SIAM J. MATH. DATA SCI.
Vol. 2, No. 1, pp. 132–157

Graph Powering and Spectral Robustness*

Emmanuel Abbe†, Enric Boix-Adserà‡, Peter Ralli§, and Colin Sandon‡

Abstract. Spectral algorithms, such as principal component analysis and spectral clustering, rely on the extremal eigenpairs of a matrix A. However, these may be uninformative without preprocessing A with a proper transformation. The reason is that the spectrum of A may be contaminated by top eigenvalues resulting from scale variations in the data, such as high-degree nodes. Designing a good ψ and establishing what good means is often challenging and model dependent. This paper proposes a simple and generic construction for sparse graphs, $\psi(A) = \mathbb{1}((I + A)^r \geq 1)$, where A denotes the adjacency matrix, r is an integer, and the indicator function is applied entrywise. We support this "graph powering" construction with the following regularization properties: (i) if the graph is drawn from the sparse Erdős–Rényi ensemble, which has no spectral gap, then graph powering produces a "maximal" spectral gap, comparable to that obtained when powering a random regular graph; (ii) if the graph is drawn from the sparse stochastic block model, graph powering achieves the fundamental limit for weak recovery (the Kesten–Stigum threshold), settling at the same time a related conjecture by Massoulié in 2013; (iii) we also demonstrate that graph powering is significantly more robust to tangles and cliques than previous spectral algorithms based on self-avoiding or nonbacktracking walk counts, using a geometric block model as our benchmark and introducing new conjectures for this model.

Key words. community detection, stochastic block models, random graphs, spectral algorithms, network data analysis, spectral embeddings

AMS subject classifications. 62M15, 05C85, 62F35, 62F15

DOI. 10.1137/19M1257135

1. Introduction.

1.1. Spectral data analysis and robustness.

A large variety of algorithms exploit the spectrum of graph operators. This includes most methods of unsupervised learning that rely on spectral decomposition, e.g., principal component analysis, clustering, or linear embeddings. The common base of spectral algorithms is to first obtain a Euclidean embedding of the data (which may a priori have no relation to a metric space) and then use this embedding for further tasks. Namely, given an n-vertex graph G with adjacency matrix A_G,

1. construct an operator $M_G = \psi(A_G)$, where $\psi : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$ is properly chosen;
2. take the top k eigenvectors of M_G to create the $n \times k$ matrix Ψ_k, and use $\Psi_k(i)$ as the

*Received by the editors April 22, 2019; accepted for publication (in revised form) December 11, 2019; published electronically February 18, 2020.
https://doi.org/10.1137/19M1257135
Funding: This work was partly supported by NSF CAREER award CCF-1552131, NSF CSOI CCF-0939370, ARO grant W911NF-14-1-0094, and NSF grant DMS-1811735.
†Mathematics and Computer and Communications Sciences, EPFL, Lausanne, CH-1015 (emmanuel.abbe@epfl.ch).
‡Electrical Engineering and Computer Science, MIT, Cambridge, MA 02139 (eboix@mit.edu, csandon@mit.edu).
§PACM (Program in Applied and Computational Sciences), Princeton University, Princeton, NJ 08544 (pralli@math.princeton.edu).

k-dimensional embedding for the data point $i \in [n]$.

In clusterings, one typically looks for k much smaller than n, cutting off a potentially significant matrix norm, and running the k-means algorithm on the embedded points to obtain clusters [45]. In word embeddings, one may preserve almost all the matrix norm in order to approximate each pair of words' co-occurrences [44].

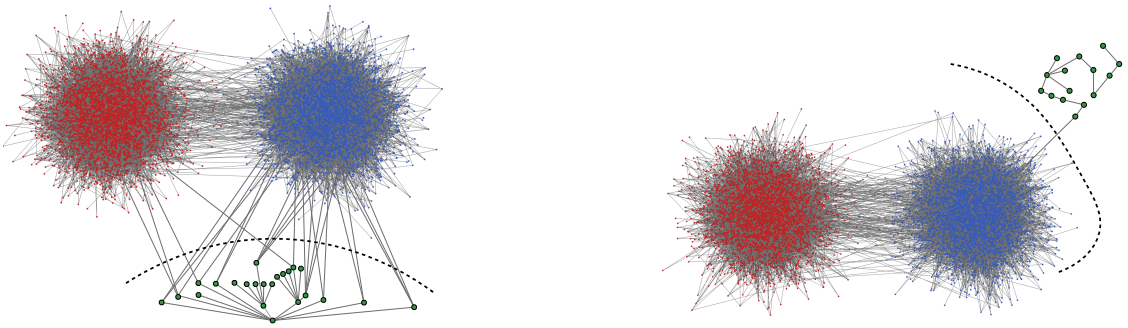
Popular choices for M_G (depending on applications) are the adjacency matrix A, the Laplacian $D - A$, the normalized Laplacian $I - D^{-1/2}AD^{-1/2}$ (or the random walk Laplacian $D^{-1}A$), and various regularized versions of the above using trimming/thresholding operations or smoothing/completion operations. More recently, operators based on self-avoiding or nonbacktracking walks have become popular in the context of block models [6, 9, 32, 36]; see further discussion of these below. The function ψ can also take specific forms such as the log-PMI function in word embeddings that uses both normalization and the application of the logarithm entrywise [11, 37, 44], or more sophisticated (nonpositive) forms such as in phase retrieval [34]. A long list of other forms is omitted here.

