A class of network models recoverable by spectral clustering

Yali Wan
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA
yaliwan@washington.edu

Marina Meilă
Department of Statistics
University of Washington
Seattle, WA 98195-4322, USA
mmp@stat.washington.edu

Abstract

Finding communities in networks is a problem that remains difficult, in spite of the amount of attention it has recently received. The Stochastic Block-Model (SBM) is a generative model for graphs with "communities" for which, because of its simplicity, the theoretical understanding has advanced fast in recent years. In particular, there have been various results showing that simple versions of spectral clustering using the Normalized Laplacian of the graph can recover the communities almost perfectly with high probability. Here we show that essentially the same algorithm used for the SBM and for its extension called Degree-Corrected SBM, works on a wider class of Block-Models, which we call Preference Frame Models, with essentially the same guarantees. Moreover, the parametrization we introduce clearly exhibits the free parameters needed to specify this class of models, and results in bounds that expose with more clarity the parameters that control the recovery error in this model class.

1 Introduction

There have been many recent advances in the recovery of communities in networks, under "block-model" assumptions [19, 18, 9]. In particular, advances in recovering communities by spectral clustering algorithms. These have been extended to models including node-specific propensities. In this paper, we argue that one can further expand the model class for which recovery by spectral clustering is possible, and describe a model that subsumes a number of existing models, which we call the PFM. We show that under the PFM model, the communities can be recovered with small error, w.h.p. Our results correspond to what [6] termed the "weak recovery" regime, in which w.h.p. the fraction of nodes that are mislabeled is $o(1)$ when $n \to \infty$.

2 The Preference Frame Model of graphs with communities

This model embodies the assumption that interactions at the community level (which we will also call *macro* level) can be quantified by meaningful parameters. This general assumption underlies the (p, q) and the related parameterizations of the SBM as well. We define a *preference frame* to be a graph with K nodes, one for each community, that encodes the connectivity pattern at the community level by a (non-symmetric) *stochastic matrix* R. Formally, given $[K] = \{1, \dots K\}$, a $K \times K$ matrix R ($\det(R) \neq 0$) representing the transition matrix of a *reversible* Markov chain on $[K]$, the weighted graph $\mathcal{H} = ([K], R)$, with edge set $\operatorname{supp} R$ (edges correspond to entries in R not being 0) is called a K-*preference frame*. Requiring reversibility is equivalent to requiring that there is a set of *symmetric* weights on the edges from which R can be derived ([17]). We note that without the reversibility assumption, we would be modeling *directed* graphs, which we will leave for future work. We denote by ρ the left principal eigenvector of R, satisfying $\rho^T R = \rho^T$. W.l.o.g. we can assume the eigenvalue 1 or R has multiplicity 1[1] and therefore we call ρ the *stationary distribution* of R.

We say that a deterministic weighted graph $\mathcal{G} = (\mathcal{V}, S)$ with weight matrix S (and edge set $\operatorname{supp} S$) *admits a K-preference frame* $\mathcal{H} = ([K], R)$ if and only if there exists a partition $\mathcal{C}$ of the nodes $\mathcal{V}$ into K clusters $\mathcal{C} = \{C_1, \dots C_k\}$ of sizes $n_1, \dots, n_K$, respectively, so that the Markov chain on $\mathcal{V}$ with transition matrix P determined by S satisfies the linear constraints

$$\sum_{j \in C_m} P_{ij} \; = \; R_{lm} \quad \text{for all } i \in C_l, \text{ and all cluster indices } l, m \in \{1, 2, \dots k\}. \tag{1}$$

The matrix P is obtained from S by the standard row-normalization $P = D^{-1}S$ where $D = \operatorname{diag}\{d_{1:n}\}$, $d_i = \sum_{i=1}^n S_{ij}$.

A random graph family over node set $\mathcal{V}$ *admits a K-preference frame* $\mathcal{H}$, and is called a *Preference Frame Model* (PFM), if the edges i, j, $i < j$ are sampled independently from Bernoulli distributions with parameters S_{ij}. It is assumed that the edges obtained are undirected and that $S_{ij} \leq 1$ for all pairs $i \neq j$. We denote a realization from this process by A. Furthermore, let $\hat{d}_i = \sum_{j \in \mathcal{V}} A_{ij}$ and in general, throughout this paper, we will denote computable quantities derived from the observed A with the same letter as their model counterparts, decorated with the "hat" symbol. Thus, $\hat{D} = \operatorname{diag} \hat{d}_{1:n}$, $\hat{P} = \hat{D}^{-1}A$, and so on.

