The Annals of Statistics
2020, Vol. 48, No. 1, 183–204
https://doi.org/10.1214/18-AOS1797

OPTIMAL RATES FOR COMMUNITY ESTIMATION IN THE WEIGHTED STOCHASTIC BLOCK MODEL

BY MIN XU[1], VARUN JOG[2,*] AND PO-LING LOH[2,**]

[1]*Department of Statistics, Rutgers University, mx76@stat.rutgers.edu*
[2]*Departments of ECE & Statistics, University of Wisconsin, Madison, *vjog@wisc.edu; **ploh@stat.wisc.edu*

Community identification in a network is an important problem in fields such as social science, neuroscience and genetics. Over the past decade, stochastic block models (SBMs) have emerged as a popular statistical framework for this problem. However, SBMs have an important limitation in that they are suited only for networks with unweighted edges; in various scientific applications, disregarding the edge weights may result in a loss of valuable information. We study a weighted generalization of the SBM, in which observations are collected in the form of a weighted adjacency matrix and the weight of each edge is generated independently from an unknown probability density determined by the community membership of its endpoints. We characterize the optimal rate of misclustering error of the weighted SBM in terms of the Renyi divergence of order 1/2 between the weight distributions of within-community and between-community edges, substantially generalizing existing results for unweighted SBMs. Furthermore, we present a computationally tractable algorithm based on discretization that achieves the optimal error rate. Our method is adaptive in the sense that the algorithm, without assuming knowledge of the weight densities, performs as well as the best algorithm that knows the weight densities.

1. Introduction. The recent explosion of network datasets has created a need for new statistical methodology [13, 16, 24, 33]. One active area of research with diverse scientific applications pertains to community detection and estimation, where observations take the form of edges between nodes in a graph, and the goal is to partition the nodes into disjoint groups based on their relative connectivity [14, 21, 29, 35, 36, 39].

A standard model assumption in community recovery problems is that—conditioned on the community labels of the nodes of the graph—each edge is generated independently according to a distribution governed solely by the community labels of its endpoints. This is the setting of the stochastic block model (SBM) [23]. Community recovery may also be viewed as estimating the latent cluster memberships of the nodes a random graph generated by an SBM. The last decade has seen great progress on this problem, beginning with the seminal conjecture of Decelle et al. [12] (see, e.g., the survey paper by Abbe [1]). Various algorithms for community recovery have been devised with guaranteed optimality properties, measured in terms of correlated recovery [28, 30, 32], exact recovery [3–5] and minimum misclustering error rate [15, 42].

However, an important shortcoming of SBMs is that all edges are assumed to be binary. In contrast, the edges appearing in many real-world networks possess weights reflecting a diversity of strengths or characteristics [10, 34]: Edges in social or cellular networks may quantify the frequency of interactions between pairs of individuals [9, 38]. Similarly, edges in gene co-expression networks are assigned weights corresponding to the correlation between expression levels of pairs of genes [43]; and in brain networks, edge weights may indicate the

Received June 2017; revised September 2018.
MSC2010 subject classifications. Primary 62H30, 91D30; secondary 62C20, 90B15.
Key words and phrases. Stochastic block models, network analysis, Renyi divergence, nonparametric estimation, optimal estimation rates.

level of neuronal activity between corresponding regions in the brain [37]. Although an unweighted adjacency matrix could be constructed by disregarding the edge weight data, this might result in a loss of valuable information that could be used to recover hidden communities.

This motivates the *weighted* stochastic block model, which we study in this paper. Each edge is generated from a Bernoulli(p) or Bernoulli(q) distribution, depending on whether its endpoints lie in the same community, and then each edge is assigned an edge weight generated from one of two arbitrary densities, $p(\cdot)$ or $q(\cdot)$. We study the problem of community estimation based on observations of the edge weights in the network, *without* assuming knowledge of p, q, $p(\cdot)$ or $q(\cdot)$. Since $p(\cdot)$ and $q(\cdot)$ are allowed to be continuous, our model strictly generalizes the discrete labeled SBMs considered in previous literature [22, 25, 27], as well as the censored SBM [2, 17, 18].

We emphasize key differences between the weighted SBM framework and the setting of other clustering problems involving continuous edge weights [7, 20]. First, we do not assume that between-cluster edges tend to have heavier weights than within-cluster edges (e.g., in mean-separation models). Such an assumption is critical to many algorithms for weighted networks, since it allows existing algorithms for unweighted SBMs, such as spectral clustering, to be applied in relatively straightforward ways. In contrast, the algorithms in this paper allow us to exploit other potential differences in $p(\cdot)$ and $q(\cdot)$, such as differences in variance or shape. This is crucial to achieve optimal performance. Second, our setting is *nonparametric* in the sense that the densities $p(\cdot)$ and $q(\cdot)$ may be arbitrary and are only required to satisfy mild regularity conditions, whereas previous approaches generally assume that $p(\cdot)$ and $q(\cdot)$ belong to a specific parametric family. Nonparametric density estimation is itself a difficult problem, made even more difficult in the case of weighted SBMs, since we do not know a priori which edge weights have been drawn from which densities.

Our main theoretical contribution is to characterize the *optimal rate of misclustering error* in the weighted SBM. On one side, we derive an information-theoretic lower bound for the performance of any community recovery algorithm for the weighted SBM. Our lower bound applies to all parameters in the parameter space (thus is not minimax) and all algorithms that produce the same output on isomorphic networks—a property that we call *permutation equivariance*. On the other side, we present a computationally tractable algorithm with a rate of convergence that matches the lower bound. Our results show that the optimal rate for community estimation in a weighted SBM is governed by the Renyi divergence of order $\frac{1}{2}$ between two mixed distributions, capturing the discrepancy between the edge probabilities and edge weight densities for between-community and within-community connections. This provides a natural but highly nontrivial generalization of the results in Zhang and Zhou [42] and Gao et al. [15], which show that the optimal rate of the unweighted SBM is characterized by the Renyi divergence of order $\frac{1}{2}$ between two Bernoulli distributions corresponding only to edge probabilities.

Remarkably, our rate-optimal algorithm is fully *adaptive* and does not require prior knowledge of $p(\cdot)$ and $q(\cdot)$. Thus, even in cases where the densities belong to a parametric family, it is possible—*without making any parametric assumptions*—to obtain the same optimal rate as if one imposes the true parametric form. This is in sharp contrast to most nonparametric estimation problems in statistics, where nonparametric methods usually lead to a slower rate of convergence than parametric methods if a specific parametric form is known. The apparent discrepancy is explained by the simply stated observation that in weighted SBMs, one does *not* need to estimate edge densities well in order to recover communities to desirable accuracy. This intuition is also reflected in the work of Abbe and Sandon [5] for the exact recovery problem and Gao et al. [15] for the unweighted SBM. Our proposed recovery algorithm hinges on a careful discretization technique: When the edge weights are bounded, we discretize the distribution via a uniformly spaced binning to convert the weighted SBM into an instance of a *labeled* SBM, where each edge possesses a label from a discrete set with finite (but divergent) cardinality; we then perform community recovery in the labeled SBM by extending a coarse-to-fine clustering algorithm that computes an initialization through spectral clustering [11, 26] and then performs refinement through nodewise likelihood maximization [15]. When the edge weights are unbounded, we reduce the problem to the bounded case by first applying an appropriate transformation to the edge weight distributions.

The remainder of our paper is organized as follows: Section 2 introduces the mathematical framework of the weighted SBM, defines the community recovery problem, and formalizes the notion of permutation equivariance. Section 3 provides an informal summary of our results, later formalized in Section 5. Section 4 outlines our proposed community estimation algorithm. The key technical components of our proofs are highlighted in Section 6, and Section 7 reports the results of various simulations.

Notation. For a positive integer n, we write $[n]$ to denote the set $\{1, \dots, n\}$ and S_n to denote the set of permutations of $[n]$. For two real numbers a and b, we write $a \vee b$ to denote $\max(a, b)$ and write $a \wedge b$ to denote $\min(a, b)$.

2. Model and problem formulation.

We begin with a formal definition of the homogeneous weighted SBM and a description of the community recovery problem.

2.1. *Weighted stochastic block model.*

Let n denote the number of nodes in the network and let $K \geq 2$ denote the number of communities. A *clustering* σ is a function $[n] \to [K]$. For each node $u \in [n]$, we refer to $\sigma(u)$ as the cluster of node u.

DEFINITION 2.1. For a positive number $\beta \geq 1$, we define $\mathcal{C}(\beta, K)$ as the set of clusterings with minimum cluster size is at least $\frac{n}{\beta K}$, that is, $\sigma \in \mathcal{C}(\beta, K)$ if and only if $|\sigma^{-1}(k)| \geq \frac{n}{\beta K}$ for all $k \in [K]$. We refer to β as the *cluster-imbalance constant.*

We first define the homogeneous unweighted SBM, which is characterized by the following distribution over adjacency matrices $A \in \{0, 1\}^{n \times n}$.

DEFINITION 2.2 (Homogeneous unweighted SBM). Let $\sigma_0 \in \mathcal{C}(\beta, K)$ and $p, q \in [0, 1]$. We say that a random binary-valued matrix A has the distribution $\mathrm{SBM}(\sigma_0, p, q)$ if for all $u < v$, the entries of A are generated independently according to

$$A_{uv} \sim \begin{cases} \mathrm{Ber}(p) & \text{if } \sigma_0(u) = \sigma_0(v), \\ \mathrm{Ber}(q) & \text{if } \sigma_0(u) \neq \sigma_0(v). \end{cases}$$

Thus, the parameters p and q correspond to the within-cluster and between-cluster edge probabilities. The more general *heterogenous* unweighted SBM is characterized by a matrix $P \in \mathbb{R}^{K \times K}$ of probabilities instead of two scalars p and q, and edges are generated independently according to $A_{uv} \sim \mathrm{Ber}(P_{\sigma_0(u),\sigma_0(v)})$.

A homogeneous weighted SBM is parametrized by $\sigma_0 \in \mathcal{C}(\beta, K)$, the edge *absence* probabilities P_0 and Q_0, and the edge weight probability densities $p(\cdot)$ and $q(\cdot)$ supported on $S \subset \mathbb{R}$, where S may be $[0, 1]$, $[0, \infty)$, or $\mathbb{R}$. The weighted SBM is then characterized by a distribution over symmetric matrices $A \in S^{n \times n}$ in the following manner.

DEFINITION 2.3 (Homogeneous weighted SBM). Let $\sigma_0 \in \mathcal{C}(\beta, K)$. A random real-valued matrix A has the distribution $\mathrm{WSBM}(\sigma_0, (P_0, p), (Q_0, q))$ if for all $u < v$,

$$A_{uv} \sim \begin{cases} P_0\delta_0(\cdot) + (1 - P_0)p(\cdot) & \text{if } \sigma_0(u) = \sigma_0(v), \\ Q_0\delta_0(\cdot) + (1 - Q_0)q(\cdot) & \text{if } \sigma_0(u) \neq \sigma_0(v), \end{cases} \tag{1}$$

where $P_0\delta_0(\cdot) + (1 - P_0)p(\cdot)$ denotes a probability distribution whose singular part (with respect to the Lebesgue measure) is a point mass at 0 with probability P_0 and whose continuous part has $(1 - P_0)p(\cdot)$ as its Radon–Nikodym derivative with respect to the Lebesgue measure; and $Q_0\delta_0(\cdot) + (1 - Q_0)q(\cdot)$ is defined analogously.