Why is it important to apply a transformation ψ? Consider graph clustering; if one takes $M_G = A$ directly, the top eigenvector is likely to localize on the largest degree vertex of the graph, which is not the type of macroscopic structure that clustering aims to extract. This is a well-known issue, illustrated in Figure 1(a) using the spectral algorithm on A to cluster the stochastic block model (SBM) in the sparse regime. Moreover, pruning (also called trimming or thresholding) the largest degree nodes as done in [10, 12, 17, 23, 29, 46] does not get around this issue in the sparse SBM [25]. Similarly, in word embeddings, without mitigating the most popular words, such as "the," the embedding assigns full dimension to these. On the flip side, if one takes normalized Laplacians for clustering, one may overcorrect the large degree nodes and output clusters that are now at the periphery of the graph, such as "tails"; see Figure 1(b) for an example on the SBM. These are discussed in more detail in section SM1 of the supplementary materials, linked from the main article webpage. In particular, neither the pruned adjacency matrix nor the normalized Laplacian achieves the threshold for weak recovery in the SBM. In general, transformations ψ crucially serve to "regularize" the graph in order to obtain useful embeddings that capture "macroscopic" structures.

The robustness of spectral methods and semidefinite programming methods has been studied in various contexts, mainly for objective-based clustering, using adversarial corruptions [14, 19, 27], obtaining certificates for balanced cuts of given conductance [39, 41], or studying noise perturbations [8, 24, 30, 45], among others. The goal of this paper is to formalize the notion of robustness in the context of community detection/clustering for probabilistic graph models, and to obtain a transformation ψ for spectral methods that is as simple and as robust as possible. The derived graph powering transformation can also be viewed as a method to extract spectral gaps in graphs that have a spectral gap hiding under local irregularities, such as the Erdős–Rényi (ER) random graph.

A concurrent work [43] analyzes the robustness under adversarial perturbations of clustering the sparse SBM based on the spectrum of the graph's distance matrix. The graph distance matrix is closely related to the graph powering matrix that we analyze in our work.

1.2. Our benchmarks. We refer to the next section for the formal statements and provide here some informal motivations. Our goal is to find an "optimal" ψ that is efficiently computable



(a) The top eigenvectors of the adjacency matrix are used to cluster. One community corresponds to the neighborhood of a high-degree vertex, and all other vertices are put in the second community.

(b) The top eigenvectors of the Laplacian matrix are used to cluster. One community corresponds to a "tail" of the graph (i.e., a small region connected by a single edge to the giant component), and all other vertices are put in the second community. The same outcome takes place for the normalized Laplacian.

Figure 1. *Clustering a sparse symmetric SBM* ($n = 100000, a = 2.2, b = 0.06$).

and that satisfies the following objectives:

(1) achieving a "maximal" spectral gap on ER random graphs, e.g., comparable to that of random regular graphs, as evidence that ψ is a good graph regularizer;
(2) achieving the Kesten–Stigum (KS) threshold for weak recovery in the SBM, as evidence that ψ is good for community detection;
(3) achieving the threshold for weak recovery in a geometric block model (GBM), as evidence that ψ is robust to cliques and tangles (common in clustering problems).

We provide further explanation on these points:

(1) *Maximal gap.* The first point refers to the following: a random regular graph of degree d has with high probability the top two eigenvalues $\lambda_1 = d$ and $\lambda_2 \sim 2\sqrt{d-1}$ [15]. This is a "maximal" gap in the sense that Alon–Boppana shows that any regular graph must have a second eigenvalue larger than $(1 - o_{\mathrm{diam}(G)}(1))2\sqrt{d-1}$. Ramanujan families of graphs are defined as families achieving the lower bound of $2\sqrt{d-1}$, and so random regular graphs are almost Ramanujan with high probability.