One question we will study is under what conditions the PFM model can be estimated from a given A by a standard spectral clustering algorithms. Evidently, the difficult part in this estimation problem is recovering the partition $\mathcal{C}$. If this is obtained correctly, the remaining parameters are easily estimated in a Maximum Likelihood framework.

But another question we elucidate refers to the parametrization itself. It is known that in the SBM and Degree Corrected-SBM (DC-SBM) [18], in spite of their simplicity, there are dependencies between the community level "intensive" parameters and the graph level "extensive"parameters, as we will show below. In the parametrization of the PFM , we can explicitly show which are the free parameters and which are the dependent ones.

Several network models in wide use admit a preference frame. For example, the SBM(B) model, which we briefly describe here. This model has parameters the cluster sizes $(n_{1:K})$ and the *connectivity matrix* $B \in [0, 1]^{K \times K}$. For two nodes $i, j \in \mathcal{V}$, the probability of an edge (i, j) is B_{kl} iff $i \in C_k$ and $j \in C_l$. The matrix B needs not be symmetric. When $B_{kk} = p, B_{kl} = q$ for $k, l \in [K]$, $k \neq l$, the model is denoted SBM(p, q). It is easy to verify that the SBM admits a preference frame. For instance, in the case of SBM(p, q), we have

$$d_i = p(n_l - 1) + q(n - n_l) \equiv d_{C_l}, \text{ for } i \in C_l,$$

$$R_{l,m} = \frac{q n_m}{d_{C_l}} \text{ if } l \neq m, \; R_{l,l} = \frac{p(n_l - 1)}{d_{C_l}}, \text{ for } l, m \in \{1, 2, \dots, k\}.$$

In the above we have introduced the notation $d_{C_l} = \sum_{j \in C_l} d_i$. One particular realization of the PFM is the *Homogeneous K-Preference Frame* model (HPFM). In a HPFM, each node $i \in \mathcal{V}$ is characterized by a *weight, or propensity to form ties* w_i. For each pair of communities l, m with $l \leq m$ and for each $i \in C_l, j \in C_m$ we sample A_{ij} with probability S_{ij} given by

$$S_{ij} = \frac{R_{ml} w_i w_j}{\rho_l}. \tag{2}$$

This formulation ensures detail balance in the edge expectations, i.e. $S_{ij} = S_{ji}$. The HPFM is virtually equivalent to what is known as the "degree model" [8] or "DC-SBM", up to a reparameterization[2]. Proposition 1 relates the node weights to the expected node degrees d_i. We note that the main result we prove in this paper uses independent sampling of edges only to prove the concentration of the laplacian matrix. The PFM model can be easily extended to other graph models

[1] Otherwise the networks obtained would be disconnected.

[2] Here we follow the customary definition of this model, which does not enforce $S_{ii} = 0$, even though this implies a non-zero probability of self-loops.

with dependent edges if one could prove concentration and eigenvalue separation. For example, when R has rational entries, the subgraph induced by each block of A can be represented by a random d-regular graph with a specified degree.

Proposition 1 *In a HPFM* $d_i = w_i \sum_{l=1}^{K} R_{kl} \frac{w_{C_l}}{\rho_l}$ *whenever* $i \in C_k$ *and* $k \in [K]$.

Equivalent statements that the expected degrees in each cluster are proportional to the weights exist in [7, 19] and they are instrumental in analyzing this model. This particular parametrization immediately implies in what case the degrees are globally proportional to the weights. This is, obviously, the situation when $w_{C_l} \propto \rho_l$ for all $l \in [K]$.

As we see, the node degrees in a HPFM are not directly determined by the propensities w_i, but depend on those by a multiplicative constant that varies with the cluster. This type of interaction between parameters has been observed in practically all extensions of the Stochastic Block-Model that we are aware of, making parameter interpretation more difficult. Our following result establishes what are the free parameters of the PFM and of their subclasses. As it will turn out, these parameters and their interactions are easily interpretable.

