Submitted to the Annals of Statistics

OPTIMAL RATES FOR COMMUNITY ESTIMATION IN THE WEIGHTED STOCHASTIC BLOCK MODEL

By Min Xu[‡] , Varun Jog[§] and Po-Ling Loh[§]

University of Pennsylvania[‡] and University of Wisconsin - Madison[§]

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 principled, computationally tractable algorithm based on discretization that achieves the optimal error rate without assuming knowledge of the weight densities.

1. Introduction. The recent explosion of network datasets has created a need for new statistical methodology [34, 13, 25, 17]. 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, 22, 37, 40, 30, 36].

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) [24]. 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 excellent

MSC 2010 subject classifications: Primary 62H30, 91D30; secondary 62C20, 90B15

Keywords and phrases: Stochastic block models, Network analysis, Renyi divergence, Nonparametric estimation, Optimal estimation rates

survey paper by Abbe [1]). Various algorithms for community recovery have been devised with guaranteed optimality properties, measured in terms of correlated recovery [31, 32, 29], exact recovery [3, 4, 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 [35, 10]: Edges in social or cellular networks may quantify the frequency of interactions between pairs of individuals [39, 9]. 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 level of neuronal activity between corresponding regions in the brain [38]. 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, $p(x)$ or $q(x)$. We study the problem of community estimation based on observations of the edge weights in the network, *without* assuming knowledge of p, q, $p(x)$ or $q(x)$. Since $p(x)$ and $q(x)$ are allowed to be continuous, our model strictly generalizes the discrete labeled SBMs considered in previous literature [23, 28, 26], as well as the censored SBM [2, 18, 19].

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 $p(x)$ and $q(x)$, 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 $p(x)$ and $q(x)$ may be arbitrary and are only required to satisfy mild regularity conditions, whereas previous approaches generally assume that $p(x)$ and $q(x)$ 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 $p(x)$ and $q(x)$. 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, 27] 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 concludes the paper with further implications and open questions.

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 $[n]$. We write $o(1)$ to denote a sequence indexed by n that tends to 0 as $n \to \infty$, and write $\Theta(1)$ to denote a sequence indexed by n that is bounded away from 0 and ∞ as $n \to \infty$. 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. Throughout the paper, we will assume that the communities are approximately balanced, so there exists a *cluster-imbalance constant* β such that the cluster size n_k satisfies $\frac{n}{\beta K} \leq n_k \leq \frac{\beta n}{K}$ for each $1 \leq k \leq K$.

DEFINITION 2.1. *Let $\sigma_0 : [n] \to [K]$ denote the true cluster assignment; i.e., $\sigma_0(u) \in \{1, 2, \dots, K\}$ denotes the cluster label of node u. We say that two cluster assignments σ and σ' are* equivalent *if a permutation $\tau \in S_K$ exists such that $\tau \circ \sigma = \sigma'$.*

We first define the homogeneous unweighted SBM, which is parametrized by (n, K, β, σ_0), the between-cluster edge probability p, and the within-cluster edge probability q. This is characterized by the following probability distribution over adjacency matrices $A \in \{0, 1\}^{n \times n}$:

DEFINITION 2.2 (Homogeneous unweighted SBM). *For all $u < v$, the entries of A are generated independently according to*

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

Note that 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 $A_{uv} \sim Ber(P_{\sigma_0(u), \sigma_0(v)})$.

A homogeneous weighted SBM is parametrized by (n, K, β, σ_0), the edge *absence* probabilities P_0 and Q_0, and the edge weight probabilities $p(x)$ and $q(x)$. Let S denote the support of $p(x)$ and $q(x)$, where $S = [0, 1]$, $S = \mathbb{R}$, or $S = \mathbb{R}^+$. The weighted SBM is then characterized by a distribution over adjacency matrices $A \in S^{n\times n}$ in the following manner:

DEFINITION 2.3 (Homogeneous weighted SBM). *For all $u < v$, we first independently generate edge presence indicator variables:*

$$Z_{uv} \sim \begin{cases} Ber(1 - P_0) & \text{if } \sigma_0(u) = \sigma_0(v), \\ Ber(1 - Q_0) & \text{if } \sigma_0(u) \neq \sigma_0(v), \end{cases}$$

and then independently generate edge weights:

$$A_{uv} \sim \begin{cases} 0 & \text{if } Z_{uv} = 0, \\ p(x) & \text{if } Z_{uv} = 1 \text{ and } \sigma_0(u) = \sigma_0(v), \\ q(x) & \text{if } Z_{uv} = 1 \text{ and } \sigma_0(u) \neq \sigma_0(v). \end{cases}$$

Note that $\mathbb{E}(A)$ may not exhibit the familiar block structure found in unweighted SBMs, since our model includes the case where $P_0 = Q_0$ and $p(x)$ and $q(x)$ have the same mean.

We may define the weighted SBM more succinctly by defining the probability measures P and Q to be mixed distributions, where the singular part of P is a point mass at 0 with probability P_0, and the continuous part of P has $(1 - P_0)p(x)$ as its Radon-Nikodym derivative with respect to the Lebesgue measure (and likewise for Q):

DEFINITION 2.4 (Homogeneous weighted SBM). *For all $u < v$, the entries of A are generated independently according to*

$$A_{uv} \sim \begin{cases} P & \text{if } \sigma_0(u) = \sigma_0(v), \\ Q & \text{if } \sigma_0(u) \neq \sigma_0(v). \end{cases} \tag{1}$$

If P and Q are Bernoulli distributions without any continuous portions, the weighted SBM reduces to the unweighted version. It is possible to generalize the weighted SBM to a *weighted and labeled* SBM by allowing the singular parts of P and Q to possess additional point masses; the theory derived in this paper extends to simple cases of those models, as well.

Our definition treats an edge with weight 0 as a missing edge, but it is straightforward to distinguish these two notions by defining P, Q as probability measures over $S \cap \{*\}$ where the symbol $*$ denotes a missing edge.

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 2.5. *We define the* misclustering error *to be*

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

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 $(n, K, \beta, \sigma_0, P_0, Q_0, p(x), q(x))$.

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 2.6. *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 $\pi \in S_n$,*

$$\hat{\sigma}(\pi A) \circ \pi \text{ is equivalent to } \hat{\sigma}(A). \tag{2}$$

In other words, $\hat{\sigma}$ is permutation equivariant if a permutation $\tau \in S_K$ exists satisfying $\hat{\sigma}(\pi A) \circ \pi = \tau \circ \hat{\sigma}(A)$. 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 2.6 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. Our results are asymptotic in n, and we treat $P_0, Q_0, p(x), q(x), \sigma_0$ as varying with n; in various places, we do not explicitly state this dependence in order to simplify our notation. We hold K and β to be fixed with respect to n. We now preview our main results.

Let P be the probability measure induced by $(P_0, p(x))$, and let Q be defined analogously. The optimal rate of misclustering error naturally depends on the extent to which P and Q are different. This is quantified by the Renyi divergence of order $\frac{1}{2}$, which we denote by I:

$$-2\log\int\left(\frac{dP}{dQ}\right)^{1/2}dQ = -2\log\left(\sqrt{P_0Q_0}+\int_S\sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right).$$

Since P and Q depend on n, the Renyi divergence I also depends on n. Note that in the analysis of unweighted SBMs, various dependencies on n in the edge probabilities are also assumed (e.g., $p, q \to 0$ as $n \to \infty$ [32, 4], or $|p-q| \to 0$ [42]). We focus on the setting where P and Q tend toward each other in the sense that $I \to 0$; this setting encompasses the sparse graph setting where $P_0, Q_0 \to 1$. If we view I as the signal level in the community estimation problem, the setting $I \to 0$ is more interesting and challenging than the setting $I = \Omega(1)$, since it corresponds to a weaker signal.

When $I = o(1)$, we may characterize I in terms of the Hellinger distance (cf. Lemma I.1):

$$\begin{aligned} I &= \left((\sqrt{P_0}-\sqrt{Q_0})^2+\int_S(\sqrt{(1-P_0)p(x)}-\sqrt{(1-Q_0)q(x)})^2dx\right)(1+o(1)) \\ &= \Big((\sqrt{P_0}-\sqrt{Q_0})^2+(\sqrt{1-P_0}-\sqrt{1-Q_0})^2 \\ &\qquad +\sqrt{(1-P_0)(1-Q_0)}\int_S(\sqrt{p(x)}-\sqrt{q(x)})^2dx\Big)\cdot(1+o(1)). \end{aligned}$$

The above equation shows that I decomposes into two terms, the first of which captures the divergence between the edge presence probabilities (and also appears in the analysis of unweighted SBM), 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 very dense (e.g., $1 - P_0$ and $1 - Q_0$ are fixed at some constant level), it is still possible for the signal level I to be weak so long as $\int_S (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \to 0$. Furthermore, if the network is completely dense in the sense that $1-P_0 = 1-Q_0 = 0$, a nonzero signal may still exist if $p(x)$ and $q(x)$ are different. Our results apply simultaneously to such 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 informal statements of our main results:

THEOREM. **(Informal statement)** *Let A be generated from a weighted SBM with parameters $(n, K, \beta, \sigma_0, P_0, Q_0, p(x), q(x))$. Under regularity conditions on the densities $p(x)$ and $q(x)$, any permutation equivariant estimator $\hat{\sigma}$ satisfies the lower bound*

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

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

$$\lim_{n\to\infty} P\left(l(\hat{\sigma}(A), \sigma_0) \leq \exp\left(-(1 + o(1))\frac{nI}{\beta K}\right)\right) \to 1.$$

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 $\exp\left(-(1 + o(1))\frac{nI}{\beta K}\right)$. If $\frac{nI}{\beta K \log n} > 1$, we have $\exp\left(-(1 + o(1))\frac{nI}{\beta K}\right) < \frac{1}{n}$ for large enough n, so $\lim_{n\to\infty} P\left(l(\hat{\sigma}(A), \sigma_0) = 0\right) \to 1$ (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 $\exp\left(-(1 + o(1))\frac{nI_{Ber}}{\beta K}\right)$ for the unweighted SBM, where $I_{Ber} = \sqrt{pq} + \sqrt{(1-p)(1-q)}$. Note that if $P \sim \text{Bernoulli}(p)$ and $Q \sim \text{Bernoulli}(q)$, we have $I = I_{Ber}$. 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 H.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 [26] propose a similar weighted block model and show the exact recovery threshold to be dependent on Renyi divergence. They focus on the setting where the distributions are discrete and known whereas we consider continuous densities that are unknown. Finally, 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 model considered in Hajek et al. [21] is also similar to the one we propose in our paper, except it only involves a single communities and assumes knowledge of the distributions P and Q. Weighted networks have also received some attention in the physics community [35, 8] 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 [29, 33].

4. Estimation algorithm. A natural approach to community estimation is to estimate the edge weight densities $p(x)$ and $q(x)$, but this is hindered by the fact that we do not know whether an edge weight observation originates from $p(x)$ or $q(x)$. An alternative approach of applying spectral

clustering directly to the weighted adjacency matrix A will be ineffective if $p(x)$ and $q(x)$ have the same mean and $P_0 = Q_0$, 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 the 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 outlining 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.

1. **Transformation & discretization.** We take as input a weighted 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 $[0,1]$, $\mathbb{R}^+$, $\mathbb{R}$) to obtain a matrix $\Phi(A)$ with weights between 0 and 1. Next, we divide the interval $[0,1]$ into L equally-spaced subintervals. We replace the real-valued weight entries $\Phi(A)$ with a categorical label $l \in \{1, \ldots, L\}$. We denote the labeled adjacency matrix by A_L.
2. **Add noise.** For a fixed constant $c > 0$, let $\delta = \frac{cL}{n}$. We perform the following process on every edge of the labeled graph, independently of other edges: With probability $1 - \delta$, keep an edge as it is, and with probability δ, erase the edge and replace it with an edge with label uniformly drawn from the set of L labels. We continue to denote the modified adjacency matrix as A_L.
3. **Initialization parts 1 & 2.** For each label l, we create a sub-network by including only edges of label l. For each sub-network, we perform spectral clustering and output the label l^* that induces the maximally separated spectral clustering. Let A_{l^*} be the adjacency matrix for label l^*. For each $u \in \{1, \ldots, n\}$, we perform spectral clustering on $A_{l^*} \setminus \{u\}$, which denotes the adjacency matrix with vertex u removed. We output n clusterings $\tilde{\sigma}_1, \ldots, \tilde{\sigma}_n$.
4. **Refinement & consensus.** From each $\tilde{\sigma}_u$, we generate a clustering $\hat{\sigma}_u$ on $\{1, 2, \ldots, n\}$ that retains the assignments specified by $\tilde{\sigma}_u$ for $\{1, 2, \ldots, n\} \setminus \{u\}$, 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.

Our algorithm is summarized pictorially in Figure 2:

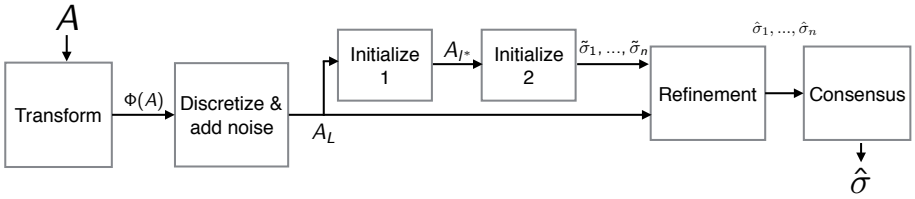

FIG 1. *Pipeline for the our proposed algorithm*

4.2. *Transformation and discretization.* These two steps are straightforward: in the transformation step, we apply an invertible CDF $\Phi : S \to [0,1]$ as the transformation function on all edge weights, so 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 is assigned the label l if the weight of that edge lies in bin l.

Algorithm 4.1 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 bins, labeled $\text{bin}_1, \ldots, \text{bin}_l$.
for every edge (u,v) **do**
  Let l be the bin in which $\Phi(A_{uv})$ falls
  Give the edge (u,v) the label l in the labeled network A_L
end for
Output A_L

4.3. *Add noise.* This step is required for technical reasons. 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 within- and between-community edge labels, but has the desirable effect of ensuring that all edge labels occur with probability at least $\frac{c}{n}$. This property is crucial to our analysis in subsequent steps of the algorithm. In the description of the algorithm below, 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 $\{1, \ldots, L\}$. The goal of the initialization procedure is

Algorithm 4.2 Add noise

Input: A labeled network A_L with $L+1$ labels and a constant c
Output: A labeled network A_L with $L+1$ labels

for every edge (u, v) **do**
With probability $1 - \frac{c(L+1)}{n}$, do nothing
With probability $\frac{c(L+1)}{n}$, replace the edge label with a label drawn uniformly at random from $\{0, 1, 2, \ldots, L\}$
end for
Output A_L

to create a rough clustering $\tilde{\sigma}$ that is suboptimal but still consistent. As outlined in Algorithm 4.3, 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 $\{\tilde{\sigma}_u\}_{u=1,\ldots,n}$, where each $\tilde{\sigma}_u : [n-1] \to [K]$ is a clustering of a network of $n-1$ nodes where node u has been removed. The clusterings $\{\tilde{\sigma}_u\}$ will later be combined into a single clustering algorithm.

Algorithm 4.3 Initialization

Input: A labeled network A_L with L labels
Output: A set of clusterings $\{\tilde{\sigma}_u\}_{u=1,\ldots,n}$, where $\tilde{\sigma}_u$ is a clustering on $\{1, 2, \ldots, n\} \setminus \{u\}$

Separate A_L into L networks $\{A_l\}_{l=1,\ldots,L}$ ▷ Stage 1
for each label l **do**
Compute $\bar{d} = \frac{1}{n}\sum_{u=1}^{n} d_u$ as the average degree
Perform spectral clustering with $\tau = C_\tau \bar{d}$ and $\mu \geq C\beta$ to obtain $\tilde{\sigma}_l$, where C and C_τ are apprpriate large constants
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 $\hat{Q}_l = \frac{\sum_{u \neq v\,:\,\tilde{\sigma}_l(u)\neq\tilde{\sigma}_l(v)} (A_l)_{uv}}{|u \neq v\,:\,\tilde{\sigma}_l(u)\neq\tilde{\sigma}_l(v)|}$
$\hat{I}_l \leftarrow \frac{(\hat{P}_l - \hat{Q}_l)^2}{\hat{P}_l \vee \hat{Q}_l}$
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 on $A_{l^*} \setminus \{u\}$ to obtain $\tilde{\sigma}_u$
end for
Output the set of clusterings $\{\tilde{\sigma}_u\}_{u=1,\ldots,n}$

Spectral clustering. Note that Algorithm 4.3 involves several applications of SPECTRAL CLUSTERING. We describe the spectral clustering algorithm used as a subroutine in Algorithm 4.4 below. Importantly, note that we may always choose the parameter μ sufficiently large such that Algo-

rithm 4.4 generates a set S with $|S| = K$.

Algorithm 4.4 SPECTRAL CLUSTERING

Input: An unweighted network A, trim threshold τ, number of communities K, tuning parameter μ
Output: A clustering σ

For each node u whose degree $d_u \geq \tau$, set $A_{uv} = 0$ to get $T_\tau(A)$
Let $\hat{A}$ be the best rank-K approximation to $T_\tau(A)$ in spectral norm
For each node u, define the neighbor set $N(u) = \{v : \|\hat{A}_u - \hat{A}_v\|_2^2 \leq \mu K^2 \frac{\bar{d}}{n}\}$
Initialize $S \leftarrow 0$
Select node u with the most neighbors and add u into S as $S[1]$
for $i = 2, \ldots, K$ **do**
    Among all u such that $|N(u)| \geq \frac{n}{\mu K}$, select $u^* = \arg\max_u \min_{v \in S} \|\hat{A}_u - \hat{A}_v\|_2$
    Add u^* into S as $S[i]$
end for
for $u = 1, \ldots, n$ **do**
    Take $\arg\min_i \|\hat{A}_u - \hat{A}_{S[i]}\|_2$ and assign $\sigma(u) = i$
end for

4.5. *Refinement and consensus.* This step parallels Gao et al. [15]. In the refinement step, we use the set of initial clusterings $\{\tilde{\sigma}_u\}_{u=1,\ldots,n}$ to generate a more accurate clustering for the labeled network by locally maximizing an approximate log-likelihood for each node u. The consensus step then resolves a cluster label consistency problem arising after the refinement stage.

Algorithm 4.5 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

$$\hat{\sigma}_u(u) = \arg\max_k \sum_{v\,:\,\tilde{\sigma}_u(v)=k,\, v\neq u} \sum_l \log \frac{\hat{P}_l}{\hat{Q}_l} \mathbf{1}(A_{uv} = l)$$

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

$$\hat{\sigma}(u) = \arg\max_k |\{v : \hat{\sigma}_1(v) = k\} \cap \{v : \hat{\sigma}_u(v) = \hat{\sigma}_u(u)\}|$$

end for
Output $\hat{\sigma}$

5. Optimal misclustering error. We now provide formal statements of our results. For clarity, we explicitly use a subscript to denote all quantities that vary with n. The following assumption states that the probability of within-community edges has the same order as the probability of between-community edges and is standard in the unweighted SBM literature.

Assumption A0:. There exists an absolute constant $c_0 > 0$ such that $\frac{1}{c_0} \leq \frac{1-P_{0,n}}{1-Q_{0,n}} \leq c_0$. If $P_{0,n} \vee Q_{0,n} > 0$, we also assume $\frac{1}{c_0} \leq \frac{P_{0,n}}{Q_{0,n}} \leq c_0$.

Let $\mathcal{P}$ be the set of all densities on S and let

$$(\mathcal{P} \times \mathcal{P})^\infty := \left\{ (p_n, q_n)_{n \geq 1} \, : \, p_n, q_n \geq 0, \, \int_S p_n d\lambda = \int_S q_n d\lambda = 1 \right\},$$

where λ denotes the Lebesgue measure. Our conditions on p_n and q_n define a subset of $(\mathcal{P} \times \mathcal{P})^\infty$, which we formally describe in the next two subsections.

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 (a) $\int_S \phi(x)^s dx < \infty$ for any $0 < s < 1$, and (b) $\left| \frac{\phi'(x)}{\phi(x)} \right| < \infty$.*

For $S = [0,1]$, we always take Φ to be the identity. For $S = \mathbb{R}$ or $\mathbb{R}^+$, the function Φ must be chosen so that ϕ has all finite moments in order satisfy (a) and a subexponential tail in order to satisfy (b). However, the specific choice of Φ is not crucial. We will use the following definitions:

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

The expressions are similar to a generalized normal density, modified so that $\left| \frac{\phi'(x)}{\phi(x)} \right|$ is bounded. It is easy to verify that $\Phi(x) = \int_0^x \phi(t)dt$ (respectively, $\Phi(x) = \int_{-\infty}^x \phi(t)dt$) is a valid transformation function.

The function Φ induces a probability measure on S, and we let $\Phi\{\cdot\}$ denote the Φ-measure of a set. The next definition is a shape constraint, which appears in our regularity conditions and ensures that we do not lose too much information through discretization.

DEFINITION 5.2. *We say that a nonnegative function $f(x)$ is (c_{s1}, c_{s2}, C_s)-*bowl-shaped *if $f(x)$ is nonincreasing for all $x \leq c_{s1}$, nondecreasing for all $x \geq c_{s2}$, and bounded by C_s for all $x \in [c_{s1}, c_{s2}]$.*

Let $H_n = \int_S (\sqrt{p_n(x)} - \sqrt{q_n(x)})^2 dx$ be the Hellinger distance between the edge weight densities $p_n(x), q_n(x)$. We have different conditions depending on whether $H_n = \Theta(1)$ or $H_n = o(1)$; when $H_n = \Theta(1)$, the regularity conditions are straightforward but when $H_n = o(1)$, the conditions become more complicated. Intuitively, this is because when $|p_n - q_n| \to 0$, we need additional conditions to ensure that $|p_n' - q_n'| \to 0$ at the same rate. The same algorithm applies to both these cases however.

5.1.1. *The case* $H_n = \Theta(1)$. In this case, which encompasses the sparse graph setting where $P_0, Q_0 \to 0$ very quickly, the densities $p_n(x)$ and $q_n(x)$ do not tend to each other. Let $\mathcal{G}_\Phi^{\text{upper}} \subset (\mathcal{P} \times \mathcal{P})^\infty$ be the sequences of densities satisfying $H_n = \Theta(1)$, as well as the following regularity conditions:

A1 $p_n(x), q_n(x) > 0$ on the interior of S, $\sup_{n,x} \{p_n(x) \vee q_n(x)\} < \infty$, and $\sup_{n,x} \left\{ \frac{p_n(x) \vee q_n(x)}{\phi(x)} \right\} < \infty$.

A2 $\sup_n \int_S \left| \log \frac{p_n(x)}{q_n(x)} \right|^r \phi(x) dx < \infty$ for an absolute constant $r > 4$.

A3 There exists a function $h_n(x)$ such that

(a) $h_n(x) \geq \max \left\{ \left| \frac{q_n'(x)}{q_n(x)} \right|, \left| \frac{p_n'(x)}{p_n(x)} \right| \right\}$,

(b) $h_n(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped for some absolute constants c_{s1}, c_{s2}, and C_s, and

(c) For some constant t such that $\frac{4}{r} < 2t < 1$, we have

$$\sup_n \int_S |h_n(x)|^{2t} \phi(x) dx < \infty.$$

A4 $(\log p_n)'(x), (\log q_n)'(x) \geq (\log \phi)'(x)$ for all $x < c_{s1}$, and $(\log p_n)'(x), (\log q_n)'(x) \leq (\log \phi)'(x)$ for all $x > c_{s2}$.

Note that we allow $c_{s1} = 0$ when $S = \mathbb{R}^+$, and we allow $c_{s1} = 0$ and $c_{s1} = 1$ when $S = [0, 1]$. The above conditions depend on the choice of Φ, but it generally suffices to choose Φ such that its derivative ϕ as a heavy-tailed density where all moments exist. In particular, we show in Section 5.1.5 that choosing Φ according to equations (3) allows $\mathcal{G}_\Phi^{\text{upper}}$ to encompass Gaussian, Laplace, and other broad classes of densities. We then have our upper bound:

THEOREM 5.1. *Suppose* $I_n \to 0$ *and* $nI_n \to \infty$. *Let* $\hat{\sigma}$ *be the algorithm described in Section 4 with transformation function* Φ *and discretization level* L_n *chosen such that* $L_n \to \infty$ *and* $\frac{nI_n}{L_n \exp(L_n^{1/r})} \to \infty$. *Suppose* $P_{0,n}$ *and* $Q_{0,n}$ *satisfy Assumption A0 and* $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{upper}$. *Then*

$$\lim_{n \to \infty} P \left\{ l(\hat{\sigma}(A), \sigma_0) \leq \exp\left(-\frac{nI_n}{\beta K} (1 + o(1)) \right) \right\} \to 1.$$

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

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

For the proof of Theorem 5.1, see Appendix F.2. Since $nI_n \to \infty$, it is always possible to choose $L_n \to \infty$ satisfying the conditions of the theorem.

5.1.2. *The case* $H_n = o(1)$. In this case, the densities $p_n(x)$ and $q_n(x)$ converge to each other, encompassing the setting of completely dense graphs where $P_0 = Q_0 = 0$. Let $\mathcal{G}_\Phi^{\text{upper}\prime} \subset (\mathcal{P} \times \mathcal{P})^\infty$ be the sequences of densities satisfying $H_n = o(1)$, as well as the following regularity conditions:

A1' $p_n(x), q_n(x) > 0$ on the interior of S, $\sup_{n,x} \{p_n(x) \vee q_n(x)\} < \infty$, and $\sup_{n,x} \left\{\frac{p_n(x) \vee q_n(x)}{\phi(x)}\right\} < \infty$.

A2' There exists an absolute constant ρ and subintervals $R_n \subseteq S$ such that

(a) $\frac{1}{\rho} \leq \frac{p_n(x)}{q_n(x)} \leq \rho$ for $x \in R_n$, and

(b) $\Phi\{R_n^c\} = o(H_n)$ where $R_n^c \equiv S \backslash R_n$

A3' Denoting $\alpha_n^2 = \int_{R_n} q_n(x) \left(\frac{p_n(x) - q_n(x)}{q_n(x)}\right)^2 dx$ and $\gamma_n(x) = \frac{q_n(x) - p_n(x)}{\alpha_n}$, there exists an absolute constant $r > 4$ such that

$$\sup_n \int_{R_n} q_n(x) \left|\frac{\gamma_n(x)}{q_n(x)}\right|^r dx < \infty,$$

A4' There exists a function $h(x)$ such that

(a) $h_n(x) \geq \max\left\{\left|\frac{\gamma_n'(x)}{q_n(x)}\right|, \left|\frac{q_n'(x)}{q_n(x)}\right|, \left|\frac{\gamma_n(x)}{q_n(x)}\right|\right\}$.

(b) $h_n(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped for absolute constants c_{s1}, c_{s2}, and C_s, and

(c) for an absolute constant t where $\frac{4}{r} < 2t < 1$, we have

$$\sup_n \int_{R_n} |h_n(x)|^{2t} \phi(x) dx < \infty.$$

A5' $(\log p_n)'(x), (\log q_n)'(x) \geq (\log \phi)'(x)$ for all $x \leq c_{s1}$, and $(\log p_n)'(x), (\log q_n)'(x) \leq (\log \phi)'(x)$ for all $x \geq c_{s2}$.

Again, we allow $c_{s1} = 0$ when $S = \mathbb{R}^+$, and we allow $c_{s1} = 0$ and $c_{s1} = 1$ when $S = [0, 1]$. We have the following result:

THEOREM 5.2. *Suppose $I_n \to 0$ and $nI_n \to \infty$. Let $\hat{\sigma}$ be the algorithm described in Section 4 with transformation Φ and discretization level L_n chosen such that $L_n \to \infty$, $L_n = o(\frac{1}{H_n})$, and $L_n = o(nI_n)$. Suppose $P_{0,n}$ and $Q_{0,n}$ satisfy Assumption A0 and $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{upper\prime}$. Then*

$$\lim_{n\to\infty} P\left\{ l(\hat{\sigma}(A), \sigma_0) \leq \exp\left(-\frac{nI_n}{\beta K}(1 + o(1)) \right) \right\} \to 1.$$

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

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

The proof of Theorem 5.2 is outlined in Appendix F.1. Again, we note that because $\frac{1}{H_n}$ and nI_n both diverge by assumption, it is always possible to choose a sequence $L_n \to \infty$ satisfying the conditions of the theorem.

5.1.3. *Additional discussion of conditions.* It is crucial to note that our algorithm does *not* require prior knowledge of the form of $p_n(x)$ and $q_n(x)$; the same algorithm and guarantees apply so long as $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{\text{upper}}$ or $(p_n(x), q_n(x))_n \in \mathcal{G}_\Phi^{\text{upper}\prime}$. To aid the reader, we now provide a brief, non-technical 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 $p_n(x)$ and $q_n(x)$. Condition A2 requires the likelihood ratio to be integrable. It is analogous to a bounded likelihood ratio condition, but much weaker. Condition A3 controls the smoothness of the derivatives of $\log p_n(x)$ and $\log q_n(x)$. We add a mild bowl-shape constraint for technical reasons related to the analysis of binning. Condition A4 is a mild shape constraint on $p_n(x)$ and $q_n(x)$. When $S = \mathbb{R}$, this condition essentially requires p_n and q_n to be monotonically increasing in x for $x \to -\infty$, and decreasing in x for $x \to \infty$.

An analogous interpretation may be used to describe to the conditions A1'–A5'. However, we pay additional attention to A2'–A4': In condition A2', we require the likelihood ratio $\frac{p_n(x)}{q_n(x)}$ be bounded except on a set of diminishing Φ-measure. Combined with the assumption $H_n \to 0$, this allows us to say that the χ^2-distance between p_n and q_n, denoted by α_n, converges to 0. In condition A3', since $\alpha_n \to 0$, the function $\gamma_n(x)$ is of constant order in the sense that $\int_S q_n(x) \left(\frac{\gamma_n(x)}{q_n(x)}\right)^2 dx = 1$. Requirements on $\gamma_n(x)$ translate into convergence statements on $|p_n - q_n|$: For instance, an L_∞-bound on γ_n implies almost uniform convergence (with respect to Φ) of $|p_n - q_n|$ to 0. The integrability condition we impose on $\gamma_n(x)$ in condition A3' is analogous to

an L_∞-bound, but much weaker. Condition A4', in imposing an integrability condition on $\gamma_n'(x)$, intuitively requires $|p_n' - q_n'| \to 0$, as well.

5.1.4. *Examples for $S = [0,1]$.* When $S = [0,1]$, we 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. When $H_n = \Theta(1)$, the simplest example of $p_n(x)$ and $q_n(x)$ that satisfy conditions A1–A4 is when $p_n(x)$ and $q_n(x)$ are bounded away from 0 and ∞, uniformly in n, and $|p_n'(x)|$ and $|q_n'(x)|$ are also bounded uniformly in n. However, A1–A4 allow p_n and q_n to be 0 and p_n' and q_n' to be ∞ at the boundary points $\{0,1\}$. When $H_n = o(1)$, the densities p_n and q_n satisfy A1'–A5' if they satisfy the boundedness conditions described above and $x \mapsto \frac{p_n(x)-q_n(x)}{\|p_n - q_n\|_2}$ is uniformly bounded away from 0 and ∞ and has uniformly bounded first derivative.

5.1.5. *Examples for $S = \mathbb{R}$ or $\mathbb{R}^+$.* We begin with a proposition that allows us to generate large classes of examples. Let $f_\theta : S \to \mathbb{R}$ be a class of function indexed by a parameter vector $\theta \in \Theta \subseteq \mathbb{R}^{d_\Theta}$. Suppose Θ is compact and the class $\{f_\theta\}_{\theta\in\Theta}$ satisfies the following conditions with respect to Φ:

B1 $\exp(f_\theta(x))$ is a density for all θ and $\inf_{\theta,x}\{\log\phi(x) - f_\theta(x)\} > -\infty$.

B2 The Fisher information matrix $G_\theta := \int_S (\nabla_\theta f_\theta(x))(\nabla_\theta f_\theta(x))^{\mathrm{T}} \exp(f_\theta(x))dx$ is full-rank:

$$0 < c_{\min} < \inf_{\theta\in\Theta} \lambda_{\min}(G_\theta) \le \sup_{\theta\in\Theta} \lambda_{\max}(G_\theta) < c_{\max} < \infty,$$

for absolute constants $c_{\min}$ and $c_{\max}$.

B3 There exist functions $g_1(x) \ge \sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(x)\|$ and $g_{2,\theta}(x) \ge \max\{\|\nabla_\theta f_\theta'(x)\|, |f_\theta'(x)|\}$ such that g_1 and $g_{2,\theta}$ are $(c_{s1}, c_{s2}, \tilde{C}_s)$-bowl-shaped, and

$$\int_S g_1(x)^r \phi(x)dx < \infty, \quad \text{and} \quad \sup_{\theta\in\Theta} \int_S g_{2,\theta}(x)^{4t}\phi(x)dx < \infty,$$

where t and r are constants satisfying $\frac{8}{r} \le 4t < 1$.

B4 $\inf_{\theta\in\Theta} f_\theta'(x) \ge (\log\phi)'(x)$ for all $x \le c_{s1}$, and $\sup_{\theta\in\Theta} f_\theta'(x) \le (\log\phi)'(x)$ for all $x \ge c_{s2}$.

We then have the following result:

PROPOSITION 5.1. *Let $\{f_\theta\}_{\theta\in\Theta}$ be a class of functions that satisfy Assumptions B1–B5 with respect to a transformation function Φ. Let $\theta_{1,n}, \theta_{0,n}$ be two sequences of parameters in Θ. Let $p_n(x) = \exp(f_{\theta_{1,n}}(x))$ and $q_n(x) = \exp(f_{\theta_{0,n}}(x))$. Then*

(a) If $\|\theta_{1,n} - \theta_{0,n}\|_2 = \Theta(1)$, *then* $(p_n, q_n) \in \mathcal{G}_\Phi^{upper}$.
(b) If $\|\theta_{1,n} - \theta_{0,n}\|_2 = o(1)$, *then* $(p_n, q_n) \in \mathcal{G}_\Phi^{upper\prime}$.

Proposition 5.1 is useful for generating various examples of densities belonging to $\mathcal{G}_\Phi^{\text{upper}}$ or $\mathcal{G}_\Phi^{\text{upper}\prime}$. For all the examples we show, it suffices to choose the transformation function as in equations (3). Details are provided in Appendix G.

EXAMPLE 5.1 (Location-scale family over $\mathbb{R}$). *Let* $\exp(f(x))$ *be a positive base density over* $\mathbb{R}$. *Define* $\theta = (\mu, \sigma)$ *and* $\Theta = [-C_\mu, C_\mu] \times \left[\frac{1}{c_\sigma}, c_\sigma\right]$ *for some absolute constants* C_μ *and* c_σ, *and let* $f_\theta = f\left(\frac{x-\mu}{\sigma}\right) - \log\sigma$. *If*

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

then $\{f_\theta\}_{\theta\in\Theta}$ *satisfy Assumptions B1–B4 when* ϕ *is chosen according to equation* (3). *For a sequence* $\{(\mu_{1,n}, \sigma_{1,n}), (\mu_{0,n}, \sigma_{0,n})\}_n \subset (\Theta \times \Theta)^\infty$, *define*

$$p_n(x) = \frac{1}{\sigma_{1,n}} \exp\left(f\left(\frac{x-\mu_{1,n}}{\sigma_{1,n}}\right)\right), \text{ and } q_n(x) = \frac{1}{\sigma_{0,n}} \exp\left(f\left(\frac{x-\mu_{0,n}}{\sigma_{0,n}}\right)\right),$$

where f *satisfies assumptions* (a) *and* (b) *above.*

As a direct consequence of Proposition 5.1, we have the following:

(i) if $|\mu_{1,n} - \mu_{0,n}| + |\sigma_{1,n} - \sigma_{0,n}| = \Theta(1)$, *then* $(p_n, q_n)_n \in \mathcal{G}_\Phi^{upper}$;
(ii) if $|\mu_{1,n} - \mu_{0,n}| + |\sigma_{1,n} - \sigma_{0,n}| = o(1)$, *then* $(p_n, q_n)_n \in \mathcal{G}_\Phi^{upper\prime}$.

The above assumptions on $f(x)$ *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* $f(x) = -|x| - \log 2$.

We emphasize that we make no assumption on any separation between $\mu_{1,n}$ *and* $\mu_{0,n}$: *Our results apply even when* $\mu_{1,n} = \mu_{0,n}$ *for all* n, *which implies that* $p_n(x)$ *and* $q_n(x)$ *have the same mean.*

EXAMPLE 5.2 (Scale family over $\mathbb{R}^+$). *If the base density* $\exp(f(x))$ *is supported and positive on* $\mathbb{R}^+$, *we can define a scale family parametrized by* σ *as* $\exp(f(\frac{x}{\sigma}))$. *Let* $\theta = (\sigma)$ *and* $\Theta = \left[\frac{1}{c_\sigma}, c_\sigma\right]$ *for some absolute constant* c_σ, *and let* $f_\theta = f\left(\frac{x}{\sigma}\right) - \log\sigma$. *Again, if* $f(x)$ *satisfies conditions*

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

then $\{f_\theta\}_{\theta\in\Theta}$ *satisfies Assumptions B1–B4, when* ϕ *is chosen according to equation* (3). *For a sequence* $\{\sigma_{1,n},\sigma_{0,n}\}_n \subset (\Theta\times\Theta)^\infty$, *define*

$$p_n(x)=\frac{1}{\sigma_{1,n}}\exp\left(f\left(\frac{x}{\sigma_{1,n}}\right)\right),\quad \textit{and}\quad q_n(x)=\frac{1}{\sigma_{0,n}}\exp\left(f\left(\frac{x}{\sigma_{0,n}}\right)\right).$$

As a direct consequence of Proposition 5.1, we have the following:

(i) if $|\sigma_{1,n}-\sigma_{0,n}|=\Theta(1)$, *then* $(p_n,q_n)_n\in\mathcal{G}_\Phi^{upper}$;
(ii) if $|\sigma_{1,n}-\sigma_{0,n}|=o(1)$, *then* $(p_n,q_n)_n\in\mathcal{G}_\Phi^{upper\prime}$.