In contrast, an ER random graph of expected constant degree d has the top two eigenvalues both of order $\sqrt{\log(n)/\log\log(n)}$ with high probability, and these correspond to eigenvectors localized on the vertices of largest degree. So ER graphs are not almost Ramanujan (naively applying the above Ramanujan definition for irregular graphs). Are there more general definitions of "maximal gaps" that are suitable to irregular graphs? Some proposals are based on the universal cover [18] and the nonbacktracking operator.

In particular, it is shown in [21] that a regular graph is Ramanujan if and only if its nonbacktracking matrix has no eigenvalue of magnitude within $(\sqrt{\rho}, \rho)$, where ρ is the Perron–Frobenius eigenvalue of the nonbacktracking matrix. This gives a possibility to extend the definition of Ramanujan to irregular graphs, requiring no eigenvalue of magnitude within $(\sqrt{\rho}, \rho)$ for its nonbacktracking matrix, which we refer to as NB-Ramanujan. In this context,

it was proved in [9] that ER(n, c) is NB-Ramanujan with high probability. Further, in [9] it also shows that NB achieves our objective (2). However, the nonbacktracking matrix has some downsides: (i) it is asymmetric, and some of the intuition from spectral graph theory is lost; (ii) it is not robust to cliques and tangles and fails to achieve point (3) as discussed below.

Thus we will look for other solutions than the NB operator that achieve (2) and (3), but this will require us to revisit what a "maximal" spectral gap means for objective (1). An intuitive notion is that $\psi(A)$ is a good graph regularizer if it gives the same type of spectral gap whether A is random regular or from the ER ensemble.

(2) *SBM clustering.* The SBM is a random graph with planted clusters, and the weak recovery problem consists of recovering the clusters with a positive correlation. In the case of two symmetric clusters, the fundamental limit for efficient algorithms is the KS threshold. Recently, a few algorithms have been proved to achieve this threshold [6, 9, 32, 36]. However, these are not robust to cliques and tangles as we will discuss next.

(3) *GBM clustering.* Diverse versions of GBMs have recently been studied to integrate the presence of small cycles in the graph [2, 4, 5, 16] (see [1] for further references). In this paper, we consider a Gaussian-mixture block model (see section 2.4). This model has the advantage of having a simple expression for the weak recovery threshold (conjectured here), allowing us to evaluate whether the proposed methods are optimal or not. In particular, this simple model breaks the algorithms of [6, 7, 9, 32, 36], making eigenvectors concentrate possibly on tangles (i.e., dense regions with cliques intertwined); see Figure 2.

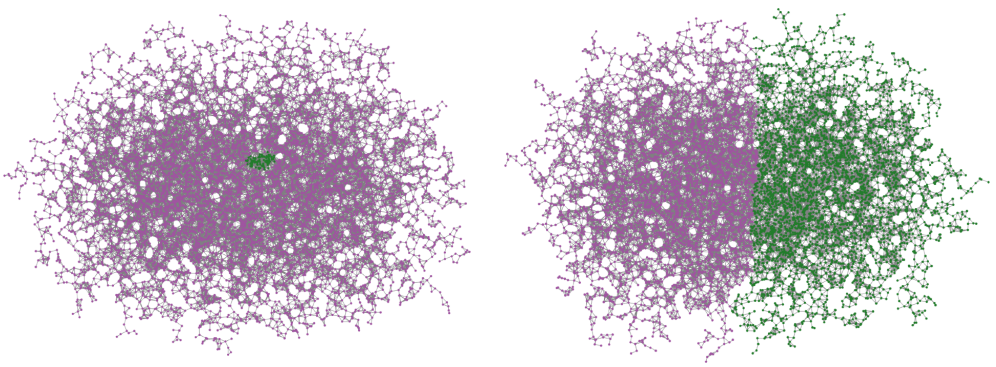


Figure 2. *A graph drawn from the GBM(n, t, s) (see Definition 2.4 and section 2.4), where $n/2$ points are sampled i.i.d. from a Gaussian in dimension 2 centered at $(-s/2, 0)$ and $n/2$ points are sampled i.i.d. from a Gaussian in dimension 2 centered at $(s/2, 0)$, and any points at distance less than $t/\sqrt{n}$ are connected (here $n = 10000$, $s = 2$, and $t = 10$). The spectral algorithm on the NB matrix gives the left plot, which puts a small fraction of densely connected vertices (a tangle) in a community and all other vertices in the second community. Graph powering gives the right plot, as desired.*

1.3. Graph powering.

The solution proposed in this paper is as follows.