Proposition 2 *Let* $(n_1, \dots n_K)$ *be a partition of* n *(assumed to represent the cluster sizes of* $\mathcal{C} = \{C_1, \dots C_K\}$ *a partition of node set* $\mathcal{V}$*),* R *a non-singular* $K \times K$ *stochastic matrix,* ρ *its left principal eigenvector, and* $\pi_{C_1} \in [0,1]^{n_1}, \dots \pi_{C_K} \in [0,1]^{n_K}$ *probability distributions over* $C_{1:K}$*. Then, there exists a PFM consistent with* $\mathcal{H} = ([K], R)$*, with clustering* $\mathcal{C}$*, and whose node degrees are given by*

$$d_i = d_{tot} \rho_k \pi_{C_k, i}, \tag{3}$$

whenever $i \in C_k$*, where* $d_{tot} = \sum_{i \in \mathcal{V}} d_i$ *is a user parameter which is only restricted above by Assumption 2.*

The proof of this result is constructive, and can be found in the extended version.

The parametrization shows to what extent one can specify independently the degree distribution of a network model, and the connectivity parameters R. Moreover, it describes the pattern of connection of a node i as a composition of a macro-level pattern, which gives the total probability of i to form connections with a cluster l, and the micro-level distribution of connections between i and the members of C_l. These parameters are meaningful on their own and can be specified or estimated separately, as they have no hidden dependence on each other or on n, K.

The PFM enjoys a number of other interesting properties. As this paper will show, almost all the properties that make SBM's popular and easy to understand hold also for the much more flexible PFM. In the remainder of this paper we derive recovery guarantees for the PFM. As an additional goal, we will show that in the frame we set with the PFM, the recovery conditions become clearer, more interpretable, and occasionally less restrictive than for other models.

As already mentioned, the PFM includes many models that have been found useful by previous authors. Yet, the PFM class is much more flexible than those individual models, in the sense that it allows other unexplored degrees of freedom (or, in other words, achieves the same advantages as previously studied models with fewer constraints on the data). Note that there is an infinite number of possible random graphs $\mathcal{G}$ with the same parameters $(d_{1:n}, n_{1:k}, R)$ satisfying the constraints (1) and Proposition 2, yet for reliable community detection we do not need to control S fully, but only aggregate statistics like $\sum_{j \in C} A_{ij}$.

3 Spectral clustering algorithm and main result

Now, we address the community recovery problem from a random graph $(\mathcal{V}, A)$ sampled from the PFM defined as above. We make the standard assumption that K is known. Our analysis is based on a very common spectral clustering algorithm used in [13] and described also in [14, 21].

```
Input  : Graph $(\mathcal{V}, A)$ with $|\mathcal{V}| = n$ and $A \in \{0,1\}^{n\times n}$, number of clusters $K$
Output: Clustering $\mathcal{C}$
1. Compute $\hat{D} = \mathrm{diag}(\hat{d}_1, \cdots, \hat{d}_n)$ and Laplacian
       $\hat{L} = \hat{D}^{-1/2} A \hat{D}^{-1/2}$        (4)
2. Calculate the $K$ eigenvectors $\hat{Y}_1, \cdots, \hat{Y}_K$ associated with the $K$ eigenvalues $|\hat{\lambda}_1| \geq \cdots \geq |\hat{\lambda}_K|$ of $\hat{L}$. Normalize the eigenvectors to unit length. We denote them as the first $K$ eigenvectors in the following text;
3. Set $\hat{V}_i = \hat{D}^{-1/2}\hat{Y}_i$, $i = 1, \cdots, K$. Form matrix $\hat{V} = [\hat{V}_1 \cdots \hat{V}_K]$;
4. Treating each row of $\hat{V}$ as a point in $K$ dimensions, cluster them by the K-means algorithm to obtain the clustering $\hat{\mathcal{C}}$.
```

Algorithm 1: Spectral Clustering

Note that the vectors $\hat{V}$ are the first K eigenvectors of P. The K-means algorithm is assumed to find the global optimum. For more details on good initializations for K-means in step 4 see [16].

We quantify the difference between $\hat{\mathcal{C}}$ and the true clusterings $\mathcal{C}$ by the mis-clustering rate p_{err}, which is defined as

$$p_{err} = 1 - \frac{1}{n} \max_{\phi:[K]\to[K]} \sum_k |C_{\phi(k)} \cap \hat{C}_k|. \tag{5}$$

Theorem 3 (Mis-clustering rate bound for HPFM and PFM) *Let the $n \times n$ matrix S admit a PFM, and $w_{1:n}, R, \rho, P, A, d_{1:n}$ have the usual meaning, and let $\lambda_{1:n}$ be the eigenvalues of P, with $|\lambda_i| \geq |\lambda_{i+1}|$. Let $d_{min} = \min d_{1:n}$ be the minimum expected degree, $\hat{d}_{min} = \min \hat{d}_i$, and $d_{max} = \max_{ij} nS_{ij}$. Let $\gamma \geq 1$, $\eta > 0$ be arbitrary numbers. Assume:*