An example of a function f *that satisfies condition (a) and (b) is* $f(x)=-x$, *which forms the base density of the* ***exponential distributions****.*

We end this section with the family of Gamma distributions on $\mathbb{R}^+$, which falls outside the above example but still satisfies Assumptions B1–B4.

EXAMPLE 5.3 (Gamma distribution). *Let* $\theta=(\alpha,\beta)$ *and* $\Theta=[\frac{1}{C},C]^2$ *for some absolute constant* C, *and let*

$$f_\theta(x)=(\alpha-1)\log x-\beta x+\alpha\log\beta-\log\Gamma(\alpha),$$

where $\Gamma(\cdot)$ *is the Gamma function. If we choose* ϕ *as in equation* (3), *then* $\{f_\theta\}_{\theta\in\Theta}$ *satisfies Assumptions B1–B4.*

For a sequence $\{(\alpha_{1,n},\beta_{1,n}),(\alpha_{0,n},\beta_{0,n})\}_n\subset(\Theta\times\Theta)^\infty$, *let*

$$p_n(x)=\frac{\beta_{1,n}^{\alpha_{1,n}}}{\Gamma(\alpha_{1,n})}x^{\alpha_{1,n}-1}e^{-\beta_{1,n}x},\quad \textit{and}\quad q_n(x)=\frac{\beta_{0,n}^{\alpha_{0,n}}}{\Gamma(\alpha_{0,n})}x^{\alpha_{0,n}-1}e^{-\beta_{0,n}x},$$

defined over $\mathbb{R}^+$. *As a consequence of Proposition 5.1, we have the following:*

(i) if $|\alpha_{1,n}-\alpha_{0,n}|+|\beta_{1,n}-\beta_{0,n}|=\Theta(1)$, *then* $(p_n,q_n)_n\in\mathcal{G}_\Phi^{upper}$;
(ii) if $|\alpha_{1,n}-\alpha_{0,n}|+|\beta_{1,n}-\beta_{0,n}|=o(1)$, *then* $(p_n,q_n)_n\in\mathcal{G}_\Phi^{upper\prime}$.

Note that our results apply even when $\frac{\alpha_{1,n}}{\beta_{1,n}}=\frac{\alpha_{0,n}}{\beta_{0,n}}$ *for all* n, *implying that* $p_n(x)$ *and* $q_n(x)$ *have the same mean.*

5.2. *Lower bound.* As with the upper bound analysis, the condition required for our lower bound depends on whether $H_n=o(1)$ or $H_n=\Theta(1)$. We capture these conditions by defining the sets $\mathcal{G}_\Phi^{\text{lower}}$ and $\mathcal{G}_\Phi^{\text{lower}\prime}$. Let $\mathcal{G}^{\text{lower}}\subset(\mathcal{P}\times\mathcal{P})^\infty$ denote the set of sequences such that $H_n=\Theta(1)$ and

$$\sup_n\int_S p_n(x)\left|\log\frac{p_n(x)}{q_n(x)}\right|^2dx<\infty,\ \text{and}\ \sup_n\int_S q_n(x)\left|\log\frac{p_n(x)}{q_n(x)}\right|^2dx\ <\infty.$$

A comparison of these conditions with Assumptions A1'–A4' shows that $\mathcal{G}^{\text{lower}} \supseteq \mathcal{G}_\Phi^{\text{upper}}$ for any Φ, since by Assumption A1', $\left|\log \frac{p_n(x)}{q_n(x)}\right|^2$ must be integrable with respect to $\phi(x)$.

When $H_n = o(1)$, we impose a much stronger condition on p_n and q_n—we require the likelihood ratio to be bounded. Define $\mathcal{G}^{\text{lower}\prime} \subset (\mathcal{P} \times \mathcal{P})^\infty$ to be the set of sequences satisfying $H_n = o(1)$ and $\sup_{n,x} \left|\log \frac{p_n(x)}{q_n(x)}\right| < \infty$.The bounded likelihood ratio condition that defines $\mathcal{G}^{\text{lower}\prime}$ is generally more restrictive than Assumptions A1–A5. However, $\mathcal{G}^{\text{lower}\prime}$ still has significant overlap with $\mathcal{G}_\Phi^{\text{upper}\prime}$ for Φ defined as in equations (3).

THEOREM 5.3. *Suppose we have K clusters, of which at least one has size $\frac{n}{\beta K}$ and at least one has size $\frac{n}{\beta K} + 1$. Let σ_0 denote the true clustering. Suppose $I_n \to 0$ and $P_{0,n}$ and $Q_{0,n}$ satisfy Assumption A0, and let $(p_n(x), q_n(x))$ be a sequence of densities either in $\mathcal{G}^{lower}$ or in $\mathcal{G}^{lower\prime}$. Then any permutation equivariant algorithm $\hat{\sigma}$ satisfies the following:*

(i) If $nI_n \to \infty$, then $\mathbb{E}l(\hat{\sigma}(A), \sigma_0) \geq \exp\left(-(1 + o(1))\frac{nI_n}{\beta K}\right)$.
(ii) If $\sup_n nI_n < \infty$, then $\liminf_{n\to\infty} \mathbb{E}l(\hat{\sigma}(A), \sigma_0) > 0$.

Theorem 5.3 shows that if $nI_n \to \infty$, the misclustering risk of any permutation equivariant algorithm is at least $\exp\left(-(1 + o(1))\frac{nI_n}{\beta K}\right)$. If $nI_n = O(1)$, any permutation invariant algorithm is inconsistent.

REMARK 5.1. *Rather than being a minimax lower bound involving a supremum over a parameter space, Theorem 5.3 applies to any collection of parameters $(p_n(x), q_n(x), P_0, Q_0, K, \beta, \sigma_0)$ that satisfy the assumptions. This is possible because the permutation equivariance condition excludes the trivial case where $\hat{\sigma} = \sigma_0$.*

Proof sketch of Theorem 5.3. The full proof of the theorem is provided in Appendix H; we highlight key points here. Although the proof is inspired by the work from Yun and Proutiere [41] and Zhang and Zhou [42], it contains significant novel elements: For instance, our result holds for any parameters in the parameter space, rather than adopting a minimax framework or assuming a prior on σ_0.

We first formalize what it means for a node to be *misclustered*. Let

$$S_K[\hat{\sigma}(A), \sigma_0] := \arg\min_{\rho \in S_K} d_H(\rho \circ \hat{\sigma}(A), \sigma_0).$$

It is straightforward to define a set of misclustered nodes when $S_K[\hat{\sigma}(A), \sigma_0]$ is a singleton, but we must be more careful when $S_K[\hat{\sigma}(A), \sigma_0]$ contains

multiple elements. We define the set of *misclustered nodes* as

$$\mathcal{E}[\hat{\sigma}(A), \sigma_0] := \Big\{ v : (\rho \circ \hat{\sigma}(A))(v) \neq \sigma_0(v), \text{ for some } \rho \in S_K[\hat{\sigma}(A), \sigma_0] \Big\}. \tag{4}$$

To see an example of the subtlety that arises when $S_K[\hat{\sigma}(A), \sigma_0]$ is not a singleton, note that the "for some $\rho \in S_K[\hat{\sigma}(A), \sigma_0]$" qualifier in the definition of $\mathcal{E}[\hat{\sigma}(A), \sigma_0]$ cannot be replaced with "for all $\rho \in S_K[\hat{\sigma}(A), \sigma_0]$." Otherwise, in the case where the clusters in σ_0 all have the same size, and where $\hat{\sigma}(A)$ is a trivial algorithm that maps all nodes to cluster 1, the set $S_K[\hat{\sigma}(A), \sigma_0]$ equals S_K and $\mathcal{E}[\hat{\sigma}(A), \sigma_0]$ would be empty.

The definition of $\mathcal{E}[\hat{\sigma}(A), \sigma_0]$ allows us to formalize certain symmetry properties of permutation equivariant estimators. For example, if A is distributed according to a weighted SBM with σ_0 as the true cluster assignment, and if nodes u and v lie in clusters of equal sizes, then $P(u \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]) = P(v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0])$ for any permutation equivariant $\hat{\sigma}$ (cf. Corollary H.1).

Without loss of generality, let cluster 1 and 2 be the clusters in σ_0 that have sizes $\frac{n}{\beta K} + 1$ and $\frac{n}{\beta K}$, respectively, and let node 1 belong to cluster 1. Let σ^* be a random cluster assignment where $\sigma^*(u) = \sigma_0(u)$ for all $u \neq 1$, and $\sigma^*(1)$ is 1 or 2 with probability $\frac{1}{2}$ each. Let $P_{SBM}(A \,|\, \sigma^*)$ denote the distribution on A induced by the random cluster assignment σ^*. We perturb $P_{SBM}(A \,|\, \sigma^*)$ to define a new distribution $P_\Psi(A)$, where under P_Ψ, the A_{uv}'s are independent; and if $u = 1$ and v is in cluster 1 or 2, then A_{uv} is distributed according to a new probability measure Y^* instead of P or Q. If $u \neq 1$ or if v is not in cluster 1 or 2, then A_{uv} is distributed according to $P_{SBM}(A \,|\, \sigma^*)$. Y^* is the arginf of $\inf_Y \max \left\{ \int_S \log \frac{dY}{dP} dY, \int_S \log \frac{dY}{dQ} dY \right\}$ and is constructed to be similar to both P and Q (cf. Lemma H.2).

Under $P_\Psi(A)$, we can show it is impossible to consistently cluster node 1 with respect to σ^* as the true cluster assignment. We then use the fact that $P_\Psi(A)$ and $P_{SBM}(A \,|\, \sigma^*)$ have similar likelihood to deduce the difficulty of correctly clustering node 1 under $P_{SBM}(A \,|\, \sigma^*)$. Permutation equivariance translates this to a result on the number of misclustered nodes. Finally, we finish the proof by using permutation equivariance again to say that the misclustering error is the same regardless of whether $\sigma^*(1) = 1$ or $\sigma^*(1) = 2$.

5.3. *Adaptivity.* Theorem 5.3, in conjunction with Theorems 5.1 and 5.2, directly implies the following corollary:

COROLLARY 5.1. *Let $\mathcal{F}_{p.e.}$ be the class of permutation equivariant estimators. Suppose $P_{0,n}$ and $Q_{0,n}$ satisfy Assumption A0 and $\{(p_n, q_n)\}_n \in \mathcal{G}_\Phi^{upper} \cap \mathcal{G}^{lower}$ or $\mathcal{G}_\Phi^{upper\prime} \cap \mathcal{G}^{lower\prime}$, for some transformation function Φ. Sup-*

pose we have K clusters, at least one of which has size $\frac{n}{\beta K}$ and at least one of which has size $\frac{n}{\beta K}+1$. Suppose $I_n \to 0$.

(i) If $\frac{nI_n}{\beta K \log n} \le 1$ and $nI_n \to \infty$, then

$$\inf_{\hat{\sigma}\in\mathcal{F}_{p.e.}} \mathbb{E}l(\hat{\sigma}(A),\sigma_0) = \exp\left(-(1+o(1))\frac{nI_n}{\beta K}\right).$$

(ii) If $\liminf_{n\to\infty} \frac{nI_n}{\beta K \log n} > 1$, then $\inf_{\hat{\sigma}\in\mathcal{F}_{p.e.}} P(l(\hat{\sigma},\sigma_0)\neq 0) \to 0$.
(iii) If $\sup_n nI_n < \infty$, then $\liminf_{n\to\infty} \inf_{\hat{\sigma}\in\mathcal{F}_{p.e.}} \mathbb{E}l(\hat{\sigma}(A),\sigma_0) > 0$.

The algorithm $\hat{\sigma}$ described in Section 4.1 with discretization level $L_n \to \infty$ achieves the optimal rate in part (i) for *any* $(P_{0,n}, Q_{0,n}, p_n, q_n)$ that satisfy the conditions of Corollary 5.1 and the additional conditions $nI_n = \omega(e^{L_n})$ and $H_n = \omega(1/L_n)$. Thus, $\hat{\sigma}$ adapts to the edge probabilities $P_{0,n}$ and $Q_{0,n}$ and the edge weight densities $p_n(x)$ and $q_n(x)$: Although $\hat{\sigma}$ has no knowledge of the parameters $(P_{0,n}, Q_{0,n}, p_n(x), q_n(x))$, it achieves the same optimal rate as if $(P_{0,n}, Q_{0,n}, p_n(x), q_n(x))$ 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 $o(1)$ term in the exponent and is thus of lower order. As an illustrative example, suppose $1-P_{0,n} = 1-Q_{0,n} = \frac{\log n}{n}$, and the densities $p_n(x)$ and $q_n(x)$ are $N(\mu_1,\sigma_1^2)$ and $N(\mu_0,\sigma_0^2)$, respectively. Then $I_n = (1+o(1))\frac{\log n}{n}\theta$, where $\theta = 2\left(1-\sqrt{\frac{2\sigma_1^2\sigma_0^2}{\sigma_1^2+\sigma_0^2}}e^{-\frac{1}{4}\frac{(\mu_1-\mu_0)^2}{\sigma_1^2+\sigma_0^2}}\right)$, 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 $1-P_{0,n} = 1-Q_{0,n} = \frac{\log n}{n}$ and the densities $p_n(x)$ and $q_n(x)$ are $\mathrm{Exp}(\lambda_1)$ and $\mathrm{Exp}(\lambda_0)$, respectively, then $I_n = (1+o(1))\frac{\log n}{n}\theta'$, where $\theta' = 2\left(1-\sqrt{\frac{\lambda_1\lambda_0}{\lambda_1+\lambda_0}}\right)$. The optimal rate $n^{-(1+o(1))\frac{2\theta'}{\beta K}}$ is again achieved by $\hat{\sigma}$.

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.

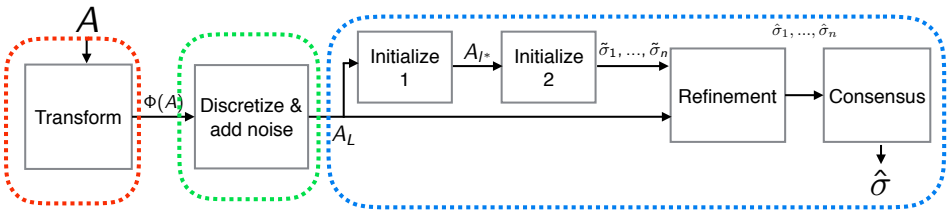

FIG 2. *Analysis of the right-most blue region is in subsection 6.1, of the middle green region in subsection 6.2, and of the left-most red region in subsection 6.3*

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 2) for recovering communities on a network where the edges have discrete labels $l = 1, \ldots, L_n$. The following proposition characterizes the rate of convergence of the output of the subroutine, where within-community edges are assigned edge labels with probabilities $\{P_{l,n}\}$, and between-community edges are assigned edge labels according to $\{Q_{l,n}\}$. For convenience, if an edge does not exist between u and v, we assign the label 0 to A_{uv}, so $P_{0,n}, Q_{0,n}$ are the edge absence probabilities.

PROPOSITION 6.1. *Suppose* $\frac{1}{\rho_n} \leq \frac{P_{l,n}}{Q_{l,n}} \leq \rho_n$, *for a sequence* $\rho_n = \Omega(1)$ *for all* $l = 0, 1, \ldots, L_n$. *Define* $I_{L_n} = -2 \log \sum_{l=0}^{L_n} \sqrt{P_{l,n} Q_{l,n}}$ *and suppose* $I_{L_n} \to 0$. *Suppose* $L_n = \Omega(1)$ *and* $\frac{nI_{L_n}}{L_n \rho_n^4} \to \infty$. *Then*

$$\lim_{n\to\infty} P\left(l(\hat{\sigma}(A), \sigma_0) \leq \exp\left(-\frac{nI_{L_n}}{\beta K}(1 + o(1)) \right) \right) \to 1.$$

Furthermore, if $\frac{nI_{L_n}}{\beta K \log n} \leq 1$, *then*

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

Note that I_{L_n} is the Renyi divergence of order $\frac{1}{2}$ between the discrete distributions defined by $\{P_{l,n}\}_{l=0,\ldots,L_n}$ and $\{Q_{l,n}\}_{l=0,\ldots,L_n}$.

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 H.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 2). The input to this step is the weighted network $\Phi(A)$ in which all the edge weights are in $[0,1]$. We use $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$ for $z \in [0,1]$ to denote the densities of the transformed edge weights; the next section shows the relationship between $p_{\Phi,n}(z)$ and $p_n(x)$ and $q_{\Phi,n}(z)$ and $q_n(x)$. 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 $P_{l,n} = (1-P_{0,n})\int_{a_l}^{b_l} p_{\Phi,n}(z)dz$ or $Q_{l,n} = (1-Q_{0,n})\int_{a_l}^{b_l} q_{\Phi,n}(z)dz$. A missing edge is assigned the label 0. Define

$$I_{\Phi,n} = -2\log\left(\sqrt{P_{0,n}Q_{0,n}} + \int_0^1 \sqrt{(1-P_{0,n})(1-Q_{0,n})p_{\Phi,n}(z)q_{\Phi,n}(z)}dz\right)$$

to be the Renyi divergence between $(P_{0,n}, p_{\Phi,n})$ and $(Q_{0,n}, q_{\Phi,n})$. It is easy to show that discretization always leads to a loss of information; i.e., $I_{L_n} \le I_{\Phi,n}$. Propositions 6.2 and 6.3 show that if $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$ are sufficiently regular in that they can be well-approximated via discretization, then $|I_{L_n} - I_{\Phi,n}| = o(I_{\Phi,n})$.

Again, we have two different sets of conditions depending on whether $H_{\Phi,n} := \int_0^1 (\sqrt{p_{\Phi,n}(z)} - \sqrt{q_{\Phi,n}(z)})^2 dz = \Theta(1)$ or $o(1)$. The following proposition shows that I_{L_n} approximates $I_{\Phi,n}$ well when $H_{\Phi,n} = \Theta(1)$ and is useful for proving Theorem 5.1:

PROPOSITION 6.2. *Suppose $P_{0,n}$ and $Q_{0,n}$ satisfy Assumption A0 and $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$ are supported on $[0,1]$. Suppose $H_{\Phi,n} = \Theta(1)$ and*

C1 $p_{\Phi,n}(z), q_{\Phi,n}(z) > 0$ on $(0,1)$, and $\sup_n \sup_z \{p_{\Phi,n}(z) \vee q_{\Phi,n}(z)\} < \infty$.

C2 $\sup_n \int_0^1 \left|\log \frac{p_{\Phi,n}(z)}{q_{\Phi,n}(z)}\right|^r dz < \infty$ for an absolute constant $r > 4$.

C3 There exists $h_{\Phi,n}(z)$ such that

(a) $h_{\Phi,n}(z) \ge \max\left\{\left|\frac{p'_{\Phi,n}(z)}{p_{\Phi,n}(z)}\right|, \left|\frac{q'_{\Phi,n}(z)}{q_{\Phi,n}(z)}\right|\right\}$,

(b) $h_{\Phi,n}(z)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped, and

(c) $\sup_n \int_0^1 |h_{\Phi,n}(z)|^t dz < \infty$ *for some constant* t *such that* $\frac{2}{r} \leq t \leq 1$.

C4 $p'_{\Phi,n}(z), q'_{\Phi,n}(z) \geq 0$ *for all* $z < c'_{s1}$, *and* $p'_{\Phi,n}(z), q'_{\Phi,n}(z) \leq 0$ *for all* $z > c'_{s2}$.

Then

$$\left| \frac{I_{\Phi,n} - I_{L_n}}{I_{\Phi,n}} \right| = o(1),$$

and $\frac{1}{4c_0} \exp(-L_n^{1/r}) \leq \frac{P_{l,n}}{Q_{l,n}} \leq 4c_0 \exp(L_n^{1/r})$, *for all* l, *where* c_0 *is the constant in Assumption A0.*

We prove Proposition 6.2 in Section E of the Appendix.

The following proposition handles the case where $H_{\Phi,n} = o(1)$, and is useful for proving Theorem 5.2:

PROPOSITION 6.3. *Suppose* $P_{0,n}, Q_{0,n}$ *satisfy assumption A0 and* $p_{\Phi,n}(z)$ *and* $q_{\Phi,n}(z)$ *are supported on* $[0,1]$. *Suppose* $H_{\Phi,n} = o(1)$. *Let* L_n *be a sequence such that* $L_n \to \infty$ *and suppose* $L_n \leq \frac{2}{H_{\Phi,n}}$. *Suppose the following assumptions are satisfied:*

C1' $p_{\Phi,n}(z), q_{\Phi,n}(z) > 0$ *on* $(0,1)$, *and* $\sup_n \sup_z \{p_{\Phi,n}(z) \vee q_{\Phi,n}(z)\} < \infty$.

C2' There exists a subinterval $R_{\Phi,n} \subseteq [0,1]$ *such that*

(a) $\frac{1}{\rho} \leq \left| \frac{p_{\Phi,n}(z)}{q_{\Phi,n}(z)} \right| \leq \rho$ *for all* $z \in R_{\Phi,n}$, *where* ρ *is an absolute constant.*

(b) $\mu\{R^c_{\Phi,n}\} = o(H_{\Phi,n})$, *where* μ *is the Lebesgue measure and* $R^c_{\Phi,n} := [0,1] \backslash R_{\Phi,n}$.

C3' Let $\alpha^2_{\Phi,n} = \int_{R_{\Phi,n}} \frac{(p_{\Phi,n}(z) - q_{\Phi,n}(z))^2}{q_{\Phi,n}(z)} dz$ *and* $\gamma_{\Phi,n}(z) = \frac{q_{\Phi,n}(z) - p_{\Phi,n}(z)}{\alpha}$. *There exists a constant* $r > 4$ *such that* $\sup_n \int_{R_{\Phi,n}} q_{\Phi,n}(z) \left| \frac{\gamma_{\Phi,n}(z)}{q_{\Phi,n}(z)} \right|^r dz < \infty$.

C4' There exists $h_{\Phi,n}(z)$ *such that*

(a) $h_{\Phi,n}(z) \geq \max \left\{ \left| \frac{\gamma'_{\Phi,n}(z)}{q_{\Phi,n}(z)} \right|, \left| \frac{q'_{\Phi,n}(z)}{q_{\Phi,n}(z)} \right| \right\}$, *and*

(b) $h_{\Phi,n}(z)$ *is* (c'_{s1}, c'_{s2}, C'_s)*-bowl-shaped for absolute constants* c'_{s1}, c'_{s2}, *and* C'_s, *and*

(c) $\sup_n \int_{R_{\Phi,n}} |h_{\Phi,n}(z)|^t dz < \infty$ *for an absolute constant* $\frac{2}{r} < t < 1$.

C5' $p'_{\Phi,n}(z), q'_{\Phi,n}(z) \geq 0$ *for all* $z < c'_{s1}$, *and* $p'_{\Phi,n}(z), q'_{\Phi,n}(z) \leq 0$ *for all* $z > c'_{s2}$.

Then

$$\left| \frac{I_{\Phi,n} - I_{L_n}}{I_{\Phi,n}} \right| = o(1)$$

and $\frac{1}{4\rho c_0} \leq \frac{P_{l,n}}{Q_{l,n}} \leq 4\rho c_0$, *for all* l, *where* c_0 *is the constant in A0.*

We prove Proposition 6.3 in Section C of the Appendix.

6.3. *Analysis of the transformation function.* Propositions 6.2 and 6.3 consider densities supported on $[0,1]$. This suffices, because once we transform the densities by an application of Φ, the new densities are compactly supported and, importantly, the Renyi divergence I_n and the Hellinger divergence H_n are invariant with respect to the transformation Φ.

To see this, let $p_n(x)$ and $q_n(x)$ denote densities over S, and let $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$ denote the transformed densities over $[0,1]$. It is easy to see that $p_{\Phi,n}(z) = \frac{p_n(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $q_{\Phi,n}(z) = \frac{q_n(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$. Therefore, via the change of variables $z = \Phi^{-1}(x)$, we have the following relations:

$$\int_S \sqrt{p_n(x)q_n(x)}dx = \int_0^1 \sqrt{p_{\Phi,n}(z)q_{\Phi,n}(z)}dz,$$
$$\int_S \left(\sqrt{p_n(x)} - \sqrt{q_n(x)}\right)^2 dx = \int_0^1 \left(\sqrt{p_\Phi(z)} - \sqrt{q_\Phi(z)}\right)^2 dz.$$

Therefore, the Renyi and Hellinger divergences between $p_n(x)$ and $q_n(x)$ are equal to the divergences between $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$; i.e.,

$$I_{\Phi,n} = I_n \qquad H_{\Phi,n} = H_n$$

To prove Theorems 5.1 and 5.2, it remains to show that if $p_n(x)$ and $q_n(x)$ satisfy Assumptions A1–A4 (or A1'–A5'), the transformed densities $p_{\Phi,n}(z)$ and $q_{\Phi,n}(z)$ satisfy Assumptions C1–C4 (or C1'–C5') in Proposition 6.2 (or Proposition 6.3). This is done in Propositions F.1 and F.2.

7. Conclusion. We have provided a rate-optimal community estimation algorithm for the homogeneous weighted stochastic block model. Our algorithm includes a preprocessing step consisting of transforming and discretizing the (possibly) continuous edge weights to obtain a simpler graph with edge weights supported on a finite, discrete set. This approach may be useful for other network data analysis problems involving continuous distributions, where discrete versions of the problem are simpler to analyze.

Our paper provides a significant step toward understanding the weighted SBM under the same mathematical framework that has been exceptionally fruitful in the case of unweighted models. It is far from comprehensive, however, and many open questions remain. We describe a few here:

1. An important extension is the *heterogenous* stochastic block model, where edge weight distributions depend on the exact community assignments of both endpoints. In such a setting, Abbe and Sandon [4]

and Yun and Proutiere [41] have shown that a generalized information divergence—the CH divergence—governs the intrinsic difficulty of community recovery. We believe that a similar discretization-based approach should lead to analogous results in the case of a heterogeneous weighted SBM. The key challenge would be to show that discretization does not lose much information with respect to the CH-divergence.

2. Real-world networks often have nodes with very high degrees, which may adversely affect the accuracy of recovery methods for the stochastic block model. To solve this problem, degree-corrected SBMs [44, 16] have been proposed as an effective alternative to regular SBMs. It is straightforward to extend the concept of degree-correction to the weighted SBM, but it is unclear whether our discretization-based approach would be effective in obtaining optimal error rates.
3. It is easy to extend our results to the weighted *and* labeled SBMs if the number of labels is finite or assumed to be slowly growing. However, this excludes some interesting cases such as when edge labels represent counts from a Poisson distribution. We suspect that in such a situation, it is possible to combine low-probability labels in a clever way to obtain a discretization amenable to our approach.

Acknowledgments. The authors would like to thank Zongming Ma for extensive and enlightening discussions in the earlier stages of this project.

References.

[1] E. Abbe. Community detection and stochastic block models: Recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

[2] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014.

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

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

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

[6] C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, page cnu026, 2014.

[7] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962, 2011.

[8] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
[10] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.
[11] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Proceedings of The 28th Conference on Learning Theory*, pages 391–423, 2015.
[12] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011.
[13] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.
[14] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.
[15] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.
[16] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*, 2016.
[17] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, February 2010.
[18] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.
[19] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.
[20] B. Hajek, Y. Wu, and J. Xu. Submatrix localization via message passing. *arXiv preprint arXiv:1510.09219*, 2015.
[21] B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. *IEEE Transactions on Information Theory*, 2017.
[22] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6):175–181, 2000.
[23] S. Heimlicher, M. Lelarge, and L. Massoulié. Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*, 2012.
[24] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
[25] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2010.
[26] V. Jog and P. Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.
[27] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.
[28] M. Lelarge, L. Massoulié, and J. Xu. Reconstruction in the labeled stochastic block model. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE, 2013.
[29] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703. ACM, 2014.
[30] F. McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.
[31] E. Mossel, J. Neeman, and A. Sly. Stochastic Block Models and Reconstruction.

arXiv preprint arXiv:1202.1499, 2012.
[32] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.
[33] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.
[34] M. Newman, A.-L. Barabasi, and D. J. Watts. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, NJ, USA, 2006.
[35] M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.
[36] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
[37] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.
[38] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. Computational Models of the Brain.
[39] D.S. Sade. Sociometrics of Macaca mulatta: I. Linkages and cliques in grooming matrices. *Folia Primatologica*, 18(3–4):196–223, 1972.
[40] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.
[41] S. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.
[42] A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block model. *arXiv preprint arXiv:1507.05313*, 2015.
[43] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.
[44] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

SUPPLEMENTARY MATERIAL

Supplement to: Optimal rates for community estimation in the weighted stochastic block model
(doi: 10.1214/00-AOASXXXXSUPP; .pdf). We provide detailed proofs of the theorems and propositions in the main paper.

Submitted to the Annals of Statistics

SUPPLEMENT TO: OPTIMAL RATES FOR COMMUNITY ESTIMATION IN THE WEIGHTED STOCHASTIC BLOCK MODEL

By Min Xu[‡] , Varun Jog[§] and Po-Ling Loh[§]

University of Pennsylvania[‡] and University of Wisconsin - Madison[§]

This supplement contains supporting appendices for the main manuscript, "Optimal rates for community estimation in the weighted stochastic block model."

APPENDIX A: PROOF OF PROPOSITION 6.1

We structure the proof according to the flow of our algorithm. Since this proposition addresses the case of discrete labels, we do not need to consider the "transformation and discretization" step. We begin by analyzing Algorithm 4.2, where we deliberately add noise to the graph by changing edge colors at random. Although adding random noise destroys information and increases the difficulty of community recovery, Lemma B.2 in Appendix B.5 shows that the process does not significantly affect the Renyi divergence I_L. Furthermore, the new probabilities of edge labels are at least $\frac{c}{n}$, which is important for our later analysis. To simplify notation, we continue to refer the new edge label probabilities as P_l and Q_l throughout the proof.

Next, our algorithm performs spectral clustering using only the edges with label l, and calculates $\hat{I}_l := \frac{(\hat{P}_l - \hat{Q}_l)^2}{\hat{P}_l \vee \hat{Q}_l}$, where $\hat{P}_l$ and $\hat{Q}_l$ are the estimated probabilities obtained by using the output of spectral clustering:

$$\hat{P}_l = \frac{\sum_{u \neq v \,:\, \tilde{\sigma}_l(u) = \tilde{\sigma}_l(v)} (A_l)_{uv}}{|u \neq v \,:\, \tilde{\sigma}_l(u) = \tilde{\sigma}_l(v)|}, \quad \text{and} \quad \hat{Q}_l = \frac{\sum_{u \neq v \,:\, \tilde{\sigma}_l(u) \neq \tilde{\sigma}_l(v)} (A_l)_{uv}}{|u \neq v \,:\, \tilde{\sigma}_l(u) \neq \tilde{\sigma}_l(v)|}.$$

We then select $l^* \in \arg\max_{1 \leq l \leq L} \hat{I}_L$. Note that if $|\hat{P}_l - P_l|$ and $|\hat{Q}_l - Q_l|$ are small, $\hat{I}_l$ provides a measure of how "good" a color is for clustering: larger values of $\hat{I}_l$ correspond to greater separation between P_l and Q_l. Naturally, the accuracy of the estimated edge probabilities $\hat{P}_l$ and $\hat{Q}_l$ depends on the accuracy of the spectral clustering step. Proposition B.1 below makes this statement rigorous. Before stating the proposition, we define the set of "good" labels as follows:

$$L_1 = \left\{ l : \frac{n(P_l - Q_l)^2}{P_l \vee Q_l} := \frac{\Delta_l^2}{P_l \vee Q_l} \geq 1 \right\}.$$

We bound the difference between the estimated and true probabilities for good and bad colors in Proposition B.1, the formal statement and proof of which are contained in Appendix B.1.

PROPOSITION B.1. *Suppose σ is a clustering with error rate at most γ; i.e., $l(\sigma, \sigma_0) \leq \gamma$, for sufficiently small $\gamma \geq \frac{1}{n}$. With probability at least $1 - Ln^{-(3+\delta_p)}$, for a small $\delta_p > 0$, the following hold:*

1. *For $l \in L_1$, we have $|\hat{P}_l - P_l| \leq \eta\Delta_l$ and $|\hat{Q}_l - Q_l| \leq \eta\Delta_l$.*
2. *For $l \in L_1^c$, we have $|\hat{P}_l - P_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}$ and $|\hat{Q}_l - Q_l| \leq \eta\sqrt{\frac{P_l \vee Q_l}{n}}$.*

In both cases, $\eta = C\sqrt{\gamma \log \frac{1}{\gamma}}$, for an absolute constant C.

We now work toward obtaining an initial clustering with small error rate γ. In Proposition B.2, we show that if the edge probabilities for a particular label are well-separated, the spectral clustering output of Algorithm 4.4 is reasonably accurate. We provide a rough statement of the proposition here, and refer to Appendix B.2 for the precise statement and proof.

PROPOSITION B.2. *If P_l and Q_l satisfy $C_1 \frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2} \leq 1$ for an absolute constant C_1, the output σ^l of Algorithm 4.4 satisfies the inequality*

$$l(\sigma^l, \sigma_0) \leq C_2 \frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2},$$

for a constant C_2, with probability at least $1 - n^{-4}$.

Thus, if we want to cluster to an arbitrary degree of accuracy, we need $\frac{(P_l \vee Q_l)}{n(P_l - Q_l)^2} \to 0$ for at least one well-separated label l. We show that the label l^* selected in Algorithm 4.3 satisfies $\frac{(P_{l^*} \vee Q_{l^*})}{n(P_{l^*} - Q_{l^*})^2} \to 0$ in the following proposition. A more detailed restatement and proof is contained in Appendix B.3.

PROPOSITION B.4. *With probability at least $1 - 2Ln^{-(3+\delta_p)}$, we have $\frac{n(P_{l^*} - Q_{l^*})^2}{\rho_L^4 (P_{l^*} \vee Q_{l^*})} \to \infty$.*

Now let E_1 denote the high-probability event that the label l^* is chosen according to Proposition B.4. Our algorithm performs spectral clustering n times, omitting one vertex and clustering on the remaining graph each time, and we denote the resulting community assignments by $\{\tilde{\sigma}_u\}_{u=1,\ldots,n}$.

Let E_2 denote the event that all the clusterings$\{\tilde{\sigma}\}_{u=1,\ldots,n}$ have error rate bounded by $\gamma := C\frac{P_{l^*} \vee Q_{l^*}}{n(P_{l^*} - Q_{l^*})^2}$, for some constant C. Proposition B.2,

together with a simple union bound, implies that E_2 occurs with probability at least $1 - n^{-3}$.

On $E_1 \cap E_2$, we then have $\gamma\rho_L^4 \to 0$. Thus, we may apply Proposition B.1 on each clustering $\tilde{\sigma}_u$ to show that the conclusion of proposition B.1 holds simultaneously for all $\tilde{\sigma}_u$'s, with probability at least $1 - Ln^{-(2+\delta_p)}$. Furthermore, the η in proposition B.1 satisfies $\eta = \Theta\left(\sqrt{\gamma \log \frac{1}{\gamma}}\right)$, so $|\eta\rho_L^2| \to 0$. We denote this last event by E_3.

We now construct the clustering $\hat{\sigma}_u$ which is identical to $\tilde{\sigma}_u$ except that node u is possibly re-assigned using the relation from Algorithm 4.5.

In Proposition B.5, we show that with high probability, the assignment $\hat{\sigma}_u(u)$ is "correct". We provide a rough statement of the proposition here, and defer the exact statement and proof to Appendix B.4:

Define $S_K[\hat{\sigma}_u, \sigma_0] := \arg\min_{\rho \in S_K} d_H(\rho \circ \tilde{\sigma}, \sigma_0)$ to be set of all cluster label permutations minimizing the Hamming distance between $\tilde{\sigma}_u$ and σ_0.

PROPOSITION B.5. *Let $\pi_u \in S_K[\hat{\sigma}_u, \sigma_0]$. Conditioned on $E_1 \cap E_2 \cap E_3$, with probability at least $1 - (K-1)\exp\left(-(1-o(1))\frac{n}{\beta K} I_L\right)$, we have $\pi_u^{-1}(\sigma_0(u)) = \hat{\sigma}_u(u)$.*

Before proceeding, we need to show that on event E_2, the set $S_K[\hat{\sigma}_u, \sigma_0]$ is a singleton for any u. By construction, the error rate of $\hat{\sigma}_u$ is at most $\gamma + \frac{1}{n}$, so $l(\sigma_0, \hat{\sigma}_u) < \frac{1}{8\beta K}$ for sufficiently small γ. Now note that for any u, the minimum cluster size of the clustering $\hat{\sigma}_u$ is at least $\frac{n}{\beta K} - (n\gamma + 1) \geq \frac{n(1-\beta K\gamma - \beta K/n)}{\beta K} \geq \frac{n}{2\beta K}$, for small γ. A simple argument (cf. Lemma B.8) shows that there is a unique permutation obtaining such a small error rate.

Let us still condition on E_2. Define π_1 and π_u as the unique element in $S_K[\hat{\sigma}_1, \sigma_0]$ and $S_K[\hat{\sigma}_u, \sigma_0]$ respectively. We now show that $\pi_1^{-1} \circ \pi_u$ is the only element in $S_K[\hat{\sigma}_u, \hat{\sigma}_1]$. We know that $d_H(\sigma_0, \pi_1 \circ \hat{\sigma}_1) < \frac{1}{8\beta K}$ and $d(\sigma_0, \pi_u \circ \hat{\sigma}_u) < \frac{1}{8\beta K}$. Thus, the triangle inequality implies

$$d(\hat{\sigma}_1, \pi_1^{-1} \circ \pi_u \circ \hat{\sigma}_u) = d(\pi_1 \circ \hat{\sigma}_1, \pi_u \circ \hat{\sigma}_u) \leq d(\sigma_0, \pi_1 \circ \hat{\sigma}_1) + d(\sigma_0, \pi_u \circ \hat{\sigma}_u) < \frac{1}{4\beta K}.$$

Since the minimum cluster size of both $\hat{\sigma}_1$ and $\hat{\sigma}_u$ is $\frac{n}{2\beta K}$, Lemma B.8 implies that $\pi_1^{-1} \circ \pi_u$ is the only element in $S_K[\hat{\sigma}_u, \hat{\sigma}_1]$.

