* π_0-proper *if* $\min_{1 \le k \le K} |\{i \in \mathcal{V} : g_i = k\}| \ge \pi_0 n/K$.

We make the following assumptions.

(A1) The maximum entry of B_0 is bounded by 1, and the minimum l_2 difference between two rows of B_0 is at least $\gamma = \gamma(K) > 0$:

(A2) The true community membership g is π_0-proper for some constant $\pi_0 \in (0, 1]$.

(A3) The initial community recovery algorithm $\mathcal{S}$, with high probability, has recovery error at most $n/f(n\alpha_n, K)$ when $\alpha_n \ge \log n/n$, where f may depend on π_0 and B_0.

Assumption A1 puts a lower bound on pairwise difference between the rows of B_0, which is a minimum requirement for the communities to be distinguishable. The largest possible value of $\gamma(K)$ is $\sqrt{K}$. Assumption A2 puts a lower bound on the minimum community size. These are mainly for simplicity, so that we can focus on the dependence on the network sparsity. Our argument does allow for some mild generalizations so that the minimum community size can change with n in a non-trivial manner, as discussed in Section 6. Assumption A3 puts a requirement on the accuracy of the initial community recovery algorithm. Thus, with high probability the initial algorithm $\mathcal{S}$ correctly recovers the membership of all but a vanishing proportion of nodes, as the expected node degrees grow at $\Omega(\log n)$ rate or faster. The function f specifies how fast the proportion of mis-clustered nodes decays as the average degree increases. This assumption can be satisfied by some simple and practical methods. For example, the spectral clustering method, which applies k-means to the rows of leading eigenvectors of the adjacency matrix, satisfies Assumption A3 with $f(n\alpha_n, K) = c(\pi_0)\lambda^2_{\min}(B_0)n\alpha_n/K^2$ for some function $c > 0$ independent of n (Lei and Rinaldo, 2015).

In the following analysis, we consider Algorithms 1 and 2 with subroutine $\mathcal{D}$ being minimum spanning tree clustering, which constructs a minimum spanning tree on the input data and removes the $K - 1$ largest edges.

Theorem 1 (Exact recovery using sample splitting)**.** *Consider a membership vector g and connectivity matrix $B = \alpha_n B_0$ satisfying Assumptions A1, A2. Let $\widehat{g}$ be the output of* V-Clust *(Algorithm 2) with input matrix A generated by the corresponding stochastic block model and subroutine $\mathcal{S}$ satisfying Assumption A3. There exists a constant C, depending only on π_0 and V, such that if*

$f(n\alpha_n/2, K)\gamma(K) \geq CK^{5/2}$, $\alpha_n \geq CK^3 \log n/(\gamma^2(K)n)$, and $Cn \geq K^3$, then with high probability we have $\widehat{g} = g$.

Theorem 1 also imposes some requirements on the rate α_n. In the case of simple spectral clustering, the conditions are satisfied if $\alpha_n \geq CK^3 \max\{K^{3/2}, \log n\}/n$, provided the minimum singular value of $B_{0,K}$ is uniformly bounded away from 0 and $\gamma(K) \asymp 1$. This is the case in the Planted Partition Model (Condon and Karp (2001); McSherry (2001)), one of the most commonly assumed settings of stochastic block models, where the diagonal entries of B_0 are 1 and off-diagonal entries of B_0 are $\theta \in (0, 1)$.

Lemma 1 (Accuracy of CrossClust)**.** *Suppose A is an adjacency matrix generated by a stochastic block model satisfying Assumptions A1 and A2, and let $\mathcal{V}_1$ be a subset with π_0-proper membership vector $g^{(1)}$ and $|\mathcal{V}_1| \geq n/2$. Let $\widehat{g}^{(1)}$ be an estimated membership vector on $\mathcal{V}_1$ independent of the edges between $\mathcal{V}_1$ and $\mathcal{V}_2$, with recovery error at most $|\mathcal{V}_1|/f(|\mathcal{V}_1|\alpha_n, K)$. Then under the assumptions of Theorem 1, with high probability the output of Algorithm 1 satisfies $\widehat{g}^{(2)} = g^{(2)}$.*

In the context of V-fold cross clustering, $\mathcal{V}_1$ and $\mathcal{V}_2$ in Lemma 1 correspond to $\mathcal{V}^{(-j)}$ and $\mathcal{V}^{(j)}$ in Algorithm 2, respectively. Lemma 1 ensures that the subroutine CrossClust produces exact recovery on $\mathcal{V}_2$ with high probability. The probabilistic claim in Lemma 1 is indeed conditional given $\widehat{g}^{(1)}$. Here we do not emphasize the conditional nature of this result as the randomness is from edges between $\mathcal{V}_1$ and $\mathcal{V}_2$, and hence is independent of $\widehat{g}^{(1)}$ by assumption. The proof of Lemma 1, as detailed in Section S1, is based on a careful decomposition of estimation error $|\widehat{h}_{v,k} - B(g_v, k)|$ followed by concentration inequalities.