Note that if $p(\cdot)$ and $q(\cdot)$ are Dirac delta masses at 1, the weighted SBM reduces to the unweighted version. We make a few additional remarks about the definition of the weighted SBM. First, we observe that $\mathbb{E}(A)$ may not exhibit the familiar block structure found in unweighted SBMs, since our model includes the case where $(P_0, p(\cdot))$ and $(Q_0, q(\cdot))$ have the same mean. Second, our definition treats an edge with weight 0 as a missing edge, but it is straightforward to distinguish the two notions by defining P and Q as probability measures over $S \cap \{*\}$, where the symbol $*$ denotes a missing edge. Lastly, it is possible to generalize the weighted SBM to a *weighted and labeled* SBM with the model

$$A_{uv} \sim \begin{cases} P & \text{if } \sigma_0(u) = \sigma_0(v), \\ Q & \text{if } \sigma_0(u) \neq \sigma_0(v), \end{cases}$$

where P and Q are general probability distributions over S (and the labels correspond to a discrete part). The theory derived in this paper extends in a straightforward fashion to the cases where the discrete portion of P and Q has finite support.

2.2. *Community estimation.* Given an observation $A \in S^{n\times n}$ generated from a weighted SBM, the goal of community estimation is to recover the true cluster membership structure σ_0. We assume throughout our paper that the number of clusters K is known.

We evaluate the performance of a community recovery algorithm in terms of its misclustering error. For a clustering algorithm $\hat{\sigma}$, let $\hat{\sigma}(A) : [n] \to [K]$ denote the clustering produced by $\hat{\sigma}$ when provided with the input A. We have the following definition.

DEFINITION 2.4. We define the *misclustering error* to be

$$l(\hat{\sigma}(A), \sigma_0) := \min_{\pi \in S_K} \frac{1}{n} d_H(\pi \circ \hat{\sigma}(A), \sigma_0),$$

where $d_H(\cdot, \cdot)$ denotes the Hamming distance. The *risk* of $\hat{\sigma}$ is defined as $R(\hat{\sigma}, \sigma_0) := \mathbb{E}l(\hat{\sigma}(A), \sigma_0)$, where the expectation is taken with respect to both the random network A and any potential randomness in the algorithm $\hat{\sigma}$.

The goal of this paper is to characterize the minimal achievable risk for community recovery on the weighted SBM in terms of the parameters $(n, \beta, K, (P_0, p), (Q_0, q))$.

2.3. *Permutation equivariance.* Since the cluster structure in a network does not depend on how the nodes are labeled, it is natural to focus on estimation algorithms that output equivalent clusterings when provided with isomorphic inputs. We formalize this property in the following definition.

DEFINITION 2.5. For an $n \times n$ matrix A and a permutation $\pi \in S_n$, let πA denote the $n \times n$ matrix such that $A_{uv} = [\pi A]_{\pi(u),\pi(v)}$. Let $\hat{\sigma}$ be a deterministic clustering algorithm. Then $\hat{\sigma}$ is *permutation equivariant* if, for any A and any $\pi \in S_n$,

$$\tau \circ \hat{\sigma}(\pi A) \circ \pi = \hat{\sigma}(A) \quad \text{for some } \tau \in S_K. \tag{2}$$

Note that $\hat{\sigma}(\pi A)$ by itself is not equivalent to $\hat{\sigma}(A)$, since the nodes in πA are labeled with respect to the permutation π. It is straightforward to extend Definition 2.5 to randomized algorithms by requiring condition (2) to hold almost everywhere in the probability space that underlies the algorithmic randomness. Permutation equivariance is a natural property satisfied by all the clustering algorithms studied in literature except algorithms that leverage extra side information in addition to the given network. In Section 5.2, we study permutation equivariance in detail and provide some properties of permutation equivariant estimators.

3. Overview of main results. The difficulty of community recovery depends on the extent to which (P_0, p) and (Q_0, q) are different; it is clearly impossible to have a consistent clustering algorithm if (P_0, p) and (Q_0, q) are equal. We show in this paper that a natural measure of discrepancy between (P_0, p) and $(Q_0, q))$ which governs the optimal rate of convergence is the Renyi divergence of order $\frac{1}{2}$.

Given any probability distributions P and Q that are absolutely continuous with respect to each other, the Renyi divergence of order $\frac{1}{2}$ is defined as $I(P, Q) := -2\log\int(\frac{dP}{dQ})^{1/2}\,dQ$. For our setting, the Renyi divergence takes the special form

$$I\big((P_0, p), (Q_0, q)\big) = -2\log\biggl(\sqrt{P_0Q_0} + \int\sqrt{(1-P_0)(1-Q_0)p(x)q(x)}\,dx\biggr).$$

If $I((P_0, p), (Q_0, q))$ is bounded above by a universal constant, the Renyi divergence is of the same order as the Hellinger distance (cf. Lemma H.2):

$$\begin{aligned}
&I\big((P_0, p), (Q_0, q)\big)\\
&\quad\asymp (\sqrt{P_0} - \sqrt{Q_0})^2 + \int_S\bigl(\sqrt{(1-P_0)p(x)} - \sqrt{(1-Q_0)q(x)}\bigr)^2\,dx\\
&\quad= (\sqrt{P_0} - \sqrt{Q_0})^2 + (\sqrt{1-P_0} - \sqrt{1-Q_0})^2\\
&\qquad + \sqrt{(1-P_0)(1-Q_0)}\int_S\bigl(\sqrt{p(x)} - \sqrt{q(x)}\bigr)^2\,dx.
\end{aligned}$$

Thus, we can think of $I((P_0, p), (Q_0, q))$ as having two components, the first of which captures the divergence between the edge presence probabilities (and also appears in the analysis of unweighted SBMs), and the second of which captures the divergence between the edge weight densities.

The presence of the second term illustrates how the weighted SBM behaves quite differently from its unweighted counterpart—in particular, *dense* networks may be interesting in a weighted setting. For example, even if the weighted network is completely dense in the sense that $1 - P_0 = 1 - Q_0 = 1$, a nonzero signal I may still exist if $p(\cdot)$ and $q(\cdot)$ are sufficiently different. Our results apply simultaneously to dense and sparse settings; it is important to note that dense weighted networks arise in real-world settings, such as gene co-expression data.

We now provide an informal overview of our main results.

THEOREM (Informal statement). *Let A be generated from a weighted SBM. Under regularity conditions on $((P_0, p), (Q_0, q))$, any permutation equivariant estimator $\hat{\sigma}$ satisfies the lower bound*

$$\mathbb{E}l\big(\hat{\sigma}(A), \sigma_0\big) \geq \exp\biggl(-(1 + o(1))\frac{n}{\beta K}I\big((P_0, p), (Q_0, q)\big)\biggr).$$

THEOREM (Informal statement). *Under regularity conditions on the parameters $((P_0, p), (Q_0, q))$, there exists a permutation equivariant algorithm $\hat{\sigma}$ achieving the following misclustering error rate*:

$$\lim_{n\to\infty} P\biggl(l\big(\hat{\sigma}(A), \sigma_0\big) \leq \exp\biggl(-(1 + o(1))\frac{n}{\beta K}I\big((P_0, p), (Q_0, q)\big)\biggr)\biggr) = 1.$$

Furthermore, if $\frac{nI}{\beta K \log n} \leq 1$, *we have*

$$\mathbb{E}l(\hat{\sigma}(A), \sigma_0) \leq \exp\left(-(1+o(1))\frac{nI}{\beta K}\right).$$

Taken together, the theorems imply that in the regime where $\frac{nI}{\beta K \log n} \leq 1$, the optimal risk is tightly characterized by the quantity $\exp(-(1+o(1))\frac{nI}{\beta K})$. On the other hand, if $\frac{nI}{\beta K \log n} > 1$, we have $\exp(-(1+o(1))\frac{nI}{\beta K}) < \frac{1}{n}$ for large enough n, so $\lim_{n\to\infty} P(l(\hat{\sigma}(A), \sigma_0) = 0) \to 1$ (since $l(\hat{\sigma}(A), \sigma_0) < \frac{1}{n}$ implies $l(\hat{\sigma}(A), \sigma_0) = 0$). Thus, the regime where $\frac{nI}{\beta K \log n} > 1$ is in some sense an easier problem, since we can guarantee perfect recovery with high probability.

3.1. *Relation to previous work.* Our result generalizes the work of Zhang and Zhou [42], which establishes the minimax rate of

$$\exp\left(-(1+o(1))\frac{n}{\beta K} I(\mathrm{Ber}(p), \mathrm{Ber}(q))\right)$$

for the unweighted SBM, where

$$I(\mathrm{Ber}(p), \mathrm{Ber}(q)) = -2\log(\sqrt{pq} + \sqrt{(1-p)(1-q)}).$$

The optimal algorithm proposed in Zhang and Zhou [42] is intractable, but a computationally feasible version was developed by Gao et al. [15]; the latter algorithm is a building block for the estimation algorithm proposed in this paper.

Our result should also be viewed in comparison to Yun and Proutiere [41], who studied the optimal risk for the heterogenous labeled SBM with finitely many labels, with respect to a prior on the cluster assignment σ_0. They characterize the optimal rate under a notion of divergence that reduces to the Renyi divergence of order $\frac{1}{2}$ between two discrete distributions over a fixed finite number of labels in the homogeneous setting (cf. Lemma G.2). Since the discussion is somewhat technical, we provide a more detailed comparison of our work to the results of Yun and Proutiere in Section 6.1.

Jog and Loh [25] proposed a similar weighted block model and show the exact recovery threshold to be dependent on the Renyi divergence. They focus on the setting where the distributions are discrete and known, whereas we consider continuous densities that are unknown. Aicher et al. [6] introduced a version of a weighted SBM that is a special case of the setting discussed in this paper, where the densities P and Q in equation (1) are drawn from a known exponential family. Notably, the definition of Aicher et al. [6] cannot incorporate sparsity. The weighted SBM considered in Hajek et al. [19] is also similar to the one we propose in our paper, except it only involves a single hidden community and assumes knowledge of the distributions P and Q. Weighted networks have also received some attention in the physics community [8, 34], and various ad hoc methods have been proposed; since theoretical properties are generally unknown, we do not explore these connections in our paper.

Other notions of recovery. A closely related problem is that of finding the exact recovery threshold. We say that the unweighted SBM has an *exact recovery threshold* if a function $\theta(p, q, n, K, \beta, \sigma_0)$ exists such that exact recovery is asymptotically almost always impossible if $\theta < 1$, and almost always possible if $\theta > 1$. For the homogeneous unweighted SBM, Abbe et al. [3] show that when $\beta = 1$, $K = 2$, $1 - P_0 = \frac{a \log n}{n}$, and $1 - Q_0 = \frac{b \log n}{n}$, for some constants a and b, the exact recovery threshold is $\sqrt{a} - \sqrt{b}$. This result was later generalized to multiple communities with heterogenous edge probabilities in Abbe and Sandon [4], where a notion of CH-divergence was shown to characterize the threshold for exact recovery. A notion of weak recovery, corresponding to a detection threshold, has also been considered [28, 31].

4. Estimation algorithm. A natural approach to community estimation is to first estimate the edge weight densities $p(\cdot)$ and $q(\cdot)$, but this is hindered by the fact that we do not know whether an edge weight observation originates from $p(\cdot)$ or $q(\cdot)$. An alternative approach of applying spectral clustering directly to the weighted adjacency matrix A will also be ineffective if (P_0, p) and (Q_0, q) have the same mean, so $\mathbb{E}(A)$ does not exhibit any cluster structure. A third idea is to output the clustering that maximizes the Kolmogorov–Smirnov distance (or another nonparametric two-sample test statistic) between the empirical CDFs of within-cluster edge weights and between-cluster edge weights. This idea, though feasible, is computationally intractable, since it involves searching over all possible clusterings. Our approach is appreciably different from the methods suggested above, and consists of combining the idea of discretization from nonparametric density estimation with clustering techniques for unweighted SBMs.

4.1. *Outline of algorithm.* We begin by describing the main components of our algorithm. The key ideas are to convert the edge weights into a finite set of labels by discretization, and then cluster nodes on the labeled network. Our algorithm is summarized pictorially in Figure 1.

(1) *Transformation and discretization.* We take a weighted adjacency matrix A and apply an invertible transformation function $\Phi : S \to [0, 1]$ (recall S is the support of the edge weights and can be $[0, 1]$, $[0, \infty)$ or $\mathbb{R}$) on the nonzero edges to obtain a matrix $\Phi(A)$ with weights in $[0, 1]$. Next, we divide the interval $[0, 1]$ into L equally-spaced subintervals and replace the entries of $\Phi(A)$ with categorical labels in $[L]$. We denote the labeled adjacency matrix by A_L.