Definition 1.1 (graph powering). *We give two equivalent definitions:*

- *Given a graph G and a positive integer r, define the rth graph power of G as the graph $G^{(r)}$ with the same vertex set as G and where two vertices are connected if there exists a path of length $\leq r$ between them in G.*

- *If A denotes the adjacency matrix of G, then the adjacency matrix $A^{(r)}$ of $G^{(r)}$ is defined by*

$$A^{(r)} = \mathbb{1}((I + A)^r \geq 1), \tag{1.1}$$

where the indicator function is applied entrywise to $(I + A)^r$. Note also that $(I + A)^r$ has the same spectrum as A (up to the transformation $t \mapsto (1 + t)^r$), but the action of the nonlinearity $\mathbb{1}(\cdot \geq 1)$ gives the key modification to the spectrum.

Definition 1.2. *For a graph G, an r-power-cut in G corresponds to a cut in $G^{(r)}$, i.e.,*

$$\partial^{(r)}(S) = \{(u, v) \in S \times S^c : (A^{(r)})_{u,v} = 1\}, \quad S \subseteq V(G). \tag{1.2}$$

Note that powering is mainly targeted for sparse graphs, and is useful only if the power r is not too small and not too large. If it is too small, the powered graph may not be sufficiently different from the underlying graph. If it is too large, say $r \geq \text{diameter}(G)$, then powering turns any graph to a complete graph, which destroys all the information. However, powering with r less than the diameter but large enough will be effective on the benchmarks (1), (2), and (3). As a rule of thumb, one may take $r = \sqrt{\text{diam}(G)}$. We now discuss the main insights behind graph powering.

Power-cuts as Bayes-like cuts. The spectral algorithms based on A, L, or L_{norm} can be viewed as relaxations of the MAP estimator, i.e., the min-bisection:

$$\max_{x \in \{-1,+1\}^n,\, x^T 1^n = 0} x^T A x. \tag{1.3}$$

However, the MAP estimator is the right benchmark only when aiming to maximize the probability of recovering the entire communities. It is not the right objective in the regime where one can only partially recover the communities, which is the sparse regime of interest to us in this paper. We illustrate this distinction on the following example; see also Figure 3.

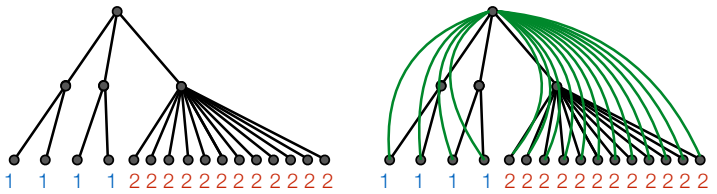


Figure 3. *In the left graph, assumed to come from* $\text{SBM}(n, 2, 3/n, 2/n)$*, the root vertex is labelled community* 1 *from the ML estimator given the leaf labels, which corresponds to the min-cut around that vertex. In contrast, the Bayes optimal estimator puts the root vertex in community* 2*, as the belief of its right descendent towards community* 2 *is much stronger than the belief of its two left descendents towards community* 1*. This corresponds in fact to the min-power-cut obtained from the right graph, where* 2*-power edges are added by graph powering (note that only a subset of relevant edges are added in the figure).*

Imagine that a graph drawn from $\text{SBM}(n, 2, 3/n, 2/n)$ contained the following induced subgraph: v_0 is adjacent to v_1, v_2, and v_3; v_1 and v_2 are each adjacent to two outside vertices that are known to be in community 1, and v_3 is adjacent to a large number of vertices that

are known to be in community 2. Then the single most likely scenario is that v_0, v_1, and v_2 are in community 1 while v_3 is in community 2. This puts v_0 in the community that produces the sparsest cut (one edge in the cut vs. two edges in the other case). However, v_3 is almost certain to be in community 2, while if we disregard any evidence provided by their adjacency to v_0, the vertices v_1 and v_2 are each only about 69% likely to be in community 1. As a result, v_0 is actually slightly more likely to be in community 2 than it is to be in community 1.

Power-cuts precisely help with getting feedback from vertices that are further away, making the cuts more "Bayes-like" and less "MAP-like," as seen in the previous example where v_1 is now assigned to community 2 using 2-power-cuts rather than community 1 using standard cuts. Note that this is also the case when using self-avoiding or nonbacktracking walk counts, but these tend to overcount in the presence of small cycles. For example, in the graph of Figure 4, the count of NB walks is doubled around the 4-cycle; in contrast, graph powering projects the count back to 1, thanks to the nonlinearity $\mathbb{1}(\cdot \geq 1)$.