4. Extension to Degree Corrected Block Models

The degree corrected block model (Karrer and Newman (2011)) extends the stochastic block model by introducing additional node level degree heterogeneity. In addition to the membership vector g and community-wise connectivity matrix B, the degree corrected block model incorporates a parameter $\psi \in (0, 1]^n$ to model the node level activeness. Then the edge A_{ij} between nodes i and j is an independent Bernoulli variable with parameter $\psi_i\psi_j B_{g_i g_j}$. For identifiability, we assume $\max_{i:g_i=k} \psi_i = 1$, for all $1 \leq k \leq K$. The parameter ψ_i reflects the relative activeness of node i in its community. The degree corrected block model is able to model a much wider range of network data and is more realistic than the regular stochastic block model. There are relatively fewer results on exact recovery for degree corrected block models. Zhao, Levina and Zhu (2012) extended the result

Algorithm 1': Spherical Cross Clustering (CrossClustSphere)

Input: adjacency matrix A; subset of nodes $\mathcal{V}_1$; subset of nodes $\mathcal{V}_2$; membership vector $g^{(1)}$ on $\mathcal{V}_1$.

Require subroutine: distance based clustering algorithm $\mathcal{D}$.

1. For each $v \in \mathcal{V}_2$, let $\widehat{h}_v = (\widehat{h}_{v,1}, \ldots, \widehat{h}_{v,K})$ be the same as give in Step 1 of Algorithm 1 (CrossClust).
2. Output $\widehat{g}^{(2)} = \mathcal{D}(\{\widehat{h}_v/\|\widehat{h}_v\| : v \in \mathcal{V}_2\}, K)$.

of Bickel and Chen (2009), showing that the profile likelihood estimator can recover exactly when $\alpha_n = \Omega(\log n/n)$. Chaudhuri, Chung and Tsiatas (2012) extended the method of McSherry (2001) to a special case of degree corrected models with a stronger requirement on the decay rate of α_n. In the following we consider the more general setting where K may grow with n, and establish error probability bounds for a variation of V-fold cross clustering under degree corrected block models. When K is fixed, our result implies that the simple sample splitting method can be successful under general degree corrected block models when $\alpha_n \geq C \log n/n$ for sufficiently large constant C.

Under the degree corrected block model, we need to modify the CrossClust algorithm so that the effect of nuisance parameter ψ is cancelled out by a normalization step. To this end, we introduce the spherical cross clustering algorithm in Algorithm 1'.

The exact recovery property of the sample splitting approach can be established for degree corrected block models under slightly stronger conditions, with a modified community separation condition and an additional condition on the nuisance parameter ψ. Recalling that $B_0(k, \cdot)$ denotes the k-th row of B_0, we assume the following.

(A1') The minimum l_2 difference between two normalized-rows of B_0 is at least $\widetilde{\gamma} = \widetilde{\gamma}(K) > 0$, and the minimum l_2 norm of rows of B_0 is at least $L = L(K) > 0$:

$$\min_{1 \leq k < k' \leq K} \left\| \frac{B_0(k, \cdot)}{\|B_0(k, \cdot)\|} - \frac{B_0(k', \cdot)}{\|B_0(k', \cdot)\|} \right\| := \widetilde{\gamma}(K) > 0\,,$$
$$\min_{1 \leq k \leq K} \|B_0(k, \cdot)\| := L(K) > 0\,.$$

(A3') The initial community recovery algorithm $\mathcal{S}$, with high probability, has recovery error at most $n/f(n\alpha_n, K)$ when $\alpha_n \geq \log n/n$, where f may depend on π_0 and B_0.

(A4) $\min_{1 \leq i \leq n} \psi_i \geq \psi_0$ for some constant $\psi_0 \in (0, 1]$.

Algorithm 3': Spherical Merge V-fold Labels (MergeSphere)

Input: adjacency matrix A; the nodes and estimated community membership labels in V folds $\{\mathcal{V}^{(j)}, \widehat{g}^{(j)}\}_{j=1,\ldots,V}$.

1. For each $j = 1, \cdots, V$, calculate $\widehat{B}_*^{(j)}$ as
$$\widehat{B}_*^{(j)}(k,\cdot) \leftarrow \frac{\widehat{B}^{(j)}(k,\cdot)}{\|\widehat{B}^{(j)}(k,\cdot)\|},$$
where $\widehat{B}^{(j)}$ is calculated as in step 1 of Algorithm 3 (Merge).
2. For each $j = 2, \cdots, V$:
$$\widehat{\sigma}_j = \arg\min_{\sigma} \sum_{1\le k\le K} \left\| \widehat{B}_*^{(j)}(\sigma(k),\cdot) - \widehat{B}_*^{(1)}(k,\cdot) \right\|^2,$$
where σ denotes a permutation over $\{1, \cdots, K\}$.
3. Output $\widehat{g} = (\widehat{g}^{(1)}, \widehat{\sigma}_2(\widehat{g}^{(2)}), \cdots, \widehat{\sigma}_V(\widehat{g}^{(V)}))$.