(2) *Add noise.* We perform the following process on every edge of the labeled graph, independently of other edges: With probability $1 - \delta$ where $\delta = \frac{2(L+1)}{n}$, keep an edge as it is, and with probability δ, replace the edge label with a new label drawn uniformly from the set of labels. We continue to denote the modified adjacency matrix as A_L.

(3) *Initialization parts 1 and 2.* For each label l, we create a sub-network by including only edges of label l. We then perform spectral clustering on all subnetworks, and output the label l^* that induces the maximally separated spectral clustering. Let A_{l^*} be the adjacency matrix for label l^*. For each $u \in \{1, \dots, n\}$, we perform spectral clustering on $A_{l^*} \setminus \{u\}$, which denotes the adjacency matrix with vertex u removed. We output n clusterings $\tilde{\sigma}_1, \dots, \tilde{\sigma}_n$.

(4) *Refinement and consensus.* From each $\tilde{\sigma}_u$, we generate a clustering $\hat{\sigma}_u$ on $\{1, 2, \dots, n\}$ that retains the assignments specified by $\tilde{\sigma}_u$ for $\{1, 2, \dots, n\} \setminus \{u\}$, and assigns $\hat{\sigma}_u(u)$ by maximizing the likelihood taking into account only the neighborhood of u. We then align the cluster assignments made in the previous step.

4.2. *Transformation and discretization.* In the transformation step, we apply an invertible CDF $\Phi : S \to [0, 1]$ as the transformation function on all the edge weights, so that each entry of $\Phi(A)$ lies in $[0, 1]$. In the discretization step, we divide the interval $[0, 1]$ into L equally-spaced bins of the form $[a_l, b_l]$, where $a_1 = 0, b_L = 1$, and $b_l - a_l = \frac{1}{L}$. An edge

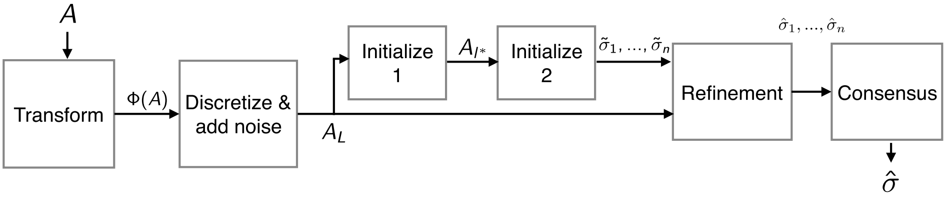


FIG. 1. *Pipeline for our proposed algorithm.*

Algorithm 1 Transformation and discretization

```
Input: A weighted network $A$, a positive integer $L$ and an invertible function $\Phi : S \to [0, 1]$
Output: A labeled network $A_L$ with $L$ labels
  Divide $[0, 1]$ into $L$ equally-spaced bins, labeled $\text{bin}_1, \dots, \text{bin}_L$
  for every edge $(u, v)$ such that $A_{uv} \neq 0$ do
      Let $l$ be the bin in which $\Phi(A_{uv})$ falls
      Assign the edge $(u, v)$ the label $l$ in the labeled network $A_L$
  end for
```

is assigned the label l if the weight of that edge lies in bin l. These steps are sumarized in Algorithm 1.

4.3. *Add noise.* For technical reasons, we inject noise into the network as a form of regularization. As detailed in the proof of Proposition 6.1 in Appendix A, deliberately forming a noisy version of the graph barely affects the separation between the distributions of the within-community and between-community edge labels, but has the desirable effect of ensuring that all edge labels occur with probability at least $\frac{2}{n}$. This property is crucial to our analysis in subsequent steps of the algorithm. In the description of Algorithm 2, we treat the label 0 (i.e., an empty edge) as a separate label, so we have a network with $L + 1$ labels.

4.4. *Initialization.* The initialization procedure takes as input a network with edges labeled $\{0, 1, \dots, L\}$. The goal of the initialization procedure is to create a rough clustering $\tilde{\sigma}$ that is consistent but not necessarily optimal. As outlined in Algorithm 3, the rough clustering is based on a single label l^*, selected based on the maximum value of the estimated Renyi divergence between within-community and between-community distributions for the unweighted SBMs based on individual labels.

For technical reasons, we actually create n separate rough clusterings $\{\tilde{\sigma}_u\}_{u=1,\dots,n}$, where each $\tilde{\sigma}_u : [n-1] \to [K]$ is a clustering of a network of $n - 1$ nodes with u removed. The clusterings $\{\tilde{\sigma}_u\}$ will later be combined into a single clustering algorithm. In practice, it is sufficient to create a single rough clustering (see Remark 4.2 below).

REMARK 4.1. The initialization procedure that we propose is based on choosing a single best label l^* and deriving an initial clustering from the unweighted network associated with l^*. This is sufficient in theory, but a better initial clustering may be gained in practice by aggregating information from all labels. Such an aggregation must, however, be performed with care, so that uninformative labels do not dilute the information content of the informative labels.

Algorithm 2 Add noise

```
Input: A labeled network $A_L$ with $L + 1$ labels
Output: A labeled network $A_L$ with $L + 1$ labels
  for every edge $(u, v)$ do
      With probability $1 - \frac{2(L+1)}{n}$, do nothing
      With probability $\frac{2(L+1)}{n}$, replace the edge label with a label drawn uniformly at random
  from $\{0, 1, 2, \dots, L\}$
  end for
```

Algorithm 3 Initialization

```
Input: A labeled network $A_L$ with $L$ labels
Output: A set of clusterings $\{\tilde{\sigma}_u\}_{u=1,\dots,n}$, where $\tilde{\sigma}_u$ is a clustering on $\{1,2,\dots,n\} \setminus \{u\}$
Separate $A_L$ into $L$ networks $\{A_l\}_{l=1,\dots,L}$, where $A_{l,uv} = 1$ if $A_{L,uv} = l$ and $A_{l,uv} = 0$ otherwise    ▷ Stage 1
for each label $l$ do
    Perform SPECTRAL CLUSTERING (Algorithm 4) with $\tau = 40K\bar{d}$ and $\mu = 4\beta$, where $\bar{d} = \frac{1}{n}\sum_{u=1}^{n} d_u$ is the average degree, to obtain $\tilde{\sigma}_l$
    Estimate $\hat{P}_l = \frac{\sum_{u \neq v: \tilde{\sigma}_l(u) = \tilde{\sigma}_l(v)} (A_l)_{uv}}{|\{u \neq v: \tilde{\sigma}_l(u) = \tilde{\sigma}_l(v)\}|}$ and $\hat{Q}_l = \frac{\sum_{u \neq v: \tilde{\sigma}_l(u) \neq \tilde{\sigma}_l(v)} (A_l)_{uv}}{|\{u \neq v: \tilde{\sigma}_l(u) \neq \tilde{\sigma}_l(v)\}|}$
    Compute $\hat{I}_l \leftarrow \frac{(\hat{P}_l - \hat{Q}_l)^2}{\hat{P}_l \vee \hat{Q}_l}$
end for
Choose $l^* = \arg\max_l \hat{I}_l$
for each node $u$ do    ▷ Stage 2
    Create network $A_{l^*} \setminus \{u\}$ by removing node $u$ from $A_{l^*}$
    Perform SPECTRAL CLUSTERING (with the same parameter setting as stage 1) on $A_{l^*} \setminus \{u\}$ to obtain $\tilde{\sigma}_u$
end for
Output the set of clusterings $\{\tilde{\sigma}_u\}_{u=1,\dots,n}$
```

SPECTRAL CLUSTERING. Note that Algorithm 3 involves several applications of SPECTRAL CLUSTERING. We describe the spectral clustering algorithm used as a subroutine in Algorithm 4 below. Importantly, note that we may always choose the parameter μ sufficiently large such that Algorithm 4 generates a set S with $|S| = K$.

4.5. *Refinement and consensus.* These steps, as outlined in Algorithm 5, parallel Gao et al. [15]. In the refinement step, we use the set of initial clusterings $\{\tilde{\sigma}_u\}_{u=1,\dots,n}$ to generate a more accurate clustering for the labeled network by locally maximizing an approximate log-likelihood for each node u. The consensus step resolves any cluster label inconsistencies present after the refinement stage.

REMARK 4.2. In our simulation studies, we find that it is sufficient to output a single clustering $\tilde{\sigma}$ on the whole of A_{l^*} in the initialization stage. In the refinement stage, we simply estimate $\{\hat{P}_l, \hat{Q}_l\}_{l \in \{0,\dots,L\}}$ based on $\tilde{\sigma}$, assign $\hat{\sigma}(u) = \arg\max_{k \in [K]} \sum_{v: \tilde{\sigma}(v) = k, v \neq u} \sum_{l=0}^{L} \log \frac{\hat{P}_l}{\hat{Q}_l} \times \mathbf{1}(A_{uv} = l)$, and then output $\hat{\sigma}$ directly. We also note that one could in practice use a discretization level for the refinement stage that is different from that of the initialization stage (see discussions in Section 6).

5. Optimal misclustering error.

We analyze the rate of convergence of the estimation algorithm from Section 4 in Section 5.1. In Section 5.2, we provide a matching information-theoretic lower bound. In both sections, we let $\mathcal{P}$ denote the set of probability distributions on S whose singular part is a point mass at 0.

5.1. *Upper bound.* We begin by stating a condition on the function Φ.

DEFINITION 5.1. Let S be $[0,1]$, $\mathbb{R}$, or $\mathbb{R}^+$. We say that $\Phi : S \to [0,1]$ is a *transformation function* if it is a differentiable bijection and $\phi := \Phi'$ satisfies $|\frac{\phi'(x)}{\phi(x)}| < \infty$.

Algorithm 4 SPECTRAL CLUSTERING

Input: An unweighted network A with columns $\{A_u\}$, trim threshold τ, number of communities K and tuning parameter μ

Output: A clustering σ

```
For each node $u$ with degree $d_u \geq \tau$, set $A_u = 0$ and $(A^\top)_u = 0$ to obtain $T_\tau(A)$
Compute $\hat{A} := \arg\min_{\tilde{A}:\text{rank}(\tilde{A})\leq K} \|\tilde{A} - T_\tau(A)\|_2$ by SVD
For each node $u$, index the other nodes by $v_{(1)}, \ldots, v_{(n-1)}$ such that

    $\|\hat{A}_u - \hat{A}_{v_{(1)}}\|_2 \leq \|\hat{A}_u - \hat{A}_{v_{(2)}}\|_2 \leq \cdots \leq \|\hat{A}_u - \hat{A}_{v_{(n-1)}}\|_2,$

and define

    $D(u) := \|\hat{A}_u - \hat{A}_{v_{(\lceil n/\mu K\rceil)}}\|_2$

Initialize $S \leftarrow 0$
Select node $u_1 := \arg\min_u D(u)$ and add $u_1$ to $S$ as $S[1]$
for $i = 2, \ldots, K$ do
    Among all $u$ such that $|D(u)| \leq (1 - \frac{1}{\mu K})$-quantile$\{D(v) : v \in [n]\}$, select

        $u_i = \arg\max_u \min_{v\in\{S[1],\ldots,S[i-1]\}} \|\hat{A}_u - \hat{A}_v\|_2$

    Add $u_i$ to $S$ as $S[i]$
end for
for $u = 1, \ldots, n$ do
    Assign $\sigma(u) = \arg\min_i \|\hat{A}_u - \hat{A}_{S[i]}\|_2$
end for
```

Algorithm 5 Refinement

Input: A labeled network A_L and a set of clusterings $\{\tilde{\sigma}_u\}_{u=1,\ldots,n}$, where $\tilde{\sigma}_u$ is a clustering on the set $\{1, 2, \ldots, n\} \setminus \{u\}$, for each u