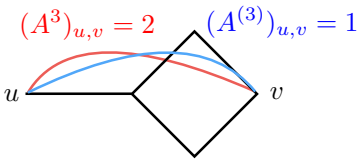


Figure 4. *Classical powering A^3 vs. graph powering $A^{(3)}$. In this example, the number of nonbacktracking walks of length 3 between u and v is 2 and not 1 as for graph powering.*

Powering to homogenize the graph. Powering helps to mitigate degree variations, and more generally density variations in the graph, both with respect to high and low densities. Since the degree of all vertices is raised with powering, density variations in the regions of the graph do not contrast as much. Large degree vertices (as in Figure 1(a)) do not stick out as much and tails (as in Figure 1(b)) are thickened, and the more macroscopic properties of the graph can prevail.

One could probably look for a nonlinearity function that is "optimal" (say for the agreement in the SBM) rather than $\mathbb{1}(\cdot \geq 1)$; however, this is likely to be model dependent, while our choice seems both natural and generic. A downside of powering is that it densifies the graph, so one would ideally combine graph powering with degree normalizations to reduce the power or some graph sparsification (such as in [42]).

Note that powering and sparsifying do not counter each other: powering adds edges to "complete the graph" in sparse regimes where edges should be present, while sparsifying prunes down the graph by adding weights on representative edges. Finally, one may also peel out leaves and paths, and use powered Laplacians to further reduce the powering order; see section SM3 in the supplementary materials for further discussions on the implementations.

2. Main results.

2.1. Overview. We believe that the recent algorithms [6, 9, 32, 36] proved to achieve objective (2) of clustering sparse SBMs fail on objective (3) of being robust to tangles and

cliques in the graphs; this is backed in section SM1 of the supplementary materials. While semidefinite programs are likely to succeed on objective (3)—they were already shown to be robust to a vanishing fraction of edge perturbations in [35]—they do not achieve objective (2) [22, 33, 35] and are also more demanding computationally.[1] Moreover, it is also unclear how appropriate the definition of "weak Ramanujan" from [32] for the matrix of self-avoiding walk counts is; i.e., could the spectral gap be larger for a matrix of self-avoiding walk counts? Thus objective (1) also remains unsettled in this case.

We achieve all three objectives (1), (2), and (3) via a robust clustering algorithm based on graph powering. In particular, for (1) we give an upper bound in Theorem 2.12 on the spectral gap of random regular graphs after powering, which allows us to infer that graph powering on ER graphs achieves a maximal spectral gap, up to a potential exponent offset on the logarithmic factor. For (2), we show in Theorem 2.6 that graph powering allows us to cluster the SBM whenever possible. And for (3), we give evidence in section 2.4 that powering clusters the GBMs down to the optimal threshold, without providing formal proofs.[2] We also implement the different algorithms and in section 3 compare them on SBMs and GBMs, as well as hybrid block models (HBMs) that mix SBMs with GBMs.

Finally, in section SM1 of the supplementary materials, we provide a principled derivation of the graph powering algorithm, starting from the Bayes-optimal estimator and developing graph powering as a correction to the nonbacktracking linearization of belief propagation (BP). And in section SM3 of the supplementary materials we discuss details of implementing graph powering in practice.

We now state the results formally.

2.2. Weak recovery in block models with hidden clusters. We consider different models of random graphs with planted clusters. In each case, an ensemble $M(n)$ provides a distribution on a pair of random variables (X, G), where X is an n-dimensional random vector with i.i.d. components, corresponding to the community labels of the vertices, and G is an n-vertex random graph, connecting the vertices in $V = [n]$ depending on their community labels. The goal is to recover X from G, i.e., to reconstruct the communities by observing the connections. The focus of this paper is on the sparse regime and the weak recovery problem, defined below.

Definition 2.1. *In the case of k communities, an algorithm $\hat{X} : 2^{\binom{[n]}{2}} \to [k]^n$ recovers communities with accuracy $f(n)$ in $M(n)$ if, for $(X, G) \sim M(n)$ and $\Omega_i := \{v \in [n] : X_v = i\}$, $i \in [k]$,*

$$\mathbb{P}\left\{\max_{\pi \in S_k} \frac{1}{k} \sum_{i=1}^{k} \frac{|\{v \in \Omega_i : \pi(\hat{X}_v) = i\}|}{|\Omega_i|} \geq f(n)\right\} = 1 - o(1). \tag{2.1}$$

In the case of roughly equally sized communities, $|\Omega_i| = n/k + o(n)$, an algorithm solves weak recovery if it recovers with accuracy $1/k + \Omega(1)$. This will be the case for this paper.