Assumption A1' modifies Assumption A1 to account for the normalization step in CrossClustSphere, which is necessary for degree corrected block models because two rows in B differing only by a constant scaling are indistinguishable due to the node activeness parameter. Assumption A4 prevents any node from being too inactive, otherwise there are too few edges for that node, making exact recovery unlikely. Under these assumptions, the spherical spectral clustering method described and analyzed in Lei and Rinaldo (2015) satisfies Assumption A3' with $f(n\alpha_n, K) \propto \psi_0 \lambda_{\min}(B_0)\sqrt{n\alpha_n}/K$, provided B_0 has full rank and the communities have balanced sizes.

The row scaling indistinguishability problem also requires a different merging algorithm to combine the communities cross different folds. Here we introduce the spherical merging algorithm MergeSphere (Algorithm 3').

Theorem 2 (Exact recovery for degree corrected block models)**.** *Let A be an adjacency matrix generated from a degree corrected block model with membership vector g, connectivity matrix $B = \alpha_n B_0$, and node activeness vector ψ satisfying Assumptions A1′, A2, and A4. Let $\widehat{g}$ be the output of V-Clust (Algorithm 2) using subroutine CrossClustSphere (Algorithm 1′) and MergeSphere (Algorithm 3′), with initial recovery algorithm $\mathcal{S}$ satisfying Assumption A3′. There exists a constant $C = C(\pi_0, \psi_0, V)$ such that if $f(\alpha_n n/2, K)\widetilde{\gamma}(K)L(K) \ge CK^{5/2}$, $\alpha_n \ge CK^3 \log n/(\widetilde{\gamma}^2(K)^2 L^2(K) n)$, and $Cn \ge K^3$, then $\widehat{g} = g$ with high probability.*

In the case of sphere spectral clustering, the conditions required in Theorem 2 are satisfied if $\alpha_n \ge CK^2 \max\{K^4, \log n\}/n$ in the common situation that

$\lambda_{\min}(B_0) \asymp 1$, $\widetilde{\gamma}(K) \asymp 1$, and $L(K) \asymp \sqrt{K}$.

The proof of Theorem 2 is similar to that of Theorem 1, and uses the following analogous result of Lemma 1. Proofs of these results are given in the Supplementary Material (Lei and Zhu (2016)).

Lemma 2 (Accuracy of CrossClustSphere)**.** *Suppose A is an adjacency matrix generated by a degree corrected block model satisfying Assumptions A1' and A2, and let $\mathcal{V}_1$ be a subset with π_0-proper membership vector $g^{(1)}$. Let $\widehat{g}^{(1)}$ be an estimated membership vector on $\mathcal{V}_1$ independent of the edges between $\mathcal{V}_1$ and $\mathcal{V}_2$, with recovery error at most $|\mathcal{V}_1|/f(|\mathcal{V}_1|\alpha_n, K)$. There exists a constant C such that if $\alpha_n \geq CK^3 \log n/(\widetilde{\gamma}^2(K)L^2(K)n)$, then with high probability, $\widehat{g}^{(2)}$, the output of CrossClustSphere (Algorithm 1'), satisfies $\widehat{g}^{(2)} = g^{(2)}$.*

Remark 2. As seen in our simulations, for regular stochastic block models Algorithm 1 usually outperforms Algorithm 1' when the signal is very weak, because the latter tends to introduce extra error in the redundant normalization step. Moreover, Theorem 2 generally requires a higher average degree than Theorem 1, where the main difference is in their dependence on K, the number of communities.

Remark 3. As in the case for regular stochastic block models when more information is available about the model, there are other practical and simple alternatives to Algorithm 3'. If within-community edge probability is higher than between-community edge probability, the same heuristic merging method given in Remark 1 applies to the degree-corrected model.

5. Numerical Examples

5.1. Simulation 1: a diagonal dominant SBM

We considered a stochastic block model (SBM) with $K = 2$, $|\mathcal{I}_1| = |\mathcal{I}_2| = 500$, and

$$B = \begin{pmatrix} a & b \\ b & a \end{pmatrix}, \quad \text{for some } a, b \in (0,1), a > b. \tag{5.1}$$

Our focus was the community separation measured by $a - b$, and overall sparsity measured by a. In particular we considered combinations of (a, b) over a 40 by 40 grid in $(0, 1)^2$. To reduce redundancy, we only report combinations of (a, b) such that $b \in [a - 0.3, a)$ so that community recovery is less trivial. We considered two implementations of our proposed Algorithm 2 (V-Clust).

Method I uses simple spectral clustering to obtain initial community recovery and CrossClust (Algorithm 1) for refinement. This method is designed for SBM's and not suitable for degree corrected block models (DCSBM).

Method II uses spherical spectral clustering (Lei and Rinaldo, 2015) to obtain initial community recovery and CrossClustSphere (Algorithm 1') for refinement. This method works for both SBM's and DCSBM's.

Both implementations use the heuristic merge method described in Remark 1, since we always have $a > b$.