Output: A clustering $\hat{\sigma}$ over the whole network

```
for each node $u$ do
    Estimate $\{\hat{P}_l, \hat{Q}_l\}_{l=0,\ldots,L}$ from $\tilde{\sigma}_u$
    Let $\hat{\sigma}_u : [n] \to [K]$, where $\hat{\sigma}_u(v) = \tilde{\sigma}_u(v)$ for all $v \neq u$ and

        $\hat{\sigma}_u(u) = \arg\max_{k\in[K]} \sum_{v:\tilde{\sigma}_u(v)=k, v\neq u} \sum_{l=0}^{L} \log \frac{\hat{P}_l}{\hat{Q}_l} \mathbf{1}(A_{uv} = l)$

end for
Let $\hat{\sigma}(1) = \hat{\sigma}_1(1)$                                  ▷ Consensus Stage
for each node $u \neq 1$ do

    $\hat{\sigma}(u) = \arg\max_{k\in[K]} \left|\{v : \hat{\sigma}_1(v) = k\} \cap \{v : \hat{\sigma}_u(v) = \hat{\sigma}_u(u)\}\right|$

end for
Output $\hat{\sigma}$
```

For $S = [0, 1]$, we always take Φ to be the identity. For $S = \mathbb{R}$ or $[0, \infty)$, we choose the function Φ so that all moments exist and ϕ has a subexponential tail. The specific choice of Φ is not crucial, and we will use the following definitions:

$$\phi(x) = \frac{e^{1-\sqrt{x+1}}}{4} \quad \text{if } S = [0, \infty), \qquad \phi(x) = \frac{e^{1-\sqrt{|x|+1}}}{8} \quad \text{if } S = \mathbb{R}. \tag{3}$$

These expressions are similar to a generalized normal density, modified so that $|\frac{\phi'(x)}{\phi(x)}|$ is bounded. It is easy to verify that $\Phi(x) = \int_0^x \phi(t)\,dt$ (resp., $\Phi(x) = \int_{-\infty}^x \phi(t)\,dt$) is a valid transformation function. We let $\Phi\{\cdot\}$ denote the probability measure induced by Φ.

We describe our regularity conditions by defining an appropriate subset of $\mathcal{P}^2$. For $C \in [1, \infty)$, $c_1, c_2 \in \operatorname{int}(S)$, $r > 2$, and $t \in (2/r, 1)$, we define $\mathcal{G}_{\Phi,C,c_1,c_2,r,t} \subset \mathcal{P}^2$ such that $((P_0, p), (Q_0, q)) \in \mathcal{G}_{\Phi,C,c_1,c_2,r,t}$ if and only if:

A0 We have $\frac{1}{C} \le \frac{1-P_0}{1-Q_0} \le C$ and $\frac{1}{C} \le \frac{P_0}{Q_0} \le C$.

A1 For all x in the interior of S, we have $0 < p(x), q(x) \le C\phi(x)$.

A2 There exists a quasi-convex function $g : S \to [0, \infty)$ such that $g(x) \ge |\log \frac{p(x)}{q(x)}|$ and $\int_S g(x)^r \phi(x)\,dx \le C$.

A3 Denoting $\alpha^2 := \int_S (\sqrt{p(x)} - \sqrt{q(x)})^2\,dx$ and $\gamma(x) := \frac{p(x)-q(x)}{\alpha}$, we have

$$\int_S \left(\frac{\gamma(x)}{p(x)+q(x)}\right)^r \bigl(p(x)+q(x)\bigr)\,dx \le C.$$

A4 There exists a quasi-convex function $h : S \to [0, \infty)$ such that

$$h(x) \ge \frac{1}{\phi(x)} \max\left\{\left|\frac{\gamma(x)}{p(x)+q(x)}\right| \left|\frac{\gamma'(x)}{p(x)+q(x)}\right| \left|\frac{q'(x)}{q(x)}\right|, \left|\frac{p'(x)}{p(x)}\right|\right\}$$

and $\int_S |h(x)|^t \phi(x)\,dx \le C$.

A5 We have[1]

$$(\log p)'(x), (\log q)'(x) \ge (\log \phi)'(x) \quad \text{for all } x < c_1 \quad \text{and}$$
$$(\log p)'(x), (\log q)'(x) \le (\log \phi)'(x) \quad \text{for all } x > c_2.$$

The above conditions depend on the choice of Φ, but it generally suffices to choose Φ such that its derivative ϕ is a heavy-tailed density where all moments exist. In particular, we show in Section 5.1.3 that choosing Φ according to equation (3) allows $\mathcal{G}_\Phi$ to encompass Gaussian, Laplace and other broad classes of densities. We also provide an intuitive discussion of the regularity conditions in Section 5.1.1 below.

We now state our upper bound. For a given $((P_0, p), (Q_0, q)) \in \mathcal{P}^2$ and clustering σ_0, let $A \sim \mathrm{WSBM}(\sigma_0, (P_0, p), (Q_0, q))$.

THEOREM 5.1. *Let $\sigma_0 \in \mathcal{C}(\beta, K)$. Let $C \ge 1$, $c_1, c_2 \in \operatorname{int}(S)$, $r > 2$, and $t \in (2/r, 1)$, and let Φ be a transformation function. Define $\mathcal{G}_\Phi := \mathcal{G}_{\Phi,C,c_1,c_2,r,t}$. Let $\{I_n, I_n'\}_{n\in\mathbb{N}}$ be arbitrary sequences such that $I_n \to 0$ and $nI_n' \to \infty$. Let L_n be a sequence such that $\frac{nI_n'}{L_n \exp(L_n^{2/r})} \to \infty$. Let $\hat{\sigma}_{\Phi,L_n}$ be the algorithm described in Section 4 with transformation function Φ and discretization level L_n. Then there exists a sequence of real numbers $\zeta_n \to 0$ such that*

$$\lim_{n\to\infty} \sup_{\substack{((P_0,p),(Q_0,q))\in\mathcal{G}_\Phi:\\ I_n' \le I((P_0,p),(Q_0,q)) \le I_n}} \mathbb{P}_{\substack{(P_0,p),\\(Q_0,q)}}\biggl\{ l\bigl(\hat{\sigma}_{\Phi,L_n}(A), \sigma_0\bigr)$$
$$\le \exp\biggl(-(1-\zeta_n)\frac{n}{\beta K} I\bigl((P_0, p), (Q_0, q)\bigr)\biggr)\biggr\} = 1.$$

[1] If $S = [0, \infty)$ and g is nondecreasing, we only need $(\log p)'(x), (\log q)'(x) \le (\log \phi)'(x)$ for all $x > c_2$.

Furthermore, if $\frac{nI_n}{\beta K \log n} \leq 1$, *we have*

$$\sup_{\substack{((P_0,p),(Q_0,q)) \in \mathcal{G}_\Phi: \\ I'_n \leq I((P_0,p),(Q_0,q)) \leq I_n}} \mathbb{E}_{(P_0,p),(Q_0,q)}\big[l(\hat{\sigma}_{\Phi,L_n}(A), \sigma_0)\big] \\ \times \exp\biggl((1-\zeta_n)\frac{n}{\beta K} I\big((P_0,p),(Q_0,q)\big)\biggr) \leq 1.$$

We relegate the full proof of Theorem 5.1 to Appendix D.1, but we provide a proof overview in Section 6. Since Theorem 5.1 involves many technical details, we first make a few high-level remarks to illustrate its implications.

REMARK 5.1. It is important to observe that the supremum over $\mathcal{G}_\Phi$ appears *after* the limit. Thus, an equivalent way to understand the theorem is to think of a sequence $((P_{0,n}, p_n), (Q_{0,n}, q_n))$, each term of which is a member of $\mathcal{G}_\Phi$. If $I((P_{0,n}, p_n), (Q_{0,n}, q_n))$ is $o(1)$ but $\omega(L_n \exp(L_n^{2/r}) n^{-1})$, Theorem 5.1 states that $\mathbb{P}\{l(\hat{\sigma}(A), \sigma_0) \leq \exp(-(1 + o(1))\frac{n}{\beta K} I((P_{0,n}, p_n), (Q_{0,n}, q_n)))\} \to 1$. Theorem 5.1 thus applies to the so-called *sparse* setting where $P_0, Q_0 \to 1$. In particular, suppose there are constants $a, b > 0$ such that $P_{0,n} = 1 - \frac{a \log n}{n}$ and $Q_0 = 1 - \frac{b \log n}{n}$. Then Theorem 5.1 states that perfect recovery is achievable if $(\sqrt{a} - \sqrt{b})^2 + \sqrt{ab} \int_S (\sqrt{p_n(x)} - \sqrt{q_n(x)})^2 \, dx > \beta K$; this generalizes the previously known result that perfect recovery for unweighted SBMs when $p = 1 - \frac{a \log n}{n}$ and $q = 1 - \frac{b \log n}{n}$ is possible if $(\sqrt{a} - \sqrt{b})^2 > \beta K$.

REMARK 5.2. The assumption that there exist sequences $I_n \to 0$ and $I'_n = \omega(1/n)$ such that $I'_n \leq I((P_0, p), (Q_0, q)) \leq I_n$ is very mild. As shown by our information-theoretic lower bound (cf. Section 5.2), estimation consistency is impossible if a sequence $I'_n = \omega(1/n)$ such that $I((P_0, p), (Q_0, q)) \geq I'_n$ does not exist. Moreover, we observe that if $I((P_0, p), (Q_0, q)) > \beta K \frac{\log n}{n}$, then $\mathbb{P}(l(\hat{\sigma}(A), \sigma_0) = 0) \to 1$, and we are able to perfectly recover the clustering with high probability. Since the estimation problem is intrinsically easier when $I((P_0, p), (Q_0, q))$ is larger, we expect the same perfect recovery guarantee to hold in the case when $I((P_0, p), (Q_0, q))$ is positively bounded away from 0.

REMARK 5.3. Since $nI'_n \to \infty$, it is always possible to choose a sequence $L_n \to \infty$ satisfying the conditions of the theorem. Note that L_n must grow very slowly to satisfy the condition that $\frac{nI'_n}{L_n \exp(L_n^{2/r})} \to \infty$; indeed, our simulation studies (cf. Section 7) confirm that we should choose the discretization level to be very small in order to achieve good performance. We note that L_n has a second-order effect on the rate and appears in the ζ_n term.

5.1.1. *Additional discussion of the conditions.* It is crucial to note that our algorithm does *not* require prior knowledge of the form of $p(\cdot)$ and $q(\cdot)$; the same algorithm and guarantees apply so long as $((P_0, p), (Q_0, q)) \in \mathcal{G}_{\Phi,C,c_1,c_2,r,t}$ for some universal constants C, c_1, c_2, r and t. To aid the reader, we now provide a brief, nontechnical interpretation of the regularity conditions described above.

Condition A1 is simple; the last part states that ϕ must have a tail at least as heavy as that of $p(\cdot)$ and $q(\cdot)$. Condition A2 requires that the likelihood ratio be integrable. It is analogous to a bounded likelihood ratio condition, but much weaker; we add a mild quasi-convexity constraint for technical reasons related to the analysis of binning. In condition A3, the function $\gamma(\cdot)$ is of constant order in the sense that $\int_S (\frac{\gamma(x)}{p(x)+q(x)})^r (p(x) + q(x)) \, dx \leq C$. Requirements on $\gamma(\cdot)$ translate into convergence statements on $|p - q|$: For instance, an L_∞-bound on

γ implies almost uniform convergence (with respect to Φ) of $|p-q|$ to 0. The integrability condition we impose on $\gamma(\cdot)$ in condition A3 is analogous to an L_∞-bound, but much weaker.

Condition A4 controls the smoothness of the derivatives of $\log p(\cdot)$ and $\log q(\cdot)$. Condition A5 is a mild shape constraint on $p(\cdot)$ and $q(\cdot)$. When $S=\mathbb{R}$, this condition essentially requires $p(\cdot)$ and $q(\cdot)$ to be monotonically increasing in x for $x\to-\infty$, and decreasing in x for $x\to\infty$.