Definition 2.2. *Let n be a positive integer (the number of vertices), k be a positive integer (the number of communities), $p = (p_1, \ldots, p_k)$ be a probability vector on $[k] := \{1, \ldots, k\}$ (the prior*

[1] It would be interesting to see how the SOS-based algorithm of [20] performs; while its complexity is superpolynomial for achieving the KS threshold for weak recovery, it should afford some robustness to cliques.

[2] Formal proofs would require percolation estimates that depart from the focus of the current paper.

on the k communities), and W be a $k \times k$ symmetric matrix with entries in $[0,1]$ (the connectivity probabilities). The pair (X,G) is drawn under $\mathrm{SBM}(n,p,W)$ *if X is an n-dimensional random vector with i.i.d. components distributed under p, and G is an n-vertex simple graph where vertices i and j are connected with probability W_{X_i,X_j}, independently of other pairs of vertices. We also define the community sets by $\Omega_i = \Omega_i(X) := \{v \in [n] : X_v = i\}$, $i \in [k]$.*

In this paper, we refer to the above as the "general stochastic block model" and use "stochastic block model" for the version with two symmetric and sparse communities, as follows.

Definition 2.3. *(X,G) is drawn under* $\mathrm{SBM}(n,a,b)$ *if $k = 2$, if W takes value a/n on the diagonal and b/n off the diagonal, and if the community prior is $p = \{1/2, 1/2\}$. The ER random graph* $\mathrm{ER}(n,d)$ *with expected degree d arises as a special case when $a = b = d$.*

We also define a Gaussian-mixture GBM with two hidden clusters.

Definition 2.4. *Given a positive integer n and $s,t \geq 0$, we define* $\mathrm{GBM}(n,s,t)$ *as the probability distribution over ordered pairs (X,G) as follows. First, each vertex $v \in V$ is independently assigned a community $X_v \in \{-1,1\}$ with equal probability. Then, each vertex v is independently assigned a location U_v in $\mathbb{R}^2$ according to $\mathcal{N}(X_v \cdot s/2, 0), I_2)$. Finally, we add an edge between every pair of vertices u,v such that $\|U_u - U_v\| \leq t/\sqrt{n}$.*

The parameter s is the separation between the two isotropic Gaussian means, centered at $(-s/2,0)$ and $(s/2,0)$. The scaling $t/\sqrt{n}$ gives a sparse graph with one or two giant components as long as t is a sufficiently large constant [40], which we assume it is.

Note that the expected adjacency matrix of the GBM, conditioned on the vertex labels, has the same rank-2 block structure as the SBM: it takes value $\mathbb{E}(A_{ij}|X_i = X_j)$ for all i,j on the diagonal blocks and $\mathbb{E}(A_{ij}|X_i \neq X_j)$ for all i,j on the off-diagonal block. However, the sampled realizations are very different for the GBM.

The main point of introducing the GBM is to have a simple model that accounts for having many more small cycles than the SBM does. The SBM gives a good framework to understand how to cluster sparse graphs with some degree variations, where "abstract" edges occur frequently, i.e., when connection is not necessarily based on metric attributes (x can be friends with y for a certain reason and y can be friends with z for a different reason, while x and z have nothing in common). These abstract edges turn the SBM into a sparse graph with small diameter, which is an important feature in various applications, sometimes referred to as the "small-world" phenomenon. However, the opposite effect also takes place when connections are based on metric attributes; i.e., if x and y are close, and y and z are close too, then x and z must also be close to some extent. This creates many more small cycles than in the SBM.

One possible way to study the effect of small cycles is to consider adversaries that can modify the graph, for example, monotone adversaries that can add edges inside clusters and remove edges across clusters, or budgeted adversaries that can only alter a certain number of edges. A drawback of such adversarial models is that they typically ensure robustness to many fewer cliques than observed in applications, because the possible worst-case obstructions may not be typical in applications. The goal of the GBM is to have a simple model for geometric

connections and tangles, albeit with the usual downsides of generative models.[3]

Note that for both the SBM and the GBM, defining $X, \hat{X}$ to take values in $\{-1,+1\}^n$, we have that weak recovery is solvable if and only if $|\langle X, \hat{X}(G)\rangle| = \Omega(n)$ with high probability.

2.3. Graph powering for SBM clustering. In order to prove that graph powering achieves weak recovery for the SBM, we will prove that it yields a "spectral separator," defined below.