Figure 1 summarizes the results for the SBM, where we plot average proportion of correctly clustered nodes over 100 repetitions. The top row shows results for Method I, and the bottom row is for Method II where we treat the SBM as a DCBM. For each method we also compare 2-fold cross clustering (left column), 10-fold cross clustering (middle column), and self-cross clustering (right column). Here self-cross clustering is like an n-fold cross clustering, except that the initial clustering is obtained from the entire network. We see a clear phase transition pattern in all plots, where the algorithm achieves exact recovery when $a - b$ is sufficiently large. We also observe that 10-fold and self-cross clustering have slightly better performances. In this simple case, CrossClustSphere (Algorithm 1') gives almost identical performance as CrossClust (Algorithm 1).

5.2. Simulation 2: a more general SBM

Our second experiment was also conducted under an SBM, but without the diagonal dominant structure as in Simulation 1. We considered community-wise edge probability matrices of the form

$$B = \begin{pmatrix} a & b \\ b & b \end{pmatrix} \tag{5.2}$$

with the same grid of (a, b) values as in Simulation 1. The merging was implemented using Algorithm 3 for Method I and Algorithm 3' for Method II, as here the heuristic merging method described in Remark 1 is not applicable.

The results summarized in Figure 2 shows something different than in the previous case. Here both methods are less accurate due to the reduced signal strength. But spherical cross clustering is significantly less accurate than cross clustering, because the normalization step becomes much more noisy when the minimum singular value of B is very small. This confirms our claim that when the true model is SBM, it is preferable to use methods designed for SBM's, rather than those for more general DCSBM's.

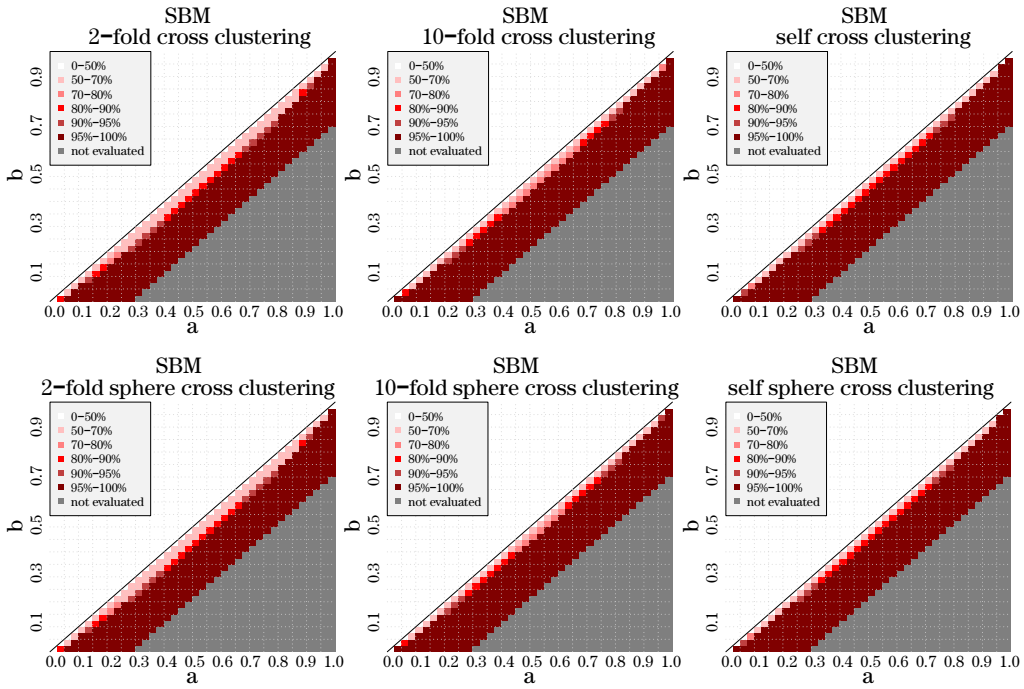

Figure 1. Average accuracy over 100 repetitions under regular stochastic block models with $K = 2$, within-community edge probability a and between-community edge probability b. Each community has size 500. Top row: Cross clustering; bottom row: sphere cross clustering. Left column: 2-fold; middle column: 10-fold; right column: self cross clustering.

5.3. Simulation 3: degree corrected block models

We also examined the performance of both methods in a DCSBM model with node heterogeneity. We assumed the same matrix B as in (5.1), and $P(A_{ij} = 1) = \psi_i \psi_j B_{g_i g_j}$, where the node activeness parameter ψ_i's were independently generated from a uniform distribution on $(0.5, 1)$. We applied Method I and Method II, each with 2-fold, 10-fold, and self cross clustering. As in Simulation 1, heuristic merging was used here. Figure 3 shows that Method I fails in DCSBM, while Method II achieves much better accuracy.