Let $\hat{\sigma}$ be the cluster assignment created after the consensus stage. By Lemma B.7, we have that $\hat{\sigma}(u) = (\pi_1^{-1} \circ \pi_u \circ \hat{\sigma}_u)(u)$ where $\pi_1^{-1} \circ \pi_u$ is the unique element in $S_K[\hat{\sigma}_u, \hat{\sigma}_1]$. Therefore, we have that $\hat{\sigma}_u(u) \neq (\pi_u^{-1} \circ \sigma_0)(u)$ if and only if $(\pi_1 \circ \hat{\sigma})(u) \neq \sigma_0(u)$.

Restating Proposition B.5, we then have

$$P\left(\hat{\sigma}(u) \neq \pi_1^{-1}(\sigma_0(u)), \exists \pi_1 \in S_K[\hat{\sigma}_1, \sigma_0] \middle| E_1 \cap E_2 \cap E_3\right) \leq \exp\left(-(1+o(1))\frac{nI_L}{\beta K}\right). \tag{5}$$

Therefore,

$$\begin{aligned}
\mathbb{E}l(\hat{\sigma}, \sigma_0) &\leq \mathbb{E}\Big[l(\hat{\sigma}, \sigma_0) \,|\, E_1 \cap E_2 \cap E_3\Big] + P(E_1^c) + P(E_2^2) + P(E_3^c) \\
&\leq \mathbb{E}\Big[l(\hat{\sigma}, \sigma_0) \,|\, E_1 \cap E_2 \cap E_3\Big] + 2Ln^{-3+\delta_p} + n^{-3} + Ln^{-(2+\delta_p)} \\
&\leq \mathbb{E}\Big[l(\hat{\sigma}, \sigma_0) \,|\, E_1 \cap E_2 \cap E_3\Big] + Cn^{-(2+\delta_p)}
\end{aligned}$$

Now we bound the first term:

$$\begin{aligned}
&\mathbb{E}\Big[l(\hat{\sigma}, \sigma_0) \,|\, E_1 \cap E_2 \cap E_3\Big] \\
&\quad= \mathbb{E}\left[\min_{\tau \in S_K} \frac{1}{n}\sum_{u=1}^{n} \mathbf{1}\{(\tau \circ \hat{\sigma})(u) \neq \sigma_0(u)\} \,\middle|\, E_1 \cap E_2 \cap E_3\right] \\
&\quad\leq \mathbb{E}\left[\min_{\tau \in S_K[\hat{\sigma}_1, \sigma_0]} \frac{1}{n}\sum_{u=1}^{n} \mathbf{1}\{(\tau \circ \hat{\sigma})(u) \neq \sigma_0(u)\} \,\middle|\, E_1 \cap E_2 \cap E_3\right] \\
&\quad\leq \mathbb{E}\left[\frac{1}{n}\sum_{u=1}^{n} \mathbf{1}\{(\tau \circ \hat{\sigma})(u) \neq \sigma_0(u),\, \exists \tau \in S_K[\hat{\sigma}_1, \sigma_0]\} \,\middle|\, E_1 \cap E_2 \cap E_3\right] \\
&\quad\leq \frac{1}{n}\sum_{u=1}^{n} P\big\{(\tau \circ \hat{\sigma})(u) \neq \sigma_0(u),\, \exists \tau \in S_K[\hat{\sigma}_1, \sigma_0] \,\big|\, E_1 \cap E_2 \cap E_3\big\} \\
&\quad\leq \exp\left(-(1+o(1))\frac{nI_L}{\beta K}\right)
\end{aligned}$$

Altogether, we conclude that

$$\mathbb{E}l(\hat{\sigma}, \sigma_0) \leq \exp\left(-(1+\eta')\frac{nI_L}{\beta K}\right) + n^{-(2+\delta_p)}$$

where $\eta' = o(1)$. If $\frac{nI_L}{\beta K \log n} \leq 1$, then $\exp\left(-(1+\eta')\frac{nI_L}{\beta K}\right) \geq n^{-(1-\eta')} \geq Cn^{-(2+\delta_p)}$ for small enough η'. Therefore, if $\frac{nI_L}{\beta K \log n} \leq 1$, we have

$$\mathbb{E}l(\hat{\sigma}, \sigma_0) \leq 2\exp\left(-(1+\eta')\frac{nI_L}{\beta K}\right) \leq \exp\left(-(1+o(1))\frac{nI_L}{\beta K}\right),$$

where the 2 is absorbed into the $o(1)$ term by the assumption that $nI_L \to \infty$.

Now we prove the probability bound. First suppose $\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) \geq n^{-(1+\delta_p)}$. Defining $\eta'' = \eta' + \beta\sqrt{\frac{K}{nI_L}} = o(1)$, we have

$$P\left\{l(\hat{\sigma},\sigma_0) > \exp\left(-(1-\eta'')\frac{nI_L}{\beta K}\right)\right\} \leq \frac{\mathbb{E}l(\hat{\sigma},\sigma_0)}{\exp\left(-(1-\eta'')\frac{nI_L}{\beta K}\right)}$$
$$\leq \exp\left\{-(\eta''-\eta')\frac{nI_L}{\beta K}\right\} + \frac{Cn^{-(2+\delta_p)}}{\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right)} = o(1).$$

On the other hand, if $\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) \leq n^{-(1+\delta_p)}$, we have

$$\begin{aligned}
P\left\{l(\hat{\sigma},\sigma_0) > \exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right)\right\} &\leq P(l(\hat{\sigma},\sigma_0) > 0) \\
&\leq P\left(\min_{\tau\in S_K} d_H(\tau\circ\hat{\sigma},\sigma_0) > 0\right) \\
&\leq P\left(\min_{\tau\in S_K[\hat{\sigma}_1,\sigma_0]} d_H(\tau\circ\hat{\sigma},\sigma_0) > 0\right) \\
&\leq \sum_{u=1}^{n} P\Big((\tau\circ\hat{\sigma})(u) \neq \sigma_0(u),\, \exists\tau\in S_K[\hat{\sigma}_1,\sigma_0]\Big) \\
&\overset{(a)}{\leq} n\exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) + Cn^{-(1+\delta_p)} = o(1),
\end{aligned}$$

where (a) follows because

$$\begin{aligned}
&P\Big((\tau\circ\hat{\sigma})(u) \neq \sigma_0(u),\, \exists\tau\in S_K[\hat{\sigma}_1,\sigma_0]\Big) \\
&\leq P\left(\hat{\sigma}(u) \neq \tau^{-1}(\sigma_0(u)), \exists\tau\in S_K[\hat{\sigma}_1,\sigma_0]\middle| E_1\cap E_2\cap E_3\right) \\
&\quad + P(E_1^c) + P(E_2^c) + P(E_3^c) \\
&\leq \exp\left(-(1-\eta')\frac{nI_L}{\beta K}\right) + Cn^{-(2+\delta_p)},
\end{aligned}$$

using inequality (5). This completes the proof of Proposition 6.1.

APPENDIX B: SUPPORTING RESULTS FOR PROPOSITION 6.1

We now provide proofs for the supporting results stated in Appendix A.

B.1. Analysis of estimation error of $\hat{P}_l$ and $\hat{Q}_l$.

PROPOSITION B.1. *Let A be the adjacency matrix of a labeled network with true clustering assignment σ_0. Suppose σ is a random initial clustering satisfying $l(\sigma, \sigma_0) \leq \gamma$. Let $\hat{P}_l = \frac{\sum_{u \neq v\,:\,\sigma(u)=\sigma(v)} \mathbf{1}(A_{uv}=l)}{|\{u \neq v\colon \sigma(u)=\sigma(v)\}|}$ and $\hat{Q}_l = \frac{\sum_{u \neq v\,:\,\sigma(u)\neq\sigma(v)} \mathbf{1}(A_{uv}=l)}{|\{u \neq v\colon \sigma(u)=\sigma(v)\}|}$ be the MLE of P_l and Q_l based on σ. Let δ_p be a positive, fixed, and arbitrarily small real number, and let $c > 0$ be an absolute constant. Then with probability at least $1 - Ln^{-(3+\delta_p)}$, the following hold for all sufficiently small γ:*

1. *For all l such that $P_l \vee Q_l \geq \frac{c}{n}$, if $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq 1$, then*

$$|\hat{P}_l - P_l| \leq \eta \Delta_l, \quad \text{and} \quad |\hat{Q}_l - Q_l| \leq \eta \Delta_l.$$

2. *For all l such that $P_l \vee Q_l \geq \frac{c}{n}$, if $\frac{n\Delta_l^2}{P_l \vee Q_l} \leq 1$, then*

$$|\hat{P}_l - P_l| \leq \eta \sqrt{\frac{P_l \vee Q_l}{n}}, \quad \text{and} \quad |\hat{Q}_l - Q_l| \leq \eta \sqrt{\frac{P_l \vee Q_l}{n}}.$$

In both cases, $\eta = C\sqrt{\gamma \log \frac{1}{\gamma}}$, for an absolute constant C.

PROOF. Our proof proceeds by showing that for any fixed assignment σ with error rate bounded by γ, the event described in Proposition B.1 holds with high probability. For a fixed assignment σ, we call a random graph a "bad graph for σ" if the event does not hold. For each σ, we upper-bound the probability that a randomly chosen graph lies in the set of bad graphs for σ; we then use a union bound over all choices of σ and all L colors to show that the probability of choosing a bad graph is bounded by $Ln^{-(3+\delta_p)}$.

We begin by bounding the bias of $\hat{P}_l$. We have

$$\begin{aligned} \mathbb{E}(\hat{P}_l) &= \frac{\sum_{u \neq v\,:\,\sigma(u)=\sigma(v)} \{\mathbf{1}(\sigma_0(u) = \sigma_0(v))P_l + \mathbf{1}(\sigma_0(u) \neq \sigma_0(v))Q_l\}}{|\{u \neq v :\ \sigma(u) = \sigma(v)\}|} \\ &= (1-\lambda)P_l + \lambda Q_l = P_l + \lambda(Q_l - P_l), \end{aligned} \tag{6}$$

where $\lambda := \frac{\sum_{u \neq v\,:\,\sigma(u)=\sigma(v)} \mathbf{1}(\sigma_0(u) \neq \sigma_0(v))}{|\{u \neq v\colon \sigma(u)=\sigma(v)\}|}$. Thus, $|\mathbb{E}(\hat{P}_l) - P_l| \leq \lambda |Q_l - P_l|$. Furthermore, if $\hat{n}_k$ denotes the number of vertices in cluster k according to

σ, we have

$$\begin{aligned}\lambda &= \frac{\sum_{u\neq v\,:\,\sigma(u)=\sigma(v)} \mathbf{1}(\sigma_0(u)\neq\sigma_0(v))}{|\{u\neq v:\ \sigma(u)=\sigma(v)\}|} \\ &= \frac{\sum_k \sum_{u\neq v\,:\,\sigma(u)=\sigma(v)=k} \mathbf{1}(\sigma_0(u)\neq\sigma_0(v))}{\sum_k \hat{n}_k(\hat{n}_k-1)} \\ &\leq \frac{\sum_k \sum_{u\neq v\,:\,\sigma(u)=\sigma(v)=k} \mathbf{1}(\neg(\sigma_0(u)=\sigma_0(v)=k))}{\sum_k \hat{n}_k(\hat{n}_k-1)} \\ &\leq \frac{\sum_k \sum_{u\neq v\,:\,\sigma(u)=\sigma(v)=k} \{\mathbf{1}(\sigma_0(v))\neq k)+\mathbf{1}(\sigma_0(u)\neq k)\}}{\sum_k \hat{n}_k(\hat{n}_k-1)}.\end{aligned}$$

Define $\gamma_k = \frac{1}{n}\sum_{u\,:\,\sigma(u)=k}\mathbf{1}(\sigma_0(u)\neq k)$ to be the error rate within the estimated cluster k. Then $\sum_k \gamma_k \leq \gamma$ and $\sum_{u\,:\,\sigma(u)=k}\sum_{v\,:\,\sigma(v)=k}\mathbf{1}(\sigma_0(v)\neq k) = \gamma_k n\hat{n}_k$, implying that

$$\lambda \leq \frac{\sum_k 2\gamma_k n\hat{n}_k}{\sum_k \hat{n}_k(\hat{n}_k-1)} = \frac{n}{\sum_k \hat{n}_k(\hat{n}_k-1)}\sum_k 2\gamma_k\hat{n}_k \overset{(a)}{\leq} \frac{K}{n-K}\sum_k 2\gamma_k\hat{n}_k \overset{(b)}{\leq} 4\gamma K,$$

where (a) uses the fact that

$$\sum_k \frac{\hat{n}_k}{n}(\hat{n}_k-1) = n\sum_k\left(\frac{\hat{n}_k}{n}\right)^2 - 1 \geq \frac{n}{K}-1, \tag{7}$$

and (b) uses the assumption $K<\frac{n}{2}$. Altogether, we conclude that $|\mathbb{E}(\hat{P}_l)-P_l|\leq 4\gamma K\Delta_l$. A similar calculation may be performed for $\hat{Q}_l$, so

$$\max\left\{|\mathbb{E}(\hat{P}_l)-P_l|,\ |\mathbb{E}(\hat{Q}_l)-Q_l|\right\}\leq C_2\gamma\Delta_l,$$

for a constant $C_2>0$. To simplify presentation, we define $\eta_1 = C_2\gamma$, so

$$\max\left\{|\mathbb{E}(\hat{P}_l)-P_l|,\ |\mathbb{E}(\hat{Q}_l)-Q_l|\right\}\leq \eta_1\Delta_l. \tag{8}$$

We now turn to bounding $|\hat{P}_l-P_l|$ and $|\hat{Q}_l-Q_l|$. Denoting $\tilde{A}_{uv}=\mathbf{1}(A_{ij}=l)$ and using Bernstein's inequality, we have

$$P\left(\left|\sum_{u,v\,:\,\sigma(u)=\sigma(v)}(\tilde{A}_{uv}-\mathbb{E}\tilde{A}_{uv})\right|>t\right)\leq 2\exp\left(-\frac{t^2}{2\sum_{u,v\,\sigma(u)=\sigma(v)}\mathbb{E}\tilde{A}_{uv}+\frac{2}{3}t}\right).$$

By equation (6), we have

$$\sum_{u,v\,\sigma(u)=\sigma(v)}\mathbb{E}\tilde{A}_{uv} = \sum_k \hat{n}_k(\hat{n}_k-1)\mathbb{E}\hat{P}_l \leq (P_l\vee Q_l)\sum_k \hat{n}_k(\hat{n}_k-1),$$

implying that

$$P\left(\left|\sum_{u,v\,:\,\sigma(u)=\sigma(v)}(\tilde{A}_{uv}-\mathbb{E}\tilde{A}_{uv})\right|>t\right) \leq 2\exp\left(-\frac{t^2}{2(P_l\vee Q_l)\sum_k \hat{n}_k(\hat{n}_k-1)+\frac{2}{3}t}\right).$$

Let

$$t^2 = 4\left\{\left(2(P_l\vee Q_l)\sum_k \hat{n}_k(\hat{n}_k-1)\right)\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)\right\} \vee 4\left\{\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)^2\right\},$$

for a constant C_1 to be defined later. Let

$$A = 2(P_l\vee Q_l)\sum_k \hat{n}_k(\hat{n}_k-1), \quad \text{and} \quad B = C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n.$$

We split into two cases:

1. Suppose $A \geq B$. Then $t^2 = 4AB$, and the probability term is at most
$$2\exp\left(-\frac{4AB}{A+\frac{4}{3}\sqrt{AB}}\right) \leq 2\exp\left(-\frac{4AB}{A+\frac{4}{3}A}\right) \leq 2\exp(-B).$$
2. Suppose $A \leq B$. Then $t^2 = 4B^2$, and the probability term is at most
$$2\exp\left(-\frac{4B^2}{A+\frac{4}{3}B}\right) \leq 2\exp\left(-\frac{4B^2}{B+\frac{4}{3}B}\right) \leq 2\exp(-B).$$

Hence, with probability at least $1-2\exp\left(-\left(C_1\gamma n\log\frac{1}{\gamma}+(3+\delta_p)\log n\right)\right)$,

$$|\hat{P}_l-\mathbb{E}(\hat{P}_l)| = \frac{\sum_{u\neq v\,\sigma(u)=\sigma(v)}(\tilde{A}_{uv}-\mathbb{E}\tilde{A}_{uv})}{\sum_{u\neq v}\mathbf{1}(\sigma(u)=\sigma(v))} \leq \frac{t}{\sum_{u\neq v}\mathbf{1}(\sigma(u)=\sigma(v))}.$$

We now derive a more manageable upper bound for t. Using the notation from above, we have $t^2 = \max(4AB, 4B^2) \leq 4(\sqrt{AB}+B)^2$. Since $\gamma \geq \frac{1}{n}$, we

have $C_1\gamma n \log\frac{1}{\gamma} + (3+\delta_p)\log n \le \tilde{C}_1\gamma n\log\frac{1}{\gamma}$, implying that

$$\begin{aligned}\frac{t}{\sum_{u\neq v}\mathbf{1}(\sigma(u)=\sigma(v))} &\le 2\frac{\sqrt{2(P_l\vee Q_l)\tilde{C}_1\gamma n\log\frac{1}{\gamma}}}{\sqrt{\sum_k \hat{n}_k(\hat{n}_k-1)}} + 2\frac{\tilde{C}_1\gamma n\log\frac{1}{\gamma}}{\sum_k \hat{n}_k(\hat{n}_k-1)}\\ &\overset{(a)}{\le} 2\frac{\sqrt{2(P_l\vee Q_l)}\sqrt{\tilde{C}_1\gamma K\log\frac{1}{\gamma}}}{\sqrt{n-K}} + 2\frac{\tilde{C}_1\gamma K\log\frac{1}{\gamma}}{n-K}\\ &\overset{(b)}{\le} 4\sqrt{\frac{P_l\vee Q_l}{n}}\sqrt{\tilde{C}_1 K\gamma\log\frac{1}{\gamma}} + 4\frac{\tilde{C}_1 K\gamma\log\frac{1}{\gamma}}{n},\end{aligned}$$

where (a) uses inequality (7) and (b) uses the assumption $n-K\ge\frac{n}{2}$.

To simplify the expression, note that $P_l\vee Q_l \ge \frac{c}{n}$ implies $\frac{1}{n}\le\frac{1}{\sqrt{c}}\sqrt{\frac{P_l\vee Q_l}{n}}$, so with probability at least $1-\exp(-C_1\gamma n\log\frac{1}{\gamma}-(3+\delta_p)\log n)$, we have

$$|\hat{P}_l - \mathbb{E}(\hat{P}_l)| \le \sqrt{\frac{P_l\vee Q_l}{n}}\left(C_1'\sqrt{\gamma\log\frac{1}{\gamma}} + C_2'\gamma\log\frac{1}{\gamma}\right), \tag{9}$$

for suitable constants C_1' and C_2'. Using a similar calculation, we may show that there exist suitable constants C_3' and C_4' such that

$$|\hat{Q}_l - \mathbb{E}(\hat{Q}_l)| \le \sqrt{\frac{P_l\vee Q_l}{n}}\left(C_3'\sqrt{\gamma\log\frac{1}{\gamma}} + C_4'\gamma\log\frac{1}{\gamma}\right).$$

When γ is sufficiently small, the first term dominates, so we may take choose the right-hand sides to be $\eta_2\sqrt{\frac{P_l\vee Q_l}{n}}$, where $\eta_2 = C_3\sqrt{\gamma\log\frac{1}{\gamma}}$.

Finally, note that there are at most $\binom{n}{\gamma n}K^{\gamma n}$ possible σ's satisfying the error bound. We have

$$\begin{aligned}\log\left(\binom{n}{\gamma n}K^{\gamma n}\right) &\le \log\left(\frac{n^{\gamma n}e^{\gamma n}}{(\gamma n)^{\gamma n}}\frac{1}{\sqrt{2\pi\gamma n}}\right) + \gamma n\log K\\ &\le \log\left(\frac{e^{\gamma n}}{\gamma^{\gamma n}}\right) - \frac{1}{2}\log 2\pi\gamma n + \gamma n\log K\\ &\le \gamma n\log\frac{e}{\gamma} + \gamma n\log K = \gamma n\log\frac{Ke}{\gamma} \le C_1\gamma n\log\frac{1}{\gamma},\end{aligned}$$

for a suitable constant C_1. Taking a union bound across all cluster assignments, we then conclude that the probability of inequality (9) holding simultaneously for all labels l is at least $1-Ln^{-(3+\delta_p)}$.

Combining inequalities (8) and (9), we arrive at the bound

$$\max\left\{|P_l-\hat{P}_l|,\ |Q_l-\hat{Q}_l|\right\} \le \eta_1\Delta_l + \eta_2\sqrt{\frac{P_l\vee Q_l}{n}}.$$

If $\frac{n\Delta_l^2}{P_l \vee Q_l} \geq 1$, we therefore have

$$\max\left\{|P_l - \hat{P}_l|,\ |Q_l - \hat{Q}_l|\right\} \leq \eta_1 \Delta_l + \eta_2 \Delta_l = (\eta_1 + \eta_2)\Delta_l,$$

whereas if $\frac{n\Delta_l^2}{P_l \vee Q_l} < 1$, we have

$$\max\left\{|P_l - \hat{P}_l|,\ Q_l - \hat{Q}_l|\right\} \leq \eta_1 \sqrt{\frac{P_l \vee Q_l}{n}} + \eta_2 \sqrt{\frac{P_l \vee Q_l}{n}} = (\eta_1 + \eta_2)\sqrt{\frac{P_l \vee Q_l}{n}}.$$

Since $\eta_2 = C_3\sqrt{\gamma \log \frac{1}{\gamma}}$ dominates $\eta_1 = C_2\gamma$ for sufficiently small γ, the desired bounds follow. □

B.2. Analysis of spectral clustering.

PROPOSITION B.2. *Suppose an unweighted adjacency matrix A is drawn from a homogeneous stochastic block model with probabilities p and q and cluster imbalance factor β. Suppose $p, q \geq \frac{c}{n}$. Then there exist constants C and C_τ such that if $256\mu\beta C^2 K^3 \frac{(p \vee q)}{n(p-q)^2} \leq 1$, the output σ of Algorithm 4.4 with parameters $\mu \geq 32C^2\beta$ and $\tau = C_\tau \bar{d}$ satisfies*

$$l(\sigma, \sigma_0) \leq 64C^2\beta \frac{K^2(p \vee q)}{n(p-q)^2},$$

with probability at least $1 - n^{-C'}$, where $C' > 4$.

PROOF. Note that τ is a random variable, since the average degree $\bar{d}$ is random. Since the community sizes are bounded by $\frac{n}{\beta K}$, we may find constants $C_{d_1} < C_{d_2} = 1$, depending only on K and β, such that

$$C_{d_1} n(p \vee q) \leq \mathbb{E}[\bar{d}] \leq C_{d_2} n(p \vee q).$$

Using Hoeffding's inequality, we conclude that with probability at least $1 - \exp(-nC_{\bar{d}})$, for some constant $C_{\bar{d}}$, we have

$$\frac{C_{d_1}}{2} n(p \vee q) \leq \bar{d} \leq 2C_{d_2} n(p \vee q). \tag{10}$$

We now apply the following lemma:

LEMMA B.1. *(Lemma 5 of Gao et al. [15]) Let $P \in [0,1]^{n \times n}$ be a symmetric matrix, and let $p_{max} := \max_{u \geq v} P_{uv}$. Let A be an adjacency matrix*

such that $A_{uu} = 0$ *and* $A_{uv} \sim Ber(P_{uv})$ *for* $u < v$. *For any* $C' > 0$ *and for all large enough constants* $C_1 < C_2$, *there exists some* $C > 0$ *such that*

$$\|T_\tau(A) - P\|_2 \leq C\sqrt{np_{max} + 1}, \qquad \forall \tau \in [C_1(np_{max} + 1), C_2(np_{max} + 1)],$$

with probability at least $1 - n^{-C'}$.

REMARK B.1. *Lemma 5 from Gao et al. [15] is stated slightly differently: For any* $C' > 0$, *there exist* c, C_1, *and* C_2 *such that the result holds with probability at least* $1 - n^{-C'}$. *However, our restatement follows immediately from slight modifications of the proof. Furthermore, note that the statement of Lemma B.1 refers to the output of spectral clustering with respect to a fixed trim parameter, but we will apply it in a setting where* τ *is random.*

We choose a large enough constant C_τ so that we may use Lemma B.1 with a fixed $C' > 4$ and constants $C_1 = C_\tau \frac{C_{d_1}}{2}$ and $C_2 = 2C_\tau C_{d_2}$, we conclude that there exists a constant $C > 0$ such that

$$\|T_\tau(A) - P\|_2 \leq C\sqrt{n(p \vee q)}, \qquad \forall \tau \in [C_1, C_2], \tag{11}$$

with probability at least $1 - n^{-C'}$, for the random choice $\tau = C_\tau \bar{d}$, by inequality (10). Furthermore, we may assume $C \geq 1$, since inequality (11) holds with C replaced by $\max(1, C)$. Thus,

$$\begin{aligned}\|\hat{A} - P\|_2 &\leq \|T_\tau(A) - P\|_2 + \|\hat{A} - T_\tau(A)\|_2 \\ &\overset{(a)}{\leq} 2\|T_\tau(A) - P\|_2 \leq 2C\sqrt{n(p \vee q)},\end{aligned}$$

where (a) follows because $\hat{A}$ is the best rank-K approximation of $T_\tau(A)$ and $\text{rank}(P) = K$, so $\|T_\tau(A) - \hat{A}\|_2 \leq \|T_\tau(A) - P\|_2$ by the Eckart-Young-Mirsky Theorem. This implies that

$$\sum_{u=1}^{n} \|\hat{A}_u - P_u\|_2^2 = \|\hat{A} - P\|_F^2 \leq K\|\hat{A} - P\|_2^2 \leq 4KC^2 n(p \vee q).$$

We now denote the K distinct rows of P by $\{\mathcal{Z}_i\}_{1 \leq i \leq K}$, and for a vertex u, denote the row P_u by $\mathcal{Z}(u)$. Note that

$$\|\mathcal{Z}_i - \mathcal{Z}_j\|_2^2 \geq \frac{2n}{\beta K}(p - q)^2, \qquad \forall i \neq j,$$

since each cluster contains at least $\frac{n}{\beta K}$ vertices.

A vertex u is considered *valid* if $\|\hat{A}_u - \mathcal{Z}_i\|_2^2 \le \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$ for some $\mathcal{Z}_i$; otherwise, u is *invalid*. Also define

$$\mathcal{Z}^*(u) := \arg\min_{\mathcal{Z}_i} \|\hat{A}_u - \mathcal{Z}_i\|_2^2,$$

so $\mathcal{Z}^*(u)$ is the row of P closest to $\hat{A}_u$. Note that if u is valid, then $\|\hat{A}_u - \mathcal{Z}^*(u)\|_2^2 \le \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$.

We show that the set S constructed in Algorithm 4.4 satisfies the following properties:

Claim 1: S contains only valid points.
Claim 2: For every pair of distinct nodes $u, v \in S$, we have $\mathcal{Z}^*(u) \neq \mathcal{Z}^*(v)$.

We first prove that the proposition follows from the claims. We denote the rows of $\hat{A}$ corresponding to the members of S by $\mathcal{S}_i$, assigning indices so that $\mathcal{S}_i$ is the surrogate for $\mathcal{Z}_i$ (i.e., both $\mathcal{S}_i$ and $\mathcal{Z}_i$ are associated to a common vertex u). In particular, note that

$$\|\mathcal{S}_i - \mathcal{Z}_i\|_2 \le \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n, \qquad \forall 1 \le i \le K,$$

since S only contains valid points. Let $\mathcal{S}(u)$ be the surrogate of $\mathcal{Z}(u)$, and denote $\mathcal{S}^*(u) = \arg\min_{\mathcal{S}_i} \|\hat{A}_u - \mathcal{S}_i\|^2$; i.e., the member of S that is closest to $\hat{A}_u$. We say that a valid point u is *misclassified* if $\mathcal{S}^*(u) \neq \mathcal{S}(u)$. The number of mistakes we make is thus bounded by the number of invalid points plus the number of misclassified valid points. Note that if u is invalid, we have $\|\hat{A}_u - P_u\|_2^2 \ge \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. We claim that the same inequality holds for any misclassified valid point u.

Consider such a point u. Since u is valid, there exists $\mathcal{Z}_i$ such that

$$\|\hat{A}_u - \mathcal{Z}_i\|^2 \le \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n.$$

We claim that $\mathcal{S}^*(u) = \mathcal{S}_i$. For any $j \neq i$, we have

$$\begin{aligned}
\|\hat{A}_u - \mathcal{S}_j\|_2 &\ge \|\mathcal{Z}_i - \mathcal{Z}_j\|_2 - \|\mathcal{Z}_j - \mathcal{S}_j\|_2 - \|\mathcal{Z}_i - \hat{A}_u\|_2 \\
&\ge \sqrt{\frac{2}{\beta K}(p-q)^2 n} - 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} \\
&> 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n}.
\end{aligned}$$

Furthermore,

$$\|\hat{A}_u - \mathcal{S}_i\|_2 \leq \|\hat{A}_u - \mathcal{Z}_i\|_2 + \|\mathcal{S}_i - \mathcal{Z}_i\|_2 \leq 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n}.$$

Thus, for any $j \neq i$, we have $\|\hat{A}_u - \mathcal{S}_j\|_2 > \|\hat{A}_u - \mathcal{S}_i\|_2$, implying that $\mathcal{S}^*(u) = \mathcal{S}_i$.

Since u is also misclassified, we have $\mathcal{S}(u) \neq \mathcal{S}^*(u) = \mathcal{S}_i$. Let $\mathcal{S}(u) = \mathcal{S}_j$ and $\mathcal{Z}(u) = \mathcal{Z}_j$. We have the following sequence of inequalities:

$$\begin{aligned}\|\hat{A}_u - \mathcal{Z}(u)\|_2 = \|\hat{A}_u - \mathcal{Z}_j\|_2 &\geq \|\hat{A}_u - \mathcal{S}_j\|_2 - \|\mathcal{S}_j - \mathcal{Z}_j\|_2 \\ &\geq 2\sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} - \sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n} \\ &= \sqrt{\frac{1}{16}\frac{1}{\beta K}(p-q)^2 n},\end{aligned}$$

which is the bound we wanted to prove.

Finally, we conclude that the number of mistakes incurred by algorithm is bounded by

$$\frac{\sum_{u=1}^n \|\hat{A}_u - P_u\|_2^2}{\frac{1}{16\beta K}(p-q)^2 n} \leq \frac{4KC^2 n(p \vee q)}{\frac{1}{16\beta K}(p-q)^2 n} \leq \frac{64\beta K^2 C^2 (p \vee q)}{(p-q)^2},$$

as wanted.

Proof of Claim 1:. Recall the notation $N(u) = \{v : \|\hat{A}_u - \hat{A}_v\|_2^2 \leq \mu K^2 \frac{\bar{d}}{n}\}$. Furthermore, by a Chernoff bound, we have $\bar{d} \leq 2(p \vee q)n$ with probability at least $1 - \exp(-C_{\bar{d}} n)$. We condition on this event so that if $v \in N(u)$, then $\|\hat{A}_u - \hat{A}_v\|^2 \leq 2\mu K^2 (p \vee q)$. We prove the claim by showing that an invalid point u cannot have $\frac{1}{\mu}\frac{n}{K}$ neighbors.

By the definition of invalidity, $\|\hat{A}_u - \mathcal{Z}_i\|_2^2 \geq \frac{1}{16\beta K}(p-q)^2 n$, for any $\mathcal{Z}_i$. Let v be a neighbor of u. By the triangle inequality, we then have

$$\begin{aligned}\|\hat{A}_v - \mathcal{Z}(v)\|_2 &\geq \|\hat{A}_u - \mathcal{Z}(v)\|_2 - \|\hat{A}_u - \hat{A}_v\|_2 \\ &\geq \sqrt{\frac{1}{16\beta K}(p-q)^2 n} - \sqrt{2\mu K^2 (p \vee q)} \\ &\overset{(a)}{\geq} \sqrt{\frac{1}{16\beta K}(p-q)^2 n} - \sqrt{\frac{1}{64\beta K}(p-q)^2 n} = \sqrt{\frac{1}{64\beta K}(p-q)^2 n},\end{aligned}$$

where (a) follows from our assumption coupled with the choice of $C \geq 1$, which essentially states that

$$2\mu K^2 (p \vee q) \leq \frac{1}{128\beta K}(p-q)^2 n < \frac{1}{64\beta K}(p-q)^2 n.$$

Thus, for every neighbor v of u, we must have $\|\hat{A}_v - P_v\|_2^2 \geq \frac{1}{64\beta K}(p-q)^2 n$. The number of neighbors of u may be bounded by

$$\frac{\sum_{v=1}^n \|\hat{A}_v - P_v\|_2^2}{\frac{1}{64\beta K}(p-q)^2 n} \leq \frac{4KC^2 n(p \vee q)}{\frac{1}{64\beta K}(p-q)^2 n} \leq \frac{256\beta K^2 C^2 (p \vee q)}{(p-q)^2}.$$

By assumption, this quantity is less than $\frac{1}{\mu}\frac{n}{K}$.

Proof of Claim 2:. We first claim that in every cluster, at least half the points u satisfy $\|\hat{A}_u - P_u\|_2^2 \leq \frac{1}{4}\mu K^2(p \vee q)$. This is because the total error is bounded by $\sum_{u=1}^n \|\hat{A}_u - P_u\|_2^2 \leq 4KC^2 n(p \vee q)$, so the total number of points that violate the condition is at most $\frac{4KC^2 n(p\vee q)}{\frac{1}{4}\mu K^2(p\vee q)} \leq \frac{n}{2\beta K}$, using the assumption that $\mu \geq 32C^2\beta$.

For two points u and v in the same cluster satisfying $\|\hat{A}_w - P_w\|_2^2 \leq \frac{1}{4}\mu K^2(p \vee q)$ for $w \in \{u, v\}$, we also have $\|\hat{A}_u - \hat{A}_v\|_2^2 \leq \mu K^2(p \vee q)$, by the triangle inequality. Thus, every cluster contains a point u such that $N(u) \geq \frac{n}{2\beta K} \geq \frac{1}{\mu}\frac{n}{K}$, since $\mu \geq 32C^2\beta > 2\beta$ by our choice of $C > 1$.

Suppose that at iteration r, the set S consists of points $s_1, \ldots, s_r$, where $1 \leq r < K$, and suppose for a contradiction that s_{r+1} is such that $\mathcal{Z}(s_{r+1}) = \mathcal{Z}(s_i)$ for some $1 \leq i \leq r$. Since s_i and s_{r+1} are both valid points, the triangle inequality implies $\|\hat{A}_{s_{r+1}} - \hat{A}_{s_i}\| \leq \frac{1}{4\beta K}(p-q)^2 n$.

On the other hand, since S does not yet have cardinality K, some $\mathcal{Z}_j$ must exist without a surrogate in S. The cluster that corresponds to $\mathcal{Z}_j$ must, by our neighborhood size analysis, contain a node u such that $N(u) \geq \frac{1}{\mu}\frac{n}{K}$ and

$$\|\hat{A}_u - \mathcal{Z}_j\|_2^2 \leq \frac{1}{4}\mu K^2(p \vee q) \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n,$$

where the second inequality follows by assumption. Since $\mathcal{Z}_j \neq \mathcal{Z}(s_i)$ for any $1 \leq i \leq r$, we have $\|\mathcal{Z}_j - \mathcal{Z}(s_i)\|_2^2 \geq 2\frac{1}{\beta K}(p-q)^2 n$, for all $1 \leq i \leq r$. By Claim 1, all s_i's are valid, so $\|\hat{A}_{s_i} - \mathcal{Z}(s_i)\| \leq \frac{1}{16}\frac{1}{\beta K}(p-q)^2 n$. Hence, by the triangle inequality, we have $\|\hat{A}_u - \hat{A}_{s_i}\|_2^2 \geq \frac{1}{\beta K}(p-q)^2 n$, for all $s_i \in S$. This is a contradiction because u is further from every point in S than s_{r+1}, so our assumption that $\mathcal{Z}(s_{r+1}) = \mathcal{Z}(s_i)$ must be incorrect. □

B.3. Choosing the label l^*. First, we show that for sufficiently well-separated labels, $\hat{I}_l$ is close to $\frac{(P_l - Q_l)^2}{P_l \vee Q_l}$. If the probabilities are not well-separated, we claim that $\hat{I}_l$ is negligibly small.