Definition 2.5 (r-spectral separator for the SBM). *Suppose (X, G) is drawn from* $\mathrm{SBM}(n, a, b)$ *with $(a+b)/2 > 1$. We say that a matrix $M(G)$ with eigenvalues $|\lambda_1(M)| \geq |\lambda_2(M)| \geq \cdots \geq |\lambda_n(M)|$ and corresponding eigenvectors $\phi_1(M), \ldots, \phi_n(M)$ is an r-spectral separator for* $\mathrm{SBM}(n, a, b)$ *if the following hold with high probability:*

A. *If $\left(\frac{a+b}{2}\right) < \left(\frac{a-b}{2}\right)^2$, then*
   1. $\lambda_1(M) \asymp \left(\frac{a+b}{2}\right)^r$,
   2. $\lambda_2(M) \asymp \left(\frac{a-b}{2}\right)^r$,
   3. $|\lambda_3(M)| \leq \left(\frac{a+b}{2}\right)^{r/2} \log(n)^{O(1)}$.

B. *If $\left(\frac{a+b}{2}\right) > \left(\frac{a-b}{2}\right)^2$, then*
   1. $\lambda_1(M) \asymp \left(\frac{a+b}{2}\right)^r$,
   2. $|\lambda_2(M)| \leq \left(\frac{a+b}{2}\right)^{r/2} \log(n)^{O(1)}$.

And rounding $\phi_2(M)$ by the median, dividing the rows into those with above- and below-median sums of the entries (similarly to [32]), achieves weak recovery whenever $\left(\frac{a+b}{2}\right) < \left(\frac{a-b}{2}\right)^2$, i.e., down to the KS threshold.

The following theorem states that graph powering gives a spectral separator and therefore that rounding its second eigenvector solves weak recovery for the SBM.

Theorem 2.6. *Suppose $a, b \geq 0$ are such that $(a+b)/2 \geq 1$, and $r = \varepsilon \log(n)$ is such that $\varepsilon = \Omega(1)$, $\varepsilon \log(a+b)/2 < 1/4$. Let $A^{(r)}$ be the adjacency matrix of the rth graph power of G. Then $A^{(r)}$ is an r-spectral separator for* $\mathrm{SBM}(n, a, b)$.

We illustrate Theorem 2.6 in Figure 5, which demonstrates the effect of graph powering on the spectrum of an SBM for various choices of r.

As a step in our proof, we prove the following conjecture of [32]. A similar proof appears in the concurrent work [43].

Theorem 2.7 (conjectured in [32]). *Suppose $a, b \geq 0$ are such that $(a+b)/2 \geq 1$, and $r = \varepsilon \log(n)$ is such that $\varepsilon = \Omega(1)$, $\varepsilon \log(a+b)/2 < 1/4$. Let $A^{[r]}(G)$ be the r-distance matrix of G ($A^{[r]}_{ij} = 1$ if and only if $d_G(i,j) = r$). Then $A^{[r]}$ is an r-spectral separator for* $\mathrm{SBM}(n, a, b)$.

Note that combining graph powering with cleaning and normalizing, one may also take smaller values of r than in Theorem 2.6, as long as $r = \omega(\log\log(n))$ to surpass the effect of the high degree vertices. We refer the reader to section SM3 of the supplementary materials for further discussion.

[3] Note that this version of the GBM is different from the version studied recently in [2]. The model of [2] introduces the geometry in a different manner: each vertex has a geometric attribute that does not have any community bias, e.g., a point process on $\mathbb{R}^2$ or points uniformly drawn on the torus, and each vertex has an independent equiprobable community label; then the probability that two vertices connect is a function of both their geometric distance and their abstract community labels. In contrast, in our GBM, vertices connect only based on their locations, and these locations are encoding the communities. This makes the GBM simpler to analyze than the model of [2].

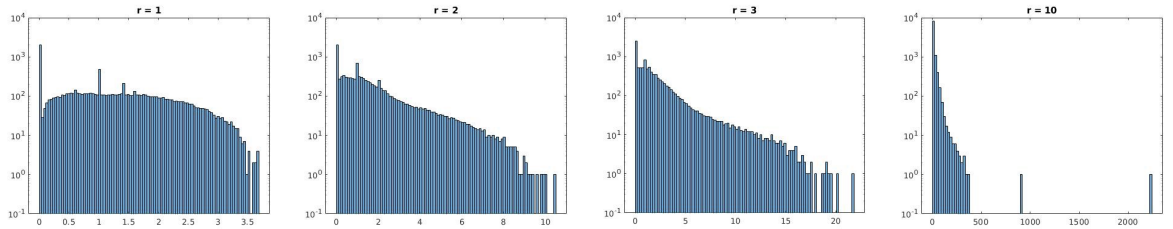


Figure 5. *Spectrum for the absolute eigenvalues of a two-community SBM with $n = 10000$, $a = 4$, $b = 0.2$ for $r = 1, 2, 3, 10$ powers of the graph. For $r = 1, 2$, the second eigenvector is localized, and the best eigenvector for community detection (maximizing agreement) is not the second eigenvector but the third and seventh, respectively. For $r = 3$ onwards, the second eigenvector is optimal.*