We also checked how the initial recovery accuracy on $\mathcal{V}_1$ affects the cross-clustering accuracy on $\mathcal{V}_2$ in Algorithm 1 (CrossClust). Specifically, in Method I with 2-fold cross clustering under regular SBM, we recorded the accuracy of the initial community recovery using simple spectral clustering on the first half and the accuracy of the cross-clustering on the independent second half. Figure 4 shows the results from 500 trials randomly selected from all repetitions. It is

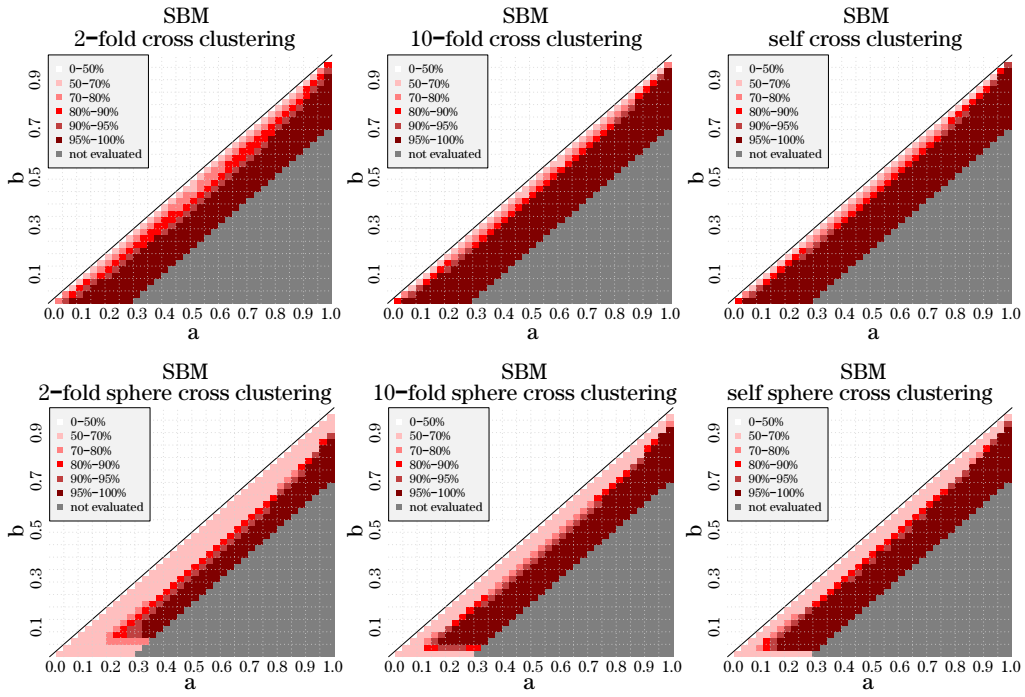

Figure 2. Average accuracy over 100 repetitions under regular stochastic block models with $K = 2$, and community-wise edge probability given in (5.2). Each community has size 500. Top row: Cross clustering; bottom row: sphere cross clustering. Left column: 2-fold; middle column: 10-fold; right column: self cross clustering.

obvious that the cross-clustering step boosts the accuracy in most cases when the output $\widehat{g}^{(1)}$ of the initial algorithm has high accuracy.

5.4. Simulation 4: networks with more than two communities

Here we investigate the performance of cross clustering for larger values of K. For simplicity we focused on regular stochastic block models with similar settings as in Section 5.1, except that K was larger than 2 and n was increased to 5,000, because n needs to grow faster than K as indicated by the theory.

Figure 5 shows that the 2-fold cross clustering works reasonably well when $K = 5, 10$, for stochastic block models with 5,000 nodes. The results for 10-fold cross-clustering are similar and hence are omitted. It is worth noting that when K gets larger, the merge algorithm is computationally demanding as it needs to search over all $K!$ label permutations. Here we use the fact that the edge density is higher within-community than between-community, which leads to a faster greedy merge method. Using the notation in Algorithm 3, let $\mathcal{V}^{(1)}$, $\mathcal{V}^{(2)}$