Assumption 1 *S admits a HPFM model and (2) holds.*

Assumption 2 $S_{ij} \leq 1$

Assumption 3 $\hat{d}_{min} \geq \log(n)$

Assumption 4 $d_{min} \geq \log(n)$

Assumption 5 $\exists \varkappa > 0$, $d_{max} \leq \varkappa \log n$

Assumption 6 *$g_{row} > 0$, where g_{row} is defined in Proposition 4.*

Assumption 7 *$\lambda_{1:K}$ are the eigenvalues of R, and $|\lambda_K| - |\lambda_{K+1}| = \sigma > 0$.*

We also assume that we run Algorithm 1 on S and that K-means finds the optimal solution. Then, for n sufficiently large, the following statements hold with probability at least $1 - e^{-\gamma}$.
PFM *Assumptions 2 - 7 imply*

$$p_{err} \leq \frac{K d_{tot}}{n d_{min} g_{row}} \left[\frac{C_0 \gamma^4}{\sigma^2 \log n} + \frac{4(\log n)^\eta}{\hat{d}_{min}} \right] \tag{6}$$

HPFM *Assumptions 1 - 6 imply*

$$p_{err} \leq \frac{K d_{tot}}{n d_{min} g_{row}} \left[\frac{C_0 \gamma^4}{\lambda_K^2 \log n} + \frac{4(\log n)^\eta}{\hat{d}_{min}} \right] \tag{7}$$

where C_0 is a constant depending on κ and γ.

Note that p_{err} decreases at least as $1/\log(n)$ when $\hat{d}_{min} = d_{min} = \log(n)$. This is because $\hat{d}_{min}$ and d_{min} help with the concentration of L. Using Proposition 4, the distances between rows of V, i.e, the true centers of the k-means step, are lower bounded by g_{row}/d_{tot}. After plugging in the assumptions for $d_{min}, \hat{d}_{min}, d_{max}$, we obtain

$$p_{err} \leq \frac{K\varkappa}{g_{row}} \left[\frac{C_0 \gamma^4}{\sigma^2 \log n} + \frac{4}{(\log n)^{(1-\eta)}} \right]. \tag{8}$$

When n is small, the first component on the right hand side dominates because of the constant C_0, while the second part dominates when n is very large. This shows that p_{err} decreases almost as $1/\log n$. Of the remaining quantities, κ controls the spread of the degrees d_i. Notice that λ_K and σ are eigengaps in HPFM model and PFM model respectively and depend only on the preference frame, and likewise for g_{row}. The eigengaps ensure the stability of principal spaces and the separation from the spurious eigenvalues, as shown in Proposition 6. The term containing $(\log n)^\eta$ is designed to control the difference between d_i and $\hat{d}_i$ with η a small positive constant.

3.1 Proof outline, techniques and main concepts

The proof of Theorem 3 (given in the extended version of the paper) relies on three steps, which are to be found in most results dealing with spectral clustering. First, concentration bounds of the empirical Laplacian $\hat{L}$ w.r.t L are obtained. There are various conditions under which these can be obtained, and ours are most similar to the recent result of [9]. The other tools we use are Hoeffding bounds and tools from linear algebra. Second, one needs to bound the perturbation of the eigenvectors Y as a function of the perturbation in L. This is based on the pivotal results of Davis and Kahan, see e.g [18]. A crucial ingredient in these type of theorems is the size of the eigengap between the invariant subspace Y and its orthogonal complement. This is a condition that is model-dependent, and therefore we discuss the techniques we introduce for solving this problem in the PFM in the next subsection.