PROPOSITION B.3. *Suppose* $\frac{1}{\rho_L} \le \frac{P_l}{Q_l} \le \rho_L$ *for all* l. *Let* σ^l *be the output of spectral clustering based on* $\tilde{A}_{ij} = \mathbf{1}(A_{ij} = l)$, *and let* $\hat{P}_l$ *and* $\hat{Q}_l$ *be estimates of* P_l *and* Q_l *constructed from* σ^l. *There exist constants* C_{test}, C_1, C_2, C, *and* δ_p *such that, with probability at least* $1 - Ln^{-3+\delta_p}$:

1. *For all labels* l *satisfying* $P_l \vee Q_l > \frac{c}{n}$ *and* $\Delta_l \ge \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$,

$$C_1 \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}} \le \frac{|\hat{P}_l - \hat{Q}_l|}{\sqrt{\hat{P}_l \vee \hat{Q}_l}} \le C_2 \frac{|P_l - Q_l|}{\sqrt{P_l \vee Q_l}}. \tag{12}$$

2. *For all labels satisfying* $P_l \vee Q_l > \frac{c}{n}$ *and* $\Delta_l < \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$,

$$\frac{|\hat{P}_l - \hat{Q}_l|}{\sqrt{\hat{P}_l \vee \hat{Q}_l}} \le C\sqrt{\frac{1}{n}}. \tag{13}$$

PROOF. Recall from Proposition B.1 that given a clustering with error rate γ, and under the assumptions $P_l \vee Q_l > \frac{c}{n}$ and $\Delta_l^2 \ge \frac{P_l \vee Q_l}{n}$, the estimated probabilities $\hat{P}_l$ and $\hat{Q}_l$ satisfy

$$|\hat{P}_l - \hat{P}| \le \eta\Delta_l, \quad \text{and} \quad |\hat{Q}_l - \hat{Q}| \le \eta\Delta_l,$$

with probability at least $1 - n^{-(3+\delta_p)}$. We first pick a value of γ such that $\eta < \frac{1}{4}$. We now ensure that the error rate obtained from Proposition B.2 matches our choice of γ. Recall that Proposition B.2 states that if $P_l \vee Q_l > \frac{c}{n}$ and $C_1 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2} \le 1$, we have $l(\sigma, \sigma_0) \le C_2 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2}$, for appropriate constants C_1 and C_2. In particular, for $C_{test} \ge 1$ sufficiently large,

$$C_1 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2} \le \frac{C_1}{C_{test}} < 1, \quad \text{and}$$
$$l(\sigma, \sigma_0) \le C_2 \frac{P_l \vee Q_l}{n(P_l - Q_l)^2} \le \frac{C_2}{C_{test}} < \gamma,$$

for all labels l such that $\Delta_l \ge \sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}$. Next, note that

$$|\hat{P}_l - \hat{Q}_l| \le |\hat{P}_l - P_l| + |P_l - Q_l| + |\hat{Q}_l - Q_l| \le 2\eta\Delta_l + \Delta_l \le \frac{3}{2}\Delta,$$
$$|\hat{P}_l - \hat{Q}_l| \ge |P_l - Q_l| - |\hat{Q}_l - Q_l| - |\hat{P}_l - P_l| \ge \Delta_l - 2\eta\Delta_l \ge \frac{1}{2}\Delta_l.$$

Furthermore,

$$\hat{P}_l \vee \hat{Q}_l \leq (P_l \vee Q_l) + \eta\Delta_l \leq (P_l \vee Q_l) + \eta(P_l \vee Q_l) \leq \frac{5}{4}(P_l \vee Q_l),$$

and

$$\hat{P}_l \vee \hat{Q}_l \geq (P_l \vee Q_l) - \eta\Delta_l \geq \frac{3}{4}(P_l \vee Q_l).$$

We conclude that

$$\frac{1}{\sqrt{5}}\frac{\Delta_l}{P_l \vee Q_l} \leq \frac{|\hat{P}_l - \hat{Q}_l|}{\sqrt{\hat{P}_l \vee \hat{Q}_l}} \leq \frac{3}{\sqrt{3}}\frac{\Delta_l}{P_l \vee Q_l}.$$

Now suppose $\Delta_l^2 < C_{test}\frac{P_l \vee Q_l}{n}$. Note that this does not necessarily imply $\Delta_l^2 \leq \frac{P_l \vee Q_l}{n}$, since $C_{test} \geq 1$. However, we may take the maximum of the bounds in Proposition B.1, so with probability at least $1 - Ln^{-(3+\delta_p)}$,

$$|\hat{P}_l - P_l| \leq \eta\left(\Delta_l \vee \sqrt{\frac{P_l \vee Q_l}{n}}\right) \leq \frac{\sqrt{C_{test}}}{4}\sqrt{\frac{P_l \vee Q_l}{n}},$$

using the choice $\eta \leq \frac{1}{4}$. An analogous bound holds for $|\hat{Q}_l - Q_l|$. Hence,

$$\begin{aligned}|\hat{P}_l - \hat{Q}_l| &\leq \Delta_l + |\hat{P}_l - P_l| + |\hat{Q}_l - Q_l| \\ &\leq \Delta_l + \frac{\sqrt{C_{test}}}{2}\sqrt{\frac{P_l \vee Q_l}{n}} \leq \frac{3}{2}\sqrt{C_{test}}\sqrt{\frac{P_l \vee Q_l}{n}}.\end{aligned}$$

Now note that

$$\hat{P}_l \vee \hat{Q}_l \geq (P_l \vee Q_l) - \frac{\sqrt{C_{test}}}{4}\sqrt{\frac{P_l \vee Q_l}{n}} \geq C'(P_l \vee Q_l).$$

for some constant C'. It follows that $\frac{|\hat{P}_l - \hat{Q}_l|}{\sqrt{\hat{P}_l \vee \hat{Q}_l}} \leq C\sqrt{\frac{1}{n}}$. □

We apply Proposition B.3 to conclude that Algorithm 4.3 succeeds in choosing a color l^* for which $\frac{n(P_{l^*} - Q_{l^*})^2}{P_{l^*} \vee Q_{l^*}}$ is arbitrarily large:

PROPOSITION B.4. *Suppose* $a_n := \frac{nI_L}{L\rho_L^4} \to \infty$. *For sufficiently large* n, *with probability at least* $1 - 2Ln^{-(3+\delta_p)}$, *we have* $\frac{n(P_{l^*} - Q_{l^*})^2}{(P_{l^*} \vee Q_{l^*})\rho_L^4} \geq Ca_n$, *for some constant* C.

PROOF. Let C_{test} be the constant in Proposition B.3. By Lemma B.6, we know that I_L is of the same order as $\sum_{l\in L_1}\frac{\Delta_l^2}{P_l\vee Q_l}$, implying the existence of a label l such that $\Delta_l \geq C_{test}\sqrt{\frac{P_l\vee Q_l}{n}}$ and $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq C\frac{nI_L}{L} = Ca_n\rho_L^4$, for a constant C. Suppose the event of Proposition B.3 holds, which happens with probability at least $1 - Ln^{-(3+\delta_p)}$.

Step 1. We claim that l^* satisfies $\Delta_{l^*} \geq C_{test}\sqrt{\frac{P_l\vee Q_l}{n}}$. Let l be a label such that $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq Ca_n\rho_L^4$, and suppose the claim is false. By Proposition B.3 and the maximality of l^*, we have

$$\frac{|\hat{P}_l - \hat{Q}_l|}{\sqrt{\hat{P}_l\vee\hat{Q}_l}} \overset{(a)}{\leq} \frac{|\hat{P}_{l^*} - \hat{Q}_{l^*}|}{\sqrt{\hat{P}_{l^*}\vee\hat{Q}_{l^*}}} \leq C\sqrt{\frac{1}{n}},$$

Proposition B.3 also implies that

$$\frac{|\hat{P}_l - \hat{Q}_l|}{\sqrt{\hat{P}_l\vee\hat{Q}_l}} \geq C'\frac{|P_l - Q_l|}{\sqrt{P_l\vee Q_l}} \geq C''\sqrt{\frac{a_n\rho_L^4}{n}}.$$

However, this is a contradiction, since $a_n \to \infty$ and $\rho_L \geq 1$.

Step 2: Again, let l be a label such that $\frac{n\Delta_l^2}{P_l\vee Q_l} \geq Ca_n\rho_L^4$. By Proposition B.3, we then have

$$\frac{|P_{l^*} - Q_{l^*}|}{\sqrt{P_{l^*}\vee Q_{l^*}}} \geq C\frac{|\hat{P}_{l^*} - \hat{Q}_{l^*}|}{\sqrt{\hat{P}_{l^*}\vee\hat{Q}_{l^*}}} \geq C\frac{|\hat{P}_l - \hat{Q}_l|}{\sqrt{\hat{P}_l\vee\hat{Q}_l}} \geq C'\frac{|P_l - Q_l|}{\sqrt{P_l\vee Q_l}} \geq C''\sqrt{\frac{a_n\rho_L^4}{n}},$$

implying the desired result. □

B.4. Analysis of error probability for a single node.

PROPOSITION B.5. *Let u be an arbitrary fixed node, and let $\tilde{\sigma}_u$ be the output of Algorithm 4.3. Suppose $\pi_u \in S_K$ satisfies*

$$l(\sigma_0, \tilde{\sigma}_u) = d(\sigma_0, \pi_u(\tilde{\sigma}_u)),$$

where both l and d are taken with respect to the set $\{1, 2, \ldots, n\}\setminus\{u\}$. Conditioned on the events that the error rate γ of $\tilde{\sigma}_u$ satisfies $\gamma\rho_L^4 \to 0$, and also the event that the result of Proposition B.1 holds for a sequence η satisfying $\eta\rho_L^2 \to 0$, we have

$$\pi_u^{-1}(\sigma_0(u)) = \arg\max_k \sum_{v:\tilde{\sigma}_u(v)=k}\sum_l \log\frac{\hat{P}_l}{\hat{Q}_l}\mathbf{1}(A_{uv} = l),$$

with probability at least $1 - (K-1)\exp\left(-(1-o(1))\frac{n}{\beta K}I_L\right)$.

PROOF. Throughout the proof, we assume that n is large enough so $\frac{1}{2}\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2 \le \frac{1}{2}$. Suppose without loss of generality that $\sigma_0(u)=1$ and π_u is the identity. We misclassify u into community k if

$$\sum_{v\,:\,\tilde{\sigma}_u(v)=k}\sum_l \log\frac{\hat{P}_l}{\hat{Q}_l}\mathbf{1}(A_{uv}=l) \ge \sum_{v\,:\,\tilde{\sigma}_u(v)=1}\sum_l \log\frac{\hat{P}_l}{\hat{Q}_l}\mathbf{1}(A_{uv}=l),$$

or equivalently,

$$\sum_{v\,:\,\tilde{\sigma}_u(v)=k}\bar{A}_{uv} - \sum_{v\,:\,\tilde{\sigma}_u(v)=1}\bar{A}_{uv} \ge 0, \tag{14}$$

where $\bar{A}_{uv} \equiv \sum_l \log\frac{\hat{P}_l}{\hat{Q}_l}\mathbf{1}(A_{uv}=l)$. Note that the edges from u are independent of the clustering $\tilde{\sigma}_u$, since this clustering was obtained by running the algorithm with vertex u excluded.

Define $m_1 = |\{v : \tilde{\sigma}_u(v)=1\}|$ and $m_k = |\{v : \tilde{\sigma}_u(v)=k\}|$, and let $m_1' = \{v : \tilde{\sigma}_u(v)=1, \sigma_0(v)=1\}$ be the points correctly clustered by σ_u. Let $m_k' = \{v : \tilde{\sigma}_u(v)=k, \sigma_0(v)=k\}$ denote the points correctly classified by $\tilde{\sigma}_u$ in community k. With these definitions, the probability of the bad event in equation (14) is the probability of the event

$$\left(\sum_{i=1}^{m_k'}\tilde{Y}_i + \sum_{i=1}^{m_k-m_k'}\tilde{X}_i+\right) - \left(\sum_{i=1}^{m_1'}\tilde{X}_i + \sum_{i=1}^{m_1-m_1'}\tilde{Y}_i\right) \ge 0,$$

where $\tilde{X}_i = \log\frac{\hat{P}_l}{\hat{Q}_l}$ with probability P_l and $\tilde{Y}_i = \log\frac{\hat{P}_l}{\hat{Q}_l}$ with probability Q_l. (For simplicity, we abuse notation and write $\tilde{Y}_i$ and $\tilde{X}_i$ in both bracketed terms. These random variables are not the same, but are independent and identical copies.) This is equal to the probability of the event

$$\exp\left(t\left(\sum_{i=1}^{m_k'}\tilde{Y}_i + \sum_{i=1}^{m_k-m_k'}\tilde{X}_i - \sum_{i=1}^{m_1'}\tilde{X}_i - \sum_{i=1}^{m_1-m_1'}\tilde{Y}_i\right)\right) \ge 1.$$

We further bound this probability as follows:

$$
\begin{aligned}
&P\left(\exp\left(t\left(\sum_{i=1}^{m_k'}\tilde{Y}_i+\sum_{i=1}^{m_k-m_k'}\tilde{X}_i-\sum_{i=1}^{m_1'}\tilde{X}_i-\sum_{i=1}^{m_1-m_1'}\tilde{Y}_i\right)\right)\geq 1\right)\\
&\leq \mathbb{E}\left[\exp\left(t\left(\sum_{i=1}^{m_k'}\tilde{Y}_i+\sum_{i=1}^{m_k-m_k'}\tilde{X}_i-\sum_{i=1}^{m_1'}\tilde{X}_i-\sum_{i=1}^{m_1-m_1'}\tilde{Y}_i\right)\right)\right]\\
&=\mathbb{E}[\exp(t\tilde{Y}_i)]^{m_k'}\mathbb{E}[\exp(t\tilde{X}_i])^{m_k-m_k'}\mathbb{E}[\exp(-t\tilde{X}_i)]^{m_1'}\mathbb{E}[\exp(-t\tilde{Y}_i)]^{m_1-m_1'}\\
&=\left(\sum_l e^{t\log\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{m_k'}\left(\sum_l e^{t\log\frac{\hat{P}_l}{\hat{Q}_l}}P_l\right)^{m_k-m_k'}\\
&\quad\cdot\left(\sum_l e^{-t\log\frac{\hat{P}_l}{\hat{Q}_l}}P_l\right)^{m_1'}\left(\sum_l e^{-t\log\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{m_1-m_1'}.
\end{aligned}
$$

We will set $t=\frac{1}{2}$, in which case

$$
\left(\sum_l\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{m_k'}\left(\sum_l\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}P_l\right)^{m_k-m_k'}\left(\sum_l\sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}Q_l\right)^{m_1-m_1'}\left(\sum_l\sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)^{m_1'} \tag{15}
$$

$$
=\left(\frac{\sum_l\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}P_l}{\sum_l\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l}\right)^{m_k-m_k'}\left(\frac{\sum_l\sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}Q_l}{\sum_l\sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l}\right)^{m_1-m_1'}\left(\sum_l\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{m_k}\left(\sum_l\sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)^{m_1}.
$$

Loosely speaking, we will show that the first pair of terms is bounded in magnitude by $\exp(o(I_L)\frac{n}{K})$, and the second pair of terms is bounded by $\exp(-\frac{n}{\beta K}(1+o(1))I_L)$.

Bound for first pair. We derive a number of separate lemmas bounding various intermediate terms in the computation. In particular, we use the bounds from Lemmas B.5, B.4, and B.6 in the following sequence of

inequalities:

$$\begin{aligned}
\left|1-\frac{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}P_l}{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l}\right| &= \left|\frac{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}(P_l-Q_l)}{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l}\right| \overset{(a)}{\leq} \frac{8}{\sum_l \sqrt{P_l Q_l}}\left|\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}(P_l-Q_l)\right| \\
&\overset{(b)}{\leq} 16\left|\sum_l \left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}-1\right)(P_l-Q_l)\right| \\
&\leq 16\left|\sum_{l\in L_1} \left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}-1\right)(P_l-Q_l)\right| + 16\sum_{l\notin L_1}\left|\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}-1\right||P_l-Q_l| \\
&\overset{(c)}{\leq} 16\sum_{l\in L_1}\frac{\Delta_l^2}{Q_l}(1+\eta') + \sum_{i\notin L_1} 32\rho_L\frac{\Delta_l}{\sqrt{n(P_l\vee Q_l)}} \\
&\leq 16\rho_L\sum_{l\in L_1}\frac{\Delta_l^2}{P_l\vee Q_l}(1+\eta') + \sum_{l\notin L_1} 32\rho_L\frac{\Delta_l}{\sqrt{n(P_l\vee Q_l)}} \\
&\overset{(d)}{\leq} CI_L\rho_L(1+\eta') + C'\rho_L\frac{L}{n} \overset{(e)}{\leq} C\rho_L I_L.
\end{aligned}$$

In (a), we have used Lemma B.5. In (b), we have used the fact $\sum_l \sqrt{P_l Q_l} \to 1$, so this sum exceeds $\frac{1}{2}$ when n is sufficiently large. In (c), we have employed Lemma B.4, which appropriately bounds the term $\left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}-1\right)$. Here, $\eta' = o(1)$. Inequality (d) follows from Lemma B.6. Finally, inequality (e) follows from the assumption that $\frac{I_L n}{L\rho_L^4} \to \infty$ (note that $\rho_L \geq 1$) and by appropriately redefining η'). Identical analysis shows that $\left|1-\frac{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}Q_l}{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l}\right| \leq C\rho_L I_L$. Finally, note that $|x| \leq \exp(|1-x|)$, so we have the bound

$$\begin{aligned}
&\left(\frac{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}P_l}{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l}\right)^{m_k-m_k'}\left(\frac{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}Q_l}{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l}\right)^{m_1-m_1'} \\
&\qquad \leq \exp\left(C\rho_L I_L(m_k-m_k'+m_1-m_1')\right) \leq \exp(CI_L\rho_L\gamma n).
\end{aligned}$$

Since $\gamma\rho_L = o(1)$, we obtain the desired bound $\exp\left(\frac{n}{K}o(I_L)\right)$.

Bound for second pair. Let $\hat{I} = -\log\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)\left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)$. With this definition, we have

$$\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{m_k}\left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)^{m_1}$$
$$= \exp(-\hat{I})^{\frac{m_k+m_1}{2}}\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{\frac{m_k-m_1}{2}}\left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)^{\frac{m_1-m_k}{2}}.$$

We claim that the following statements are true:

1. $m_1, m_k \geq \frac{n}{\beta K}(1-\beta K\gamma)$.
2. $\hat{I} \geq I_L(1+o(1))$.
3. $\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{\frac{m_k-m_1}{2}}\left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)^{\frac{m_1-m_k}{2}} = \exp\left(\frac{n}{K}o(I_L)\right)$.

Let us first assume these statements are true and derive the bound. We have

$$\begin{aligned}
&\exp(-\hat{I})^{\frac{m_1+m_k}{2}}\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)^{\frac{m_k-m_1}{2}}\left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)^{\frac{m_1-m_k}{2}}\\
&\quad\leq \exp\left(-I_L(1+o(1))\frac{n}{\beta K}\cdot(1-\beta K\gamma)+\frac{n}{K}o(I_L)\right)\\
&\quad\leq \exp\left(-(1+o(1))\frac{n}{\beta K}I_L\right),
\end{aligned}$$

where the last inequality holds since $\gamma = o(1)$. It remains to prove the claims.
Claim 1: This is straightforward. The labeling $\tilde{\sigma}_u$ has at most γn errors, so $m_1 \geq m_1' \geq \frac{n}{\beta K} - \gamma n$. A similar argument works for m_k.
Claim 2: We begin by writing

$$\hat{I} - I_L = -\log\frac{\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l\right)\left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l\right)}{\left(\sum_l \sqrt{P_lQ_l}\right)^2}.$$

Let us first consider the numerator:

$$\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} Q_l\right)\left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} P_l\right) = \left(\sum_l \sqrt{P_l Q_l}\sqrt{\frac{\hat{P}_l}{P_l}\frac{Q_l}{\hat{Q}_l}}\right)\left(\sum_l \sqrt{P_l Q_l}\sqrt{\frac{P_l}{\hat{P}_l}\frac{\hat{Q}_l}{Q_l}}\right)$$
$$= \sum_l P_l Q_l + 2\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}} + \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)$$
$$= \left(\sum_l \sqrt{P_l Q_l}\right)^2 + \sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right),$$

where $T_{l,l'} := \frac{\hat{P}_l}{P_l}\frac{Q_l}{\hat{Q}_l}\frac{P_{l'}}{\hat{P}_{l'}}\frac{\hat{Q}_{l'}}{Q_{l'}}$. Furthermore, assuming $\sum_l \sqrt{P_l Q_l} \geq 1/2$, we have

$$\hat{I} - I_L = -\log\left(1 + \frac{\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)}{\left(\sum_l \sqrt{P_l Q_l}\right)^2}\right)$$
$$\geq -\log\left(1 + 4\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right)\right)$$
$$\geq -4\sum_{l<l'} \sqrt{P_l Q_l P_{l'} Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right). \tag{16}$$

We now bound $|T_{l,l'} - 1|$:

$$|T_{l,l'} - 1| = \left|\frac{\hat{P}_l}{P_l}\frac{Q_l}{\hat{Q}_l}\frac{P_{l'}}{\hat{P}_{l'}}\frac{\hat{Q}_{l'}}{Q_{l'}} - 1\right|$$
$$= \left|\left(1 - \frac{P_l - \hat{P}_l}{P_l}\right)\left(1 - \frac{\hat{Q}_l - Q_l}{\hat{Q}_l}\right)\left(1 - \frac{\hat{P}_{l'} - P_{l'}}{\hat{P}_{l'}}\right)\left(1 - \frac{Q_{l'} - \hat{Q}_{l'}}{Q_{l'}}\right) - 1\right|$$
$$\overset{(a)}{\leq} 2\left(\frac{|P_l - \hat{P}_l|}{P_l} + \frac{|\hat{Q}_l - Q_l|}{\hat{Q}_l} + \frac{|\hat{P}_{l'} - P_{l'}|}{\hat{P}_{l'}} + \frac{|Q_{l'} - \hat{Q}_{l'}|}{Q_{l'}}\right)$$
$$\overset{(b)}{\leq} 4\left(\frac{|P_l - \hat{P}_l|}{P_l} + \frac{|\hat{Q}_l - Q_l|}{Q_l} + \frac{|\hat{P}_{l'} - P_{l'}|}{P_{l'}} + \frac{|Q_{l'} - \hat{Q}_{l'}|}{Q_{l'}}\right),$$

where (a) and (b) follow from Lemma B.3. Since we only work with pairs (l, l') such that $l' > l$, we may choose any ordering we like. Thus, suppose the l's are ordered in decreasing order of $\frac{|\hat{P}_l - P_l|}{P_l} + \frac{|\hat{Q}_l - Q_l|}{Q_l}$. For all pairs $l < l'$, we then have $|T_{l,l'} - 1| \leq 8\left(\frac{|\hat{P}_l - P_l|}{P_l} + \frac{|\hat{Q}_l - Q_l|}{Q_l}\right)$. By Proposition B.1, we have

$$\frac{|P_l - \hat{P}_l|}{P_l} + \frac{|\hat{Q}_l - Q_l|}{Q_l} \leq \eta\Delta_l\left(\frac{1}{P_l} + \frac{1}{Q_l}\right) \leq \frac{\eta\Delta_l}{P_l \vee Q_l} \cdot 2\rho_L \leq \eta' \frac{\Delta_l}{P_l \vee Q_l},$$

for any $l \in L_1$. For $l \notin L_1$, we have

$$\frac{|P_l - \hat{P}_l|}{P_l} \;\le\; \eta\sqrt{\frac{P_l \vee Q_l}{nP_l^2}} \;=\; \eta\frac{P_l \vee Q_l}{P_l}\sqrt{\frac{1}{n(P_l \vee Q_l)}} \;\le\; \eta'\sqrt{\frac{1}{n(P_l \vee Q_l)}},$$

and similarly for the $\frac{|\hat{Q}_l - Q_l|}{Q_l}$ term. Plugging these bounds into the previous derivation, we obtain

$$|T_{l,l'} - 1| \le \begin{cases} \eta'\frac{\Delta_l}{P_l \vee Q_l}, & \text{for } l \in L_1, \\ \eta'\frac{1}{\sqrt{n(P_l \vee Q_l)}}, & \text{for } l \notin L_1. \end{cases}$$

Using the Taylor approximation $\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2 = \frac{1}{4}(T_{l,l'}-1)^2 + O(T_{l,l'}-1)^3$ and continuing the bound (16), we obtain

$$\begin{aligned}
\hat{I} - I_L &\ge -4\sum_{l<l'}\sqrt{P_lQ_lP_{l'}Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right) \\
&\ge -4\sum_{l\in L_1}\sum_{l'>l}\sqrt{P_lQ_lP_{l'}Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right) \\
&\qquad -4\sum_{l\notin L_1}\sum_{l'>l}\sqrt{P_lQ_lP_{l'}Q_{l'}}\left(\sqrt{T_{l,l'}} + \frac{1}{\sqrt{T_{l,l'}}} - 2\right) \\
&\ge -\sum_{l\in L_1}\sum_{l'>l}\sqrt{P_lQ_lP_{l'}Q_{l'}}\eta'\left(\frac{\Delta_l}{P_l \vee Q_l}\right)^2 - \sum_{l\notin L_1}\sum_{l'>l}\sqrt{P_lQ_lP_{l'}Q_{l'}}\eta'\frac{1}{n(P_l \vee Q_l)} \\
&\ge -\eta'\left(\sum_{l\in L_1}\frac{\Delta_l^2\sqrt{P_lQ_l}}{(P_l \vee Q_l)^2}\right)\left(\sum_{l'}\sqrt{P_{l'}Q_{l'}}\right) - \eta'\left(\sum_{l\notin L_1}\frac{\sqrt{P_lQ_l}}{n(P_l \vee Q_l)}\right)\left(\sum_{l'}\sqrt{P_{l'}Q_{l'}}\right) \\
&\ge -\eta'\left(\sum_{l\in L_1}\frac{\Delta_l^2}{P_l \vee Q_l}\right)\left(\sum_{l'}\sqrt{P_{l'}Q_{l'}}\right) - \eta'\left(\sum_{l\notin L_1}\frac{1}{n}\right)\left(\sum_{l'}\sqrt{P_{l'}Q_{l'}}\right) \\
&\overset{(a)}{=} -o(I_L),
\end{aligned}$$

where (a) follows from $\sum_{l'}\sqrt{P_{l'}Q_{l'}} \le 1$, the statement $\sum_{l\in L_1}\frac{\Delta_l^2}{P_l \vee Q_l} = \Theta(I_L)$ from Lemma B.6, and our assumption that $\sum_{l\notin L_1}\frac{1}{n} \le \frac{L}{n} = o(I_L)$. This proves the claim.

Claim 3. We rewrite the term in claim 3 as follows:

$$\left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} Q_l\right)^{\frac{m_k-m_1}{2}} \left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} P_l\right)^{\frac{m_1-m_k}{2}}$$
$$= \left(\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} Q_l\right)^{\frac{m_k-m_1}{2}} \left(\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} P_l\right)^{\frac{m_1-m_k}{2}} \left(\frac{\sum_l \sqrt{\hat{P}_l \hat{Q}_l}}{\sum_l \sqrt{\hat{P}_l \hat{Q}_l}}\right)^{\frac{m_1-m_k}{2}}$$
$$= \left(\frac{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} Q_l}{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} \hat{Q}_l}\right)^{\frac{m_k-m_1}{2}} \left(\frac{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} P_l}{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} \hat{P}_l}\right)^{\frac{m_1-m_k}{2}}.$$

Assume $m_k \geq m_1$. The reverse case may be analyzed in an identical manner. We may rewrite the term as

$$\left(1 + \frac{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} (Q_l - \hat{Q}_l)}{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} \hat{Q}_l}\right)^{\frac{m_k-m_1}{2}} \left(1 + \frac{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} (\hat{P}_l - P_l)}{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} P_l}\right)^{\frac{m_k-m_1}{2}}.$$

Note that $\sum_l \sqrt{P_l Q_l} \to 1$, so Lemma B.5 implies that the denominators are $\Theta(1)$. We bound the numerator, as follows:

$$\begin{aligned}
\left|\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} (Q_l - \hat{Q}_l)\right| &= \left|\sum_l \left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right)(Q_l - \hat{Q}_l)\right| \\
&\leq \left|\sum_{l \in L_1} \left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right)(Q_l - \hat{Q}_l)\right| + \left|\sum_{l \notin L_1} \left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right)(Q_l - \hat{Q}_l)\right| \\
&\overset{(a)}{\leq} \left|\sum_{l \in L_1} \left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right)\eta \Delta_l\right| + \left|\sum_{l \notin L_1} \left(\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right)\eta \sqrt{\frac{P_l \vee Q_l}{n}}\right| \\
&\overset{(b)}{\leq} \sum_{l \in L_1} \eta \frac{\Delta_l^2}{Q_l} + \sum_{l \notin L_1} \eta \rho_L \frac{1}{n} \leq \eta \rho_L \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} + \sum_{l \notin L_1} \eta \rho_L \frac{1}{n} \\
&\overset{(c)}{\leq} \eta' I_L + \eta' \frac{L}{n} \overset{(d)}{\leq} \eta' I_L.
\end{aligned}$$

In the above sequence of inequalities, step (a) follows from Proposition B.1, step (b) follows from Lemma B.4, step (c) follows from Lemma B.6 and the

assumption $\eta\rho_L \to 0$, and step (d) follows from our assumption $\frac{L}{n} = o(I_L)$. Thus, we obtain

$$\left(1 + \frac{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}(Q_l - \hat{Q}_l)}{\sum_l \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}\hat{Q}_l}\right)^{\frac{m_k - m_1}{2}} \left(1 + \frac{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}(\hat{P}_l - P_l)}{\sum_l \sqrt{\frac{\hat{Q}_l}{\hat{P}_l}}P_l}\right)^{\frac{m_k - m_1}{2}}$$
$$\leq \exp\left((m_k - m_1)\log(1 + o(I_L))\right) \leq \exp\left(\frac{n}{K}o(I_L)\right),$$

proving the claim.

Combining bounds in inequality (15): Altogether, we conclude that the probability of misclassifying u into some cluster $k \neq 1$ is at most $\exp\left((1 + o(1))\frac{nI_L}{\beta K}\right)$. Taking a union bound over all clusters $k \neq 1$ completes the proof. □

B.5. Additional lemmas for Proposition 6.1.

LEMMA B.2. *Let L, P_l, Q_l, ρ_L, and I_L satisfy the assumptions in Proposition 6.1. Define the new probabilities of edge labels as follows:*

$$P'_l := P_l(1 - \delta) + \frac{\delta}{L+1}, \quad \text{and} \quad Q'_l := Q_l(1 - \delta) + \frac{\delta}{L+1},$$

for all $0 \leq l \leq L$, where $\delta = \frac{c(L+1)}{n}$. Let I'_L denote the Renyi divergence between P'_l and Q'_l. Then for all sufficiently large n, we have $P'_l, Q'_l > \frac{c}{n}$ for all $0 \leq l \leq L$, and

$$I'_L = I_L(1 + o(1)).$$

PROOF. Clearly, $P'_l, Q'_l > \frac{c}{n}$. For the second part of the lemma, we begin by writing

$$I_L = -2\log\sum_{l=0}^{L}\sqrt{P_l Q_l} = -2\log\left(1 - \frac{1}{2}\sum_{l=0}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2\right)$$
$$= \left(\sum_{l=0}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2\right)(1 + o(1)).$$

Similarly, we have $I'_L = \left(\sum_{l=0}^{L}(\sqrt{P'_l} - \sqrt{Q'_l})^2\right)(1 + o(1))$, so it is enough to show that

$$\left(\sum_{l=0}^{L}(\sqrt{P_l} - \sqrt{Q_l})^2\right) = \left(\sum_{l=0}^{L}(\sqrt{P'_l} - \sqrt{Q'_l})^2\right)(1 + o(1)).$$

We consider two cases: $\rho_L = \omega(1)$ and $\rho_L = \Theta(1)$. If $\rho_L = \omega(1)$, we choose $a = \frac{nI_L}{\rho_L(L+1)}$. If $\rho_L = \Theta(1)$, we choose $a = o\left(\frac{nI_L}{L+1}\right)$ such that $a \to \infty$. Note that in both cases, we have $\frac{a}{\rho_L} \to \infty$ and $\frac{a(L+1)}{n} = o(I_L)$. We now break the set of labels into two groups, where G_1 contains all labels satisfying $P_l \vee Q_l \leq \frac{a}{n}$, and $G_2 = G_1^c$.

Let $\Delta_l := |P_l - Q_l|$ and $\Delta_l' := |P_l' - Q_l'|$. In G_1, we have $\Delta_l \leq \frac{a}{n}$, so

$$(\sqrt{P_l} - \sqrt{Q_l})^2 = \frac{\Delta_l^2}{(\sqrt{P}_l + \sqrt{Q}_l)^2} \leq \Delta_l \leq \frac{a}{n},$$

and

$$\left(\sqrt{P_l'} - \sqrt{Q_l'}\right)^2 = \frac{(\Delta_l')^2}{(\sqrt{P_l'} + \sqrt{Q_l'})^2} \leq \Delta_l' = (1-\delta)\Delta_l \leq (1-\delta)\frac{a}{n}.$$

Therefore,

$$\left|\sum_{l\in G_1}(\sqrt{P_l} - \sqrt{Q_l})^2 - \sum_{l\in G_1}(\sqrt{P_l'} - \sqrt{Q_l'})^2\right| \leq \frac{a(L+1)}{n} = o(I_L).$$

For labels in G_2, we may write

$$\begin{aligned}&\left|\sum_{l\in G_2}(\sqrt{P_l} - \sqrt{Q_l})^2 - \sum_{l\in G_2}(\sqrt{P_l'} - \sqrt{Q_l'})^2\right| \\ &\qquad = \sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P}_l + \sqrt{Q}_l)^2}\left|1 - (1-\delta)^2\frac{(\sqrt{P}_l + \sqrt{Q}_l)^2}{(\sqrt{P_l'} + \sqrt{Q_l'})^2}\right|.\end{aligned}$$

We analyze the term inside the absolute value as follows:

$$\begin{aligned}\frac{(\sqrt{P}_l + \sqrt{Q}_l)^2}{(\sqrt{P_l'} + \sqrt{Q_l'})^2} &= \left(\frac{\sqrt{P}_l + \sqrt{Q}_l}{\sqrt{P}_l\sqrt{\frac{P_l'}{P_l}} + \sqrt{Q}_l\sqrt{\frac{Q_l'}{Q_l}}}\right)^2 \\ &= \left(\frac{\sqrt{P}_l + \sqrt{Q}_l}{\sqrt{P}_l\sqrt{1-\delta+\frac{c}{nP_l}} + \sqrt{Q}_l\sqrt{1-\delta+\frac{c}{nQ_l}}}\right)^2.\end{aligned}$$

Since $P_l \vee Q_l > \frac{a}{n}$, we have

$$\frac{1}{n(P_l \vee Q_l)} < \frac{1}{a} = o(1), \quad \text{so} \quad \frac{1}{nP_l} \leq \frac{\rho_L}{n(P_l \vee Q_l)} < \frac{\rho_L}{a} = o(1).$$

Furthermore, since $\delta = o(1)$, we have

$$\left(\frac{\sqrt{P_l}+\sqrt{Q_l}}{\sqrt{P_l}\sqrt{1-\delta+\frac{c}{nP_l}}+\sqrt{Q_l}\sqrt{1-\delta+\frac{c}{nQ_l}}}\right)^2 = \left(\frac{\sqrt{P_l}+\sqrt{Q_l}}{\sqrt{P_l}(1+o(1))+\sqrt{Q_l}(1+o(1))}\right)^2 = 1+o(1).$$

Hence, we may conclude that

$$\sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P_l}+\sqrt{Q_l})^2}\left|1-(1-\delta)^2\frac{(\sqrt{P_l}+\sqrt{Q_l})^2}{(\sqrt{P_l'}+\sqrt{Q_l'})^2}\right| = \sum_{l\in G_2}\frac{\Delta_l^2}{(\sqrt{P_l}+\sqrt{Q_l})^2}\cdot o(1) = o(I_L).$$

Combining the results for G_1 and G_2, implies that $I_L' = (1+o(1))I_L$. □

We often use the bound $\frac{1}{2}P \le \hat{P}_l \le 2P_l$, justified in the following lemma:

LEMMA B.3. *Let l be any label and suppose $\frac{1}{\rho_L} \le \frac{P_l}{Q_l} \le \rho_L$ where $\rho_L > 1$; also suppose $P_l, Q_l \ge \frac{c}{n}$. Conditioned on the event that the conclusion of Proposition B.1 holds with a sequence η such that $\eta\rho_L^2 \to 0$, we have*

$$\max_l \frac{|\hat{P}_l - P_l|}{P_l} \to 0, \quad \textit{and} \quad \max_l \frac{|\hat{Q}_l - Q_l|}{Q_l} \to 0.$$

For sufficiently small η, we thus have $\frac{1}{2}P_l \le \hat{P}_l \le 2P_l$, and likewise for Q_l.