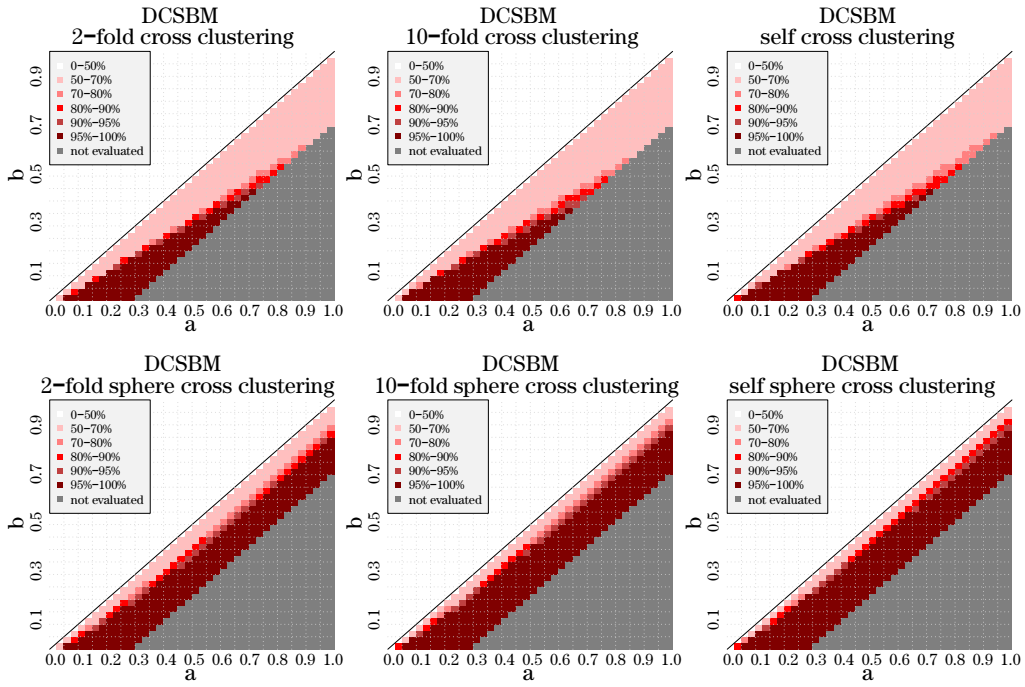

Figure 3. Average accuracy over 100 repetitions under degree corrected stochastic block models with $K = 2$, within-community egde probability a and between-community edge probability b, and node activeness $\psi_i \sim \text{Uniform}(0.5, 1)$. Each community has size 500. Top row: Method I; bottom row: Method II. Left column: 2-fold; middle column: 10-fold; right column: self cross clustering.

be two subsets of nodes with clusters $\widehat{g}^{(1)}$, $\widehat{g}^{(2)}$, respectively. For $k = 1, \ldots, K$, take $\sigma(k) = \arg\max_{1 \le l \le K} \widehat{B}^{(2)}(k, l)$, where $\widehat{B}^{(2)}$ is defined in step 1 of Algorithm 3. Then we permute the labels in $\widehat{g}^{(2)}$ by replacing k with $\sigma(k)$, and directly combine the permuted $(\mathcal{V}^{(2)}, \sigma(\widehat{g}^{(2)}))$ with $(\mathcal{V}^{(1)}, \widehat{g}^{(1)})$.

In Figure 6 we report the results for even larger values of K. We see that the performance decreases as K increases. For $K = 100$, the algorithm can pick up part of the signal for well separated values of a and b.

5.5. Political blog data

As an example, we considered the political blog data (Adamic and Glance, 2005), where the edges represent hyperlinks among 1,222 weblogs on U.S. politics in 2004. Each weblog belongs to one of the two communities recognized as "liberal" and "conservative", with sizes 586 and 636, respectively. It is widely believed that the degree corrected block model with $K = 2$ fits the data well (Zhao, Levina and Zhu (2012); Jin (2012); Yan et al. (2014)). We started with

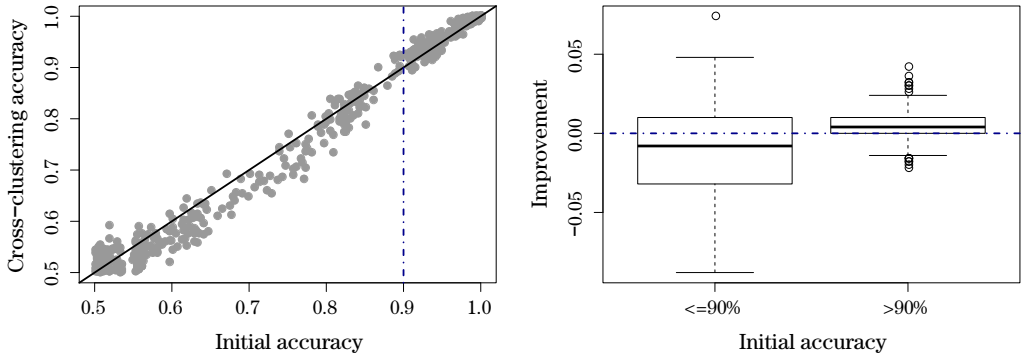

Figure 4. Initial recovery accuracy on $\mathcal{V}_1$ and the cross-clustering accuracy on $\mathcal{V}_2$ in 500 trials. Left: comparison of initial accuracy and cross-clustering accuracy. Right: the improvement, defined as the difference between the refined accuracy and the initial accuracy, when initial accuracy is low ($\leq 90\%$) and high ($> 90\%$).

an initial community recovery given by the spherical k-median spectral clustering (Lei and Rinaldo (2015)), and applied the proposed method in Algorithm 1' (CrossClustSphere) with 2-fold, 10-fold and self cross clustering, followed by heuristic merging as described in Remark 3. For 2-fold and 10-fold cross clustering implementation, we repeated the data splitting 100 times. The average proportion of correctly clustered nodes was 90.68% for 2-fold cross clustering, 94.60% for 10-fold cross clustering, and 95.17% for self-cross clustering. Directly applying spherical spectral clustering on the entire data set yielded an accuracy of 94.76%. This reflects a trade-off between computational efficiency and estimation accuracy in the cross-clustering method. If V is small, then the initial clustering is less accurate due to the reduced sample size. If V is large, then the initial accuracy improves, but the algorithm requires more computation resources for the estimation.

6. Discussion

In this paper we demonstrate that sample splitting can be combined with almost any approximately correct community recovery algorithms to obtain better clustering results for stochastic block models and degree corrected block models. Under general conditions, satisfied by spectral clustering in particular, we show that such a method can achieve exact community recovery with optimal dependence on the rate of network sparsity. Our results unify and simplify such existing works as McSherry (2001) and Vu (2014), and lead to a more general V-fold cross clustering algorithm.