The third step is to bound the error of the K-means clustering algorithm. This is done by a counting argument. The crux of this step is to ensure the separation of the K distinct rows of V. This, again, is model dependent and we present our result below. The details and proof are in the extended version. All proofs are for the PFM; to specialize to the HPFM, one replaces σ with $|\lambda_K|$

3.2 Cluster separation and bounding the spurious eigenvalues in the PFM

Proposition 4 (Cluster separation) *Let $V, \rho, d_{1:n}$ have the usual meaning and define the* cluster volume $d_{C_k} = \sum_{i \in C_k} d_i$*, and c_{max}, c_{min} as* $\max_k, min_k \frac{d_{C_k}}{n\rho_k}$*. Let $i, j \in \mathcal{V}$ be nodes belonging respectively to clusters k, m with $k \neq m$. Then,*

$$||V_{i:} - V_{j:}||^2 \geq \frac{1}{d_{tot}} \left[\frac{1}{c_{max}} \left(\frac{1}{\rho_k} + \frac{1}{\rho_m} \right) - \frac{1}{\sqrt{\rho_k \rho_m}} \left(\frac{1}{c_{min}} - \frac{1}{c_{max}} \right) \right] = \frac{g_{row}}{d_{tot}}, \tag{9}$$

where $g_{row} = \left[\frac{1}{c_{max}} \left(\frac{1}{\rho_k} + \frac{1}{\rho_m} \right) - \frac{1}{\sqrt{\rho_k \rho_m}} \left(\frac{1}{c_{min}} - \frac{1}{c_{max}} \right) \right]$. Moreover, if the columns of V are normalized to length 1, the above result holds by replacing $d_{tot} c_{max,min}$ with $\max, \min_k \frac{n_k}{\rho_k}$.

In the square brackets, $c_{max,min}$ depend on the cluster-level degree distribution, while all the other quantitities depend only of the preference frame. Hence, this expression is invariant with n, and as long as it is strictly positive, we have that the cluster separation is $\Omega(1/d_{tot})$.

The next theorem is crucial in proving that L has a constant eigengap. We express the eigengap of P in terms of the preference frame $\mathcal{H}$ and the mixing inside each of the clusters C_k. For this, we resort to *generalized stochastic matrices*, i.e. rectangular positive matrices with equal row sums, and we relate their properties to the mixing of Markov chains on bipartite graphs.

These tools are introduced here, for the sake of intuition, toghether with the main spectral result, while the rest of the proofs are in the extended version.

Given $\mathcal{C}$, for any vector $x \in \mathbb{R}^n$, we denote by x_k, $k = 1, \ldots K$, the block of x indexed by elements of cluster k of $\mathcal{C}$. Similarly, for any square matrix $A \in \mathbb{R}^{n \times n}$, we denote by $A_{kl} = [A_{ij}]_{i \in k, j \in l}$ the block with rows indexed by $i \in k$, and columns indexed by $j \in l$.

Denote by ρ, $\lambda_{1:K}$, $\nu^{1:K} \in \mathbb{R}^K$ respectively the stationary distribution, eigenvalues[3], and eigenvectors of R.

We are interested in block stochastic matrices P for which the eigenvalues of R are the principal eigenvalues. We call $\lambda_{K+1} \dots \lambda_n$ *spurious* eigenvalues. Theorem 6 below is a sufficient condition that bounds $|\lambda_{K+1}|$ whenever each of the K^2 blocks of P is ”homogeneous” in a sense that will be defined below.

When we consider the matrix $L = D^{-1/2} S D^{-1/2}$ partitioned according to $\mathcal{C}$, it will be convenient to consider the off-diagonal blocks in pairs. This is why the next result describes the properties of matrices consisting of a pair of off-diagonal blocks.

Proposition 5 (Eigenvalues for the off-diagonal blocks) *Let M be the square matrix*

$$M = \left[\begin{array}{cc} 0 & B \\ A & 0 \end{array} \right] \tag{10}$$

where $A \in \mathbb{R}^{n_2 \times n_1}$ and $B \in \mathbb{R}^{n_1 \times n_2}$, and let $x = \left[\begin{array}{c} x_1 \\ x_2 \end{array} \right]$, $x_{1,2} \in \mathbb{C}^{n_{1,2}}$ be an eigenvector of M with eigenvalue λ. Then

$$\begin{aligned} Bx_2 &= \lambda x_1 \quad & ABx_2 &= \lambda^2 x_2 \qquad (11) \\ Ax_1 &= \lambda x_2 \quad & BAx_1 &= \lambda^2 x_1 \qquad (12) \\ & & M^2 &= \left[\begin{array}{cc} BA & 0 \\ 0 & AB \end{array} \right] \qquad (13) \end{aligned}$$

Moreover, if M is symmetric, i.e $B = A^T$, then λ is a singular value of A, x is real, and $-\lambda$ is also an eigenvalue of M with eigenvector $[x_1^T \; -x_2^T]^T$. Assuming $n_2 \le n_1$, and that A is full rank, one can write $A = V \Lambda U^T$ with $V \in \mathbb{R}^{n_2 \times n_2}$, $U \in \mathbb{R}^{n_1 \times n_2}$ orthogonal matrices, and Λ a diagonal matrix of non-zero singular values.