PROOF. We prove the statement first for P_l; the same argument applies to Q_l. By Proposition B.1, we have that either $|\hat{P}_l - P_l| \le \eta\Delta_l$ or $|\hat{P}_l - P_l| \le \eta\sqrt{\frac{P_l \vee Q_l}{n}}$. First suppose $|\hat{P}_l - P_l| \le \eta\Delta_l$. Then

$$\frac{|\hat{P}_l - P_l|}{P_l} \le \eta\frac{\Delta_l}{P_l} \le \eta(1+\rho_L) \to 0.$$

If instead $|\hat{P}_l - P_l| \le \eta\sqrt{\frac{P_l \vee Q_l}{n}}$, we have

$$\frac{|\hat{P}_l - P_l|}{P_l} \le \eta\sqrt{\frac{\rho_L}{P_l n}} \le \eta\sqrt{\rho}_L\sqrt{\frac{1}{c}} \to 0,$$

where we use the fact that $\frac{P_l \vee Q_l}{P_l}$ is at most ρ_L. □

LEMMA B.4. *Suppose* $\frac{1}{\rho_L} \le \frac{P_l}{Q_l} \le \rho_L$ *and* $P_l, Q_l \ge \frac{c}{n}$ *for all* l*, where* $\rho_L > 1$*. Conditioned on the event that the conclusion of Proposition B.1 holds for a sequence* η *such that* $\eta\rho_L^2 \to 0$*:*

1. *For all* l *satisfying* $n\frac{\Delta_l^2}{P_l \vee Q_l} \ge 1$*, we have*

$$\left|\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right| \le \left|\frac{P_l - Q_l}{Q_l}\right|(1+\eta'),$$

   where $\eta' \to 0$ *and* η' *does not depend on the color* l*.*
2. *For all* l *satisfying* $n\frac{\Delta_l^2}{P_l \vee Q_l} < 1$*, we have*

$$\left|\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right| \le 2\rho_L \frac{1}{\sqrt{n(P_l \vee Q_l)}}.$$

By symmetry, the same bounds also hold for $\sqrt{\frac{\hat{Q}_l}{\hat{P}_l}} - 1$*.*

PROOF. First suppose $\frac{n\Delta_l^2}{P_l \vee Q_l} \ge 1$. By Lemma B.3, we have $\frac{\hat{Q}_l - Q_l}{Q_l} = \eta'$, where $\eta' \to 0$. In the following derivation, we use η' to denote a sequence such that $\eta' = o(1)$; the actual value of η' may change from instance to instance. We use η to denote a sequence where $\eta\rho_L = o(1)$. We have

$$\begin{aligned}
\frac{\hat{P}_l}{\hat{Q}_l} - 1 &= \frac{\hat{P}_l - P_l + P_l}{\hat{Q}_l - Q_l + Q_l} - 1 = \frac{\frac{\hat{P}_l - P_l}{Q_l} + \frac{P_l}{Q_l}}{\frac{\hat{Q}_l - Q_l}{Q_l} + 1} - 1 \\
&= \left(\frac{P_l}{Q_l} + \frac{\hat{P}_l - P_l}{Q_l}\right)\left(1 - \frac{\hat{Q}_l - Q_l}{Q_l}(1+\eta')\right) - 1 \\
&= \frac{P_l}{Q_l} + \frac{\hat{P}_l - P_l}{Q_l} - \frac{P_l}{Q_l}\frac{\hat{Q}_l - Q_l}{Q_l}(1+\eta') - \frac{\hat{P}_l - P_l}{Q_l}\frac{\hat{Q}_l - Q_l}{Q_l}(1+\eta') - 1 \\
&\overset{(a)}{=} \frac{P_l}{Q_l} + \frac{\eta\Delta_l}{Q_l} + \rho_L\frac{\eta\Delta_l}{Q_l}(1+\eta') + \frac{\eta\Delta_l}{Q_l}\eta' - 1 \\
&= \frac{P_l}{Q_l} + \eta\frac{\Delta_l}{Q_l} + \eta\rho_L\frac{\Delta_l}{Q_l} - 1 = \frac{P_l - Q_l}{Q_l}(1+\eta').
\end{aligned}$$

In (a), we have used the fact that $|\hat{P}_l - P_l| \le \eta\Delta_l$ and $|\hat{Q}_l - Q_l| \le \eta\Delta_l$ by Proposition B.1. Applying the inequality $|\sqrt{x} - 1| \le |x - 1|$ then completes the proof of the first case.

The proof of the second case is almost identical. Suppose $\frac{n\Delta_l^2}{P_l \vee Q_l} < 1$. Then

$$|\hat{P}_l - P_l| = \eta\sqrt{\frac{P_l \vee Q_l}{n}}, \quad \text{and} \quad |\hat{Q}_l - Q_l| = \eta\sqrt{\frac{P_l \vee Q_l}{n}}.$$

By Lemma B.3, we have $\frac{\hat{Q}_l - Q}{Q_l} = \eta'$, where $\eta' = o(1)$. Hence,

$$\begin{aligned}
\frac{\hat{P}_l}{\hat{Q}_l} - 1 &= \left(\frac{P_l}{Q_l} + \frac{\hat{P}_l - P_l}{Q_l}\right)\left(1 - \frac{\hat{Q}_l - Q_l}{Q_l}(1+\eta')\right) - 1 \\
&= \frac{P_l}{Q_l} + \frac{\hat{P}_l - P_l}{Q_l} - \frac{P_l}{Q_l}\frac{\hat{Q}_l - Q_l}{Q_l}(1+\eta') - \frac{\hat{P}_l - P_l}{Q_l}\frac{\hat{Q}_l - Q_l}{Q_l}(1+\eta') - 1 \\
&= \frac{P_l}{Q_l} + \eta\sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} + \rho_L\eta\sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} + \eta\sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} - 1 \\
&= \frac{P_l}{Q_l} + \rho_L\eta\sqrt{\frac{(P_l \vee Q_l)^2}{nQ_l^2(P_l \vee Q_l)}} - 1 \\
&= \frac{P_l - Q_l}{Q_l} + \eta\rho_L^2\sqrt{\frac{1}{n(P_l \vee Q_l)}} = \frac{P_l - Q_l}{Q_l} + \eta'\rho_L\sqrt{\frac{1}{n(P_l \vee Q_l)}}.
\end{aligned}$$

Using the inequality $|\sqrt{1+x} - 1| \le x$ for $x \ge 0$, we conclude that

$$\begin{aligned}
\left|\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} - 1\right| &= \left|\sqrt{1 + \frac{P_l - Q_l}{Q_l} + \eta'\rho_L\frac{1}{\sqrt{n(P_l \vee Q_l)}}} - 1\right| \\
&\le \left|\frac{P_l - Q_l}{Q_l} + \eta'\rho_L\frac{1}{\sqrt{n(P_l \vee Q_l)}}\right| \\
&\overset{(a)}{\le} 2\rho_L\frac{1}{\sqrt{n(P_l \vee Q_l)}},
\end{aligned}$$

where (a) follows because we have assumed that $\frac{n\Delta_l^2}{P_l \vee Q_l} < 1$, implying

$$\left|\frac{P_l - Q_l}{Q_l}\right| \le \sqrt{\frac{P_l \vee Q_l}{nQ_l^2}} = \sqrt{\frac{(P_l \vee Q_l)^2}{nQ_l^2(P_l \vee Q_l)}} \le \rho_L\frac{1}{\sqrt{n(P_l \vee Q_l)}}.$$

□

The following lemma provides a bound for $\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}}Q_l$:

LEMMA B.5. *Suppose*

$$\frac{|\hat{Q}_l - Q_l|}{Q_l} = \eta', \quad \textit{and} \quad \frac{|\hat{P}_l - P_l|}{P_l} = \eta',$$

where $\eta' = o(1)$. *For all sufficiently small* η, *we have*

$$\frac{1}{8}\sqrt{P_l Q_l} \leq \frac{1}{2}\sqrt{\hat{P}_l \hat{Q}_l} \leq \sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} Q_l.$$

PROOF. We have the sequence of equalities

$$\begin{aligned}\sqrt{\frac{\hat{P}_l}{\hat{Q}_l}} Q_l = \sqrt{\hat{P}_l \hat{Q}_l}\frac{Q_l}{\hat{Q}_l} &= \sqrt{\hat{P}_l \hat{Q}_l}\frac{1}{\frac{\hat{Q}_l - Q_l}{Q_l} + 1} \\ &= \sqrt{\hat{P}_l \hat{Q}_l}\left(1 - \frac{\hat{Q}_l - Q_l}{Q_l}(1+\eta')\right) = \sqrt{\hat{P}_l \hat{Q}_l}(1-\eta),\end{aligned}$$

where the penultimate equality uses the fact that $\frac{\hat{Q}_l - Q_l}{Q_l} = \eta' \to 0$. Taking η sufficiently small yields the upper bound. For sufficiently small η, we also have $\hat{P}_l \geq \frac{1}{2}P_l$ and $\hat{Q}_l \geq \frac{1}{2}Q_l$, yielding the lower bound. □

LEMMA B.6. *Define* $L_1 = \{l : \frac{n\Delta_l^2}{P_l \vee Q_l} \geq C_{test}^2\}$. *Then*

$$C_1 \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l} \leq I_L \leq C_2 \sum_{l \in L_1} \frac{\Delta_l^2}{P_l \vee Q_l}, \tag{17}$$

for some constants C_1 *and* C_2.

PROOF. Throughout the proof, let η' denote a sequence converging to 0, and let C denote a $\Theta(1)$ sequence that may change from line to line. First observe that

$$I_L = -2\log \sum_l \sqrt{P_l Q_l} = -2\log\left(1 - \frac{\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2}{2}\right).$$

Using the fact that $I_L \to 0$ as $n \to \infty$, so $\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 \to 0$, we have the following bound for sufficiently large n:

$$\frac{1}{2}\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 \leq I_L \leq 2\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2.$$

Since $\sum_l(\sqrt{P_l}-\sqrt{Q_l})^2 = \sum_l \frac{\Delta_l^2}{(\sqrt{P_l}+\sqrt{Q_l})^2}$, there exist constants $\tilde{C}_1$ and $\tilde{C}_2$ such that

$$\tilde{C}_1 \sum_l \frac{\Delta_l^2}{P_l \vee Q_l} \leq I_L \leq \tilde{C}_2 \sum_l \frac{\Delta_l^2}{P_l \vee Q_l}.$$

Now note that $\sum_{l\in L_1^c} \frac{\Delta_l^2}{P_l \vee Q_l} \leq \frac{LC_{test}^2}{n}$, and $\frac{L}{n} = o(I_L)$ by assumption. Hence, the sum over labels in L_1 must be $\Theta(I_L)$, implying inequality (17). □

The following lemmas analyze the consensus step of the algorithm:

LEMMA B.7 (Lemma 4 of Gao et al [15]). *Let σ and σ' be two clusters such that, for some constant $C \geq 1$, the minimum cluster size is at least $\frac{n}{Ck}$. Define the map $\xi : [k] \to [k]$ according to*

$$\xi(k) = \arg\max_{k'} |\{v : \sigma(v) = k\} \cap \{v : \sigma'(v) = k'\}|.$$

If $\min_{\pi\in S_k} l(\pi(\sigma), \sigma') < \frac{1}{Ck}$, then $\xi \in S_k$ and $l(\xi(\sigma), \sigma') = \min_{\pi\in S_k} l(\pi(\sigma), \sigma')$.

LEMMA B.8. *Let $\sigma, \sigma' : [n] \to [K]$ be two clusterings where the minimum cluster size of σ is T. Let $\pi, \xi \in S_K$ be such that $d(\pi(\sigma), \sigma') < \frac{T}{2n}$ and $d(\xi(\sigma), \sigma') < \frac{T}{2n}$. Then $\pi = \xi$.*

PROOF. Suppose the contrary, and choose any k such that $\pi(k) \neq \xi(k)$. We then have

$$|\{\sigma(u) = k\} \cap \{\sigma'(u) \neq \pi(k)\}| < n \cdot d(\pi(\sigma), \sigma') < \frac{T}{2},$$

implying that $|\{\sigma(u) = k\} \cap \{\sigma'(u) = \pi(k)\}| > \frac{T}{2}$. But then

$$\begin{aligned} n \cdot d(\xi(\sigma), \sigma') &\geq |\{\sigma(u) = k\} \cap \{\sigma'(u) \neq \xi(k)\}| \\ &\geq |\{\sigma(u) = k\} \cap \{\sigma'(u) = \pi(k)\}| \geq \frac{T}{2}, \end{aligned}$$

a contradiction. □

APPENDIX C: PROOF OF PROPOSITION 6.3

Note on notation: In order to simplify presentation, we use slightly different notation from the main paper. In the entirety of Appendices C, D, and E, we drop the subscript n on $P_l, Q_l, p(z), q(z), h(z), \gamma(z), I, L, H, \alpha$, and R, and make the dependence implicit. Furthermore, we drop the subscript

Φ on $p(z), q(z), h(z), \phi(z)$, and I. Lastly, we redefine $P_l = \int_{a_l}^{b_l} p(z)dz$ and $Q_l = \int_{a_l}^{b_l} q(z)dz$ and use

$$\tilde{P}_l := (1 - P_0) \int_{a_l}^{b_l} p(z)dz := (1 - P_0)P_l,$$
$$\tilde{Q}_l := (1 - Q_0) \int_{a_l}^{b_l} q(z)dz := (1 - Q_0)Q_l.$$

With the new notation, we have

$$I_L = -2 \log \left(\sqrt{P_0 Q_0} + \sum_{l=1}^{L} \sqrt{\tilde{P}_l \tilde{Q}_l} \right),$$
$$I = -2 \log \left(\sqrt{P_0 Q_0} + \int_0^1 (1 - P_0)(1 - Q_0) p(z) q(z) dz \right)$$

in place of I_{L_n} and $I_{\Phi,n}$.

PROOF. We first show that the likelihood ratio $\frac{\tilde{P}_l}{\tilde{Q}_l} = \frac{1-P_0}{1-Q_0} \frac{P_l}{Q_l}$ satisfies the claimed bounds. Consider an l such that $\text{bin}_l \cap R^c = \emptyset$. For all $z \in \text{bin}_l$, we have $\frac{1}{\rho} \leq \frac{p(z)}{q(z)} \leq \rho$ by Assumption C2'. It follows that $\frac{P_l}{Q_l} = \frac{\int_{\text{bin}l} p(z)dz}{\int_{\text{bin}_l} q(z)dz} \leq \rho$. The lower bound is derived in the same manner. Since $\frac{1-P_0}{1-Q_0} \in \left[\frac{1}{c_0}, c_0\right]$, we conclude that $\frac{1}{\rho c_0} \leq \frac{P_l}{Q_l} \leq \rho c_0$.

Now suppose $\text{bin}_l \cap R^c \neq \emptyset$. Since R is an interval and that $\mu\{R^c\} = o(H)$, and since $L \leq \frac{2}{H}$, we conclude that $\mu\{R^c\} < \frac{1}{2L}$ for all sufficiently large n. Thus, only bins $[0, \frac{1}{L}]$ and $[1 - \frac{1}{L}, 1]$ can satisfy $\text{bin}_l \cap R^c \neq 0$. Note that Assumption C5' implies both $p(z)$ and $q(z)$ are increasing in $\left[0, \frac{1}{L}\right]$ and decreasing in $\left[1 - \frac{1}{L}, 1\right]$. Define $P_l' = \int_{\text{bin}_l \cap R} p(z)dz$ and $Q_l' = \int_{\text{bin}_l \cap R} q(z)dz$, and define $P_l'' = \int_{\text{bin}_l \cap R^c} p(z)dz$ and $Q_l'' = \int_{\text{bin}_l \cap R^c} q(z)dz$. Then $P_l = P_l' + P_l''$ and $Q_l = Q_l' + Q_l''$. Note that $\frac{1}{\rho} \leq \frac{P_l'}{Q_l'} \leq \rho$ by the same argument as before. Furthermore, using the monotonic properties of $p(z)$ and $q(z)$ in the relevant intervals, we have

$$P_l' \geq \min_{z \in \text{bin}_l \cap R} \frac{p(z)}{2L} \geq \max_{z \in \text{bin}_l \cap R^c} \frac{p(z)}{2L} \geq P_l'',$$

where the first inequality follows because $\mu(R^c) \leq \frac{1}{2L}$, and the second inequality follows from Assumption C5'. Similarly, $Q_l' \geq Q_l''$. Thus,

$$\frac{P_l}{Q_l} \leq \frac{2P_l'}{Q_l'} \leq 2\rho, \quad \text{and} \quad \frac{P_l}{Q_l} \geq \frac{P_l'}{2Q_l'} \geq \frac{1}{2\rho}.$$

Using the bound on $\frac{1-P_0}{1-Q_0}$ completes the proof.

We now proceed with bounding $|I-I_L|$. Using the simple relation between Renyi divergence and Hellinger distance detailed in Lemma I.1, we have

$$\begin{aligned} I &= (1+o(1))\left\{(\sqrt{P_0}-\sqrt{Q_0})^2+\int_0^1\left(\sqrt{(1-P_0)p(z)}-\sqrt{(1-Q_0)q(z)}\right)^2dz\right\} \\ &= (1+o(1))\Bigg\{(\sqrt{P_0}-\sqrt{Q_0})^2+(\sqrt{1-P_0}-\sqrt{1-Q_0})^2 \\ &\qquad +\sqrt{(1-P_0)(1-Q_0)}\int_0^1\left(\sqrt{p(z)}-\sqrt{q(z)}\right)^2dz\Bigg\}. \end{aligned}$$

Likewise,

$$\begin{aligned} I_L &= (1+o(1))\left\{(\sqrt{P_0}-\sqrt{Q_0})^2+\sum_{l=1}^{L}(\sqrt{\tilde{P}_l}-\sqrt{\tilde{Q}_l})^2dz\right\} \\ &= (1+o(1))\Bigg\{(\sqrt{P_0}-\sqrt{Q_0})^2+(\sqrt{1-P_0}-\sqrt{1-Q_0})^2 \\ &\qquad +\sqrt{(1-P_0)(1-Q_0)}\sum_{l=1}^{L}(\sqrt{P_l}-\sqrt{Q_l})^2\Bigg\}. \end{aligned}$$

The key step in completing our proof is the following proposition, proved in Appendix D.1:

PROPOSITION C.1. *Under Assumptions C1'–C5', we have*

$$\left|\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz-\sum_{l=1}^{L}(\sqrt{P_l}-\sqrt{Q_l})^2\right|=o\left(\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz\right).$$

The claimed result follows from Proposition C.1 by noticing that

$$\begin{aligned} I_L &= (1+o(1))\Big\{(\sqrt{P_0}-\sqrt{Q_0})^2+(\sqrt{1-P_0}-\sqrt{1-Q_0})^2 \\ &\qquad +\sqrt{(1-P_0)(1-Q_0)}(1+o(1))\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz\Big\} \\ &= (1+o(1))I. \end{aligned}$$

The proof of Proposition C.1 contains a number of subparts, which we briefly outline below. Since $p(z)$ and $q(z)$ are easier to handle on the interval

R, we initially only concern ourselves with comparing

$$H_R := \int_R (\sqrt{p(z)} - \sqrt{q(z)})^2 dz, \quad \text{and} \quad H_L^R := \sum_{l=1}^{L} (\sqrt{P_l'} - \sqrt{Q_l'})^2.$$

We first notice that $\{\text{bin}_l\} \cap R$ constitute an approximately-uniform binning of R; i.e., there exist constants c_{bin} and C_{bin} such that $\frac{c_{\text{bin}}}{L} \le |\text{bin}_l \cap R| \le \frac{C_{\text{bin}}}{L}$. This is reasoned as follows: Since R is an interval, we know that $\text{bin}_l \cap R$ is an interval, as well. The inequality $|\text{bin}_l \cap R^c| \le \mu\{R^c\} \le \frac{1}{2L}$ then implies $\frac{1}{2L} \le |\text{bin}_l \cap R| \le \frac{1}{L}$.

In a series of lemmas, we show that the approximately-uniform binning of R leads to several useful bounds on H^R and H_L^R. In particular, Lemma D.1 shows that as long as L grows, we have $d_L \to \frac{1}{4}$, where $d_L := \sum_l Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)$ and $\gamma_l' = Q_l' - P_l'$. In Lemma D.2, we show that $H^R = \frac{\alpha^2}{4}(1+\eta)$, where $\eta = \Theta(\alpha)$. Similarly, Lemma D.3 establishes that $H_L^R = d_L \alpha^2 (1 + \eta_L)$. We combine the results of Lemmas D.1, D.2, and D.3 in Lemma D.4, to show that $|H^R - H_L^R| = o(H^R)$. The last step is to bound the difference between the sums and integrals over R and the entire real line. □

APPENDIX D: APPENDIX FOR PROPOSITION 6.3

D.1. Proof of Proposition C.1. Let a_L be an $o(1)$ sequence such that $\mu(R^c) \le a_L H$. We divide the set of bins into three subsets:

$$\begin{aligned} L_1 &= \{l : \text{bin}_l \cap R^c = \emptyset\}, \\ L_2 &= \{l : \text{bin}_l \cap R^c \neq \emptyset,\ P_l \vee Q_l \ge 2C a_L H\}, \\ L_3 &= \{l : \text{bin}_l \cap R^c \neq \emptyset,\ P_l \vee Q_l \le 2C a_L H\}. \end{aligned}$$

For each bin l, define $P_l' = \int_{\text{bin}_l \cap R} p(z)dz$ and $P_l'' = \int_{\text{bin}_l \cap R^c} p(z)dz$, and likewise define Q_l' and Q_l''. We now proceed in two steps:

Step 1: We first claim that for all $l \in L_2$,

$$\left|(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2\right| \le a_L H.$$

Since $\mu(R^c) \le a_L H$, we have $P_l'' = \int_{\text{bin}_l \cap R^c} p(z)dz \le C a_L H$, and likewise for Q_l''. Then

$$\begin{aligned} (\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2 &= P_l + Q_l - P_l' - Q_l' - 2\sqrt{P_l Q_l} + 2\sqrt{P_l' Q_l'} \\ &\overset{(a)}{\le} P_l'' + Q_l'' - 2\sqrt{P_l'' Q_l''} \le P_l'' + Q_l'' \le 2C a_L H. \end{aligned}$$

Here, inequality (a) holds by the following reasoning: By the AM-GM inequality, we have $2\sqrt{P_l' Q_l' P_l'' Q_l''} \le P_l' Q_l'' + P_l'' Q_l'$. Thus,

$$P_l' Q_l' + P_l'' Q_l'' + 2\sqrt{P_l' Q_l' P_l'' Q_l''} \le (P_l' + P_l'')(Q_l' + Q_l'') = P_l Q_l.$$

Taking square roots, we conclude that $\sqrt{P_l' Q_l'} + \sqrt{P_l'' Q_l''} \le \sqrt{P_l Q_l}$, yielding (a). On the other hand, we have

$$\begin{aligned}\sqrt{P_l Q_l} - \sqrt{P_l' Q_l'} &= \frac{P_l Q_l - P_l' Q_l'}{\sqrt{P_l Q_l} + \sqrt{P_l' Q_l'}} \\ &= \frac{P_l' Q_l'' + P_l'' Q_l' + P_l'' Q_l''}{\sqrt{P_l Q_l} + \sqrt{P_l' Q_l'}} \le \frac{P_l' Q_l'' + P_l'' Q_l' + P_l'' Q_l''}{2\sqrt{P_l' Q_l'}} \\ &\le Q_l'' \frac{P_l'}{2\sqrt{P_l' Q_l'}} + P_l'' \frac{Q_l'}{2\sqrt{P_l' Q_l'}} + Q_l'' \frac{P_l''}{2\sqrt{P_l' Q_l'}}.\end{aligned}$$

Note that because P_l' and Q_l' are defined on R, we have

$$\left|\frac{P_l'}{Q_l'}\right| = \left|\int_{\mathrm{bin}_l \cap R} \frac{p(z)}{Q_l'} dz\right| \le \int_{\mathrm{bin}_l \cap R} \left|\frac{p(z)}{q(z)}\right| \frac{q(z)}{Q_l'} dz \le \rho.$$

Thus, $\sqrt{\frac{P_l'}{Q_l'}} \vee \sqrt{\frac{Q_l'}{P_l'}} \le \sqrt{\rho}$, so $Q_l'' \frac{P_l'}{2\sqrt{P_l' Q_l'}} + P_l'' \frac{Q_l'}{2\sqrt{P_l' Q_l'}} \le \sqrt{\rho}(Q_l'' + P_l'')$.

We still need to bound the last term $\frac{Q_l'' P_l''}{2\sqrt{P_l' Q_l'}}$. Since $l \in L_2$, either $P_l \ge 2Ca_L H$ or $Q_l \ge 2Ca_L H$. Suppose the former inequality holds; the latter case may be handled in an identical manner. Since $P_l'' \le Ca_L H$ and $P_l \ge 2Ca_L H$, we have $P_l'' \le P_l'$, so

$$\frac{Q_l'' P_l''}{2\sqrt{P_l' Q_l'}} \le Q_l'' \frac{P_l'}{2\sqrt{P_l' Q_l'}} \le \sqrt{\rho} Q_l''.$$

Hence, $\sqrt{P_l Q_l} - \sqrt{P_l' Q_l'} \le 2\sqrt{\rho}(Q_l'' + P_l'')$, so

$$\begin{aligned}(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2 &= P_l + Q_l - P_l' - Q_l' - 2\sqrt{P_l Q_l} + 2\sqrt{P_l' Q_l'} \\ &\ge P_l'' + Q_l'' - 4\sqrt{\rho}(Q_l'' + P_l'') \\ &\ge -(4\sqrt{\rho} - 1)(P_l'' + Q_l'') \\ &\ge -(4\sqrt{\rho} - 1) \cdot 2Ca_L H.\end{aligned}$$

Combining these two bounds yields

$$\left|(\sqrt{P_l} - \sqrt{Q_l})^2 - (\sqrt{P_l'} - \sqrt{Q_l'})^2\right| \le C_{C,\rho} a_L H,$$

for an appropriate constant $C_{C,\rho}$. This completes step 1.

Step 2: In step 2, we verify that $\{\text{bin}_l\}_{l\in L_1}\cup\{\text{bin}_l\cap R\}_{l\in L_2}\cup\{\text{bin}_l\cap R\}_{l\in L_3}$ constitutes a valid approximately-uniform binning of R. First, since R is an interval, it is easy to see that $\text{bin}_l\cap R$ is also an interval. Second, we have $|\text{bin}_l\cap R^c|\le\mu\{R^c\}\le a_L H$. Since $H\le\frac{1}{L}$ by assumption, we have $\mu\{R^c\}\le\frac{a_L}{L}$, so there exists a constant C_{bin} such that $\frac{C_{\text{bin}}}{L}\le|\text{bin}_l\cap R|\le\frac{1}{L}$.

Step 3: We now turn to main step of the proof. We may bound $|H-H_L|$ as

$$
\begin{aligned}
&\left|\sum_{l=1}^{L}(\sqrt{P_l}-\sqrt{Q_l})^2-\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz\right| \\
&\quad=\Big|\sum_{l\in L_1}(\sqrt{P_l}-\sqrt{Q_l})^2+\sum_{l\in L_2}(\sqrt{P_l}-\sqrt{Q_l})^2+\sum_{l\in L_3}(\sqrt{P_l}-\sqrt{Q_l})^2 \\
&\qquad\qquad-\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz\Big| \\
&\quad\overset{(a)}{\le}\Big|\sum_{l\in L_1}(\sqrt{P_l}-\sqrt{Q_l})^2+\sum_{l\in L_2}(\sqrt{P_l}-\sqrt{Q_l})^2 \\
&\qquad\qquad-\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz\Big|+8Ca_LH \\
&\quad\overset{(b)}{\le}\Big|\sum_{l\in L_1}(\sqrt{P_l}-\sqrt{Q_l})^2+\sum_{l\in L_2}(\sqrt{P'_l}-\sqrt{Q'_l})^2 \\
&\qquad\qquad-\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz\Big|+C_{C,\rho}a_LH \\
&\quad\overset{(c)}{\le}\Big|\sum_{l\in L_1}(\sqrt{P_l}-\sqrt{Q_l})^2+\sum_{l\in L_2}(\sqrt{P'_l}-\sqrt{Q'_l})^2+\sum_{l\in L_3}(\sqrt{P'_l}-\sqrt{Q'_l})^2 \\
&\qquad\qquad-\int_0^1(\sqrt{p(z)}-\sqrt{q(z)})^2dz\Big|+C_{C,\rho}a_LH \\
&\quad\overset{(d)}{\le}\Big|\sum_{l\in L_1}(\sqrt{P_l}-\sqrt{Q_l})^2+\sum_{l\in L_2}(\sqrt{P'_l}-\sqrt{Q'_l})^2+\sum_{l\in L_3}(\sqrt{P'_l}-\sqrt{Q'_l})^2 \\
&\qquad\qquad-\int_R(\sqrt{p(z)}-\sqrt{q(z)})^2dz\Big|+C_{C,\rho}a_LH \\
&\quad\overset{(e)}{\le}C_{C,\rho}a_LH,
\end{aligned}
$$

where (a) follows because $P_l\vee Q_l\le 2Ca_LH$ for all $l\in L_3$, and $|L_3|\le 2$; (b) follows from step 1 and the fact that $|L_2|\le 2$; (c) follows because $P'_l\le P_l$, so $\sum_{l\in L_3}(\sqrt{P'_l}-\sqrt{Q'_l})^2\le 2Ca_LH$; (d) follows because $\int_{R^c}(\sqrt{p(z)}-$

$\sqrt{q(z)})^2 \leq C\mu\{R^c\} = Ca_L H$; and (e) follows by Lemma D.4. Since $a_L \to 0$, the conclusion follows.

D.2. Lemmas for Proposition 6.3.

LEMMA D.1. *Let* $d_L = \sum_l Q'_l \left(\frac{\gamma'_l}{Q_l}\right)^2$. *Suppose Assumptions C1'–C5' hold. Then* $\lim_{L\to\infty} d_L = \frac{1}{4}$.

PROOF. Let $h(z)$ be as defined in Assumption C3'; in particular, $|h(z)| \geq \left|\frac{\gamma'(z)}{q(z)}\right| \vee \left|\frac{q'(z)}{q(z)}\right|$. For a parameter $0 < \tau < 1$ to be chosen later, we call bin_l *good* if $\sup_{z\in\text{bin}_l} |h(z)| \leq L^\tau$. We first argue that the proportion of bad bins converges to 0 as $L \to \infty$. Since $h(z)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped, the set $\{z : |h(z)| \geq L^\tau\}$ is a union of at most two intervals, for all $L \geq C_s'^{1/\tau}$. Using the notation $B_l = b_l - a_l$, we have

$$\sum_{l\in\{l:|h(z)|\geq L^\tau\}} B_l \leq \mu\left(\{z : |h(z)| \geq L^\tau\}\right) + \frac{4C_{\text{bin}}}{L} \overset{(a)}{\leq} \frac{C}{L^{\tau t}} + \frac{4C_{\text{bin}}}{L} \overset{(b)}{\leq} \frac{C}{L^{\tau t}},$$

where (a) follows because $\int_R |h(z)|^t dz < \infty$ by Assumption C4'; and (b) follows because $t \leq 1$ by Assumption C4', and the fact that $\tau t < 1$ by choice. In particular, the number of bad bins may be bounded as follows:

$$\#\{l : |h(z)| \geq L^\tau\} \leq \frac{CL^{-\tau t}L}{C_{\text{bin}}} \leq CL^{1-\tau t},$$

where we redefine the constant C suitably. For a bad bin l, we may bound $Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2$ as follows:

$$\begin{aligned}
Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2 &= Q'_l \left(\frac{1}{Q'_l}\int_{\text{bin}_l} \gamma(z)dz\right)^2 = Q'_l \left(\int_{\text{bin}_l} \frac{\gamma(z)}{q(z)}\frac{q(z)}{Q'_l}dz\right)^2 \\
&\overset{(a)}{\leq} Q'_l \int_{\text{bin}_l} \frac{q(z)}{Q'_l}\left(\frac{\gamma(z)}{q(z)}\right)^2 dz \leq \int_{\text{bin}_l} q(z)\left(\frac{\gamma(z)}{q(z)}\right)^2 dz \\
&\overset{(b)}{\leq} \left(\int_{\text{bin}_l} q(z)\left|\frac{\gamma(z)}{q(z)}\right|^r dz\right)^{2/r} \left(\int_{\text{bin}_l} q(z)dz\right)^{(r-2)/r} \\
&\overset{(c)}{\leq} C(C_{\text{bin}})^{(r-2)/r} L^{-(r-2)/r} = CL^{-(r-2)/r}.
\end{aligned}$$

Here, (a) follows from Jensen's inequality, (b) follows from Hölder's inequality, and (c) follows because $\int_R q(z)\left|\frac{\gamma(z)}{q(z)}\right|^r dz < \infty$ by Assumption C3' and the fact that $\int_{\text{bin}_l} q(z)dz \leq \frac{CC_{bin}}{L}$.

We now have

$$d_L = \sum_{l=1}^{L} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2 = \sum_{l\,good} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2 + \sum_{l\,bad} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2$$
$$\leq \sum_{l\,good} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2 + CL^{-(r-2)/r}|\{l : l \text{ bad}\}|$$
$$\leq \sum_{l\,good} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2 + CL^{1-\tau t-\frac{(r-2)}{r}} = \sum_{l\,good} Q'_l \left(\frac{1}{2}\frac{\gamma'_l}{Q'_l}\right)^2 + CL^{\frac{2}{r}-\tau t}.$$

For each good bin l, define $z_l = \arg\max_{z\in \text{bin}_l} |q(z)|$. The maximum is attainable since q is continuous and bounded. Furthermore,

$$Q'_l = \int_{\text{bin}_l} q(z)dz = \int_{a_l}^{b_l} q(z)dz$$
$$= \int_{a_l}^{b_l} q(z_l) + q'(c_z)(z - z_l)dz \quad (\text{for some } c_z \in [a_l, b_l])$$
$$= B_l q(z_l) + \int_{a_l}^{b_l} q'(c_z)(z - z_l)dz = B_l q(z_l) + B_l^2 \xi_l,$$

where we define $\xi_l := \frac{1}{B_l^2}\int_{a_l}^{b_l} q'(c_z)(z - z_l)dz$. We also have

$$B_l \left|\frac{\xi_l}{q(z_l)}\right| \leq \frac{1}{B_l}\int_{a_l}^{b_l} \left|\frac{q'(c_z)}{q(z_l)}\right| |z - z_l|dz \overset{(a)}{\leq} \frac{1}{B_l}\int_{a_l}^{b_l} \left|\frac{q'(c_z)}{q(c_z)}\right| |z - z_l|dz$$
$$\overset{(b)}{\leq} \frac{1}{B_l}\int_{a_l}^{b_l} L^\tau |z - z_l|dz \leq C_{\text{bin}} L^{\tau-1},$$

where (a) follows because $q(c_z) \leq q(z_l)$, and (b) follows because l is a good bin, so $\left|\frac{q'(c_z)}{q(c_z)}\right| \leq L^\tau$. The last inequality follows because $B_l \leq \frac{C_{\text{bin}_l}}{L}$. We may perform a similar analysis on γ:

$$\gamma'_l = \int_{\text{bin}_l} \gamma(z)dz = \int_{a_l}^{b_l} \gamma(z_l) + \gamma'(c_z)(z - z_l)dz = B_l\gamma(z_l) + B_l^2\xi'_l,$$

where $\xi'_l := \frac{1}{B_l^2}\int_{a_l}^{b_l} \gamma'(c_z)(z - z_l)dz$. It is straightforward to verify that $B_l \left|\frac{\xi'_l}{q(z_l)}\right| \leq \frac{1}{2}C_{\text{bin}} L^{\tau-1}$. For any bin l, we also have $Q'_l = \int_{\text{bin}_l} q(z)dz \leq CB_l$, where C is the bound on $p(z) \vee q(z)$. Now we look at a single $Q'_l \left(\frac{\gamma'_l}{Q'_l}\right)^2$ term

for a good bin l:

$$\begin{aligned} Q_l' \left(\frac{\gamma_l'}{Q_l'}\right)^2 = \frac{\gamma_l'^2}{Q_l'} &= \frac{(B_l\gamma(z_l) + B_l^2\xi_l')^2}{B_l q(z_l) + B_l^2\xi_l} \\ &= B_l q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)} + B_l\frac{\xi_l'}{q(z_l)}\right)^2 \left(\frac{1}{1+B_l\frac{\xi_l}{q(z_l)}}\right) \\ &= B_l q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)} + B_l\frac{\xi_l'}{q(z_l)}\right)^2 \left(1 - B_l\frac{\xi_l}{q(z_l)} + \eta_l(B_l\frac{\xi_l}{q(z_l)})^2\right). \end{aligned}$$

To arrive at the last equality, we assume that $L \geq C_{\text{bin}}^{1/(1-\tau)}$. Then $\left|B_l\frac{\xi_l'}{q(z_l)}\right| \leq \frac{1}{2}$, so we may take a Taylor approximation. Here, η_l is a constant satisfying $|\eta_l| \leq 16$. Expanding the right-hand side, we have

$$\begin{aligned} Q_l' \left(\frac{\gamma_l'}{Q_l'}\right)^2 = \Bigg(&B_l q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}\right)^2 + 2B_l q(z_l)\frac{\gamma(z_l)}{q(z_l)}B_l\frac{\xi_l'}{q(z_l)} \\ &+ B_l q(z_l)\left(B_l\frac{\xi_l'}{q(z_l)}\right)^2\Bigg)\cdot\left(1 - B_l\frac{\xi_l}{q(z_l)} + \eta_l(B_l\frac{\xi_l}{q(z_l)})^2\right). \end{aligned}$$

Again, note that $\left|B_l\frac{\xi_l'}{q(z_l)}\right| \leq \frac{C_{\text{bin}}}{2}L^{\tau-1}$ and $\left|B_l\frac{\xi_l}{q(z_l)}\right| \leq \frac{C_{\text{bin}}}{2}L^{\tau-1}$. Suppose $L \geq (2C_{\text{bin}})^{1/(1-\tau)}$, so $\frac{C_{\text{bin}}}{2}L^{\tau-1} \leq \frac{1}{4}$. Then

$$\left|B_l\frac{\xi_l}{q(z_l)}\right| + \left|\eta_l(B_l\frac{\xi_l}{q(z_l)})^2\right| \leq C_{\text{bin}}L^{\tau-1}, \text{ and } \left|1 - B_l\frac{\xi_l}{q(z_l)} + \eta_l(B_l\frac{\xi_l}{q(z_l)})^2\right| \leq 2.$$

We now bound

$$\begin{aligned} &\left|Q_l'\left(\frac{\gamma_l'}{Q_l'}\right)^2 - B_l q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}\right)^2\right| \\ &\leq B_l q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}\right)^2 C_{\text{bin}}L^{\tau-1} + 2B_l q(z_l)\frac{\gamma(z_l)}{q(z_l)}C_{\text{bin}}L^{\tau-1} + B_l q(z_l)C_{\text{bin}}^2 L^{2(\tau-1)}. \end{aligned}$$

The third term is bounded by $C_1 L^{2\tau-3}$, for a suitable constant C_1. To bound the second term, we split into two cases:

Case 1: $\left|\frac{\gamma(z_l)}{q(z_l)}\right| \geq 1$. Then $q(z)\left|\frac{\gamma(z_l)}{q(z_l)}\right| \leq q(z)\left(\frac{\gamma(z_l)}{q(z_l)}\right)^2$.