5.1.2. *Examples for* $S=[0,1]$. When $S=[0,1]$, we can always take Φ to be the identity—we do not need a transformation, but we keep the same notation in order to present our results in a unified manner. The simplest example of $\mathcal{G}_\Phi$ that satisfies conditions A1–A5 is when, for all $((P_0,p),(Q_0,q))\in\mathcal{G}_\Phi$, the densities $p(\cdot)$ and $q(\cdot)$ are bounded above and below by strictly positive universal constants, and when the function $x\mapsto\frac{p(x)-q(x)}{\alpha}$ and its derivative are bounded by universal constants.

5.1.3. *Examples for* $S=\mathbb{R}$ *or* $[0,\infty)$. We begin with a proposition that characterizes conditions A1–A5 in the setting where $p(\cdot)=e^{f_{\theta_1}(\cdot)}$ and $q(\cdot)=e^{f_{\theta_0}(\cdot)}$, for some parametrized family $\{f_\theta\}_{\theta\in\Theta}$. This result allows us to generate several large classes of examples.

PROPOSITION 5.1. *Let* $C^{**}\in[1,\infty)$, $c_1,c_2\in S$, $r>2$, *and* $t\in(2/r,1/2)$. *Let* $\Theta\subset\mathbb{R}^d$ *be compact and suppose diam*$(\Theta)<1\wedge\frac{1}{2C^{**2}}$. *Let* $\{f_\theta\}_{\theta\in\Theta}$ *be a collection of functions such that* $e^{f_\theta(\cdot)}$ *is a density and*:

B1 *For all* $\theta\in\Theta$ *and all* $x\in S$, *we have* $0<e^{f_\theta(x)}\le C^*\phi(x)$.

B2 *We have* $\inf_{\theta\in\Theta}\lambda_{\min}(\int_S 2\nabla f_\theta(x)(\nabla f_\theta(x))^\top\phi(x)\,dx)\ge C^{*-1}$ *and* $\sup_{\theta\in\Theta}\int_S\lambda_{\max}(H(f_\theta)(x))^2\phi(x)\,dx\le C^*$.

B3 *There exists a quasi-convex function* $g^*:S\to[0,\infty)$ *such that* $g^*(x)\ge\sup_\theta\|\nabla f_\theta(x)\|_2$ *and* $\int_S g^*(x)^r\phi(x)\,dx\le C^*$.

B4 *There exists a quasi-convex function* $h^*:S\to[0,\infty)$ *such that*

$$h^*(x)\ge\frac{1}{\phi(x)}\max\Big\{\sup_{\theta\in\Theta}\|\nabla f_\theta(x)\|_2,\sup_{\theta\in\Theta}\|\nabla f'_\theta(x)\|_2,\sup_{\theta\in\Theta}|f'_\theta(x)|\Big\}$$

and $\int_S h^*(x)^{2t}\phi(x)\,dx\le C^*$.

B5 *For all* $x\le c_1$, *we have* $\inf_{\theta\in\Theta}f'_\theta(x)\ge(\log\phi)'(x)$, *and for all* $x\ge c_2$, *we have* $\sup_{\theta\in\Theta}f'_\theta(x)\le(\log\phi)'(x)$.[2]

Then there exists $C\in[1,\infty)$ *such that for any* $\theta_1,\theta_2\in\Theta$ *and any* $P_0,Q_0\in[0,1]$ *such that* $\frac{1}{C}\le\frac{P_0}{Q_0},\frac{1-P_0}{1-Q_0}\le C$, *we have* $((P_0,e^{f_{\theta_1}}),(Q_0,e^{f_{\theta_2}}))\in\mathcal{G}_{\Phi,C,c_1,c_2,r,t}$.

In all the examples below, we take Φ to be the transformation function defined in equation (3). The proofs of all statements in the examples are provided in Section E.2.

EXAMPLE 5.1 (Location-scale family over $\mathbb{R}$). Let $f:\mathbb{R}\to\mathbb{R}$ be a continuously differentiable function such that $\int_{-\infty}^{\infty}e^{f(x)}\,dx=1$. Suppose:

(a) $|f^{(k)}(x)|$ is bounded for some $k\ge2$, and

(b) there exist $c,M>0$ such that $f'(x)>M$ for $x<-c$ and $f'(x)<-M$ for $x>c$.

For any $\mu\in\mathbb{R}$ and $\sigma>0$, define $f_{\mu,\sigma}(x):=f(\frac{x-\mu}{\sigma})-\log\sigma$.

Then there exists $C_\mu>0$ and $c_\sigma>1$ such that, with $\Theta:=[-C_\mu,C_\mu]\times[\frac{1}{c_\sigma},c_\sigma]$, the family $\{f_{\mu,\sigma}\}_{(\mu,\sigma)\in\Theta}$ satisfies conditions B1–B5 in Proposition 5.1 with respect to ϕ defined

[2] If $S=[0,\infty)$ and g^* is nondecreasing, we only need $\sup_{\theta\in\Theta}f'_\theta(x)\le(\log\phi)'(x)$ for all $x\ge c_2$.

in equation (3), and some universal constants C^{**}, c_1, c_2, r, and t. As a direct consequence of Proposition 5.1, for some universal constant $C > 0$, if we fix any $((\mu_1, \sigma_1), (\mu_0, \sigma_0)) \in \Theta^2$ and define

$$p(x) = \frac{1}{\sigma_1} \exp\left(f\left(\frac{x - \mu_1}{\sigma_1}\right)\right) \quad \text{and} \quad q(x) = \frac{1}{\sigma_0} \exp\left(f\left(\frac{x - \mu_0}{\sigma_0}\right)\right), \tag{4}$$

then $((P_0, p), (Q_0, q)) \in \mathcal{G}_{\Phi,C,c_1,c_2,r,t}$ for any $P_0, Q_0 \in [0, 1]$ that satisfy condition A0.

These assumptions on f are satisfied for *Gaussian location-scale families*, where the base density is the standard Gaussian density with $f(x) = -x^2 - \frac{1}{2}\log 2\pi$, and *Laplace location-scale families*, where the base density is the standard Laplace density with $f(x) = -|x| - \log 2$.

EXAMPLE 5.2 (Scale family over $[0, \infty)$). Let $f : [0, \infty) \to \mathbb{R}$ be a continuously differentiable function such that $\int_0^\infty e^{f(x)}\, dx = 1$. Suppose:

(a) $|f^{(k)}(x)|$ is bounded for some $k \geq 2$, and
(b) there exist $c, M > 0$ such that $f'(x) < -M$ for $x > c$.

For any $\sigma > 0$, define $f_\sigma(x) := f(\frac{x}{\sigma}) - \log \sigma$.

Then there exists $c_\sigma > 1$ such that, with $\Theta := [\frac{1}{c_\sigma}, c_\sigma]$, the family $\{f_\sigma\}_{\sigma \in \Theta}$ satisfies conditions B1–B5 in Proposition 5.1 with respect to ϕ defined in equation (3), and some universal constants C^{**}, c_1, c_2, r and t. As a direct consequence of Proposition 5.1, for some universal constant $C > 0$, if we fix any $(\sigma_1, \sigma_0) \in \Theta^2$ and define

$$p(x) = \frac{1}{\sigma_1} \exp\left(f\left(\frac{x}{\sigma_1}\right)\right) \quad \text{and} \quad q(x) = \frac{1}{\sigma_0} \exp\left(f\left(\frac{x}{\sigma_0}\right)\right),$$

then $((P_0, p), (Q_0, q)) \in \mathcal{G}_{\Phi,C,c_1,c_2,r,t}$ for any $P_0, Q_0 \in [0, 1]$ that satisfy condition A0.

These assumptions on f are satisfied for *exponential scale families*, where the base density is the standard exponential density with $f(x) = -x$.

Proposition 5.1 also applies to the family of Gamma distributions (see Proposition E.3 in the Appendix [40]). In practice, edge weights are often discrete integers, such as counts. Although Theorem 5.1 does not apply directly to such cases, our analysis is relevant to some instances of SBMs with discrete edge weights. In Appendix F, we discuss a crude way to handle count-valued edge weights, with particular attention toward Poisson-distributed edge weights.

5.2. *Lower bound.*

Our information-theoretic lower bound applies to any permutation equivariant estimators (Definition 2.5). Before stating the result, we define an appropriate subset of $\mathcal{P}^2$ to capture the conditions we need on $((P_0, p), (Q_0, q))$. Let $C^* \in [1, \infty)$, and let $\mathcal{G}^*_{C^*} \subset \mathcal{P}^2$ be such that $((P_0, p), (Q_0, q)) \in \mathcal{G}^*$ if and only if

A0* $\frac{1}{C^*} \leq \frac{P_0}{Q_0} \leq C^*$, and
A1* $\int_S (p(x) + q(x)) |\log \frac{p(x)}{q(x)}|^2\, dx \leq C^* \int_S (p(x)^{1/2} - q(x)^{1/2})^2\, dx$.

Condition A1* is similar to A2 and A3 in the definition of the set of regular distributions $\mathcal{G}_{\Phi,C,c_1,c_2,r,t}$ that appears in the upper bound (Theorem 5.1). In fact, if $\int_S (p^{1/2}(x) - q^{1/2}(x))^2\, dx$ is bounded away from 0, then there exists C^* such that A1* is equivalent to A2. Thus, although $\mathcal{G}^*_{C^*}$ is in general not a superset of $\mathcal{G}_{\Phi,C,c_1,c_2,r,t}$, the set $\mathcal{G}^*_{C^*} \cap \mathcal{G}_{\Phi,C,c_1,c_2,r,t}$ contains important and interesting examples. For instance, any family that satisfies the conditions of Proposition 5.1 belongs to the intersection, as is verified in the proof (cf. Appendix E).

THEOREM 5.2. *Let $C^* \geq 1$ and let $\sigma_0 : [n] \to [K]$ be a clustering such that one cluster is of size $\frac{n}{\beta K}$ and another is of size $\frac{n}{\beta K} + 1$. Let I_n' be any sequence such that $nI_n' \to \infty$, and let $C = 2\log 2$. Then there exists $\zeta_n \to 0$ and $c' > 0$ such that, for any permutation equivariant algorithm $\hat{\sigma}$*,

$$\inf_{\substack{((P_0,p),(Q_0,q)) \in \mathcal{G}^*_{C*} \\ I_n' \leq I((P_0,p),(Q_0,q)) \leq C}} \mathbb{E}_{\substack{(P_0,p) \\ (Q_0,q)}}\big[\ell(\hat{\sigma}(A), \sigma_0)\big] \times \exp\biggl(\frac{n}{\beta K} I\big((P_0, p), (Q_0, q)\big)(1 + \zeta_n)\biggr) \geq c'.$$

Furthermore, for any $c > 0$, there exists $c' > 0$ such that for any permutation equivariant algorithm $\hat{\sigma}$,

$$\inf_{\substack{((P_0,p),(Q_0,q)) \in \mathcal{G}^*_{C*} \\ I((P_0,p),(Q_0,q)) \leq c/n}} \mathbb{E}_{\substack{(P_0,p) \\ (Q_0,q)}}\big[\ell(\hat{\sigma}(A), \sigma_0)\big] \geq c'.$$

Theorem 5.2 shows that if $nI_n \to \infty$, the misclustering risk of any permutation equivariant algorithm is at least $\exp(-(1 + o(1))\frac{nI((P_0,p),(Q_0,q))}{\beta K})$. If $nI((P_0, p), (Q_0, q)) = O(1)$, no permutation equivariant algorithm is consistent.

REMARK 5.4. Rather than being a minimax lower bound that applies to the worst case, Theorem 5.2 applies to *any* parameter $((P_0, q), (Q_0 q)) \in \mathcal{G}^*_{C^*}$; we thus have an infimum over the parameter space rather than a supremum. This is possible because the permutation equivariance condition excludes the trivial case where $\hat{\sigma} = \sigma_0$.