Theorem 6 (Bounding the spurious eigenvalues of L**)** *Let $\mathcal{C}, L, P, D, S, R, \rho$ be defined as above, and let λ be an eigenvalue of P. Assume that (1) P is block-stochastic with respect to $\mathcal{C}$; (2) $\lambda_{1:K}$ are the eigenvalues of R, and $|\lambda_K| > 0$; (3) λ is not an eigenvalue of R; (4) denote by λ_3^{kl} (λ_2^{kk}) the third (second) largest in magnitude eigenvalue of block M_{kl} (L_{kk}) and assume that $\frac{|\lambda_3^{kl}|}{\lambda_{max}(M_{kl})} \le c < 1$ ($\frac{|\lambda_2^{kk}|}{\lambda_{max}(L_{kk})} \le c$). Then, the spurious eigenvalues of P are bounded by c times a constant that depends only on R.*

$$|\lambda| \le c \max_{k=1:K} \left(r_{kk} + \sum_{l \ne k} \sqrt{r_{kl} r_{lk}} \right) \tag{14}$$

Remarks: The factor that multiplies c can be further bounded denoting $a = [\sqrt{r_{kl}}]_{l=1:K}^T, b = [\sqrt{r_{lk}}]_{l=1:K}^T$

$$r_{kk} + \sum_{l \ne k} \sqrt{r_{kl} r_{lk}} \;=\; a^T b \;\le ||a|| ||b|| \;=\; \sqrt{\sum_{l=1}^K r_{kl} \sum_{l=1}^K r_{lk}} \;=\; \sqrt{\sum_{l=1}^K r_{lk}} \tag{15}$$

In other words,

$$|\lambda| \le \frac{c}{2} \max_{k=1:K} \sqrt{\sum_{l=1}^K r_{lk}} \tag{16}$$

The maximum column sum of a stochastic matrix is 1 if the matrix is doubly stochastic and larger than 1 otherwise, and can be as large as $\sqrt{K}$. However, one must remember that the interesting R matrices have “large” eigenvalues. In particular we will be interested in $\lambda_K > c$. It is expected that under these conditions, the factor depending on R to be close to 1.

[3] Here too, eigenvalues will always be ordered in decreasing order of their magnitudes, with positive values preceeding negatives one of the same magnitude. Consequently, for any stochastic matrix, $\lambda_1 = 1$ always

The second remark is on the condition (3), that all blocks have small spurious eigenvalues. This condition is not merely a technical convenience. If a block had a large eigenvalue, near 1 or -1 (times its λ_{max}), then that block could itself be broken into two distinct clusters. In other words, the clustering $\mathcal{C}$ would not accurately capture the cluster structure of the matrix P. Hence, condition (3) amounts to requiring that no other cluster structure is present, in other words that within each block, the Markov chain induced by P *mixes well.*

4 Related work

Previous results we used The Laplacian concentration results use a technique introduced recently by [9], and some of the basic matrix theoretic results are based on [14] which studied the P and L matrix in the context of spectral clustering. As any of the many works we cite, we are indebted to the pioneering work on the perturbation of invariant subspaces of Davis and Kahan [18, 19, 20].

4.1 Previous related models

The configuration model for regular random graphs [4, 11] and for graphs with general fixed degrees [10, 12] is very well known. It can be shown by a simple calculation that the configuration model also admits a K-preference frame. In the particular case when the diagonal of the R matrix is 0 and the connections between clusters are given by a bipartite configuration model with fixed degrees, K-preference frames have been studied by [15] under the name “equitable graphs”; the object there was to provide a way to calculate the spectrum of the graph.

Since the PFM is itself an extension of the SBM, many other extensions of the latter will bear resemblance to PFM. Here we review only a subset of these, a series of strong relatively recent advances, which exploit the spectral properties of the SBM and extend this to handle a large range of degree distributions [7, 19, 5]. The PFM includes each of these models as a subclass[4].

In [7] the authors study a model that coincides (up to some multiplicative constants) with the HPFM. The paper introduces an elegant algorithm that achieves partial recovery or better, which is based on the spectral properties of a random Laplacian-like matrix, and does not require knowledge of the partition size K.

The PFM also coincides with the model of [1] and [8] called the *expected degree model* w.r.t the distribution of *intra-cluster* edges, but not w.r.t the ambient edges, so the HPFM is a subclass of this model.