Case 2: $\left|\frac{\gamma(z_l)}{q(z_l)}\right| \leq 1$. The second term is bounded by $2B_l C C_{\text{bin}}L^{\tau-1} \leq C_2 L^{\tau-2}$, for some constant C_2.

In either case, we have

$$\left| Q_l' \left(\frac{\gamma_l'}{Q_l'} \right)^2 - Bq(z_l) \left(\frac{\gamma(z_l)}{q(z_l)} \right)^2 \right| \leq C_3 B_l q(z_l) \left(\frac{\gamma(z_l)}{q(z_l)} \right)^2 L^{\tau-1} + C_4 L^{\tau-2}.$$

Define $d_R = \sum_{l\, good} B_l q(z_l) \left(\frac{\gamma(z_l)}{q(z_l)} \right)^2$. Then

$$\begin{aligned}
|d_L - d_R| &= \left| \sum_l Q_l' \left(\frac{\gamma_l'}{Q_l'} \right)^2 - \sum_{l\, good} B_l q(z_l) \left(\frac{\gamma(z_l)}{q(z_l)} \right)^2 \right| \\
&\leq \sum_{l\, good} \left| Q_l' \left(\frac{\gamma_l'}{Q_l'} \right)^2 - B_l q(z_l) \left(\frac{\gamma(z_l)}{q(z_l)} \right)^2 \right| + C_{M,M',K,C} L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\tau-1} + L \cdot C_4 L^{\tau-2} + C_5 L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\tau-1} + C_4 L^{\tau-1} + C_5 L^{\frac{2}{r}-\tau t} \\
&\leq C_3 d_R L^{\frac{2-rt}{r(1+t)}} + C_6 L^{\frac{2-rt}{r(1+t)}},
\end{aligned}$$

where we have made the choice $\tau = \frac{2+r}{r(1+t)}$ in the last inequality to balance $L^{\tau-1}$ and $L^{2/r-\tau t}$. Notice that $0 < \tau < 1$ by Assumption C4', since $rt > 2$. Furthermore, $|d_L - d_R| = o(d_R) + o(1)$.

In a similar manner, we bound $|d_R - d|$. We use the same definition of good and bad bins as before, and obtain

$$\begin{aligned}
d &= \int_R q(z) \left(\frac{\gamma(z)}{q(z)} \right)^2 dz = \sum_{l=1}^{L} \int_{\mathrm{bin}_l} q(z) \left(\frac{\gamma(z)}{q(z)} \right)^2 dz \\
&= \sum_{l\, good} \int_{\mathrm{bin}_l} q(z) \left(\frac{\gamma(z)}{q(z)} \right)^2 dz + \sum_{l\, bad} \int_{\mathrm{bin}_l} q(z) \left(\frac{\gamma(z)}{q(z)} \right)^2 dz \\
&\leq \sum_{l\, good} \int_{\mathrm{bin}_l} q(z) \left(\frac{\gamma(z)}{q(z)} \right)^2 dz + |\{l : l\, \mathrm{bad}\}| C L^{-\frac{2}{r}} \\
&\leq \sum_{l\, good} \int_{\mathrm{bin}_l} q(z) \left(\frac{\gamma(z)}{q(z)} \right)^2 dz + C L^{\frac{2}{r}-\tau t}.
\end{aligned}$$

The bound on the second term follows from the previous analysis. For the first term, note that for all $z \in \mathrm{bin}_l$, we have

$$q(z) = q(z_l) + q'(c_l)(z - z_l), \quad \text{and} \quad \gamma(z) = \gamma(z_l) + \gamma'(c_l')(z - z_l),$$

where $c_l, c_l' \in \text{bin}_l$ depend implicitly on z. For bin_l, we have

$$\int_{\text{bin}_l} q(z)\left(\frac{\gamma(z)}{q(z)}\right)^2 = \int_{\text{bin}_l} \frac{(\gamma(z_l)+\gamma'(c_l')(z-z_l))^2}{q(z_l)+q'(c_l)(z-z_l)}dz$$
$$= \int_{\text{bin}_l} q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}+\frac{\gamma'(c_l')}{q(z_l)}(z-z_l)\right)^2\left(\frac{1}{1+\frac{q'(c_l)}{q(z_l)}(z-z_l)}\right)dz.$$

Denote

$$T_1 = \frac{q'(c_l)}{q(z_l)}(x-z_l), \quad \text{and} \quad T_2 = \frac{\gamma'(c_l')}{q(z_l)}(x-z_l).$$

Observe that $|z-z_l| \le B_l$ and $\left|\frac{\gamma'(c_l')}{q(z_l)}\right| \le \left|\frac{\gamma'(c_l')}{q(c_l')}\right| \le L^\tau$. Similarly, $\left|\frac{q'(c_l)}{q(z_l)}\right| \le L^\tau$, so $|T_1|, |T_2| \le C_{\text{bin}}L^{\tau-1}$. Now suppose $C_{\text{bin}}L^{\tau-1} \le \frac{1}{2}$, which is satisfied if $L \ge (2C_{\text{bin}})^{\frac{1}{1-\tau}}$. We obtain

$$\int_{\text{bin}_l} q(z)\left(\frac{\gamma(z)}{q(z)}\right)^2 dz = \int_{\text{bin}_l} q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}+T_2\right)^2\left(\frac{1}{1+T_1}\right)dz$$
$$= \int_{\text{bin}_l}\left(q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}\right)^2 + q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}\right)T_2 + q(z_l)T_2^2\right)$$
$$\cdot(1-T_1+\eta T_1^2)dz,$$

where η is some function of z satisfying $|\eta| \le 16$. Thus,

$$\left|\int_{\text{bin}_l} q(z)\left(\frac{\gamma(z)}{q(z)}\right)^2 dz - B_l q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}\right)^2\right|$$
$$\le B_l q(z_l)\left(\frac{\gamma(z_l)}{q(z_l)}\right)^2 C_{\text{bin}}L^{\tau-1} + B_l q(z_l)\frac{\gamma(z_l)}{q(z_l)}C_{\text{bin}}L^{\tau-1} + B_l q(z_l)C_{\text{bin}}^2 L^{2(\tau-1)}.$$

The same analysis used to bound $|d_L - d_R|$ implies $|d-d_R| = o(d_R)+o(1)$. Since $d=\frac{1}{4}$, we have $d_R \to \frac{1}{4}$, so $d_L \to \frac{1}{4}$. This completes the proof. □

LEMMA D.2. *Let*

$$H^R = \int_R (\sqrt{p(z)}-\sqrt{q(z)})^2 dz, \qquad \delta(z) = q(z)-p(z),$$
$$\alpha^2 = \int_R q(z)\left(\frac{\delta(z)}{q(z)}\right)^2 dz, \qquad \gamma(z) = \frac{\delta(z)}{\alpha}.$$

Suppose Assumptions C1'–C5' hold. Then $H^R = \frac{\alpha^2}{4}(1+\eta)$, *where* $|\eta| \le C(\alpha+\alpha^2)$ *for some constant* C. *In particular, if* $H^R \to 0$, *then* $\alpha, \eta \to 0$.

PROOF. We write

$$H^R = \int_R (\sqrt{q(z)} - \sqrt{q(z) - \delta(z)})^2 dz = \int_R q(z) \left(1 - \sqrt{1 - \frac{\delta(z)}{q(z)}}\right)^2 dz.$$

By convention, let $\frac{\delta(z)}{q(z)} = 0$ whenever $q(z) = p(z) = 0$. Thus, we may define $\xi(z) = 1 - \frac{1}{2}\frac{\delta(z)}{q(z)} - \sqrt{1 - \frac{\delta(z)}{q(z)}}$ for $z \in [0, 1]$ and rewrite

$$\begin{aligned} H^R &= \int_R q(z) \left(1 - (1 - \frac{1}{2}\frac{\delta(z)}{q(z)} + \xi(z))\right)^2 dz \\ &= \int_R q(z) \left(\frac{1}{2}\frac{\delta(z)}{q(z)} + \xi(z)\right)^2 dz = \int_R q(z) \left(\frac{1}{2}\frac{\delta(z)}{q(z)}\right)^2 (1 + \xi_2(z))^2 \, dz, \end{aligned}$$

where $\xi_2(z) = \frac{2\xi(z)}{\delta(z)/q(z)}$ if $\delta(z) \neq 0$, and $\xi_2(z) = 0$ if $\delta(z) = 0$. Thus,

$$\int_R \left(\sqrt{p(z)} - \sqrt{q(z)}\right)^2 dz = (1 + \eta)\frac{\alpha^2}{4},$$

where

$$\begin{aligned} \eta &= \frac{\int_R q(z) \left(\frac{1}{2}\frac{\delta(z)}{q(z)}\right)^2 (\xi_2(z)^2 + 2\xi_2(z)) dz}{\alpha^2/4} \\ &= \int_R q(z) \left(\frac{\gamma(z)}{q(z)}\right)^2 (\xi_2(z)^2 + 2\xi_2(z)) dz. \end{aligned}$$

By Lemma I.3, we have $\xi_2(z) \leq 2\left|\frac{\delta(z)}{q(z)}\right|$, implying that

$$\begin{aligned} |\eta| &\leq \int_R q(z) \left(\frac{\gamma(z)}{q(z)}\right)^2 \left(4\left|\frac{\delta(z)}{q(z)}\right|^2 + 4\left|\frac{\delta(z)}{q(z)}\right|\right) dz \\ &= 4\alpha^2 \int_R q(z) \left|\frac{\gamma(z)}{q(z)}\right|^4 dz + 4\alpha \int_R q(z) \left|\frac{\gamma(z)}{q(z)}\right|^3 dz \leq C(\alpha^2 + \alpha), \end{aligned}$$

using the finiteness of integrals in Assumption C3'. □

LEMMA D.3. *Let*

$$\begin{aligned} H_L^R &= \sum_{l=1}^{L} \left(\sqrt{P_l'} - \sqrt{Q_l'}\right)^2, \qquad \delta(z) = q(z) - p(z), \\ \alpha^2 &= \int_R q(z) \left(\frac{\delta(z)}{q(z)}\right)^2 dz, \qquad \gamma(z) = \frac{\delta(z)}{\alpha} dz. \end{aligned}$$

Suppose that Assumptions C1'–C5' hold. Then $H_L^R = d_L(1+\eta_L)$, *where* $d_L = \sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)^2 dz$, $\gamma_l' = \frac{Q_l' - P_l'}{\alpha}$, *and* $\sup_L |\eta_L| \le C(\alpha + \alpha^2)$, *for some constant* C.

PROOF. Let $\delta_l = Q_l' - P_l'$. We have

$$H_L^R = \sum_{l=1}^L (\sqrt{P_l'} - \sqrt{Q_l'})^2 = \sum_{l=1}^L Q_l' \left(1 - \sqrt{\frac{P_l'}{Q_l'}}\right)^2$$
$$= \sum_{l=1}^L Q_l' \left(1 - \sqrt{1 - \frac{\delta_l}{Q_l'}}\right)^2 = \sum_{l=1}^L Q_l' \left(1 - \left(1 - \frac{1}{2}\frac{\delta_l}{Q_l'} - \xi_l\right)\right)^2,$$

where by convention, we define $\frac{\delta_l}{Q_l'} = 0$ when $Q_l', P_l' = 0$, and we use the shorthand $\xi_l = 1 - \frac{1}{2}\frac{\delta_l}{Q_l'} - \sqrt{1 - \frac{\delta_l}{Q_l'}}$. Hence,

$$H_L^R = \sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\delta_l}{Q_l'} + \xi_l\right)^2 = \sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\delta_l}{Q_l'}\right)^2 (1 + \xi_{2l})^2,$$

where $\xi_{2l} = 0$ if $\frac{\delta_l}{Q_l'} = 0$, and $\xi_{2l} = 2\xi_l \frac{Q_l'}{\delta_l}$ otherwise. Then $H_L^R = (1 + \eta_L) \sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\delta_l}{Q_l'}\right)^2$, where $\eta_L = \frac{\sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\delta_l}{Q_l'}\right)^2 (2\xi_{2l} + \xi_{2l}^2)}{\sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\delta_l}{Q_l'}\right)^2}$. By Lemma I.3, we have $|\xi_{2l}| \le 2\left|\frac{\delta_l}{Q_l'}\right|$. Therefore,

$$|\eta_L| = \left|\frac{\sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\delta_l}{Q_l'}\right)^2 (2\xi_{2l} - \xi_{2l}^2)}{\sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\delta_l}{Q_l'}\right)^2}\right| \le \frac{\sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)^2 (2|\xi_{2l}| + \xi_{2l}^2)}{\sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)^2}$$
$$\le 4\frac{\alpha \sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)^3 + \alpha^2 \sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)^4}{\sum_{l=1}^L Q_l' \left(\frac{1}{2}\frac{\gamma_l'}{Q_l'}\right)^2}.$$

The denominator tends to $\frac{1}{4}$ by Lemma D.1 and may be bounded by $1/(2C')$ for large enough L. To bound the numerator, note that for a single l, we have

$$\int_{a_l}^{b_l} \frac{q(z)}{Q_l'} \left|\frac{\gamma(z)}{q(z)}\right|^3 dz \ge \left|\int_{\text{bin}_l} \frac{q(z)}{Q_l'} \frac{\gamma(z)}{q(z)} dz\right|^3 = \left|\frac{\gamma_l'}{Q_l'}\right|^3.$$

Therefore,

$$\sum_{l=1}^{L} Q_l' \left|\frac{\gamma_l'}{Q_l'}\right|^3 \leq \int_R q(z) \left|\frac{\gamma(z)}{q(z)}\right|^3 \leq M, \text{ and}$$
$$\sum_{l=1}^{L} Q_l' \left|\frac{\gamma_l'}{Q_l'}\right|^4 \leq \int_R q(z) \left|\frac{\gamma(z)}{q(z)}\right|^4 \leq M,$$

implying that $|\eta_L| \leq (2\alpha + \alpha^2)2C'M$. □

LEMMA D.4. *Suppose Assumptions A1'–A4' hold. For any sequences* $L_n, \alpha_n \to \infty$, *we have* $H_L^R = H^R(1 + o(1))$; *i.e.,*

$$\left|\frac{\sum_l (\sqrt{P_l'} - \sqrt{Q_l})^2}{\int_R (\sqrt{p(z)} - \sqrt{q(z)})^2 dz} - 1\right| \to 0.$$

PROOF. Since $|H_L^R - H^R| = \left|d_L\alpha^2(1+\eta_L) - \frac{\alpha^2}{4}(1+\eta)\right|$ by Lemmas D.2 and D.3, we have $\left|\frac{H_L^R}{H^R} - 1\right| = \left|4d_L\frac{(1+\eta_L)}{1+\eta} - 1\right|$, where $|\eta|, |\eta_L| \leq C_1(\alpha + \alpha^2)$ for all L. Thus, $\lim_{\alpha_n \to 0} \sup_L \left|\frac{1+\eta_L}{1+\eta} - 1\right| = 0$. Furthermore, by Lemma D.1, we have $|4d_L - 1| \to 0$, uniformly for all α. Thus, $\left|\frac{H_L^R}{H^R} - 1\right| \to 0$. □

APPENDIX E: PROOF OF PROPOSITION 6.2

Note on notation: In order to simplify presentation, we use slightly different notation in Appendices C, D, and E from the main paper. See the beginning of Appendix C for the notational changes.

First, we prove the bounds on $\frac{\tilde{P}_l}{Q_l}$. Define

$$R = \left\{ z \in [0,1] : \left|\log \frac{p(z)}{q(z)}\right| \leq C(2L)^{1/r} \right\},$$

where $C = \left(\int_0^1 \left|\log \frac{p(z)}{q(z)}\right|^r dz\right)^{1/r}$ is a constant. Since $\int_0^1 \left|\log \frac{p(z)}{q(z)}\right|^r dz < \infty$, Markov's Inequality implies $\mu\{R^c\} \leq \frac{1}{2L}$.

The remainder of the proof follows the argument used to prove Proposition 6.3, except for the final step, where we need to show that

$$\left|\int_0^1 (\sqrt{p(z)} - \sqrt{q(z)})^2 dz - \sum_l (\sqrt{P_l} - \sqrt{Q_l})^2\right| = o\left(\int_0^1 (\sqrt{p(z)} - \sqrt{q(z)})^2 dz\right)$$
$$= o(1).$$

We establish this fact in the following proposition:

PROPOSITION E.1. *Let Assumptions C1 and C3 be satisfied. Let* $bin_l = [a_l, b_l]$ *be a uniform binning of* $[0,1]$*, for* $l = 1, \dots, L$*, and let* $P_l = \int_{bin_l} p(z)dz$ *and* $Q_l = \int_{bin_l} q(z)dz$*. Then*

$$\left| \int_0^1 (\sqrt{p(z)} - \sqrt{q(z)})^2 dz - \sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 \right| \to 0.$$

PROOF. We use a similar argument to the proof of Proposition D.1. First observe that

$$\begin{aligned}\int_0^1 (\sqrt{p(z)} - \sqrt{q(z)})^2 dz &= \int_0^1 p(z)dz + \int_0^1 q(z)dz - 2\int_0^1 \sqrt{p(z)q(z)}dz \\ &= 2 - 2\int_0^1 \sqrt{p(z)q(z)}dz\end{aligned}$$

and

$$\sum_l (\sqrt{P_l} - \sqrt{Q_l})^2 = \sum_l P_l + \sum_l Q_l - 2\sum_l \sqrt{P_l Q_l}.$$

Thus, we only need to show that $\left| \int_0^1 \sqrt{p(z)q(z)}dz - \sum_l \sqrt{P_l Q_l} \right| \to 0$. We have $|h(z)| \geq \left|\frac{p'(z)}{p(z)}\right| \vee \left|\frac{q'(z)}{q(z)}\right|$. Let $\tau \in (0,1)$. We call bin_l *good* if $\sup_{z \in \text{bin}_l} |h(z)| \leq L^\tau$. We argue that the proportion of bad bins converges to 0 as $L \to \infty$: Since $h(z)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped, the set $\{z : |h_n(z)| \geq L^\tau\}$ is a union of at most two intervals, for all $L \geq C_s'^{1/\tau}$. Hence,

$$\begin{aligned}\sum_{l \in \left\{l : \sup_{z \in \text{bin}_l} |h(z)| \geq L^\tau\right\}} B_l &\leq \mu\left(\left\{z : \sup_{z \in \text{bin}_l} |h(z)| \geq L^\tau\right\}\right) + 4C_{\text{bin}}L^{-1} \\ &\overset{(a)}{\leq} CL^{-\tau t} + 4C_{\text{bin}}L^{-1} \overset{(b)}{\leq} CL^{-\tau t},\end{aligned}$$

where (a) follows because $\int_R |h(z)|^t dz < \infty$ by Assumption C3; and (b) follows because $t \leq 1$, so $\tau t < 1$ and the first term dominates. We now bound the number of bad bins:

$$\#\{l : |h(z)| \geq L^\tau\} \leq \frac{CL^{-\tau t}L}{C_{\text{bin}}} \leq CL^{1-\tau t}.$$

For a bad bin, we have $P_l, Q_l \leq \frac{CC_{\text{bin}}}{L}$ and $\int_{\text{bin}_l} (\sqrt{p(z)} - \sqrt{q(z)})^2 dz \leq \frac{2CC_{\text{bin}}}{L}$.

We now consider a good bin l. Let z_l be $\arg\max_{z \in \text{bin}_l} p(z)$. The argmax is attainable since p is continuous and bounded. We have

$$P_l = \int_{a_l}^{b_l} p(z)dz = \int_{a_l}^{b_l} p(z_l) + p'(c_z)(z - z_l)dz = B_l p(z_l) + B_l^2 \xi_l,$$

where $\xi_l = \frac{1}{B_l^2}\int_{a_l}^{b_l} p'(c_z)(z - z_l)dz$. Furthermore,

$$B_l\left|\frac{\xi_l}{p(z_l)}\right| \le \frac{1}{B_l}\int_{a_l}^{b_l}\left|\frac{p'(c_z)}{p(z_l)}\right||z - z_l|dz \le \frac{1}{B_l}\int_{a_l}^{b_l}\left|\frac{p'(c_z)}{p(c_z)}\right||z - z_l|dz$$
$$\le \frac{1}{B_l}\int_{a_l}^{b_l} L^{\tau}|z - z_l|dz \le C_{\text{bin}}L^{\tau-1}.$$

Likewise, define $z_l' = \arg\max_{z\in \text{bin}_l} q(z)$. We have $Q_l = B_l q(z_l') + B_l^2\xi_l'$, where $\xi_l' := \frac{1}{B_l}\int_{a_l}^{b_l} q'(c_z)(z - z_l')dz$. We can also bound $B_l\left|\frac{\xi_l'}{q(z_l')}\right| \le C_{\text{bin}}L^{\tau-1}$. Thus,

$$\sqrt{P_lQ_l} = \sqrt{(B_lp(z_l) + B_l^2\xi_l)(B_lq(z_l') + B_l^2\xi_l')}$$
$$= \sqrt{p(z_l)q(z_l')}\sqrt{(B_l + B_l^2\frac{\xi_l}{p(z_l)})(B_l + B_l^2\frac{\xi_l'}{q(z_l')})}$$
$$= \sqrt{p(z_l)q(z_l')}B_l\sqrt{(1 + B_l\frac{\xi_l}{p(z_l)})(1 + B_l\frac{\xi_l'}{q(z_l')})}.$$

By our bounds on $B_l\frac{\xi_l}{p(z_l)}$ and $B_l\frac{\xi_l'}{q(z_l')}$, we can bound the nuisance term as

$$\sqrt{(1 + B_l\frac{\xi_l}{p(z_l)})(1 + B_l\frac{\xi_l'}{q(z_l')})} \le \sqrt{1 + C_{\text{bin}}L^{\tau-1}(1 + o(1))}$$
$$\le 1 + \frac{1}{2}L^{\tau-1}(1 + o(1)).$$

It is clear that $B_l\sqrt{p(z_l)q(z_l')} \le B_lC$. Therefore,

$$\left|\sqrt{P_lQ_l} - \sqrt{p(z_l)q(z_l')}B_l\right| \le B_lCL^{\tau-1}(1 + o(1)), \tag{18}$$

and likewise,

$$\int_{a_l}^{b_l}\sqrt{p(z)q(z)}dz = \int_{a_l}^{b_l}\sqrt{p(z)q(z)}dz$$
$$= \int_{a_l}^{b_l}\sqrt{(p(z_l) + p'(c_z)(z - z_l))(q(z_l') + q'(c_z')(z - z_l'))}dz$$
$$= \int_{a_l}^{b_l}\sqrt{p(z_l)q(z_l')}$$
$$\cdot\left(\sqrt{1 + (z - z_l)\frac{p'(c_z)}{p(z_l)} + (z - z_l')\frac{q'(c_z')}{q(z_l')} + (z - z_l)(z - z_l')\frac{p'(c_z)}{p(z_l)}\frac{q'(c_z')}{q(z_l')}}\right)dz.$$

Since

$$\left|(z-z_l)\frac{p'(c_z)}{p(z_l)}\right| \le B_l\left|\frac{p'(c_z)}{p(c_z)}\right| \le L^{\tau-1},$$
$$\left|(z-z_l)\frac{q'(c'_z)}{q(z_l)}\right| \le B_l\left|\frac{q'(c'_z)}{q(c'_z)}\right| \le L^{\tau-1},$$

we may bound the nuisance term as follows:

$$\sqrt{1+(z-z_l)\frac{p'(c_z)}{p(z_l)}+(z-z'_l)\frac{q'(c'_z)}{q(z'_l)}+(z-z_l)(z-z'_l)\frac{p'(c_z)}{p(z_l)}\frac{q'(c'_z)}{q(z'_l)}}$$
$$\le \sqrt{1+C_{\mathrm{bin}}L^{\tau-1}(1+o(1))} \le 1+\frac{1}{2}C_{\mathrm{bin}}L^{\tau-1}(1+o(1)).$$

The term $B_l\sqrt{p(z_l)q(z'_l)}$ is bounded by B_lC. Hence,

$$\left|\int_{a_l}^{b_l}\sqrt{p(z)q(z)}dz - B_l\sqrt{p(z_l)q(z'_l)}\right| \le B_lCC_{\mathrm{bin}}L^{\tau-1} \tag{19}$$

By combining inequalities (18) and (19), we have

$$\left|\sqrt{P_lQ_l}-\int_{a_l}^{b_l}\sqrt{p(z)q(z)}dz\right| \le B_lCC_{\mathrm{bin}}L^{\tau-1}.$$

Hence,

$$\begin{aligned}\left|\sum_l\sqrt{P_lQ_l}-\int_0^1\sqrt{p(z)q(z)}dz\right| &\le \sum_{l:\ l\text{ bad}} B_lC \\ &\quad+\sum_{l:\ l\text{ good}}\left|\sqrt{P_lQ_l}-\int_{a_l}^{b_l}\sqrt{p(z)q(z)}dz\right| \\ &\le CL^{-\tau t}+\sum_{l:\ l\text{ good}} B_lCC_{\mathrm{bin}}L^{\tau-1} \\ &\le CL^{-\tau t}+CC_{\mathrm{bin}}L^{\tau-1}.\end{aligned}$$

Setting $\tau=\frac{1}{1+t}$, we obtain $\left|\sum_l\sqrt{P_lQ_l}-\int_0^1\sqrt{p(z)q(z)}dz\right|\to 0$, completing the proof. □

APPENDIX F: PROOFS OF THEOREMS 5.1 AND THEOREM 5.2

We now outline the proofs of Theorems 5.1 and 5.2, with proofs of supporting propositions in the succeeding subsections. We begin with Theorem 5.2.

F.1. Main argument: Proof of Theorem 5.2. By the argument outlined in Section 6.3, the divergences I and H do not change after transforming the densities $p(x)$ and $q(x)$ according to Φ. Proposition F.1 shows that under Assumptions A1'–A5', Assumptions C1'–C5' are also satisfied.

Furthermore, our assumption that $L = o(\frac{1}{H})$ implies $L \leq \frac{2}{H}$ for sufficiently large L. Hence, Proposition 6.3 applies, and we may conclude that after transformation and discretization, the label probabilities satisfy $\frac{1}{2c_0\rho} \leq \frac{P_l}{Q_l} \leq 2c_0\rho$, for all l. Using the assumption $L = o(nI)$ and the fact that $I_L = I(1 + o(1))$ from Proposition 6.3, we also have $L = o(nI_L)$, so we may use Proposition 6.1 (with $\rho_L = 2c_0\rho$) to obtain

$$\lim_{n\to\infty} P\left(l(\hat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI_L}{\beta K}(1 + o(1))\right)\right) \to 1.$$

The probability bound in the theorem then follows from the fact that $I_L = I(1 + o(1))$. If $\frac{nI}{\beta K \log n} \leq 1$, then $\frac{nI_L}{\beta K \log n} \leq 1$ as well since $I_L \leq I$ and thus, we have that $\mathbb{E}l(\hat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI_L}{\beta K}(1 + o(1))\right)$. The expectation bound in the theorem then follows from the fact that $I_L = I(1 + o(1))$ again.

F.2. Main argument: Proof of Theorem 5.1. The proof parallels the argument for Theorem 5.2 outlined above. Proposition F.2 establishes that Assumptions A1–A4 imply Assumptions C1–C4. Hence, Proposition 6.2 applies, and we may conclude that after transformation and discretization, the label probabilities satisfy $\frac{1}{2c_0 \exp(L^{1/r})} \leq \frac{P_l}{Q_l} \leq 2c_0 \exp(L^{1/r})$, for all l, and $I_L = I(1 + o(1))$. Therefore, we may again apply Proposition 6.1 (with $\rho_L = 2c_0 \exp(L^{1/r})$) to obtain

$$\lim_{n\to\infty} P\left(l(\hat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI_L}{\beta K}(1 + o(1))\right)\right) \to 1.$$

The probability bound in the theorem follows from the fact that $I_L = I(1 + o(1))$. If $\frac{nI}{\beta K \log n} \leq 1$, then $\frac{nI_L}{\beta K \log n} \leq 1$ as well since $I_L \leq I$ and thus, we have that $\mathbb{E}l(\hat{\sigma}, \sigma_0) \leq \exp\left(-\frac{nI_L}{\beta K}(1 + o(1))\right)$. The expectation bound in the theorem then follows from the fact that $I_L = I(1 + o(1))$ again.

F.3. Transformation Analysis.

PROPOSITION F.1. *Let $p(x)$ and $q(x)$ be densities over S, where $S = \mathbb{R}$ or $S = \mathbb{R}^+$, and let $\Phi : S \to [0, 1]$ be a CDF such that $\phi = \Phi'$ is positive and continuous. Suppose Assumptions A1'–A5' hold. The following conditions are satisfied for $p_\Phi(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ and $q_\Phi(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$:*

C1' $p_\Phi(z), q_\Phi(z) > 0$ *on* $(0,1)$, *and* $\sup_n \sup_z \{p_\Phi(z) \vee q_\Phi(z)\} < \infty$.

C2' There exists a subinterval $R_\Phi \subseteq [0,1]$ *such that*

- *(a) for all* $z \in R_\Phi$, $\frac{1}{\rho} \leq \left|\frac{p_\Phi(z)}{q_\Phi(z)}\right| \leq \rho$, *where* ρ *is an absolute constant, and*
- *(b)* $\mu\{R_\Phi^c\} = o(H)$, *where* μ *is the Lebesgue measure and* $R_\Phi^c = [0,1]\backslash R$.

C3' Let $\alpha^2 = \int_R \frac{(p_\Phi(z)-q_\Phi(z))^2}{q_\Phi(z)} dz$ *and* $\gamma_\Phi(z) = \frac{q_\Phi(z)-p_\Phi(z)}{\alpha}$. *Then* $\sup_n \int_R q_\Phi(z) \left|\frac{\gamma(z)}{q_\Phi(z)}\right|^r dz < \infty$, *for an absolute constant* $r > 4$.

C4' There exists $h_\Phi(z)$ *such that*

- *(a)* $h_\Phi(z) \geq \max\left\{\left|\frac{\gamma'_\Phi(z)}{q_\Phi(z)}\right|, \left|\frac{q'_\Phi(z)}{q_\Phi(z)}\right|\right\}$,
- *(b)* $h_\Phi(z)$ *is* (c'_{s1}, c'_{s2}, C'_s)*-bowl-shaped, for absolute constants* c'_{s1}, c'_{s2}, *and* C'_s, *and*
- *(c)* $\sup_n \int_R |h_\Phi(z)|^t dz < \infty$ *for an absolute constant* $\frac{2}{r} < t < 1$.

C5' $p'_\Phi(z), q'_\Phi(z) \geq 0$ *and for all* $z < c'_{s1}$, *and* $p'_\Phi(z), q'_\Phi(z) \leq 0$ *for all* $z > c'_{s2}$.

PROOF. **C1'** follows from A1' and the condition that ϕ is positive and continuous.

To prove **C2'**, assume that A2' is true, and let R be a subinterval of S such that $\frac{1}{\rho} \leq \frac{p(x)}{q(x)} \leq \rho$. Define $R_\Phi = \left\{z \in [0,1] : \Phi^{-1}(z) \in R\right\}$, so $\mathbf{1}_{R_\Phi}(z) = \mathbf{1}_R(\Phi^{-1}(z))$. Then R_Φ is clearly an interval, and $\Phi\{R^c\} = \mu\{R_\Phi^c\}$.

C3' follows from A3' via a change of variables.

It remains to prove **C4'** and **C5'**. We first prove **C5'**. Note that

$$p'_\Phi(z) = \frac{p'(\Phi^{-1}(z)) - p(\Phi^{-1}(z))\frac{\phi'(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}}{\phi(\Phi^{-1}(z))^2}.$$

Therefore, $p'_\Phi(z) \geq 0$ if and only if $p'(x) \geq p(x)\frac{\phi'(x)}{\phi(x)}$, and likewise for $q'_\Phi(z)$.

Moving onto **C4'**, we first construct $h(z)$. For ease of presentation, let $x = \Phi^{-1}(z)$. We then have

$$\frac{q'_\Phi(z)}{q_\Phi(z)} = \frac{q'(x)}{q(x)}\frac{1}{\phi(x)} - \frac{\phi'(x)}{\phi(x)}\frac{1}{\phi(x)},$$

implying that

$$\left|\frac{q'_\Phi(z)}{q_\Phi(z)}\right| \leq \left|\frac{q'(x)}{q(x)}\frac{1}{\phi(x)}\right| + \left|\frac{\phi'(x)}{\phi(x)}\frac{1}{\phi(x)}\right| \lesssim (h(x)+1)\frac{1}{\phi(x)},$$

where the last inequality follows because $\left|\frac{\phi'(x)}{\phi(x)}\right|$ is bounded. Furthermore,

$$\begin{aligned}\frac{\gamma'_\Phi(z)}{q_\Phi(z)} &= \frac{1}{\alpha}\frac{p'(x) - p(x)\frac{\phi'(x)}{\phi(x)} - q'(x) + q(x)\frac{\phi'(x)}{\phi(x)}}{q(x)\phi(x)} \\ &= \left(\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)} - \frac{1}{\alpha}\frac{p(x) - q(x)}{q(x)}\frac{\phi'(x)}{\phi(x)}\right)\frac{1}{\phi(x)},\end{aligned}$$

so

$$\begin{aligned}\left|\frac{\gamma'_\Phi(z)}{q_\Phi(z)}\right| &\le \left|\frac{1}{\alpha}\frac{p'(x) - q'(x)}{q(x)}\right|\frac{1}{\phi(x)} + \left|\frac{1}{\alpha}\frac{p(x) - q(x)}{q(x)}\right|\left|\frac{\phi'(x)}{\phi(x)}\right|\frac{1}{\phi(x)} \\ &= \left|\frac{\gamma'(x)}{q(x)}\right|\frac{1}{\phi(x)} + \left|\frac{\gamma(x)}{q(x)}\right|\left|\frac{\phi'(x)}{\phi(x)}\right|\frac{1}{\phi(x)} \overset{(a)}{\lesssim} h(x)\frac{1}{\phi(x)},\end{aligned}$$

where (a) follows because $\left|\frac{\phi'(x)}{\phi(x)}\right|$ is bounded. We want to take $h_\Phi(z) \simeq (h(x)+1)\frac{1}{\phi(x)}$, but we use a modified upper bound to ensure that $h_\Phi(z)$ is bowl-shaped. Let $\psi(x) = \max\left\{\frac{1}{\phi(c_{s1})}, \frac{1}{\phi(c_{s2})}, \frac{1}{\phi(x)}\right\}$. We then take

$$h_\Phi(z) \simeq h(x)\psi(x) = h(\Phi^{-1}(z))\psi(\Phi^{-1}(z)).$$

Note that $(h(x)+1)$ is (c_{s1}, c_{s2}, C_s+1)-bowl-shaped, and ϕ is unimodal, so $\frac{1}{\phi(x)}$ is quasi-convex. Hence, $\psi(x)$ is quasi-convex and has a mode lying in $[c_{s1}, c_{s2}]$. Therefore, $(h(x)+1)\psi(x)$ is (c_{s1}, c_{s2}, C'_s)-bowl-shaped, where $C'_s \simeq (C_s+1)\left(\frac{1}{\phi(c_{s1})} \vee \frac{1}{\phi(c_{s2})}\right)$. This shows that $h_\Phi(z)$ is (c'_{s1}, c'_{s2}, C'_s)-bowl-shaped for $c'_{s1} = \Phi(c_{s1})$ and $c'_{s2} = \Phi(c_{s2})$.

Finally, we need to verify the integrability conditions:

$$\begin{aligned}\int_0^1 |h_\Phi(z)|^t dz &\simeq \int_0^1 (h(\Phi^{-1}(z))+1)^t \psi(\Phi^{-1}(z))^t dz \\ &\overset{(a)}{=} \int_S (h(x)+1)^t \psi(x)^t \phi(x) dx \\ &\le \left\{\int_S (h(x)+1)^{2t}\phi(x)dx\right\}^{1/2}\left\{\int_S \psi(x)^{2t}\phi(x)dx\right\}^{1/2},\end{aligned}$$

where (a) follows from a change of variables. To bound the first term, note that

$$\begin{aligned}\int_S (h(x)+1)^{2t}\phi(x)dx &\le \int_S h(x)^{2t}\phi(x)dx + \int_S \phi(x)dx \\ &\le \int_S h(x)^{2t}\phi(x)dx + 1.\end{aligned}$$

The first inequality follows since $2t < 1$. Note that $\int_S h(x)^{2t}\phi(x)dx < \infty$.

We now bound the second term:

$$\int_S \psi(x)^{2t}\phi(x)dx \leq \int_S \phi(c_{s1})^{-2t}\phi(x)dx + \int_S \phi(c_{s2})^{-2t}\phi(x)dx + \int_S \phi(x)^{-2t}\phi(x)dx.$$

The first two terms are constants. The last term is $\int_S \phi(x)^{1-2t}dx$, which is finite because $1 - 2t > 0$ and ϕ is a valid transformation function. □

PROPOSITION F.2. *Suppose Assumptions A1–A4 hold. The following conditions are satisfied for* $p_\Phi(z) = \frac{p(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$ *and* $q_\Phi(z) = \frac{q(\Phi^{-1}(z))}{\phi(\Phi^{-1}(z))}$*:*

C1 $p_\Phi(z), q_\Phi(z) > 0$ *on* $(0,1)$*, and* $\sup_n \sup_z \{p_\Phi(z) \vee q_\Phi(z)\} < \infty$.

C2 For an absolute constant $r > 4$, $\sup_n \int_0^1 \left|\log \frac{p_\Phi(z)}{q_\Phi(z)}\right|^r dz < \infty$.

C3 There exists $h_\Phi(z)$ *such that*

- *(a)* $h_\Phi(z) \geq \max\left\{\left|\frac{p'_\Phi(z)}{p_\Phi(z)}\right|, \left|\frac{q'_\Phi(z)}{q_\Phi(z)}\right|\right\}$,
- *(b)* $h_\Phi(z)$ *is* (c'_{s1}, c'_{s2}, C'_s)*-bowl-shaped, and*
- *(c)* $\sup_n \int_R |h_{\Phi,n}(z)|^t dz < \infty$*, for some constant* t *such that* $\frac{2}{r} \leq t \leq 1$.