The full proof of Theorem 5.2 is provided in Appendix G. The proof borrows elements from Yun and Proutiere [41] and Zhang and Zhou [42]. One key difference is that Theorem 5.2 holds for any parameters in the parameter space, rather than adopting a minimax framework, as in Zhang and Zhou [42], or assuming a prior on σ_0, as in Yun and Proutiere [41].

5.3. *Adaptivity.* Let $\mathcal{F}_n^{p.e.}$ be the class of permutation equivariant clustering algorithm on networks with n nodes. Theorems 5.1 and 5.2 directly imply the following corollary, which sharply characterizes the optimal performance of $\mathcal{F}_n^{p.e.}$.

COROLLARY 5.1. *Let $\sigma_0 : [n] \to [K]$, and suppose one cluster is of size $\frac{n}{\beta K}$ and another is of size $\frac{n}{\beta K} + 1$. Let $C^*, C \geq 1, c_1, c_2 > 0, r > 0$, and $t \in (2/r, 1)$, and let Φ be a transformation function. Write $\mathcal{G}_\Phi := \mathcal{G}_{\Phi,C,c_1,c_2,r,t}$, $\mathcal{G}^* := \mathcal{G}^*_{C^*}$ and $\Lambda := \{\beta, K, C^*, C, c_1, c_2, r, t, \Phi\}$. Let $((P_{0,n}, p_n), (Q_{0,n}, q_n)) \in \mathcal{G}_\Phi \cap \mathcal{G}^*$ for every $n \in \mathbb{N}$.*

(i) *If $\limsup_n I((P_{0,n}, p_n), (Q_{0,n}, q_n))\frac{n}{\beta K \log n} \leq 1$, there exists $\zeta_n \to 0$, depending only on Λ, such that*

$$\inf_{\hat{\sigma} \in \mathcal{F}_n^{p.e.}} \mathbb{E}_{\substack{(P_{0,n},p_n) \\ (Q_{0,n},q_n)}}\big[l(\hat{\sigma}(A), \sigma_0)\big] = \exp\biggl(-\frac{nI((P_{0,n}, p_n), (Q_{0,n}, q_n))}{\beta K}(1 + \zeta_n)\biggr).$$

(ii) *If $\liminf_n I((P_{0,n}, p_n), (Q_{0,n}, q_n))\frac{n}{\beta K \log n} > 1$, there exists $\zeta_n \to 0$, depending only on Λ, such that $\inf_{\hat{\sigma} \in \mathcal{F}_n^{p.e.}} \mathbb{P}_{(P_{0,n},p_n),(Q_{0,n},q_n)}(l(\hat{\sigma}(A), \sigma_0) > 0) \leq \zeta_n$.*

(iii) *If there exists* $c > 0$ *such that* $\limsup_n I((P_{0,n}, p_n), (Q_{0,n}, q_n))n < c$, *there exists* $c' > 0$ *such that*

$$\liminf_{n\to\infty} \inf_{\hat{\sigma}\in\mathcal{F}_n^{p.e.}} \mathbb{E}_{(P_{0,n},p_n),(Q_{0,n},q_n)}\big[l\big(\hat{\sigma}(A), \sigma_0\big)\big] > c'.$$

The algorithm $\hat{\sigma}$ described in Section 4.1 with discretization level L_n diverging sufficiently slowly achieves the optimal rate in part (i) and (ii) for *any* $((P_{0,n}, p_n), (Q_{0,n}, q_n)) \in \mathcal{G}_\Phi \cap \mathcal{G}^*$. Thus, $\hat{\sigma}$ adapts to the edge probabilities $P_{0,n}$ and $Q_{0,n}$ and the edge weight densities p_n and q_n: Although $\hat{\sigma}$ has no knowledge of the parameters $((P_{0,n}, p_n), (Q_{0,n}, q_n))$, it achieves the same optimal rate as if $((P_{0,n}, p_n), (Q_{0,n}, q_n))$ were known.

In particular, this implies that one does not have to pay a price for taking the nonparametric approach. This seemingly counterintuitive phenomenon arises because the cost of discretization is reflected in the lower-order ζ_n term in the exponent. As an illustrative example, suppose $1 - P_{0,n} = 1 - Q_{0,n} = a\frac{\log n}{n}$ for some $a > 0$, and the densities p_n and q_n are of $N(\mu_1, \sigma_1^2)$ and $N(\mu_0, \sigma_0^2)$, respectively. Then $I_n = (1 + o(1))\frac{a\log n}{n}\theta$, where $\theta = 2(1 - \sqrt{\frac{2\sigma_1^2\sigma_0^2}{\sigma_1^2+\sigma_0^2}}e^{-\frac{1}{4}\frac{(\mu_1-\mu_0)^2}{\sigma_1^2+\sigma_0^2}})$, and the optimal rate is $n^{-(1+o(1))\frac{2\theta}{\beta K}}$, which is attained by the nonparametric discretization estimator $\hat{\sigma}$.

Similarly, if $1 - P_{0,n} = 1 - Q_{0,n} = \frac{a\log n}{n}$ and the densities p_n and q_n are $\mathrm{Exp}(\lambda_1)$ and $\mathrm{Exp}(\lambda_0)$, respectively, then $I_n = (1 + o(1))\frac{\log n}{n}\theta'$, where $\theta' = 2(1 - \sqrt{\frac{\lambda_1\lambda_0}{\lambda_1+\lambda_0}})$. The optimal rate $n^{-(1+o(1))\frac{2\theta'}{\beta K}}$ is again achieved by the nonparametric discretization estimator $\hat{\sigma}$.

6. Proof sketch: Recovery algorithm. A large portion of the Appendix is devoted to proving that our recovery algorithm succeeds and achieves the optimal error rates. We provide an outline of the proofs here.

We divide our argument into propositions that focus on successive stages of our algorithm. A birds-eye view of our method reveals that it contains two major components: (1) convert a weighted network into a labeled network, and then (2) run a community recovery algorithm on the labeled network. The first component involves two steps, transformation and discretization. Step (1) comprises the red and green steps in Figure 2 and outputs an adjacency matrix with discrete edge weights. Step (2) is denoted in blue.

In our algorithm, we use a single discretization level L throughout for ease of presentation. In practice, one could use different discretization levels for the initialization stage and for the refinement stage. By comparing Proposition 6.1, Proposition 6.2 and Theorem 5.1, we can see that the bias introduced by discretization is a second-order effect compared to the variance, which is why the discretization level should be small in both stages. The discretization level for the initialization stage can, however, be chosen to be larger than that of the refinement stage, because the initialization stage aims to produce a consistent estimator

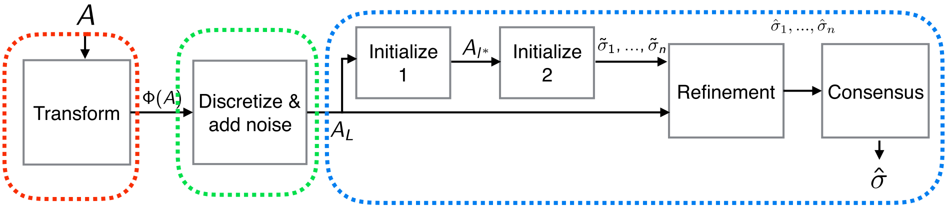


FIG. 2. *Analysis of the right-most blue region is contained in Section* 6.1, *of the middle green region in Section* 6.2, *and of the red region in Section* 6.3.

rather than an optimal one, and can thus tolerate greater variance. More precisely, the theoretical requirements on discretization for the initialization stage are $L \to \infty$ and $\frac{nI_n'}{L} \to \infty$, whereas the requirements for the refinement stage are $L \to \infty$ and $\frac{nI_n'}{Le^{L^{r/2}}} \to \infty$ (note that I_n' is defined in Theorem 5.1); L is required to be of smaller order to control the ratio $\frac{P_l}{Q_l}$ of the discretized probabilities.

6.1. *Analysis of community recovery on a labeled network.* We first examine the second component of our algorithm, which is a subroutine (right-most region in Figure 1) for recovering communities in a network where the edges have discrete labels $l = 1, \dots, L_n$. The following proposition characterizes the rate of convergence of the output of the subroutine, where within-community edges are assigned edge labels with probabilities $\{P_l\}$, and between-community edges are assigned edge labels according to $\{Q_l\}$. For convenience, if an edge does not exist between u and v, we assign the label 0 to A_{uv}, so P_0 and Q_0 are the edge absence probabilities.

Formally, for $L \in \mathbb{N}$, define $\mathcal{P}_L := \{(P_0, \dots, P_L) \in [0,1]^{L+1} : \sum_{l=1}^{L} P_l = 1\}$. For a clustering $\sigma_0 : [n] \to [K]$ and $(\{P_l\}, \{Q_l\}) \in \mathcal{P}_L^2$, we define a labeled stochastic block model $\mathrm{LSBM}(\sigma_0, \{P_l\}, \{Q_l\})$ as a distribution on $\{0, \dots, L\}^{n \times n}$ such that if $A \sim \mathrm{LSBM}(\sigma_0, \{P_l\}, \{Q_l\})$, then for any $u, v \in [n]$ such that $u > v$,

$$A_{uv} \sim \begin{cases} \{P_l\} & \text{if } \sigma_0(u) = \sigma_0(v), \\ \{Q_l\} & \text{if } \sigma_0(u) \neq \sigma_0(v). \end{cases}$$

For $\rho > 1$, let $\mathcal{G}_{L,\rho} \subset \mathcal{P}_L^2$ be such that $(\{P_l\}, \{Q_l\}) \in \mathcal{G}_{L,\rho}$ if and only if $\frac{1}{\rho} \le \frac{P_l}{Q_l} \le \rho$ for all $l = 0, \dots, L$. For a pair $(\{P_l\}, \{Q_l\}) \in \mathcal{P}_L$, we define $I(\{P_l\}, \{Q_l\}) := -2 \log \sum_{l=0}^{L} \sqrt{P_l Q_l}$.

In the next proposition, for a given clustering σ_0 and $(\{P_l\}, \{Q_l\}) \in \mathcal{P}_L^2$, we let the random network A have the distribution $\mathrm{LSBM}(\sigma_0, \{P_l\}, \{Q_l\})$.

PROPOSITION 6.1. *Let $\sigma_0 \in \mathcal{C}(\beta, K)$. Let $\{I_n, I_n', \rho_n, L_n\}_{n \in \mathbb{N}}$ be any sequences such that $I_n \to 0$, $\rho_n \ge 2$, $L_n \ge 1$, and $\frac{nI_n'}{(L_n+1)\rho_n^2 \log \rho_n} \to \infty$. Then there exists a sequence $\zeta_n \to 0$ such that*

$$\lim_{n\to\infty} \sup_{\substack{(\{P_l\},\{Q_l\}) \in \mathcal{G}_{L_n,\rho_n} \\ I_n' \le I(\{P_l\},\{Q_l\}) \le I_n}} \mathbb{P}_{(\{P_l\},\{Q_l\})}\Big(l(\hat{\sigma}(A), \sigma_0) \\ \le \exp\Big(-(1-\zeta_n)\frac{n}{\beta K} I(\{P_l\}, \{Q_l\})\Big)\Big) = 1.$$

Furthermore, if $\frac{nI_n}{\beta K \log n} \le 1$, then

$$\sup_{\substack{(\{P_l\},\{Q_l\}) \in \mathcal{G}_{L_n,\rho_n} \\ I_n' \le I(\{P_l\},\{Q_l\}) \le I_n}} \mathbb{E}\big[l(\hat{\sigma}(A), \sigma_0)\big] \exp\Big((1-\zeta_n)\frac{n}{\beta K} I(\{P_l\}, \{Q_l\})\Big) \le 1.$$

REMARK 6.1. This result resembles that of Yun and Proutiere [41], who also study an SBM where the edges carry discrete labels. They state their results using a seemingly different divergence, but it coincides with the Renyi divergence when specialized to our setting (cf. Lemma G.2). Proposition 6.1 differs critically from Yun and Proutiere [41] in two respects, however. First, they hold the number of labels L_n to be fixed and assume that the bound ρ_n on the probability ratio $\frac{P_{l,n}}{Q_{l,n}}$ is fixed, whereas we allow both L_n and ρ_n to diverge. Second, they assume that $\sum_{l=1}^{L_n} (P_{l,n} - Q_{l,n})^2$ is sufficiently large when compared to $\max_{l=1,\dots,L_n} P_{l,n}$, whereas we do not make any assumptions of this form. These generalizations are crucial in analyzing the weighted SBM, since in order to achieve consistency for continuous distributions, the discretization level L_n and the bound ρ_n must increase with n.