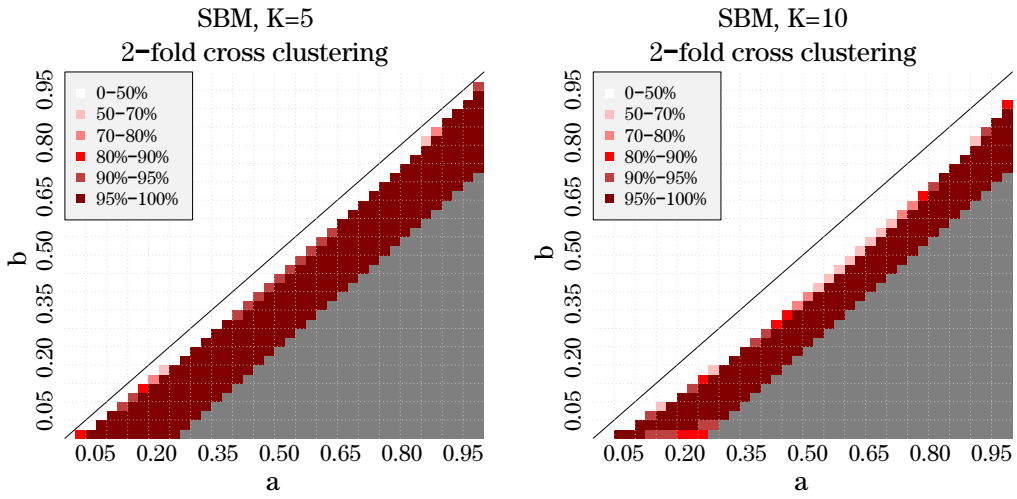

Figure 5. Average recovery accuracy over 100 repetitions for 2-fold cross clustering with $n = 5{,}000$ nodes. Left: $K = 5$; right: $K = 10$.

We investigate exact community recovery with a special focus on the overall network sparsity and number of communities. In the study of stochastic block models, the effect of other model parameters, such as community size imbalance, may also be of interest. The arguments used in here, for example Lemma 6, do keep track of these parameters and hence can be used to study the more general scenario where these parameters are also allowed to change with n, non-trivially. In particular, when $K = 2$ with $\mathcal{I}_1$, $\mathcal{I}_2$ being the two communities, one can show that the sample splitting approach succeeds with high probability when α_n is bounded away from zero and $\min(|\mathcal{I}_1|, |\mathcal{I}_2|) \geq C\sqrt{n}$ for large enough C.

Open problems The only step in our proof that requires sample splitting is the large deviation bound for the term T_1 in the proofs of Lemmas 6 and 10, where the summation of $A_{v,v'}$ is over a random set $\{v' \in \mathcal{V}_1 : \widehat{g}_{v'}^{(1)} \neq g_{v'}^{(1)}\}$. The sample splitting makes the summand $A_{v,v'}$ independent of this index set, allowing us to condition on the index set and apply Bernstein's inequality. As mentioned earlier in Section 2, a natural alternative is to obtain a preliminary community estimate for the entire set of nodes, and then cross-cluster each node using this preliminary community partition. In other words, we use $\mathcal{V}_1 = \mathcal{V}_2 = \{1, 2, \ldots, n\}$ in Algorithm 1. Such a self-cross clustering approach gives very competitive practical performance as demonstrated in our numerical examples. A heuristic explanation of its success is its close similarity to the n-fold (leave-one-out) cross-clustering, which is further refined and analyzed by Gao et al. (2015)

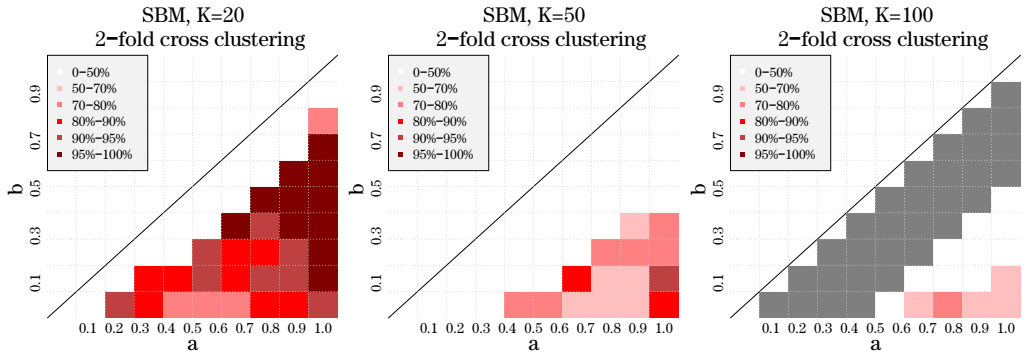

Figure 6. Average recovery accuracy over 10 repetitions for 2-fold cross clustering with $n = 5{,}000$ nodes. Left: $K = 20$; middle: $K = 50$; right: $K = 100$.

after completion of the first draft of the current paper. It would be interesting to provide rigorous performance guarantees for the self-cross clustering method.