C4 We have that $p'_\Phi(z), q'_\Phi(z) \geq 0$ *for all* $z < c'_{s1}$*, and* $p'_\Phi(z), q'_\Phi(z) \leq 0$ *for all* $z > c'_{s2}$.

PROOF. The proof is identical to that of Proposition F.1, so we omit the details. □

APPENDIX G: PROOF OF PROPOSITION 5.1

First suppose $\|\theta_1 - \theta_0\| \to 0$. In Lemma G.2, we show that $\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx \to 0$. Assumptions **A1'** and **A5'** follow directly from Assumptions B1 and B4, respectively.

We now prove **A2'**. Let ρ be a constant and define

$$R = \left\{x : g_1(x) \leq \frac{\log \rho}{\|\theta_1 - \theta_0\|}\right\}, \tag{20}$$

where $g_1(x)$ is the upper bound on $\sup_{\theta\in\Theta} \|\nabla_\theta f_\theta(x)\|$ defined in Assumption B3. Since $g_1(x)$ is bowl-shaped, we conclude that R is an interval if $\log \rho \geq$

$C_s\text{diam}(\Theta)$. Note that $\log\frac{p(x)}{q(x)} = f_{\theta_1}(x) - f_{\theta_0}(x) = (\theta_1-\theta_0)^{\mathrm{T}}\nabla_\theta f_{\bar\theta}(x)$. This implies

$$\begin{aligned}\left|\log\frac{p(x)}{q(x)}\right| &\le \|\theta_1-\theta_0\|\|\nabla_\theta f_{\bar\theta}(x)\| \le \|\theta_1-\theta_0\|\sup_\theta\|\nabla_\theta f_\theta(x)\| \\ &\le \|\theta_1-\theta_0\|\cdot g_1(x),\end{aligned}$$

where $\bar\theta$ is a convex combination of θ_0,θ_1. Therefore, for all $x\in R$, we have $\frac{1}{\rho}\le\frac{p(x)}{q(x)}\le\rho$. Since we know from Assumption B3 that $\int_S |g_1(x)|^r\phi(x)dx < \infty$, Markov's inequality gives

$$\Phi(R^c) = \Phi\left\{x\,:\,g_1(x) > \frac{\log\rho}{\|\theta_1-\theta_0\|}\right\} \le C\frac{\|\theta_1-\theta_0\|^r}{(\log\rho)^r} \overset{(a)}{=} \Theta(H^{r/2}) = o(H),$$

where (a) follows from Lemma G.2. The last equality follows from the assumption that $r>2$. This proves A2'.

Now we move on to **A3'**. By Lemma G.1, we have

$$\begin{aligned}\frac{1}{\alpha}\left|\frac{p(x)-q(x)}{q(x)}\right| &\lesssim \frac{1}{\|\theta_1-\theta_0\|}\left|\frac{p(x)}{q(x)}-1\right| \\ &= \frac{1}{\|\theta_1-\theta_0\|}\left|\exp(f_{\theta_1}(x)-f_{\theta_0}(x))-1\right| \\ &\overset{(a)}{=} \frac{1}{\|\theta_1-\theta_0\|}\left|(\theta_1-\theta_0)^{\mathrm{T}}\nabla_\theta f_{\bar\theta}(x)\right|\exp(f_{\bar\theta}(x)-f_{\theta_0}(x)) \\ &\le \|\nabla_\theta f_{\bar\theta}(x)\|\exp(f_{\bar\theta}(x)-f_{\theta_0}(x)) \\ &\overset{(b)}{=} \|\nabla_\theta f_{\bar\theta}(x)\|\exp\Big((\bar\theta-\theta_1)^{\mathrm{T}}\nabla_\theta f_{\tilde\theta}(x)\Big) \\ &\le \|\nabla_\theta f_{\bar\theta}(x)\|\exp\Big(\|\theta_1-\theta_0\|\|\nabla_\theta f_{\tilde\theta}(x)\|\Big),\end{aligned}$$

where in (a), $\bar\theta$ is a convex combination of θ_1 and θ_0, and in (b), $\tilde\theta$ is a convex combination of $\bar\theta$ and θ_0. Assumption B3 implies that both $\|\nabla_\theta f_{\bar\theta}(x)\|$ and $\|\nabla_\theta f_{\tilde\theta}(x)\|$ are upper-bounded by $g_1(x)$, so

$$\frac{1}{\alpha}\left|\frac{p(x)-q(x)}{q(x)}\right| \lesssim g_1(x)\exp\big(\|\theta_0-\theta_1\|g_1(x)\big). \tag{21}$$

Therefore,

$$\begin{aligned}\int_R\left(\frac{1}{\alpha}\left|\frac{p(x)}{q(x)}-1\right|\right)^r q(x)dx &\lesssim \int_R g_1(x)^r\exp\big(r\|\theta_0-\theta_1\|g_1(x)\big)q(x)dx \\ &\overset{(a)}{\le} \int_R g_1(x)^r\rho^r q(x)dx \overset{(b)}{\lesssim} \int_R g_1(x)^r\phi(x)dx,\end{aligned}$$

where (a) follows from the definition of R and (b) follows because $\frac{q(x)}{\phi(x)}$ is bounded. This proves A3'.

To prove **A4'**, we first construct $h(x)$. By equation 21, we have

$$\left|\frac{\gamma(x)}{q(x)}\right| \lesssim g_1(x)\exp\Big(\|\theta_0-\theta_1\|g_1(x)\Big).$$

By Assumption B3, we also have $\left|\frac{q'(x)}{q(x)}\right| = |f'_{\theta_0}(x)| \le g_{2,\theta_0}(x)$, and

$$\begin{aligned}
\left|\frac{\gamma'(x)}{q(x)}\right| &= \left|\frac{1}{\alpha}\frac{p'(x)-q'(x)}{q(x)}\right| \lesssim \frac{1}{\|\theta_0-\theta_1\|}\left|\frac{p'(x)-q'(x)}{q(x)}\right| \\
&= \frac{1}{\|\theta_0-\theta_1\|}\left|f'_{\theta_1}\frac{p(x)}{q(x)} - f'_{\theta_0}(x)\right| \\
&= \frac{1}{\|\theta_0-\theta_1\|}\left|(f'_{\theta_1}(x)-f'_{\theta_0}(x))\frac{p(x)}{q(x)} + f'_{\theta_0}\left(\frac{p(x)}{q(x)}-1\right)\right| \\
&\le \|\nabla_\theta f'_{\bar\theta}(x)\|\frac{p(x)}{q(x)} + \frac{1}{\|\theta_1-\theta_0\|}\left|\frac{p(x)}{q(x)}-1\right||f'_{\theta_0}(x)|,
\end{aligned}$$

where $\bar\theta$ is a convex combination of θ_0 and θ_1.

Using Assumption B3 and inequality (21), we have

$$\left|\frac{\gamma'(x)}{q(x)}\right| \lesssim g_{2,\bar\theta}(x)\exp\Big(\|\theta_0-\theta_1\|g_1(x)\Big) + g_1(x)\exp\Big(\|\theta_0-\theta_1\|g_1(x)\Big)g_{2,\theta_0}(x).$$

Hence, we may choose choose

$$\begin{aligned}
h(x) \simeq\ & g_{2,\bar\theta}(x)\exp\Big(\|\theta_0-\theta_1\|g_1(x)\Big) + g_1(x)\exp\Big(\|\theta_0-\theta_1\|g_1(x)\Big)g_{2,\theta_0}(x) \\
&+ g_1(x)\exp\Big(\|\theta_0-\theta_1\|g_1(x)\Big).
\end{aligned}$$

Since all the component functions are $(c_{s1}, c_{s2}, \tilde C_s)$ bowl-shaped, $h(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped, where $C_s = 3\tilde C_s^2\exp(\mathrm{diam}(\Theta)\tilde C_s)$. Furthermore,

$$\begin{aligned}
\int_R h(x)^{2t}\phi(x)dx &\overset{(a)}{\lesssim} \int_R \Big(g_{2,\bar\theta}^{2t}\rho^{2t} + g_1(x)^{2t}\rho^{2t}g_{2,\theta_0}^{2t} + g_1(x)^{2t}\rho^{2t}\Big)\phi(x)dx \\
&\lesssim \int_R g_{2,\bar\theta}(x)^{2t}\phi(x)dx + \int_R g_1(x)^{2t}g_{2,\theta_0}^{2t}\phi(x)dx + \int_R g_1(x)^{2t}\phi(x)dx,
\end{aligned}$$

where (a) follows because on R, we have $\|\theta_0-\theta_1\|g_1(x) \le \log\rho$. By Assumption B3, the first and third terms are finite, uniformly over all $\theta_1, \theta_0 \in \Theta$. It is straightforward to show that the second term is also finite, by an application of the Cauchy-Schwartz inequality.

Now suppose $\|\theta_1 - \theta_0\| = \Theta(1)$. By Lemma G.2, we have $\int_S (\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \Theta(1)$. Assumptions **A1** and **A4** follow from Assumptions B1 and B5.

To prove **A2**, note that from a previous derivation, we have

$$\left|\log \frac{p(x)}{q(x)}\right| \le \|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\| \le \|\theta_1 - \theta_0\| g_1(x).$$

Since $\|\theta_1 - \theta_0\| \le \operatorname{diam}(\Theta)$ and $\int_S g_1(x)^r \phi(x) dx < \infty$ by Assumption B3, we obtain A2.

To prove **A3**, note that $\frac{q'(x)}{q(x)} = f'_{\theta_0}(x)$ and $\frac{p'(x)}{p(x)} = f'_{\theta_1}(x)$. Therefore, $h(x) = g_{2,\theta_0}(x) + g_{2,\theta_1}(x)$ upper-bounds $\left|\frac{q'(x)}{q(x)}\right|$ and $\left|\frac{p'(x)}{p(x)}\right|$, by Assumption B3. Furthermore, $h(x)$ is (c_{s1}, c_{s2}, C_s)-bowl-shaped, where $C_s = 2\tilde{C}_s$.

To prove the last integrability condition, note that

$$\int_S h(x)^{2t}\phi(x)dx \le \int_S g_{2,\theta_0}(x)^{2t}\phi(x)dx + \int_S g_{2,\theta_1}(x)^{2t}\phi(x)dx.$$

Hence,

$$\int_S h(x)^{2t}\phi(x)dx \le 2\sup_{\theta\in\Theta}\int_S g_{2,\theta}(x)^{2t}\phi(x)dx.$$

G.1. Supporting lemmas.

LEMMA G.1. *Under Assumptions B1–B5, we have* $\alpha \asymp \|\theta_1 - \theta_0\|$.

PROOF. We write

$$\begin{aligned}\alpha^2 &= \int_R \left(\frac{p(x)}{q(x)} - 1\right)^2 q(x)dx = \int_R \left|\exp\left(f_{\theta_1}(x) - f_{\theta_0}(x)\right) - 1\right|^2 q(x)dx \\ &= \int_R \left((\theta_1 - \theta_0)^{\mathrm{T}}\nabla_\theta f_{\bar\theta}(x)\exp\left(f_{\bar\theta}(x) - f_{\theta_0}(x)\right)\right)^2 q(x)dx.\end{aligned}$$

First we show an upper bound:

$$\begin{aligned}\alpha^2 &\le \int_R \|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\bar\theta}(x)\|^2 \exp\left(f_{\bar\theta}(x) - f_{\theta_0}(x)\right)\exp(f_{\bar\theta}(x))dx \\ &\le \int_R \|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\bar\theta}(x)\|^2 \exp\left(\|\theta_1 - \theta_0\|\|\nabla_\theta f_{\tilde\theta}(x)\|\right)\exp(f_{\bar\theta}(x))dx.\end{aligned}$$

On R, we have $\|\theta_1 - \theta_0\| \sup_\theta \|\nabla_\theta f_\theta(x)\| \le \log\rho$. Hence,

$$\begin{aligned}\alpha^2 &\le \|\theta_1 - \theta_0\|^2 \int_R \|\nabla_\theta f_{\bar\theta}(x)\|^2 e^{\log\rho}\exp(f_{\bar\theta}(x))dx \\ &\le \|\theta_1 - \theta_0\|^2 \rho \int_{-\infty}^{\infty} \|\nabla_\theta f_{\bar\theta}(x)\|^2 \exp(f_{\bar\theta}(x))dx \overset{(a)}{\lesssim} \|\theta_1 - \theta_0\|^2,\end{aligned}$$

where (a) follows from Assumptions B1 and B4. We now establish a lower bound:

$$\begin{aligned}\alpha^2 &\geq \int_R \left((\theta_1-\theta_0)^{\mathrm{T}}\nabla_\theta f_{\bar\theta}(x)\right)^2 \exp\Big(-|f_{\bar\theta}(x)-f_{\theta_0}(x)|\Big)\exp(f_{\bar\theta}(x))dx \\ &\geq \int_R \left((\theta_1-\theta_0)^{\mathrm{T}}\nabla_\theta f_{\bar\theta}(x)\right)^2 \exp\Big(-\|\theta_1-\theta_0\|\|\nabla_\theta f_{\bar\theta}(x)\|\Big)\exp(f_{\bar\theta}(x))dx \\ &\overset{(a)}{\geq} \frac{1}{\rho}(\theta_1-\theta_0)^{\mathrm{T}}\left(\int_R (\nabla_\theta f_{\bar\theta}(x))(\nabla_\theta f_{\bar\theta}(x))^{\mathrm{T}}\exp(f_{\bar\theta}(x))dx\right)(\theta_1-\theta_0),\end{aligned}$$

where (a) follows from Assumption B3. Define

$$\tilde G_{\bar\theta} = \int_R (\nabla_\theta f_{\bar\theta}(x))(\nabla_\theta f_{\bar\theta}(x))^{\mathrm{T}}\exp(f_{\bar\theta}(x))dx.$$

As ρ increases, $R \to S$. Therefore, there exists an absolute constant such that for all ρ greater than or equal to this constant, we have $\lambda_{min}(\tilde G_{\bar\theta}) > \frac{1}{2}\lambda_{min}(G_{\bar\theta}) > 0$. Hence, $\alpha^2 \gtrsim \|\theta_1-\theta_0\|^2$. □

LEMMA G.2. *The Hellinger distance satisfies the bound*

$$\int_S (\sqrt{p(x)}-\sqrt{q(x)})^2 dx = c\|\theta_0-\theta_1\|_2^2,$$

where $c_{\min} \leq c \leq \frac{1}{4}c_{\max}d_\Theta$.

PROOF. Expanding the left-hand side, we have

$$\begin{aligned}\int_S (\sqrt{p(x)}-\sqrt{q(x)})^2 dx &= \int_S q(x)\left(\sqrt{\frac{p(x)}{q(x)}}-1\right)^2 dx \\ &= \int_S q(x)\left(\exp\left(\frac{f_{\theta_1}(x)}{2}-\frac{f_{\theta_0}(x)}{2}\right)-1\right)^2 dx.\end{aligned}$$

Let $h(\theta) = \exp\left(\frac{f_\theta(x)}{2}-\frac{f_{\theta_0}(x)}{2}\right)$. It is clear that $h(\theta_0)=1$ and that we wish to bound $h(\theta_1)-h(\theta_0)$. We bound this as follows:

$$\begin{aligned}|h(\theta_1)-h(\theta_0)| &= |(\theta_1-\theta_0)^{\mathrm{T}}\nabla_\theta h(\bar\theta)| \\ &= \left|\frac{1}{2}(\theta_1-\theta_0)^{\mathrm{T}}\nabla_\theta f_{\bar\theta}(x)\exp\left(\frac{f_{\bar\theta}(x)}{2}-\frac{f_{\theta_0}(x)}{2}\right)\right| \\ &\leq \frac{1}{2}\|\theta_1-\theta_0\|\|\nabla_\theta f_{\bar\theta}(x)\|\exp\left(\frac{f_{\bar\theta}(x)}{2}-\frac{f_{\theta_0}(x)}{2}\right),\end{aligned}$$

where $\bar{\theta} \in \Theta$ is some convex combination of θ_1, θ_0. Thus, we have

$$\begin{aligned}
&\int_S q(x) \left(\exp\left(\frac{f_{\theta_1}(x)}{2} - \frac{f_{\theta_0}(x)}{2} \right) - 1 \right)^2 dx \\
&\quad \le \int_S q(x) \frac{1}{4} \|\theta_1 - \theta_0\|^2 \|\nabla_\theta f_{\bar{\theta}}(x)\|^2 \exp(f_{\bar{\theta}}(x) - f_{\theta_0}(x)) dx \\
&\quad = \frac{1}{4} \|\theta_1 - \theta_0\|^2 \int_S \|\nabla_\theta f_{\bar{\theta}}(x)\|^2 \exp(f_{\bar{\theta}}(x)) dx \\
&\quad \le \frac{1}{4} \|\theta_1 - \theta_0\|^2 \operatorname{tr}(G_{\bar{\theta}}) \\
&\quad \le \frac{1}{4} \|\theta_1 - \theta_0\|^2 c_{\max} d_\Theta,
\end{aligned}$$

where $\Theta \subseteq \mathbb{R}^{d_\Theta}$. Furthermore,

$$\begin{aligned}
&\int_S q(x) \left(\left(\frac{f_{\theta_1}(x)}{2} - \frac{f_{\theta_0}(x)}{2} \right) - 1 \right)^2 dx \\
&\quad = \int_S \left((\theta_1 - \theta_0)^{\mathrm{T}} \nabla_\theta f_{\bar{\theta}}(x) \right)^2 \exp(f_{\bar{\theta}}(x)) dx \\
&\quad = (\theta_1 - \theta_0)^{\mathrm{T}} G_{\bar{\theta}} (\theta_1 - \theta_0) \ge c_{\min} \|\theta_1 - \theta_0\|^2.
\end{aligned}$$

□

G.2. Proof of examples.

PROPOSITION G.1. *Let* $\exp(f(x))$ *be a positive density over* $\mathbb{R}$*, where*

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

Let $\theta = (\mu, \sigma)$ *and* $\Theta = [-C_\mu, C_\mu] \times [\frac{1}{c_\sigma}, c_\sigma]$*, for some absolute constants* C_μ *and* c_σ*, and let* $f_\theta(x) = f\left(\frac{x-\mu}{\sigma}\right) - \log \sigma$*. Then* $\{f_\theta(x)\}_{\theta \in \Theta}$ *satisfies Assumptions B1–B4 with respect to* ϕ *defined in equation* (3)*.*

PROOF. Before we prove the claims, let us derive some useful properties of f. First, for any $x > c$, we have

$$f(x) = \int_0^x f'(t) dt = \int_0^c f'(t) dt + \int_c^x f'(t) dt \lesssim 1 - \int_c^x M dt \lesssim 1 - x.$$

Similarly, for any $x < -c$, we have $f(x) \lesssim 1 + x$. Therefore, $f(x) \lesssim 1 - |x|$.

Likewise, we have

$$f\left(\frac{x-\mu}{\sigma}\right) \lesssim 1 - \left|\frac{x-\mu}{\sigma}\right| \lesssim 1 - \left|\frac{x}{\sigma}\right| + \frac{\mu}{\sigma} \overset{(a)}{\lesssim} 1 - |x|,$$

where (a) follows because $\sigma \geq \frac{1}{c_\sigma}$ and $|\mu| \leq C_\mu$, for some absolute constants c_σ and C_μ. Thus, the density $\exp f\left(\frac{x-\mu}{\sigma}\right)$ is sub-exponential.

Since $f^{(k)}(x)$ is bounded, L'Hopital's rule implies $|f'(x)| \lesssim |x|^{k-1} + 1$ and $|f''(x)| \lesssim |x|^{k-2} + 1$. Furthermore,

$$f'\left(\frac{x-\mu}{\sigma}\right) \lesssim \left|\frac{x-\mu}{\sigma}\right|^{k-1} + 1 \overset{(a)}{\lesssim} \left|\frac{x}{\sigma}\right|^{k-1} + \left|\frac{\mu}{\sigma}\right|^{k-1} + 1 \overset{(b)}{\lesssim} |x|^{k-1} + 1, \tag{22}$$

where (a) follows because k is a constant and (b) follows because $|\mu| \leq C_\mu$ and $\sigma \geq \frac{1}{c_\sigma}$, by assumption.

Now we prove the first claim **B1**. We have

$$\begin{aligned}
\log \phi(x) - f_\theta(x) &= \log \frac{e}{8} - \sqrt{|x|+1} - f\left(\frac{x-\mu}{\sigma}\right) - \log \sigma \\
&\geq -\sqrt{|x|+1} - f\left(\frac{x-\mu}{\sigma}\right) - \log \frac{1}{c_\sigma} + \log \frac{e}{8} \\
&\geq -\sqrt{|x|+1} - C(1-|x|) - \log \frac{1}{c_\sigma} + \log \frac{e}{8} > -\infty.
\end{aligned}$$

Moving on to **B2**, we have

$$\nabla f_\theta(x) = \begin{bmatrix} -\frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \\ -\left(\frac{x-\mu}{\sigma^2}\right) f'\left(\frac{x-\mu}{\sigma}\right) - \frac{1}{\sigma} \end{bmatrix} = -\frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \begin{bmatrix} 1 \\ \frac{x-\mu}{\sigma} + 1 \end{bmatrix}.$$

To show that $\lambda_{\max}(G_\theta) < \infty$, it is sufficient to show that

$$\begin{aligned}
&\int \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)^2 \left(\frac{x-\mu}{\sigma} + 1\right) \exp f\left(\frac{x-\mu}{\sigma}\right) dx < \infty, \quad \text{and} \\
&\int \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)^2 \left(\frac{x-\mu}{\sigma} + 1\right)^2 \exp f\left(\frac{x-\mu}{\sigma}\right) dx < \infty.
\end{aligned}$$

Since $\left|f'\left(\frac{x-\mu}{\sigma}\right)\right| \lesssim \left|\frac{x-\mu}{\sigma}\right|^{k-1} + 1$ and $\exp f\left(\frac{x-\mu}{\sigma}\right)$ is sub-exponential with all moments finite, we conclude that both integrals converge.

To show that $\lambda_{\min}(G_\theta) > 0$, we need to show that $\det(G_\theta) > 0$. Let $g(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)^2 \exp f\left(\frac{x-\mu}{\sigma}\right)$, and note that g is positive and integrable. The integral of g is not 0, since $|f'(x)| \geq M$ for all $|x| > c$. Thus, g may be normalized to a density $\bar{g}$.

Showing that $\det(G_\theta) > 0$ is equivalent to showing that

$$\int g(x)dx \int \left(\frac{x-\mu}{\sigma}+1\right)^2 g(x)dx > \left(\int \left(\frac{x-\mu}{\sigma}+1\right) g(x)dx\right)^2,$$

which is equivalent to showing that

$$\mathbb{E}_{\bar{g}}\left[\left(\frac{X-\mu}{\sigma}+1\right)^2\right] - \left(\mathbb{E}_{\bar{g}}\left[\frac{X-\mu}{\sigma}+1\right]\right)^2 > 0,$$

or $\mathrm{Var}_{\bar{g}}(X) > 0$. This follows because $g(x) \neq 0$.

To verify **B3**, note that

$$\begin{aligned}\|\nabla f_\theta\| &= \left|\frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right)\right| \sqrt{1+((x-\mu)/\sigma+1)^2} \\ &\lesssim (1+|x|^{k-1})(1+|(x-\mu)/\sigma|) \lesssim 1+|x|^k.\end{aligned}$$

Thus, we set $g_1(x) = C(1+|x|^k)$ for some absolute constant C. Note that $g_1(x)$ is clearly bowl-shaped and $\int g_1(x)^r \phi(x)dx$ is finite, since all moments of ϕ are finite. To construct $g_{2,\theta}$, note that

$$f'_\theta(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right), \quad \text{and} \quad \nabla f'_\theta(x) = \begin{bmatrix} -\frac{1}{\sigma^2} f''\left(\frac{x-\mu}{\sigma}\right) \\ -\frac{1}{\sigma^2} f'\left(\frac{x-\mu}{\sigma}\right) - \frac{x-\mu}{\sigma^3} f''\left(\frac{x-\mu}{\sigma}\right) \end{bmatrix}.$$

Therefore, $|f'_\theta(x)| \lesssim 1+|x|^{k-1}$, and

$$\begin{aligned}\|\nabla f'_\theta(x)\| &\leq \frac{1}{\sigma^2}\left|f''\left(\frac{x-\mu}{\sigma}\right)\right| + \frac{1}{\sigma^2}\left|f'\left(\frac{x-\mu}{\sigma}\right)\right| + \frac{1}{\sigma^2}\left|f''\left(\frac{x-\mu}{\sigma}\right)\right|\left|\frac{x-\mu}{\sigma}\right| \\ &\overset{(a)}{\lesssim} (1+|x|^{k-2}) + (1+|x|^{k-1}) + (1+|x|^{k-2})(1+|x|) \\ &\lesssim 1+|x|^{k-1},\end{aligned}$$

where (a) follows because $|f''(x)| \lesssim 1+|x|^{k-2}$. Thus, we may take $g_{2,\theta}(x) = C(1+|x|^{k-1})$. This is clearly bowl-shaped and integrable, as well.

Finally, we prove **B4**. We have

$$(\log\phi)'(x) = \begin{cases} \frac{1}{2}\frac{1}{\sqrt{1-x}}, & \text{if } x < 0 \\ -\frac{1}{2}\frac{1}{\sqrt{1+x}}, & \text{if } x > 0. \end{cases}$$

In particular, $(\log\phi)'(x) \to 0$ as $|x| \to \infty$.

We also know that $f'(x) \geq M$ for all $x \leq -c$, and $f'(x) \leq -M$ for all $x \geq c$. If $x \leq -\frac{c}{c_\sigma} - C_\mu$, then $\frac{x-\mu}{\sigma} \leq -c$ and

$$f'_\theta(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \geq \frac{M}{c_\sigma}.$$

If $x \geq \frac{c}{c_\sigma} + C_\mu$, then $\frac{x-\mu}{\sigma} \geq c$ and

$$f'_\theta(x) = \frac{1}{\sigma} f'\left(\frac{x-\mu}{\sigma}\right) \leq -\frac{M}{c_\sigma}.$$

Thus, there exist $c_{s1} < 0$ and $c_{s2} > 0$ such that B4 holds. □

PROPOSITION G.2. *Let* $\exp(f(x))$ *be a positive density over* $\mathbb{R}^+$, *where*

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

Let $\theta = \sigma$ *and* $\Theta = [\frac{1}{c_\sigma}, c_\sigma]$ *for some absolute constant* c_σ, *and let* $f_\theta(x) = f\left(\frac{x}{\sigma}\right) - \log\sigma$. *Then* $\{f_\theta(x)\}_{\theta\in\Theta}$ *satisfies Assumptions B1–B4 with respect to* ϕ *defined in equation* (3).

The proof is almost identical to that of Proposition G.1.

PROPOSITION G.3. *Let* $\theta = (\alpha, \beta)$ *and* $\Theta = [\frac{1}{c}, c]^2$ *for some constant* c, *and let* $f_\theta = (\alpha-1)\log x - \beta x + \alpha\log\beta - \log\Gamma(\alpha)$. *Then* $\{f_\theta(x)\}_{\theta\in\Theta}$ *satisfies Assumptions B1–B4 with respect to* ϕ *defined in equation* (3).

PROOF. We first prove **B1**. We have $\log\phi(x) = \log\frac{e}{4} - \sqrt{x+1}$, so

$$\begin{aligned}\log\phi(x) - f_\theta(x) &= -\sqrt{x+1} - (\alpha-1)\log x + \beta x + \log\frac{e}{4} - \alpha\log\beta - \log\Gamma(\alpha)\\ &> -\infty.\end{aligned}$$

To prove **B2**, note that $G_\theta = \int (H_\theta f_\theta(x)) \exp f_\theta(x) dx$, where H_θ is the Hessian operator. Hence,

$$\nabla f_\theta(x) = \begin{bmatrix} \log x + \log\beta - d_\alpha \log\Gamma(\alpha) \\ -x + \frac{\alpha}{\beta} \end{bmatrix},$$

implying that

$$H_\theta f_\theta(x) = \begin{bmatrix} d_\alpha^2 \log\Gamma(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & -\frac{\alpha}{\beta^2} \end{bmatrix}.$$

Therefore, $G_\theta = H_\theta f_\theta(x)$ which is clearly full-rank.

To prove **B3**, we write

$$\begin{aligned}\|\nabla f_\theta(x)\| &\le |\log x| + x + |\log \beta| + |d_\alpha \log \Gamma(\alpha)| + \frac{\alpha}{\beta} \\ &\overset{(a)}{\le} |\log x| + x + C,\end{aligned}$$

where (a) follows because $c \ge \beta, \alpha \ge \frac{1}{c}$. Therefore, we take $g_1(x) = |\log x| + x + C$. Then

$$\begin{aligned}\int g_1(x)^r \phi(x) dx &= \int_0^\infty (|\log x| + x + C)^r \phi(x) dx \\ &\lesssim \int_0^\infty |\log x|^r \phi(x) + \int_0^\infty x^r \phi(x) dx.\end{aligned}$$

Observe that both terms are finite for our choice of ϕ. Furthermore,

$$f'_\theta(x) = \frac{(\alpha - 1)}{x} - \beta, \quad \text{and} \quad \nabla f'_\theta(x) = \begin{bmatrix} \frac{1}{x} \\ -1 \end{bmatrix},$$

implying that $|f'_\theta(x)| \lesssim 1 + x^{-1}$ and $\|\nabla f'_\theta(x)\| \lesssim 1 + x^{-1}$. Thus, $g_{2,\theta} = C(1 + x^{-1})$ satisfies

$$\int_0^\infty g_{2,\theta}(x)^{4t} \phi(x) dx \lesssim 1 + \int_0^\infty x^{-4t} \phi(x) dx.$$

Since $4t < 1$ by assumption, the integral converges.

B4 readily follows because $\beta \ge \frac{1}{c} > 0$. □

APPENDIX H: APPENDIX FOR THEOREM 5.3

We begin by defining some notation. Let $\hat{\sigma}$ denote a clustering algorithm and A denote a weighted network such that $\hat{\sigma}(A) : [n] \to [K]$ is the clustering obtained by $\hat{\sigma}$ based on the input A. Let

$$S_K[\hat{\sigma}(A), \sigma_0] := \underset{\rho \in S_K}{\arg\min}\, d_H(\rho \circ \hat{\sigma}(A), \sigma_0),$$

where $d_H(\cdot, \cdot)$ denotes the Hamming distance, and define

(23)
$$\mathcal{E}[\hat{\sigma}(A), \sigma_0] := \Big\{ v : (\rho \circ \hat{\sigma}(A))(v) \ne \sigma_0(v), \text{ for some } \rho \in S_K[\hat{\sigma}(A), \sigma_0] \Big\}.$$

When $S_K[\hat{\sigma}(A), \sigma_0]$ is a singleton, the set $\mathcal{E}[\hat{\sigma}(A), \sigma_0]$ contains all nodes misclustered by $\hat{\sigma}(A)$ in relation to σ_0. When $S_K[\hat{\sigma}(A), \sigma_0]$ contains multiple elements, we continue to call $\mathcal{E}[\hat{\sigma}(A), \sigma_0]$ the set of *misclustered* nodes.

H.1. Proof of Theorem 5.3. Throughout this proof, let C denote a $\Theta(1)$ sequence whose value may change from instance to instance. Let

$$\tilde{l}(\hat{\sigma}(A), \sigma_0) = \frac{1}{n} \sum_{v=1}^{n} \mathbf{1}\{v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]\},$$

where $\mathcal{E}[\hat{\sigma}(A), \sigma_0]$ is defined in equation (23). In particular, note that if $|S_K[\hat{\sigma}(A), \sigma_0]| = 1$, we have $\tilde{l} = l$. We have the following claims:

Claim 1: If $nI_n \to \infty$, then $\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0) \geq C \exp\left(-(1 + o(1))\frac{nI_n}{\beta K}\right)$.

Claim 2: If $nI_n \leq c < \infty$, then $\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0) \geq c' > 0$, for some constants c and c'.

We first prove the theorem from the claims. If $P\left(l(\hat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) \geq \frac{1}{2}\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0)$, we have

$$\mathbb{E}l(\hat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K} P\left(l(\hat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) \geq \frac{1}{4\beta K}\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0).$$

On the other hand, if $P\left(l(\hat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) < \frac{1}{2}\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0)$, we have

$$\begin{aligned}
\mathbb{E}l(\hat{\sigma}(A), \sigma_0) &\geq \mathbb{E}\left[l(\hat{\sigma}(A), \sigma_0) \,\middle|\, l(\hat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right] P\left(l(\hat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right) \\
&\overset{(a)}{=} \mathbb{E}\left[\tilde{l}(\hat{\sigma}(A), \sigma_0) \,\middle|\, l(\hat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right] P\left(l(\hat{\sigma}(A), \sigma_0) < \frac{1}{2\beta K}\right) \\
&= \mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0) - \mathbb{E}\left[\tilde{l}(\hat{\sigma}(A), \sigma_0) \,\middle|\, l(\hat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right] \\
&\qquad \cdot P\left(l(\hat{\sigma}(A), \sigma_0) \geq \frac{1}{2\beta K}\right) \\
&\geq \mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0) - \frac{1}{2}\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0) = \frac{1}{2}\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0),
\end{aligned}$$

where (a) holds by invoking Lemma B.8. Thus, any lower bound on $\mathbb{E}\tilde{l}(\hat{\sigma}(A), \sigma_0)$ translates into a lower bound on $\mathbb{E}l(\hat{\sigma}(A), \sigma_0)$ scaled by a suitable constant, implying the desired result.

We now focus on proving the claims. Without loss of generality, suppose clusters 1 has size $\frac{n}{\beta K} + 1$ and cluster 2 has size $\frac{n}{\beta K}$. Let $C_i = \{u : \sigma_0(u) = i\}$ denote the i^{th} cluster. We also suppose without loss of generality that $C_1 = \{1, 2, ..., \frac{n}{\beta K} + 1\}$ and $C_2 = \{\frac{n}{\beta K} + 2, ..., 2\frac{n}{\beta K} + 1\}$.

Let $\sigma_0^1 := \sigma_0$, and let $\sigma_0^2 : [n] \to [K]$ be the cluster assignment satisfying $\sigma_0^2(v) = \sigma_0(v)$ for all $v \neq 1$, and $\sigma_0^2(1) = 2$. Let σ^* be a random cluster assignment, where

$$\sigma^* = \begin{cases} \sigma_0^1, & \text{with probability } \frac{1}{2}, \\ \sigma_0^2, & \text{with probability } \frac{1}{2}. \end{cases}$$

$\sigma^*(u) = \sigma_0(u)$ for all $u \neq 1$ but $\sigma^*(1)$ is random: it is either 1 or 2 each with $\frac{1}{2}$ probability.

Let Φ denote the probability measure on $(\sigma^*, A, \hat{\sigma}(A))$, defined by

$$P_\Phi(\sigma^*, A, \hat{\sigma}(A)) = P(\sigma^*) P_{SBM}(A \,|\, \sigma^*) P_{alg}(\hat{\sigma}(A) \,|\, A),$$

where $P_{SBM}(A \,|\, \sigma^*)$ is the measure on the weighted graph defined by the weighted SBM treating σ^* as the true cluster assignment, and P_{alg} represents any randomness in the clustering algorithm. Let Ψ denote an alternative probability measure defined by

$$P_\Psi(\sigma^*, A, \hat{\sigma}(A)) = P(\sigma^*) P_\Psi(A \,|\, \sigma^*) P_{alg}(\hat{\sigma}(A) \,|\, A),$$

where $P_\Psi(A \,|\, \sigma^*)$ is defined as follows:

1. If $u, v \neq 1$, then A_{uv} is distributed just as in $P_{SBM}(A \,|\, \sigma^*)$.
2. If $v = 1$ and $u \notin C_1 \cup C_2$, then A_{uv} is distributed just as in $P_{SBM}(A \,|\, \sigma^*)$.
3. If $v = 1$ and $u \in C_1 \cup C_2$, then A_{uv} is distributed as Y^*, where Y^* is the distribution that minimizes D in Lemma H.2; i.e., $Y_0^* \propto (P_0 Q_0)^{1/2}$ and $(1 - Y_0^*) y^*(x) \propto \sqrt{(1 - P_0) p(x) (1 - Q_0) q(x)}$.

Note that $P_\Psi(A \,|\, \sigma^*) = P_\Psi(A)$ does not depend on whether $\sigma^* = \sigma_0^1$ or σ_0^2.