6.2. *Discretization of the Renyi divergence.* We now analyze the discretization step of the algorithm (green box in Figure 1). The input to this step is the weighted network $\Phi(A)$ in which all the edge weights are in $[0, 1]$. We use $\tilde{p}(z)$ and $\tilde{q}(z)$ for $z \in [0, 1]$ to denote the densities of the transformed edge weights; the next section shows the relationship between $\tilde{p}(z)$ and $p(x)$ and $\tilde{q}(z)$ and $q(x)$. The discretization step of the algorithm divides $[0, 1]$ into L_n uniform bins, denoted by $[a_l, b_l]$, for $1 \le l \le L_n$. The output is a network A_{L_n}, where each edge is assigned label $l = 1, \dots, L_n$ with probability either

$$P_l := (1 - P_0) \int_{a_l}^{b_l} \tilde{p}(z)\, dz \quad \text{or} \quad Q_l := (1 - Q_0) \int_{a_l}^{b_l} \tilde{q}(z)\, dz. \tag{5}$$

A missing edge is assigned the label 0. It is easy to show that discretization always leads to a loss of information; that is, $I(\{P_l\}, \{Q_l\}) \le I((P_0, \tilde{p}), (Q_0, \tilde{q}))$.

Let $\tilde{\mathcal{P}}$ denote the set of probability distributions on $[0, 1]$ whose singular part is a point mass at 0. Let $\tilde{C} \in (0, \infty)$, $\tilde{c}_1, \tilde{c}_2 \in (0, 1/2)$, $r > 2$, and $t > 0$, and define the set $\tilde{\mathcal{G}}_{\tilde{C},\tilde{c}_1,\tilde{c}_2,r,t} \subset \tilde{\mathcal{P}}^2$ such that $((P_0, \tilde{p}), (Q_0, \tilde{q})) \in \tilde{\mathcal{G}}$ if and only if the following hold:

C0 We have $\frac{1}{\tilde{C}} \le \frac{1-P_0}{1-Q_0} \le \tilde{C}$ and $\frac{1}{\tilde{C}} \le \frac{P_0}{Q_0} \le \tilde{C}$.

C1 For all $z \in (0, 1)$, we have $0 < \tilde{p}(z), \tilde{q}(z) \le \tilde{C}$.

C2 There exists a quasi-convex function $\tilde{g} : [0, 1] \to [0, \infty)$ such that $\tilde{g}(z) \ge |\log \frac{\tilde{p}(z)}{\tilde{q}(z)}|$ and $\int_0^1 \tilde{g}(z)^r\, dz \le \tilde{C}$.

C3 Denoting $\tilde{\alpha} := \{\int_0^1 (\sqrt{\tilde{p}(z)} - \sqrt{\tilde{q}(z)})^2\, dz\}^{1/2}$ and $\tilde{\gamma}(z) := \frac{\tilde{p}(z) - \tilde{q}(z)}{\tilde{\alpha}}$, we have

$$\int_0^1 \left\{ \frac{\tilde{\gamma}(z)}{\tilde{p}(z) + \tilde{q}(z)} \right\}^r (\tilde{p}(z) + \tilde{q}(z))\, dz \le \tilde{C}.$$

C4 There exists a quasi-convex function $\tilde{h} : [0, 1] \to [0, \infty)$ such that

$$\tilde{h}(z) \ge \max\left\{ \left| \frac{\tilde{\gamma}(z)}{\tilde{p}(z) + \tilde{q}(z)} \right|, \left| \frac{\tilde{p}'(z)}{\tilde{p}(z)} \right|, \left| \frac{\tilde{q}'(z)}{\tilde{q}(z)} \right|, \left| \frac{\tilde{\gamma}'(z)}{\tilde{p}(z) + \tilde{q}(z)} \right| \right\}$$

and $\int_0^1 \tilde{h}(z)^t\, dz < \tilde{C}$.

C5 We have $\tilde{p}'(z), \tilde{q}'(z) \ge 0$ for all $z < \tilde{c}_1$, and $\tilde{p}'(z), \tilde{q}'(z) \le 0$ for all $z > 1 - \tilde{c}_2$.[3]

PROPOSITION 6.2. *Let $\tilde{C} \in (0, \infty)$, $\tilde{c}_1, \tilde{c}_2 \in (0, 1/2)$, $r > 2$ and $t > 0$. For any $((P_0, \tilde{p}), (Q_0, \tilde{q})) \in \tilde{\mathcal{G}}_{\tilde{C},\tilde{c}_1,\tilde{c}_2,r,t}$, for any $L \in \mathbb{N}$ such that $L \ge \tilde{c}_1^{-1} \vee \tilde{c}_2^{-1}$, and for $\{P_l, Q_l\}$ defined in equation (5), we have*

$$\frac{1}{2\tilde{C} \exp((2\tilde{C}L)^{1/r})} \le \frac{P_l}{Q_l} \le 2\tilde{C} \exp\bigl((2\tilde{C}L)^{1/r}\bigr),$$

for all $l \in \{0, \dots L\}$. Furthermore,

$$\lim_{L \to \infty} \sup_{((P_0,\tilde{p}),(Q_0,\tilde{q})) \in \tilde{\mathcal{G}}_{\tilde{C},\tilde{c}_1,\tilde{c}_2,r,t}} \left| 1 - \frac{I(\{P_l\}, \{Q_l\})}{I((P_0, \tilde{p}), (Q_0, \tilde{q}))} \right| = 0.$$

We prove Proposition 6.2 in Appendix C.

6.3. *Analysis of the transformation function.* Proposition 6.2 considers densities supported on $[0, 1]$. In conjunction with Proposition 6.1, this suffices to obtain Theorem 5.1,

[3] If $\tilde{g}$ is nondecreasing, we need only $\tilde{p}'(z), \tilde{q}'(z) \le 0$ for all $z > 1 - c'_2$.

because the densities of the transformed edge weights are compactly supported and, importantly, the Renyi divergence is invariant with respect to the transformation function Φ.

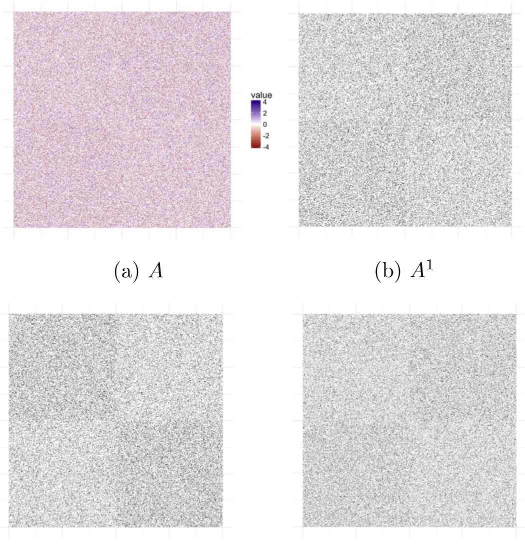
To be precise, let $p(\cdot)$ and $q(\cdot)$ denote probability densities on S, and for $X \sim p$ and $Y \sim q$, let $\tilde{p}(\cdot)$ and $\tilde{q}(\cdot)$ denote the densities of $\Phi(X)$ and $\Phi(Y)$. We then have $\tilde{p}(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $\tilde{q}(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ for $z \in [0, 1]$. Therefore, via the change of variable $z = \Phi^{-1}(x)$, we have

$$\int_S \sqrt{p(x)q(x)}\,dx = \int_0^1 \sqrt{\tilde{p}(z)\tilde{q}(z)}\,dz \quad \text{and}$$
$$I\big((P_0, p), (Q_0, q)\big) = I\big((P_0, \tilde{p}), (Q_0, \tilde{q})\big).$$

7. Simulation studies. We start with a toy example that illustrates the intuition behind our discretization-based algorithm. In this example, we have $n = 1000$ nodes, $K = 2$ clusters and $P_0 = Q_0 = 0.5$. We also set $p(\cdot)$ and $q(\cdot)$ as the normal density $N(0, 1.3^2 + 1)$ and mixture of normals $\frac{1}{2}N(-1.3, 1) + \frac{1}{2}N(1.3, 1)$, respectively. Observe that $\int_{\mathbb{R}} x\,dP = \int_{\mathbb{R}} x\,dQ = 0$ and $\int_{\mathbb{R}} x^2\,dP = \int_{\mathbb{R}} x^2\,dQ = \frac{1}{2}(1.3^2 + 1)$. The true clustering σ_0 maps the first 500 nodes to cluster 1 and the rest to cluster 2.

In Figure 3(a), we generate a random weighted network A and display the adjacency matrix *without randomly permuting the rows and columns*. It is difficult to discern the block structure because (P_0, p) and (Q_0, q) have equal mean and variance. In Figures 3(b), 3(c) and 3(d), we discretize A using the transformation $\Phi(x) = \int_{-\infty}^{x} \frac{1}{4}e^{-|t|/2}\,dt$ and $L = 3$ bins and show the discretized networks A^1, A^2 and A^3; recall that A^1 is a binary adjacency matrix, where $A^1_{uv} = 1$ if $A_{uv} \neq 0$ and $\phi(A_{uv}) \in [0, 1/3)$, and $A^1_{uv} = 0$ otherwise, and likewise for A^2 and A^3. We observe that the block structure is clearly distinguishable in A^2 because the densities $p(\cdot)$ and $q(\cdot)$ differ most around the origin; the block structure is somewhat visible in A^1 and A^3, but to a lesser extent. These figures illustrate why the discretization and initialization stages are useful.

In Figure 4(a), we test how the performance of our algorithm scales with the network size n. We use the same setting as our first simulation, except we let $n \in \{400, 600, 800, \dots, 2000\}$ and $L_n = \lfloor 0.4(\log(\log n))^4 \rfloor$. For each value of n, we perform 100 trials, where we generate



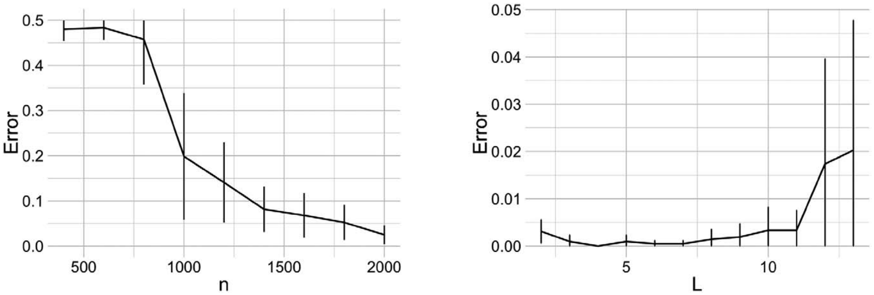
(a) A

(b) A^1

(c) A^2

(d) A^3

FIG. 3. *Effect of discretization on a weighted network.*



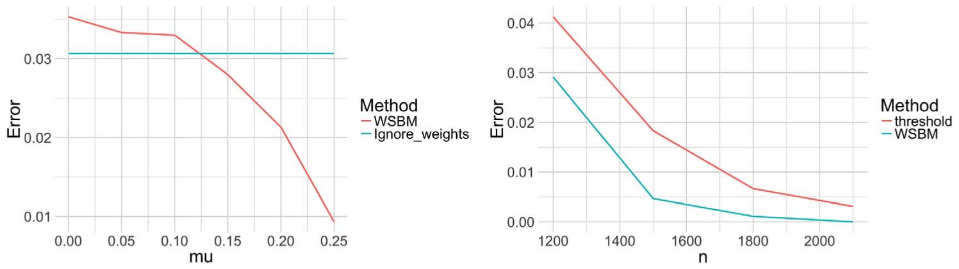
(a) Misclustering error versus n. (b) Misclustering error versus L.