A different approach to recovery The papers [5, 18, 9] propose regularizing the normalized Laplacian with respect to the influence of low degrees, by adding the scaled unit matrix τI to the incidence matrix A, and thereby they achieve recovery for much more imbalanced degree distributions than us. Currently, we do not see an application of this interesting technique to the PFM, as the diagonal regularization destroys the separation of the intracluster and intercluster transitions, which guarantee the clustering property of the eigenvectors. Therefore, currently we cannot break the $n \log n$ limit into the ultra-sparse regime, although we recognize that this is an important current direction of research.

Recovery results like ours can be easily extended to weighted, non-random graphs, and in this sense they are relevant to the spectral clustering of these graphs, when they are assumed to be noisy versions of a $\mathcal{G}$ that admits a PFM.

4.2 An empirical comparison of the recovery conditions

As obtaining general results in comparing the various recovery conditions in the literature would be a tedious task, here we undertake to do a numerical comparison. While the conclusions drawn from this are not universal, they illustrate well the stringency of various conditions, as well as the gap between theory and actual recovery. For this, we construct HPFM models, and verify numerically if they satisfy the various conditions. We have also clustered random graphs sampled from this model, with good results (shown in the extended version).

[4] In particular, the models proposed in [7, 19, 5] are variations of the DC-SBM and thus forms of the homogeneous PFM.

We generate S from the HPFM model with $K = 5$, $n = 5000$. Each w_i is uniformly generated from $(0.5, 1)$. $n_{1:K} = (500, 1000, 1500, 1000, 1000)$, $g_{row} > 0$, $\lambda_{1:K} = (1, 0.8, 0.6, 0.4, 0.2)$. The matrix R is given below; note its last row in which $r_{55} < \sum_{l=1}^{4} r_{5l}$.

$$R = \begin{pmatrix} .80 & .07 & .02 & .02 & .09 \\ .04 & .52 & .24 & .12 & .08 \\ .01 & .20 & .65 & .15 & .00 \\ .01 & .08 & .12 & .70 & .08 \\ .13 & .21 & .02 & .32 & .33 \end{pmatrix} \quad \rho = (.25, .44, .54, .65, .17). \tag{17}$$

The conditions we are verifying include besides ours, those obtained by [18], [19], [3] and [5]; since the original S is a perfect case for spectral clustering of weighted graphs, we also verify the theoretical recovery conditions for spectral clustering in [2] and [16].

Our result Theorem 3 Assumption 1 and 2 automatically hold from the construction of the data. By simulating the data, We find that $d_{min} = 77.4$, $\hat{d}_{min} = 63$, both of which are bigger than $\log n = 8.52$. Therefore Assumption 3 and 4 hold. $d_{max} = 509.3$, $g_{row} = 1.82 > 0$, thus Assumption 5 and 6 hold. After running Algorithm 1, the mis-clustering rate is $r = 0.0008$, which satisfies the theoretical bound. In conclusion, the dataset fits into both the assumptions and conclusion of Theorem 3.

Qin and Rohe[18] This paper has an assumption on the lower bound on λ_K, that is $\frac{1}{8\sqrt{3}}\lambda_K \geq \sqrt{\frac{K(ln(K/\epsilon)}{d_{min}}}$, so that the concentration bound holds with probability $(1 - \epsilon)$. We set $\epsilon = 0.1$ and obtain $\lambda_K \geq 12.3$, which is impossible to hold since λ_K is upper bounded by 1^5.

Rohe, Chatterjee, Yu[19] Here, one defines $\tau_n = \frac{d_{min}}{n}$, and requires $\tau_n^2 \log n > 2$ to ensure the concentration of L. To meet this assumption, with $n = 5000$, $d_{min} \geq 2422$. While in our case $d_{min} = 77.4$. The assumption requires a very dense graph and is not satisfied in this dataset.

Balcan, Borgs Braverman, Chayes[3]Their theorem is based on self-determined community structure. It requires all the nodes to be more connected within their own cluster. However, in our graph, 1296 out of 5000 nodes have more connections to outside nodes than to nodes in their own cluster.