Furthermore, we have

$$\begin{aligned} \mathcal{Q} := \log \frac{dP_\Psi}{dP_\Phi} &= \log \frac{dP_{SBM}(A \,|\, \sigma^*)}{dP_\Psi(A \,|\, \sigma^*)} \\ &= \sum_{u \in C_{\sigma^*(1)}} \log \frac{Y(A_{u,1})}{P(A_{u,1})} + \sum_{u \in C_1 \cup C_1 \setminus C_{\sigma^*(1)}} \log \frac{Y(A_{u,1})}{Q(A_{u,1})}, \end{aligned}$$

where we use the notation $P(A_{u,1}) = P_0$ if $A_{u,1} = 0$ and $P(A_{u,1}) = (1 - P_0) p(A_{u,1})$ if $A_{u,1} \neq 0$, and similarly for Y. Let

$$E = \left\{ 1 \notin \mathcal{E}[\hat{\sigma}(A), \sigma^*] \text{ and } \tilde{l}(\hat{\sigma}(A), \sigma^*) \leq \frac{1}{4\beta K} \right\}.$$

For an arbitrary function $f(n)$ to be defined later, we may write

$$P_\Psi(\mathcal{Q} \leq f(n)) = P_\Psi(\mathcal{Q} \leq f(n), \neg E) + P_\Psi(\mathcal{Q} \leq f(n), E). \tag{24}$$

We bound the first term as follows:

$$\begin{aligned}P_\Psi(\mathcal{Q} \le f(n), \neg E) &= \int_{\mathcal{Q}\le f(n),\neg E} dP_\Psi = \int_{\mathcal{Q}\le f(n),\neg E} \exp(\mathcal{Q}) dP_\Phi \\ &\le \exp(f(n)) P_\Phi(\mathcal{Q} \le f(n), \neg E) \le \exp(f(n)) P_\Phi(\neg E) \\ &\le \exp(f(n)) \left(P_\Phi(1 \in \mathcal{E}[\hat\sigma(A), \sigma^*]) + P_\Phi\left(\tilde l(\hat\sigma(A), \sigma^*) \ge \frac{1}{4\beta K}\right)\right).\end{aligned}$$

Furthermore,

$$\begin{aligned}\mathbb{E}_\Phi \tilde l(\hat\sigma(A), \sigma^*) &= \frac{1}{n}\sum_{v=1}^n P_\Phi(v \in \mathcal{E}[\hat\sigma(A), \sigma^*]) \ge \frac{1}{n} \sum_{v\in C_{\sigma^*(1)}} P_\Phi(v \in \mathcal{E}[\hat\sigma(A), \sigma^*]) \\ &\overset{(a)}{=} \frac{|C_{\sigma^*(1)}|}{n} P_\Phi(1 \in \mathcal{E}[\hat\sigma(A), \sigma^*]) \ge \frac{1}{\beta K} P_\Phi(1 \in \mathcal{E}[\hat\sigma(A), \sigma^*]),\end{aligned}$$

where (a) follows from Corollary H.1, and

$$\begin{aligned}\mathbb{E}_\Phi \tilde l(\hat\sigma(A), \sigma^*) &\ge \mathbb{E}_\Phi\left[\tilde l(\hat\sigma(A), \sigma^*) \,\middle|\, \tilde l(\hat\sigma(A), \sigma^*) \ge \frac{1}{4\beta K}\right] \\ &\qquad \cdot P_\Phi\left(\tilde l(\hat\sigma(A), \sigma^*) \ge \frac{1}{4\beta K}\right) \\ &\ge \frac{1}{4\beta K} P_\Phi\left(\tilde l(\hat\sigma(A), \sigma^*) \ge \frac{1}{4\beta K}\right),\end{aligned}$$

so we have the bound

$$P_\Psi(\mathcal{Q} \le f(n), \neg E) \le \exp(f(n)) \cdot 5\beta K \cdot \mathbb{E}_\Phi \tilde l(\hat\sigma(A), \sigma^*).$$

We now turn to the second term in equation (24). We have

$$\begin{aligned}P_\Psi(E) = \frac{1}{2} P_\Psi\left(1 \notin \mathcal{E}[\hat\sigma(A), \sigma_0^1] \text{ and } \tilde l(\hat\sigma(A), \sigma_0^1) \le \frac{1}{4\beta K}\right) \\ + \frac{1}{2} P_\Psi\left(1 \notin \mathcal{E}[\hat\sigma(A), \sigma_0^2] \text{ and } \tilde l(\hat\sigma(A), \sigma_0^2) \le \frac{1}{4\beta K}\right).\end{aligned} \tag{25}$$

If $l(\hat\sigma(A), \sigma_0^1) \le \tilde l(\hat\sigma(A), \sigma_0^1) \le \frac{1}{4\beta K}$, Lemma B.8 implies $S_K[\hat\sigma(A), \sigma_0^1]$ contains only one element, which we denote by ρ. Since $d_H(\sigma_0^1, \sigma_0^2) = 1$, we have $\frac{1}{n} d_H(\rho \circ \hat\sigma(A), \sigma_0^2) \le \frac{1}{4\beta K} + \frac{1}{n} \le \frac{1}{2\beta K}$, so applying Lemma B.8 again, we conclude that $\rho \in S_K[\hat\sigma(A), \sigma_0^2]$, as well. However, $(\rho \circ \hat\sigma(A))(1)$ cannot be equal to both $\sigma_0^1(1) = 1$ and $\sigma_0^2(1) = 2$. Hence, we cannot simultaneously have $1 \notin \mathcal{E}[\hat\sigma(A), \sigma_0^1]$ and $1 \notin \mathcal{E}[\hat\sigma(A), \sigma_0^2]$. In particular, the two events in

equation (25) are disjoint, so $P_\Psi(\mathcal{Q} \le f(n), E) \le P_\Psi(E) \le \frac{1}{2}$. Plugging back into equation (24), we conclude that

$$P_\Psi(\mathcal{Q} \le f(n)) \le \exp(f(n)) \cdot 5\beta K \cdot \mathbb{E}_\Phi \tilde{l}(\hat{\sigma}(A), \sigma^*) + \frac{1}{2},$$

so setting $f(n) = \log \frac{1}{20\beta K \mathbb{E}_\Phi \tilde{l}(\hat{\sigma}(A), \sigma^*)}$, we have

$$P_\Psi\left(\mathcal{Q} \le \log \frac{1}{20\beta K \mathbb{E}_\Phi l(\hat{\sigma}, \sigma^*)}\right) \le \frac{3}{4}.$$

By Chebyshev's inequality, we also have

$$P_\Psi\left(\mathcal{Q} \le \mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})}\right) \ge 4/5,$$

where $V_\Psi(Q)$ is the variance of Q under Ψ. Hence, $\log \frac{1}{20\beta K \mathbb{E}_\Phi \tilde{l}(\hat{\sigma}(A), \sigma_0)} \le \mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})}$, or equivalently,

$$\mathbb{E}_\Phi \tilde{l}(\hat{\sigma}(A), \sigma^*) \ge \frac{1}{20\beta K} \exp\Big(- (\mathbb{E}_\Psi \mathcal{Q} + \sqrt{5V_\Psi(\mathcal{Q})})\Big).$$

We now compute $\mathbb{E}_\Psi \mathcal{Q}$ and $V_\Psi(\mathcal{Q})$. Note that

$$\mathbb{E}_\Psi \mathcal{Q} = \frac{1}{2}\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^1] + \frac{1}{2}\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^2].$$

Furthermore, by Lemma H.2, we have

$$\begin{aligned}
\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^1] &= \mathbb{E}_\Psi\left[\sum_{u:\, u \ne 1,\, \sigma_0^1(u)=1} \log \frac{Y(A_{u,1})}{P(A_{u,1})} + \sum_{u:\, \sigma_0^1(u)=2} \log \frac{Y(A_{u,1})}{Q(A_{u,1})}\right] \\
&= \frac{n}{\beta K} \int \log \frac{dY}{dP} dY + \frac{n}{\beta K} \int \log \frac{dY}{dQ} dY \\
&= \frac{n}{\beta K} 2D = \frac{n}{\beta K} I.
\end{aligned}$$

Similarly, we have $\mathbb{E}_\Psi[\mathcal{Q} \,|\, \sigma^* = \sigma_0^2] = \frac{nI}{\beta K}$, so $\mathbb{E}_\Psi \mathcal{Q} = \frac{nI}{\beta K}$. We show in Lemma H.3 that the following bound holds for the variance: $\sqrt{5V_\Psi(\mathcal{Q})} \le C\sqrt{\frac{nI}{\beta K}}$.

Now note that if $\frac{nI}{\beta K} \to \infty$, we have $\sqrt{\frac{nI}{\beta K}} = o\left(\frac{nI}{\beta K}\right)$, so $\sqrt{5V_\Psi(\mathcal{Q})} = o\left(\frac{nI}{\beta K}\right)$. Therefore, $\mathbb{E}_\Phi \tilde{l}(\hat{\sigma}(A), \sigma^*) \ge C \exp\left(-(1 + o(1))\frac{nI}{\beta K}\right)$. If instead $\frac{nI}{\beta K} \le c < \infty$, then $\mathbb{E}_\Psi \mathcal{Q} = c$ and $\sqrt{5V_\Psi(\mathcal{Q})} \le C\sqrt{c}$, so $\mathbb{E}_\Phi \tilde{l}(\hat{\sigma}(A), \sigma^*) \ge c' > 0$, for some constant c'.

Now define two measures P_1, P_2 on $(A, \hat{\sigma}(A))$, as follows:

$$P_1(A, \hat{\sigma}(A)) = P_{SBM}(A \,|\, \sigma_0^1) P_{alg}(\hat{\sigma}(A) \,|\, A),$$
$$P_2(A, \hat{\sigma}(A)) = P_{SBM}(A \,|\, \sigma_0^2) P_{alg}(\hat{\sigma}(A) \,|\, A).$$

Note that $\mathbb{E}_\Phi[\tilde{l}(\hat{\sigma}, \sigma^*) \,|\, \sigma^* = \sigma_0^1] = \mathbb{E}_1 \tilde{l}(\hat{\sigma}, \sigma_0^1)$ and $\mathbb{E}_\Phi[\tilde{l}(\hat{\sigma}, \sigma^*) \,|\, \sigma^* = \sigma_0^2] = \mathbb{E}_2 \tilde{l}(\hat{\sigma}, \sigma_0^2)$, where $\mathbb{E}_1$ and $\mathbb{E}_2$ are expectations taken with respect to P_1 and P_2, respectively. We claim that $\mathbb{E}_1 \tilde{l}(\hat{\sigma}, \sigma_0^1) = \mathbb{E}_2 \tilde{l}(\hat{\sigma}, \sigma_0^2)$, in which case $\mathbb{E}_\Phi l(\hat{\sigma}, \sigma^*) = \mathbb{E}_1 \tilde{l}(\hat{\sigma}, \sigma_0^1) = \mathbb{E}\tilde{l}(\hat{\sigma}, \sigma_0)$ and the claims follow.

Define a permutation $\pi \in S_n$ that swaps $\{2, \ldots, \frac{n}{\beta K} + 1\}$ with $\{\frac{n}{\beta K} + 2, \ldots, 2\frac{n}{\beta K} + 1\}$ and satisfies $\pi(u) = u$ for $u = 1$ and $u \geq 2\frac{n}{\beta K} + 2$. Clearly, $\sigma_0^2 = \tau \circ \sigma_0^1 \circ \pi^{-1}$, where $\tau \in S_K$ swaps cluster labels 1 and 2. Now let A be fixed and let $\rho \in S_K$ be arbitrary. We have

$$\begin{aligned} d_H(\rho \circ \hat{\sigma}(A), \sigma_0^1) &= d_H(\rho \circ \hat{\sigma}(A), \tau^{-1} \circ \sigma_0^2 \circ \pi) \\ &= d_H(\rho \circ \hat{\sigma}(A) \circ \pi^{-1}, \tau^{-1} \circ \sigma_0^2) \\ &= d_H(\rho \circ \xi^{-1} \circ \hat{\sigma}(\pi A), \tau^{-1} \circ \sigma_0^2) \\ &= d_H(\tau \circ \rho \circ \xi^{-1} \circ \hat{\sigma}(\pi A), \sigma_0^2). \end{aligned}$$

Thus, $\rho \mapsto \tau \circ \rho \circ \xi^{-1}$ is a bijection between $S_K[\hat{\sigma}(A), \sigma_0^1]$ and $S_K[\hat{\sigma}(\pi A), \sigma_0^2]$. Furthermore, if v satisfies $(\rho \circ \hat{\sigma}(A))(v) \neq \sigma_0^1(v)$, we equivalently have

$$(\rho \circ \hat{\sigma}(A))(v) \neq (\tau^{-1} \circ \sigma_0^2 \circ \pi)(v) \iff (\rho \circ \hat{\sigma}(A) \circ \pi^{-1})(u) \neq (\tau^{-1} \circ \sigma_0^2)(u)$$

where $\pi(v) = u$, so $(\tau \circ \rho \circ \xi^{-1} \circ \hat{\sigma}(\pi A))(u) \neq \sigma_0^2(u)$. Thus, $v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0^1]$ if and only if $\pi(v) \in \mathcal{E}[\hat{\sigma}(\pi A), \sigma_0^2]$. Finally, we conclude that

$$\begin{aligned} \mathbb{E}_1 \tilde{l}(\hat{\sigma}(A), \sigma_0^1) &= \frac{1}{n} \sum_{v=1}^n P_1(v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0^1]) = \frac{1}{n} \sum_{v=1}^n P_1(\pi(v) \in \mathcal{E}[\hat{\sigma}(\pi A), \sigma_0^2]) \\ &\overset{(a)}{=} \frac{1}{n} \sum_{v=1}^n P_2(\pi(v) \in \mathcal{E}[\hat{\sigma}(A), \sigma_0^2]) = \mathbb{E}_2 \tilde{l}(\hat{\sigma}(A), \sigma_0^2), \end{aligned}$$

where (a) follows because $[\pi A]_{ij} = A_{\pi^{-1}(i), \pi^{-1}(j)}$, implying that if A is distributed according to $P_{SBM}(A \,|\, \sigma_0^1)$, then πA is distributed according to $P_{SBM}(A \,|\, \sigma_0^2)$. This concludes the proof.

H.2. Properties of permutation equivariant estimators. The following lemma establishes a symmetry property used to prove Theorem 5.3:

LEMMA H.1. *Let the true clustering σ_0 be arbitrary. Suppose the weight matrix A is drawn from an arbitrary probability measure and $\hat{\sigma}$ is any permutation-equivariant estimator. Let u and v be two nodes such that there exists $\pi \in S_n$ satisfying*

(1) $\pi(u) = v$,
(2) π *is measure-preserving; i.e.,* $A \stackrel{d}{=} \pi A$, *and*
(3) π *preserves the true clustering; i.e., there exists* $\tau \in S_K$ *such that* $\tau \circ \sigma_0 \circ \pi^{-1} = \sigma_0$.

Then

$$P(u \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]) = P(v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]).$$

PROOF. Since $\hat{\sigma}(A) \stackrel{d}{=} \hat{\sigma}(\pi A)$, we have

$$P\Big(v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\hat{\sigma}(\pi(A)), \sigma_0]\Big).$$

We claim that $u \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]$ if and only if $v \in \mathcal{E}[\hat{\sigma}(\pi A), \sigma_0]$, implying the desired result:

$$P\Big(u \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\hat{\sigma}(\pi A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]\Big).$$

Consider a fixed matrix A, and let $\tau \in S_K$ satisfy $\tau \circ \sigma_0 \circ \pi^{-1} = \sigma_0$. Let $\xi \in S_K$ be the permutation such that $\hat{\sigma}(\pi A) = \xi \circ \hat{\sigma}(A) \circ \pi^{-1}$. For any $\rho \in S_K$, we have

$$\begin{aligned} d_H(\rho \circ \hat{\sigma}(A),\ \sigma_0) &= d_H(\tau \circ \rho \circ \xi^{-1} \circ \xi \circ \hat{\sigma}(A) \circ \pi^{-1},\ \tau \circ \sigma_0 \circ \pi^{-1}) \\ &= d_H(\tau \circ \rho \circ \xi^{-1} \circ \hat{\sigma}(\pi A),\ \sigma_0). \end{aligned}$$

Therefore, $\rho \in S_K[\hat{\sigma}(A), \sigma_0]$ if and only if $\tau \circ \rho \circ \xi^{-1} \in S_K[\hat{\sigma}(\pi(A)), \sigma_0]$. In particular, if $v \in \mathcal{E}[\hat{\sigma}(\pi A), \sigma_0]$, we have $\tau \circ \rho \circ \xi^{-1} \circ \hat{\sigma}(\pi A)(v) \neq \sigma_0(v)$ for some $\rho \in S_K[\hat{\sigma}(A), \sigma_0]$. Then

$$\begin{aligned} \hat{\sigma}(A)(u) = \hat{\sigma}(A)(\pi^{-1}(v)) = \xi^{-1} \circ \xi \circ \hat{\sigma}(A) \circ \pi^{-1}(v) = \xi^{-1} \circ \hat{\sigma}(\pi A)(v) \\ \neq \rho^{-1} \circ \tau^{-1} \circ \sigma_0(v) = \rho^{-1} \circ \tau^{-1} \circ \sigma_0(\pi(u)) = \rho^{-1}(\sigma_0(u)). \end{aligned}$$

Thus, $u \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]$. Similar reasoning shows that if $u \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]$, then $v \in \mathcal{E}[\hat{\sigma}(\pi A), \sigma_0]$. □

COROLLARY H.1. *Let the true clustering* σ_0 *be arbitrary. Suppose the weight matrix* A *is drawn from a weighted SBM and* $\hat{\sigma}$ *is any permutation equivariant estimator. Let* u *and* v *be two nodes lying in equally-sized clusters. Then*

$$P\Big(u \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]\Big) = P\Big(v \in \mathcal{E}[\hat{\sigma}(A), \sigma_0]\Big).$$

PROOF. By Lemma H.1, it suffices to construct a permutation $\pi \in S_n$ satisfying conditions (1)–(3). First suppose u and v lie in the same cluster. It is easy to see that the conditions are satisfied when π is the permutation that swaps u and v and τ is the identity. If u and v lie in different clusters, suppose WLOG that u is in cluster 1 and v is in cluster 2, where clusters 1 and 2 have the same size. Let π be the permutation that exchanges all nodes in cluster 1 with all nodes in cluster 2. The conditions are satisfied when τ is the permutation that transposes cluster labels 1 and 2. □

H.3. Properties of Renyi divergence. We first state a lemma that provides an alternative characterization of the Renyi divergence:

LEMMA H.2. *Let P and Q be two probability measures on $\mathbb{R}$ that are absolutely continuous with respect to each other, with point masses P_0 and Q_0 at zero. The Renyi divergence satisfies $I = 2D$, where*

$$D := \inf_{Y \in \mathcal{P}} \max \left\{ \int \log \frac{dY}{dP} dY, \int \log \frac{dY}{dQ} dY \right\},$$

and $\mathcal{P}$ denotes the set of probability measures absolutely continuous with respect to both P and Q.

PROOF. First, note that D is finite by choosing $Y = P$. We claim that

$$D = \inf_{Y \in \mathcal{P}} \left\{ \int \log \frac{dY}{dP} dY : \int \log \frac{dP}{dQ} dY = 0 \right\}. \tag{26}$$

This holds because for any $Y \in \mathcal{P}$ such that $\int \log \frac{dP}{dQ} dY \neq 0$, we have $\int \log \frac{dY}{dP} dY \neq \int \log \frac{dY}{dQ} dY$. Suppose without loss of generality that the first quantity is larger. Then it is possible to take $\tilde{Y} = (1-\epsilon)Y + \epsilon P$ for ϵ small enough such that $\max \left\{ \int \log \frac{d\tilde{Y}}{dP} d\tilde{Y}, \int \log \frac{d\tilde{Y}}{dQ} d\tilde{Y} \right\}$ strictly decreases, so the infimum in the definition of D could not have been achieved. Since the new formulation (26) is convex in Y, we may solve to obtain the optimal $Y^* \in \mathcal{P}$, defined by $Y_0^* = \frac{P_0^{1/2} Q_0^{1/2}}{Z}$ and $(1 - Y_0^*) y^*(x) = \frac{((1-P_0) \cdot p(x))^{1/2} ((1-Q_0) \cdot q(x))^{1/2}}{Z}$, where $Z = P_0^{1/2} Q_0^{1/2} + \int \sqrt{(1-P_0) p(x) (1-Q_0) q(x)} dx$. Then

$$\begin{aligned} \int \log \frac{dY^*}{dP} dY^* &= \log \frac{1}{Z} \left\{ \left(\frac{Q_0}{P_0} \right)^{1/2} Y_0^* \right. \\ &\qquad \left. + \int \left(\frac{(1-P_0) p(x)}{(1-Q_0) q(x)} \right)^{1/2} (1 - Y_0^*) y^*(x) dx \right\} \\ &= \log \frac{dP}{dQ} dY^* - \log Z = -\log Z. \end{aligned}$$

It is straightforward to verify that $-2\log Z = I$. □

H.4. Bounding the variance.

LEMMA H.3. *For a suitable constant C, we have $\sqrt{5V_\Psi(\mathcal{Q})} \le C\sqrt{\frac{nI}{\beta K}}$.*

PROOF. We begin with the decomposition

$$\begin{aligned} V_\Psi(\mathcal{Q}) &= V(\mathbb{E}_\Psi[\mathcal{Q}\,|\,\sigma^*]) + E[V_\Psi(\mathcal{Q}\,|\,\sigma^*)] = E[V_\Psi(\mathcal{Q}\,|\,\sigma^*)] \\ &= \frac{1}{2}V_\Psi(\mathcal{Q}\,|\,\sigma^* = \sigma_0^1) + \frac{1}{2}V_\Psi(\mathcal{Q}\,|\,\sigma^* = \sigma_0^2). \end{aligned}$$

Then

$$\begin{aligned} V_\Psi(\mathcal{Q}\,|\,\sigma^* = \sigma_0^1) &= \sum_{u:\, u\neq 1, \sigma_0^1(u)=1} V_\Psi\left(\log\frac{Y(A_{v_1u})}{P(A_{v_1u})}\right) + \sum_{u:\,\sigma_0^1(u)=2} V_\Psi\left(\log\frac{Y(A_{v_1u})}{Q(A_{v_1u})}\right) \\ &\le \frac{n}{\beta K}\mathbb{E}_\Psi\left(\log\frac{Y(A_{v_1u})}{P(A_{v_1u})}\right)^2 + \frac{n}{\beta K}\mathbb{E}_\Psi\left(\log\frac{Y(A_{v_1u})}{Q(A_{v_1u})}\right)^2. \end{aligned}$$

We will show that $\mathbb{E}_\Psi\left(\log\frac{Y(A_{v_1u})}{P(A_{v_1u})}\right)^2 \le CI$, so $\sqrt{5V_\Psi(\mathcal{Q})} \le C\sqrt{\frac{nI}{\beta K}}$. We have

$$\begin{aligned} \mathbb{E}_\Psi\left(\log\frac{Y(A_{uv^*})}{P(A_{uv^*})}\right)^2 &= \int\left(\log\frac{dY}{dP}\right)^2 dY \\ &= Y_0\log^2\frac{Y_0}{P_0} + (1-Y_0)\int y(x)\log^2\frac{(1-Y_0)y(x)}{(1-P_0)p(x)}dx. \end{aligned} \tag{27}$$

To bound the first term, we write

$$\begin{aligned} \left|\log\frac{Y_0}{P_0}\right| &= \left|\frac{1}{2}\log\frac{Q_0}{P_0} - \log Z\right| \le \frac{1}{2}\left|\log\left(1-\frac{P_0-Q_0}{P_0}\right)\right| + \frac{I}{2} \\ &\overset{(a)}{\le} \frac{1}{2}\left|\frac{P_0-Q_0}{P_0}\right| + \left|\frac{P_0-Q_0}{P_0}\right|^2 C + \frac{I}{2} \overset{(b)}{\le} C\left|\frac{P_0-Q_0}{P_0}\right| + CI, \end{aligned}$$

where (a) follows from Lemma I.2 and the fact that $\frac{Q_0}{P_0}$ is bounded, and (b) follows from the fact that $\left|\frac{P_0-Q_0}{P_0}\right| = \left|1-\frac{Q_0}{P_0}\right| \le 1 + \left|\frac{Q_0}{P_0}\right|$ is bounded. Therefore,

$$\begin{aligned} Y_0\log^2\frac{Y_0}{P_0} &\le Y_0\left(C\frac{|P_0-Q_0|}{P_0} + CI\right)^2 \\ &\overset{(a)}{\le} Y_0\frac{|P_0-Q_0|^2}{P_0^2}C + Y_0I^2C \overset{(b)}{\le} \frac{|P_0-Q_0|^2}{P_0}C + I^2C, \end{aligned}$$

where (a) follows because $(x+y)^2 \le 2x^2+2y^2$, and (b) follows because $Y_0 = \frac{\sqrt{P_0Q_0}}{Z} = (1+o(1))CP_0$.

Since $I = o(1)$, we have $I^2 \le I$. Also,

$$I \ge (1+o(1))(\sqrt{P_0}-\sqrt{Q_0})^2 = (1+o(1))\frac{(P_0-Q_0)^2}{(\sqrt{P_0}+\sqrt{Q_0})^2}$$
$$= C(1+o(1))\frac{(P_0-Q_0)^2}{P_0},$$

from which we conclude that $Y_0 \log^2 \frac{Y_0}{P_0} \le CI$.

Now we turn our attention to the second term in equation (27). We have

$$\left|\log\frac{(1-Y_0)y(x)}{(1-P_0)p(x)}\right| \le \frac{1}{2}\left|\log\frac{(1-Q_0)q(x)}{(1-P_0)p(x)} - \log Z\right|$$
$$\le \frac{1}{2}\left|\log\frac{1-Q_0}{1-P_0}\right| + \frac{1}{2}\left|\log\frac{q(x)}{p(x)}\right| + \frac{I}{2}.$$

Therefore,

$$(1-Y_0)\int y(x)\left(\log\frac{1-Y_0}{1-P_0}\frac{y(x)}{p(x)}\right)^2 dx$$
$$\le (1-Y_0)\int y(x)\left\{\frac{1}{2}\left|\log\frac{1-Q_0}{1-P_0}\right| + \frac{1}{2}\left|\log\frac{q(x)}{p(x)}\right| + \frac{I}{2}\right\}^2 dx$$
$$\le (1-Y_0)\int y(x)\left\{C\left|\log\frac{1-Q_0}{1-P_0}\right|^2 + C\left|\log\frac{q(x)}{p(x)}\right|^2 + CI^2\right\} dx, \tag{28}$$

where we have used the fact that $(x+y+z)^2 \le 9x^2+9y^2+9z^2$ in the last inequality. Define

$$\mathcal{A} := (1-Y_0)\int y(x)\left|\log\frac{1-Q_0}{1-P_0}\right|^2 dx,\ \mathcal{B} := (1-Y_0)\int y(x)\left|\log\frac{q(x)}{p(x)}\right|^2 dx,$$
$$\mathcal{C} := (1-Y_0)\int y(x)I^2 dx.$$

We bound each term separately, beginning with $\mathcal{A}$. Note that

$$\left|\log\frac{1-Q_0}{1-P_0}\right| = \left|\log\left(1-\frac{Q_0-P_0}{1-P_0}\right)\right| \overset{(a)}{\le} \left|\frac{Q_0-P_0}{1-P_0}\right| + C\left(\frac{Q_0-P_0}{1-P_0}\right)^2$$
$$\overset{(b)}{\le} C\left|\frac{Q_0-P_0}{1-P_0}\right|,$$

where (a) follows from Lemma I.2 and the fact that $\frac{1-Q_0}{1-P_0}$ is bounded, and (b) follows from the fact that $\left|\frac{Q_0-P_0}{1-P_0}\right| = \left|1-\frac{1-Q_0}{1-P_0}\right| \le 1+\left|\frac{1-Q_0}{1-P_0}\right| \le C$. Therefore,

$$\begin{aligned}
\mathcal{A} &\le C(1-Y_0)\int y(x)\left(\frac{Q_0-P_0}{1-P_0}\right)^2 dx \\
&= C\left(\frac{Q_0-P_0}{1-P_0}\right)^2 \int \frac{\sqrt{(1-P_0)p(x)(1-Q_0)q(x)}}{Z}dx \\
&\overset{(a)}{\le} C(1+o(1))\left(\frac{Q_0-P_0}{1-P_0}\right)^2 \int \sqrt{\frac{(1-Q_0)q(x)}{(1-P_0)p(x)}}(1-P_0)p(x)dx \\
&\overset{(b)}{\le} C(1+o(1))\left(\frac{Q_0-P_0}{1-P_0}\right)^2 (1-P_0) \\
&\le C(1+o(1))\frac{(Q_0-P_0)^2}{1-P_0},
\end{aligned}$$

where in (a), we use the fact that $\frac{1}{Z} = (1+o(1))$ since $Z \to 1$, and in (b), we use the fact that $\frac{1-Q_0}{1-P_0}$ and $\frac{q(x)}{p(x)}$ are bounded. Note that

$$\begin{aligned}
I &\ge (1+o(1))(\sqrt{1-P_0}-\sqrt{1-Q_0})^2 = (1+o(1))\frac{(P_0-Q_0)^2}{(\sqrt{1-P_0}+\sqrt{1-Q_0})^2} \\
&= C(1+o(1))\frac{(P_0-Q_0)^2}{1-P_0},
\end{aligned}$$

implying that $\mathcal{A} \le C(1+o(1))I \le CI$.

Moving onto $\mathcal{B}$, first suppose $H = \Theta(1)$ and

$$\max\left\{\int p(x)\left|\log\frac{q(x)}{p(x)}\right|^2 dx,\ \int q(x)\left|\log\frac{q(x)}{p(x)}\right|^2 dx\right\} < \infty.$$

We have

$$\begin{aligned}
\mathcal{B} &\le C\int \frac{\sqrt{(1-P_0)p(x)(1-Q_0)q(x)}}{Z}\left|\log\frac{q(x)}{p(x)}\right|^2 dx \\
&\overset{(a)}{\le} C\sqrt{(1-P_0)(1-Q_0)}\int \sqrt{p(x)q(x)}\left|\log\frac{q(x)}{p(x)}\right|^2 dx \\
&\le C\sqrt{(1-P_0)(1-Q_0)}\int (p(x)+q(x))\left|\log\frac{q(x)}{p(x)}\right|^2 dx \\
&\overset{(b)}{\le} C\sqrt{(1-P_0)(1-Q_0)}H \le CI,
\end{aligned}$$

where (a) follows because $Z \to 1$ and (b) follows because $H = \Theta(1)$. Next, we make no assumption on H but assume $\left|\log \frac{q(x)}{p(x)}\right|$ is bounded. Then

$$\left|\log \frac{q(x)}{p(x)}\right| = \left|\log\left(1 - \frac{p(x)-q(x)}{p(x)}\right)\right| \overset{(a)}{\leq} \left|\frac{p(x)-q(x)}{p(x)}\right| + \left(\frac{p(x)-q(x)}{p(x)}\right)^2 C \overset{(b)}{\leq} C\left|\frac{p(x)-q(x)}{p(x)}\right|,$$

where (a) follows from Lemma I.2 and the fact that $\frac{q(x)}{p(x)}$ is bounded, and (b) follows from the fact that $\left|\frac{p(x)-q(x)}{p(x)}\right| = \left|1 - \frac{q(x)}{p(x)}\right| \leq 1 + \left|\frac{q(x)}{p(x)}\right| \leq C$. Then

$$\mathcal{B} \leq \frac{C}{Z}\int \sqrt{\frac{1-Q_0}{1-P_0}\frac{q(x)}{p(x)}}(1-P_0)p(x)\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx \overset{(a)}{\leq} \frac{C}{Z}(1-P_0)\int p(x)\left(\frac{p(x)-q(x)}{p(x)}\right)^2 dx,$$

where in (a), we use the facts that $\frac{1}{Z} = (1+o(1))$ and $\frac{1-Q_0}{1-P_0}$, and $\frac{p(x)}{q(x)}$ are both bounded by assumption. Now, note that $H = \int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx = \int \frac{(p(x)-q(x))^2}{(\sqrt{p(x)}+\sqrt{q(x)})^2} dx = C\int \frac{(p(x)-q(x))^2}{p(x)} dx$. Therefore,

$$\mathcal{B} \leq C(1-P_0)H \leq C\sqrt{(1-P_0)(1-Q_0)}H \leq CI.$$

Finally, note that $\mathcal{C} = (1-Y_0)CI^2 \leq CI$. Substituting back into inequality (28), we therefore obtain

$$(1-Y_0)\int y(x)\left(\log\frac{1-Y_0}{1-P_0}\frac{y(x)}{p(x)}\right)^2 dx \leq CI,$$

so substituting back into inequality (27), we obtain the desired bound. □

APPENDIX I: ADDITIONAL USEFUL LEMMAS

LEMMA I.1. *Let*

$$I = -2\log\left(\sqrt{P_0Q_0} + \int\sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right),$$
$$I^h = (\sqrt{P_0} - \sqrt{Q_0})^2 + \int\left(\sqrt{(1-P_0)p(x)} - \sqrt{(1-Q_0)q(x)}\right)^2 dx.$$

If $I^h < 2 - 2\epsilon$, then $I = I^h(1+\eta)$, where $|\eta| \leq \frac{I^h}{2\epsilon}$. Thus, $I \to 0$ if and only if $I^h \to 0$, in which case $I = I^h(1+o(1))$.

PROOF. We have

$$\begin{aligned} I &= -2\log\left(\sqrt{P_0Q_0} + \int \sqrt{(1-P_0)(1-Q_0)p(x)q(x)}dx\right) \\ &= -2\log\left(1 - \frac{1}{2}\left((\sqrt{P_0} - \sqrt{Q_0})^2 \right.\right. \\ &\qquad \left.\left. + \int (\sqrt{(1-P_0)p(x)} - \sqrt{(1-Q_0)q(x)})^2 dx\right)\right) \\ &= -2\log\left(1 - \frac{1}{2}I^h\right) = 2\cdot\frac{1}{2}I^h(1+\eta), \end{aligned}$$

where $|\eta| \le \frac{I^h}{2\epsilon}$. The last equality follows from Lemma I.2. □

LEMMA I.2. *Suppose* $0 < \epsilon \le 1$. *For all* $0 \le x < 1-\epsilon$, *we have* $\log(1-x) = -(1+\eta)x$, *where* $|\eta| \le \frac{x}{2\epsilon}$

PROOF. We simply Taylor expand $\log(1-x)$ around $x = 0$. □

LEMMA I.3. *Let* $f(z) = \frac{1-\frac{z}{2}-\sqrt{1-z}}{z}$, *for* $z \le 1$ *and* $z \ne 0$, *and define* $f(0) = 0$. *Then* $|f(z)| \le |z|$, *for all* $z \le 1$.

PROOF. Note that f is continuous, with derivative $f'(z) = -\frac{1}{z^2} - \frac{z-2}{2z^2\sqrt{1-z}}$. It is straightforward to check that $f'(z) \ge 0$ for all $z < 1$, and we may define $f'(0) = \frac{1}{4}$ so $f'(z)$ is continuous. Then $f(z)$ is monotonic and maximized at $z = 1$, yielding $f(1) = \frac{1}{2}$, and minimized at $\lim_{z\to-\infty} f(z) = -\frac{1}{2}$.

We now split into cases. If $z < -\frac{1}{2}$, then $|f(z)| \le \frac{1}{2} < |z|$. If $|z| \le 1/2$, a Taylor expansion gives

$$\begin{aligned} \left|\sqrt{1-z} - \left(1 - \frac{z}{2}\right)\right| &\le \frac{1}{8}(|z|^2 + |z|^3 + \cdots) \le \frac{1}{8}|z|^2(1 + |z| + |z|^2 + \cdots) \\ &\le \frac{1}{8}|z|^2\frac{1}{1-|z|} \le \frac{1}{4}|z|^2, \end{aligned}$$

implying that $|f(z)| \le \frac{1}{4}|z|$. Finally, if $z > 1/2$, we have $|f(z)| \le \frac{1}{2} < z$. □

REFERENCES

[1] E. Abbe. Community detection and stochastic block models: Recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

[2] E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer. Decoding binary node labels from censored edge measurements: Phase transition and efficient recovery. *IEEE Transactions on Network Science and Engineering*, 1(1):10–22, 2014.

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

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

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

[6] C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, page cnu026, 2014.

[7] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems*, pages 954–962, 2011.

[8] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America*, 101(11):3747–3752, 2004.

[9] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[10] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics reports*, 424(4):175–308, 2006.

[11] Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Proceedings of The 28th Conference on Learning Theory*, pages 391–423, 2015.

[12] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84:066106, Dec 2011.

[13] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, New York, NY, USA, 2010.

[14] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.

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

[16] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Community detection in degree-corrected block models. *arXiv preprint arXiv:1607.06993*, 2016.

[17] A. Goldenberg, A. X. Zheng, S. E. Fienberg, and E. M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2(2):129–233, February 2010.

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

[19] B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.

[20] B. Hajek, Y. Wu, and J. Xu. Submatrix localization via message passing. *arXiv preprint arXiv:1510.09219*, 2015.

[21] B. Hajek, Y. Wu, and J. Xu. Information limits for recovering a hidden community. *IEEE Transactions on Information Theory*, 2017.

[22] E. Hartuv and R. Shamir. A clustering algorithm based on graph connectivity. *Information Processing Letters*, 76(4–6):175–181, 2000.

[23] S. Heimlicher, M. Lelarge, and L. Massoulié. Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*, 2012.

[24] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983.
[25] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2010.
[26] V. Jog and P. Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.
[27] Jing Lei, Alessandro Rinaldo, et al. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.
[28] M. Lelarge, L. Massoulié, and J. Xu. Reconstruction in the labeled stochastic block model. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE, 2013.
[29] L. Massoulié. Community detection thresholds and the weak Ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, STOC '14, pages 694–703. ACM, 2014.
[30] F. McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.
[31] E. Mossel, J. Neeman, and A. Sly. Stochastic Block Models and Reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.
[32] E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.
[33] E. Mossel, J. Neeman, and A. Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 2014.
[34] M. Newman, A.-L. Barabasi, and D. J. Watts. *The Structure and Dynamics of Networks: (Princeton Studies in Complexity)*. Princeton University Press, Princeton, NJ, USA, 2006.
[35] M. E. J. Newman. Analysis of weighted networks. *Physical Review E*, 70(5):056131, 2004.
[36] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.
[37] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.
[38] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010. Computational Models of the Brain.
[39] D.S. Sade. Sociometrics of Macaca mulatta: I. Linkages and cliques in grooming matrices. *Folia Primatologica*, 18(3–4):196–223, 1972.
[40] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, August 2000.
[41] S. Yun and A. Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.
[42] A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block model. *arXiv preprint arXiv:1507.05313*, 2015.
[43] B. Zhang and S. Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1):1128, 2005.
[44] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, 40(4):2266–2292, 2012.

Department of Statistics
The Wharton School
3730 Walnut St
Philadelphia, PA 19104
E-mail: minx@wharton.upenn.edu

Departments of ECE & Statistics
Grainger Institute of Engineering
1415 Engineering Drive
Madison, WI 53706
E-mail: vjog@wisc.edu
E-mail: loh@ece.wisc.edu