FIG. 4. *Misclustering errors in simulation experiments.*

a random network A, perform our clustering algorithm, and calculate the misclustering error. The misclustering errors are averaged across the 100 random trials; the aggregated medians are shown, with deviations, in Figure 4(a). In Figure 4(a), we observe the same threshold behavior that arises in the unweighted setting: the misclustering error is around 0.5—equivalent to random guessing—for small n, and drops sharply to 0 as the value of n passes a threshold (around $n = 1000$ in this case). We note that for this and our next simulation study, we use a simplified version of our algorithm as described in Remark 4.2; we observed no difference in performance between the full version and the simplified version of the algorithm.

In Figure 4(b), we study the sensitivity of our algorithm to the choice of discretization level L. We let $K = 3$, $n = 2100$, $P_0 = 0.3$ and $Q_0 = 0.27$, and let $p(\cdot)$ be the density of $N(0.3, 0.8^2)$, and $q(\cdot)$ be the density of $N(0, 1)$. We let $L \in \{1, 2, 3, \dots, 12, 13\}$ and, for each setting of L, we perform 100 random trials in which we generate a random network A, perform our clustering algorithm, and calculate the misclustering error. The results are shown in Figure 4(b); the error for $L = 1$, in which we discard the edge weights, exceeds 0.56 and is thus omitted from the plot. We observe that the algorithm performs best when L is chosen to be small, though not too small, as is suggested by our theoretical analysis.

In Figure 5(a), we compare our approach against treating a weighted network as an unweighted one by discarding the edge weights. In this setting, we let $n = 1500$, $P_0 = 0.3$, $Q_0 = 0.23$ and $K = 3$. We choose $q(\cdot)$ as the density of $N(0, 1)$ and $p(\cdot)$ as the density of $N(\mu, 1)$ where we let $\mu \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25\}$. We perform 100 trials and aggregate the result in Figure 5(a). In red, we plot the misclustering error incurred by our WSBM clustering algorithm with $L = 5$; in blue, we plot the misclustering error incurred by ignoring the edge weights entirely and treating the network as an unweighted one. As we expect, when μ is close to 0, the edge weights are uninformative and it is better to ignore the edge weights. As μ increases, however, the advantage of using the weights become significant.



(a) (b)

FIG. 5. *Comparison against naive approaches.*

In Figure 5(b), we compare our algorithm against clustering an unweighted network formed by optimally thresholding the edge weights. We let $K = 3$, $P_0 = 0.3$ and $Q_0 = 0.27$, and let $p(\cdot)$ be the density of $N(0.3, 0.8)$ and $q(\cdot)$ be the density of $N(0, 1)$. For $\tau \in \mathbb{R}$, we define the thresholded network $A_\tau \in \{0, 1\}^{n\times n}$ as $A_{\tau,uv} = 1$ if $A_{uv} \neq 0$ and $A_{uv} \geq \tau$, and $A_{\tau,uv} = 0$ if $A_{uv} = 0$ or if $A_{uv} < \tau$. For each $\tau \in \{-2, -1.8, -1.6, \dots, 1.6, 1.8, 2.0\}$, we form A_τ, extract the cluster, and compute the misclustering error. We then report the lowest misclustering error among all A_τ for $\tau \in \{-2, -1.8, -1.6, \dots, 1.6, 1.8, 2.0\}$ as the red line in Figure 5(b); this approach is of course impossible to implement in practice, and we use it only for the purpose of comparison. The turquoise line is the misclustering error incurred by our algorithm, using $L_n = \lfloor 0.4(\log(\log n))^4 \rfloor$.

Acknowledgments. The authors would like to thank Zongming Ma for enlightening discussions in the earlier stages of this project. The authors would also like to thank Richard Samworth for several helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal rates for community estimation in the weighted stochastic block model" (DOI: 10.1214/18-AOS1797SUPP; .pdf). We provide detailed proofs of the theorems and propositions in the main paper.

REFERENCES

[1] ABBE, E. (2017). Community detection and stochastic block models: Recent developments. *J. Mach. Learn. Res.* **18** Paper No. 177, 86. MR3827065

[2] ABBE, E., BANDEIRA, A. S., BRACHER, A. and SINGER, A. (2014). Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Trans. Netw. Sci. Eng.* **1** 10–22. MR3349181 https://doi.org/10.1109/TNSE.2014.2368716

[3] ABBE, E., BANDEIRA, A. S. and HALL, G. (2016). Exact recovery in the stochastic block model. *IEEE Trans. Inform. Theory* **62** 471–487. MR3447993 https://doi.org/10.1109/TIT.2015.2490670

[4] ABBE, E. and SANDON, C. (2015). Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In 2015 *IEEE* 56*th Annual Symposium on Foundations of Computer Science—FOCS* 2015 670–688. IEEE Computer Soc., Los Alamitos, CA. MR3473334

[5] ABBE, E. and SANDON, C. (2015). Recovering communities in the general stochastic block model without knowing the parameters. In *Advances in Neural Information Processing Systems* 676–684.

[6] AICHER, C., JACOBS, A. Z. and CLAUSET, A. (2015). Learning latent block structure in weighted networks. *J. Complex Netw.* **3** 221–248. MR3365464 https://doi.org/10.1093/comnet/cnu026

[7] BALAKRISHNAN, S., XU, M., KRISHNAMURTHY, A. and SINGH, A. (2011). Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems* 954–962.

[8] BARRAT, A., BARTHELEMY, M., PASTOR-SATORRAS, R. and VESPIGNANI, A. (2004). The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101** 3747–3752.

[9] BLONDEL, V. D., GUILLAUME, J.-L., LAMBIOTTE, R. and LEFEBVRE, E. (2008). Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **10**.

[10] BOCCALETTI, S., LATORA, V., MORENO, Y., CHAVEZ, M. and HWANG, D.-U. (2006). Complex networks: Structure and dynamics. *Phys. Rep.* **424** 175–308. MR2193621 https://doi.org/10.1016/j.physrep.2005.10.009

[11] CHIN, P., RAO, A. and VU, V. (2015). Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Proceedings of the* 28*th Conference on Learning Theory* 391–423.

[12] DECELLE, A., KRZAKALA, F., MOORE, C. and ZDEBOROVÁ, L. (2011). Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**.

[13] EASLEY, D. and KLEINBERG, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge Univ. Press, Cambridge. MR2677125 https://doi.org/10.1017/CBO9780511761942

[14] FIENBERG, S. E., MEYER, M. M. and WASSERMAN, S. S. (1985). Statistical analysis of multiple sociometric relations. *J. Amer. Statist. Assoc.* **80** 51–67.

[15] GAO, C., MA, Z., ZHANG, A. Y. and ZHOU, H. H. (2017). Achieving optimal misclassification proportion in stochastic block models. *J. Mach. Learn. Res.* **18** Paper No. 60, 45. MR3687603

[16] GOLDENBERG, A., ZHENG, A. X., FIENBERG, S. E. and AIROLDI, E. M. (2010). A survey of statistical network models. *Found. Trends Mach. Learn.* **2** 129–233.

[17] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming. *IEEE Trans. Inform. Theory* **62** 2788–2797. MR3493879 https://doi.org/10.1109/TIT.2016.2546280

[18] HAJEK, B., WU, Y. and XU, J. (2016). Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *IEEE Trans. Inform. Theory* **62** 5918–5937. MR3552431 https://doi.org/10.1109/TIT.2016.2594812

[19] HAJEK, B., WU, Y. and XU, J. (2017). Information limits for recovering a hidden community. *IEEE Trans. Inform. Theory* **63** 4729–4745. MR3683533 https://doi.org/10.1109/TIT.2017.2653804

[20] HAJEK, B., WU, Y. and XU, J. (2017). Submatrix localization via message passing. *J. Mach. Learn. Res.* **18** Paper No. 186, 52. MR3827074

[21] HARTUV, E. and SHAMIR, R. (2000). A clustering algorithm based on graph connectivity. *Inform. Process. Lett.* **76** 175–181. MR1807676 https://doi.org/10.1016/S0020-0190(00)00142-3

[22] HEIMLICHER, S., LELARGE, M. and MASSOULIÉ, L. (2012). Community detection in the labelled stochastic block model. Preprint. Available at arXiv:1209.2910.

[23] HOLLAND, P. W., LASKEY, K. B. and LEINHARDT, S. (1983). Stochastic blockmodels: First steps. *Soc. Netw.* **5** 109–137. MR0718088 https://doi.org/10.1016/0378-8733(83)90021-7

[24] JACKSON, M. O. (2008). *Social and Economic Networks*. Princeton Univ. Press, Princeton, NJ. MR2435744

[25] JOG, V. and LOH, P. (2015). Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence. Preprint. Available at arXiv:1509.06418.

[26] LEI, J. and RINALDO, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.* **43** 215–237. MR3285605 https://doi.org/10.1214/14-AOS1274

[27] LELARGE, M., MASSOULIÉ, L. and XU, J. (2015). Reconstruction in the labelled stochastic block model. *IEEE Trans. Netw. Sci. Eng.* **2** 152–163. MR3453283 https://doi.org/10.1109/TNSE.2015.2490580

[28] MASSOULIÉ, L. (2014). Community detection thresholds and the weak Ramanujan property. In *STOC'14—Proceedings of the* 2014 *ACM Symposium on Theory of Computing* 694–703. ACM, New York. MR3238997

[29] MCSHERRY, F. (2001). Spectral partitioning of random graphs. In 42*nd IEEE Symposium on Foundations of Computer Science* (*las Vegas*, *NV*, 2001) 529–537. IEEE Computer Soc., Los Alamitos, CA. MR1948742

[30] MOSSEL, E., NEEMAN, J. and SLY, A. (2012). Stochastic block models and reconstruction. Preprint. Available at arXiv:1202.1499.

[31] MOSSEL, E., NEEMAN, J. and SLY, A. (2014). Consistency thresholds for binary symmetric block models. Preprint. Available at arXiv:1407.1591.

[32] MOSSEL, E., NEEMAN, J. and SLY, A. (2018). A proof of the block model threshold conjecture. *Combinatorica* **38** 665–708. MR3876880 https://doi.org/10.1007/s00493-016-3238-8

[33] NEWMAN, M., BARABÁSI, A.-L. and WATTS, D. J., eds. (2006). *The Structure and Dynamics of Networks. Princeton Studies in Complexity*. Princeton Univ. Press, Princeton, NJ. MR2352222

[34] NEWMAN, M. E. J. (2004). Analysis of weighted networks. *Phys. Rev. E* **70**.

[35] NEWMAN, M. E. J. and GIRVAN, M. (2004). Finding and evaluating community structure in networks. *Phys. Rev. E* **69**.

[36] PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

[37] RUBINOV, M. and SPORNS, O. (2010). Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage* **52** 1059–1069.

[38] SADE, D. S. (1972). Sociometrics of Macaca mulatta: I. Linkages and cliques in grooming matrices. *Folia Primatologica* **18** 196–223.

[39] SHI, J. and MALIK, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 888–905.

[40] XU, M., JOG, V. and LOH, P.-L (2020). Supplement to "Optimal rates for community estimation in the weighted stochastic block model." https://doi.org/10.1214/18-AOS1797SUPP.

[41] YUN, S. and PROUTIERE, A. (2016). Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems* 965–973.

[42] ZHANG, A. Y. and ZHOU, H. H. (2016). Minimax rates of community detection in stochastic block models. *Ann. Statist.* **44** 2252–2280. MR3546450 https://doi.org/10.1214/15-AOS1428

[43] ZHANG, B. and HORVATH, S. (2005). A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4** Art. 17, 45. MR2170433 https://doi.org/10.2202/1544-6115.1128