Ng, Jordan, Weiss[16] require $\lambda_2 < 1 - \delta$, where $\delta > (2 + 2\sqrt{2})\epsilon$, $\epsilon = \sqrt{K(K-1)\epsilon_1 + K\epsilon_2^2}$, $\epsilon_1 \geq \max_{i_1, i_2 \in \{1, \cdots, K\}} \sum_{j \in C_{i_1}} \sum_{k \in C_{i_2}} \frac{A_{jk}^2}{\hat{d}_j \hat{d}_k}$, $\epsilon_2 \geq \max_{i \in \{1, \cdots, K\}} \frac{\sum_{k: k \in S_i}}{\hat{d}_j} (\sum_{k,l \in S_i} \frac{A_{kl}^2}{\hat{d}_k \hat{d}_l})^{1/2}$. On the given data, we find that $\epsilon \geq 36.69$, and $\delta \geq 125.28$, which is impossible to hold since δ needs to be smaller than 1.

Chaudhuri, Chung, Tsiatas[5] The recovery theorem of this paper requires $d_i \geq \frac{128}{9} \ln(6n/\delta)$, so that when all the assumptions hold, it recovers the clustering correctly with probability at least $1 - 6\delta$. We set $\delta = 0.01$, and obtain that $d_i = 77.40$, $\frac{128}{9} \ln(6n/\delta) = 212.11$. Therefore the assumption fails as well.

For our method, the hardest condition to satisfy, and the most different from the others, was Assumption 6. We repeated this experiment with the other weights distributions for which this assumption fails. The assumptions in the related papers continued to be violated. In [Qin and Rohe], we obtain $\lambda_K \geq 17.32$. In [Rohe, Chatterjee, Yu], we still needs $d_{min} \geq 2422$. In [Balcan, Borgs Braverman, Chayes], we get 1609 points more connected to the outside nodes of its cluster. In [Balakrishnan, Xu, Krishnamurthy, Singh], we get $\sigma = 0.172$ and needs to satisfy $\sigma = o(0.3292)$. In [Ng, Jordan, Weiss], we obtain $\delta \geq 175.35$. Therefore, the assumptions in these papers are all violated as well.

5 Conclusion

In this paper, we have introduced the preference frame model, which is more flexible and subsumes many current models including SBM and DC-SBM. It produces state-of-the art recovery rates comparable to existing models. To accomplish this, we used a parametrization that is clearer and more intuitive. The theoretical results are based on the new geometric techniques which control the eigengaps of the matrices with piecewise constant eigenvectors.

We note that the main result theorem 3 uses independent sampling of edges only to prove the concentration of the laplacian matrix. The PFM model can be easily extended to other graph models with dependent edges if one could prove concentration and eigenvalue separation. For example, when R has rational entries, the subgraph induced by each block of A can be represented by a random d-regular graph with a specified degree.

[5] To make $\lambda \leq 1$ possible, one needs $d_{min} \geq 11718$.

References

[1] Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 37–54. ACM, 2012.

[2] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962, 2011.

[3] Maria-Florina Balcan, Christian Borgs, Mark Braverman, Jennifer Chayes, and Shang-Hua Teng. Finding endogenously formed communities. *arxiv preprint arXiv:1201.4899v2*, 2012.

[4] Bela Bollobas. *Random Graphs*. Cambridge University Press, second edition, 2001.

[5] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in extended planted partition model. *Journal of Machine Learning Research*, pages 1–23, 2012.

[6] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

[7] Amin Coja-Oghlan and Andre Lanka. Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics*, 23:1682–1714, 2009.

[8] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[9] Can M. Le and Roman Vershynin. Concentration and regularization of random graphs. 2015.

[10] Brendan McKay. Asymptotics for symmetric 0-1 matrices with prescribed row sums. *Ars Combinatoria*, 19A:15–26, 1985.

[11] Brendan McKay and Nicholas Wormald. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11:52–67, 1990.

[12] Brendan McKay and Nicholas Wormald. Asymptotic enumeration by degree sequence of graphs with degrees $o(n^{1/2}$. *Combinatorica*, 11(4):369–382, 1991.

[13] Marina Meilă and Jianbo Shi. Learning segmentation by random walks. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 873–879, Cambridge, MA, 2001. MIT Press.

[14] Marina Meilă and Jianbo Shi. A random walks view of spectral segmentation. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics AISTATS*, 2001.

[15] M.E.J. Newman and Travis Martin. Equitable random graphs. 2014.

[16] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[17] J.R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[18] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

[19] Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

[20] Gilbert W Stewart, Ji-guang Sun, and Harcourt Brace Jovanovich. *Matrix perturbation theory*, volume 175. Academic press New York, 1990.

[21] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.