Supplementary Materials

The Supplementary Material (Lei and Zhu (2016)) contains proofs of our theoretical results.

Acknowledgment

Jing Lei's research is partially supported by NSF grants DMS-1407771 and DMS-1553884. The authors thank the anonymous reviewers and the associate editor for their helpful comments.

References

Abbe, E., Bandeira, A. S. and Hall, G. (2014). Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*.

Abbe, E. and Sandon, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*.

Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 US election: Divided they blog. in *Proceedings of the 3rd International Workshop on Link Discovery*, ACM, 36–43.

Amini, A. A., Chen, A., Bickel, P. J. and Levina, E. (2012). Pseudo-likelihood methods for community detection in large sparse networks. *arXiv preprint arXiv:1207.2340*.

Anandkumar, A., Ge, R., Hsu, D. and Kakade, S. M. (2014). A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research* **15**, 2239–2312.

Bandeira, A. S. (2015). Random Laplacian matrices and convex relaxations. *arXiv preprint arXiv:1504.03987*.

Bickel, P., Choi, D., Chang, X. and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics* **41**, 1922–1943.

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

Celisse, A., Daudin, J.-J. and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics* **6**.

Chaudhuri, K., Chung, F. and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. *JMLR: Workshop and Conference Proceeding* 35.1–35.23.

Chen, Y., Sanghavi, S. and Xu, H. (2012). Clustering sparse graphs. in *Advances in Neural Information Processing Systems*, Pereira, F., Burges, C., Bottou, L. and Weinberger, K., eds. 2204–2212. Curran Associates, Inc.

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

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

Daudin, J.-J., Picard, F. and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing* **18**, 173–183.

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

Faust, K. and Wasserman, S. (1992). Blockmodels: Interpretation and evaluation. *Social Networks* **14**, 5–61.

Fishkind, D. E., Sussman, D. L., Tang, M., Vogelstein, J. T. and Priebe, C. E. (2013). Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM Journal on Matrix Analysis and Applications* **34**, 23–39.

Gao, C., Ma, Z., Zhang, A. Y. and Zhou, H. H. (2015). Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*.

Hajek, B., Wu, Y. and Xu, J. (2014). Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*.

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

Jin, J. (2012). Fast community detection by SCORE. arXiv:1211.5803.

Karrer, B. and Newman, M. E. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E* **83**, 016107.

Kemp, C., Tenenbaum, J. B., Griffiths, T. L., Yamada, T. and Ueda, N. (2006). Learning systems of concepts with an infinite relational model. in *AAAI*, vol. 3, p. 5.

Krzakala, F., Moore, C., Mossel, E., Neeman, J., Sly, A., Zdeborová, L. and Zhang, P. (2013). Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences* **110**, 20935–20940.

Le, C. M., Levina, E. and Vershynin, R. (2014). Optimization via low-rank approximation, with

applications to community detection in networks. *arXiv preprint arXiv:1406.0067*.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *The Annals of Statistics* **43**, 215–237.

Lei, J. and Zhu, L. (2016), Supplementary material for "Generic sample splitting for refined community recovery in degree corrected stochastic block models".

Massoulie, L. (2013). Community detection thresholds and the weak Ramanujan property. *arXiv preprint arXiv:1311.3085*.

McSherry, F. (2001). Spectral partitioning of random graphs. in *Foundations of Computer Science*, 529–537.

Mossel, E., Neeman, J. and Sly, A. (2012). Tochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*.

Mossel, E., Neeman, J. and Sly, A. (2013a). Belief propagation, robust reconstruction, and optimal recovery of block models. *arXiv preprint arXiv:1309.1380*.

Mossel, E., Neeman, J. and Sly, A. (2013b), A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*.

Mossel, E., Neeman, J. and Sly, A. (2014). Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E* **69**, 026113.

Rohe, K., Chatterjee, S. and Yu, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics* **39**, 1878–1915.

Skala, M. (2013). Hypergeometric tail inequalities: Ending the insanity. *arXiv preprint arXiv:1311.5939*.

Vu, V. (2014). A simple SVD algorithm for finding hidden partitions. *arXiv preprint arXiv:1404.3918*.

Yan, X., Shalizi, C., Jensen, J. E., Krzakala, F., Moore, C., Zdeborová, L., Zhang, P. and Zhu, Y. (2014). Model selection for degree-corrected block models. *Journal of Statistical Mechanics: Theory and Experiment*, P05007.

Yun, S.-Y. and Proutiere, A. (2014). Accurate community detection in the stochastic block model via spectral algorithms. *arXiv preprint arXiv:1412.7335*.

Zhao, Y., Levina, E. and Zhu, J. (2012). Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics* **40**, 2266–2292.

Department of Statistics Carnegie Mellon University Pittsburgh, Pennsylvania 15213, USA

E-mail: jinglei@andrew.cmu.edu

Department of Statistics Carnegie Mellon University Pittsburgh, Pennsylvania 15213, USA

E-mail: lzhu@cmu.edu

(Received August 2015; accepted August 2016)