2.4. Graph powering for GBM clustering and robustness to tangles. In order to show that spectral clustering based on graph powering is robust to models with many cliques and tangles, we provide evidence that graph powering achieves the fundamental limit for weak recovery in the Gaussian-mixture GBM. We also give a possible proof sketch but leave the result as a conjecture. First, we conjecture the fundamental limit for weak recovery.

Conjecture 2.8. *Let $s, t \geq 0$ such that $\mathrm{GBM}(n, s, t)$ has a giant component. Weak recovery is efficiently solvable in $\mathrm{GBM}(n, s, t)$ if and only if $s > 0$.*

Justification. Obviously, weak recovery is not solvable if $s = 0$, so the claim follows by showing that weak recovery is efficiently solvable if $s > 0$.

The difficult case is if there is a single giant component. In this case, assign uniformly random labels to nongiant vertices. Then, take a vertex uniformly at random among all n' vertices of the giant and assign that vertex to community 1; then assign the $n'/2$ closest vertices in the giant to community 1 and the rest to community 2. With probability $1/2 + \Omega(1)$, this will already solve weak recovery. To succeed with probability $1 - o(1)$, one can pick two anchor vertices in the giant that are at maximal distance, assign each to a different community, and assign the rest of the vertices to the same community as their closest anchor vertex. As soon as $s > 0$, these anchors will not be aligned with the y-axis, giving the result. This requires, however, concentration theorems (such as those found in [48]) for the graph distances in random geometric graphs. These techniques are, however, out of scope for the current paper. ■

We conjecture that graph powering achieves weak recovery on the GBM whenever possible.

Conjecture 2.9. *Let $s, t \geq 0$, and let t be such that $\mathrm{GBM}(n, s, t)$ has a giant component. Taking the second largest eigenvector of the powered adjacency matrix $A^{(r)}$ with $r = \varepsilon \cdot \mathrm{diam}(G)$, for $\varepsilon > 0$ small enough, and rounding it by the median (dividing the rows into those with above- and below-median sums of the entries) solves weak recovery in $\mathrm{GBM}(n, s, t)$ whenever possible.*

We refer the reader to Figure 6 in section 3 for numerical evidence supporting this conjecture, to section SM2 in the supplementary materials for a proof plan, and to section SM3 in the supplementary materials for discussions on how to reduce r, such as to $r = \sqrt{\mathrm{diam}(G)}$.

2.5. Alon–Boppana for graph powering. Finally, we investigate the size of the spectral gap produced by graph powering on ER random graphs. Note first the following statement,

which follows from Theorem 2.6 by setting $a = b = d$.

Corollary 2.10. *Let G be drawn from $ER(n, d)$, and let $A^{(r)}$ be the adjacency matrix of the rth graph power of G and $r = \varepsilon \log(n)$ with $\varepsilon > 0$, $\varepsilon \log(a+b)/2 < 1/4$. Then, with high probability,*

1. $\lambda_1(A^{(r)}) \asymp d^r$;
2. $|\lambda_2(A^{(r)})| \leq \sqrt{d}^r \log(n)^{O(1)}$.

Let us compare this gap to the one of Ramanujan graphs. Recall first that the Alon–Boppana result [38] for d-regular graphs gives

1. $\lambda_1(A) = d$;
2. $\lambda_2(A) \geq (1 - o_{\mathrm{diam}(g)}(1))2\sqrt{d-1}$.

And Ramanujan families of graphs achieve the lower bound of $2\sqrt{d-1}$. Their existence is known from [31], and Friedman [15] proved that random d-regular graphs are almost Ramanujan, i.e., with high probability their second largest eigenvalue satisfies $\lambda_2(A) \leq 2\sqrt{d-1} + o(1)$.

To argue that powering turns $\mathrm{ER}(n, d)$ into a graph of maximal spectral gap (factoring out its irregularity), we first need to understand how large of a spectral gap the powering of any regular graph can have. Powering regular graphs may not necessarily produce regular graphs, so we cannot apply directly the Alon–Boppana bound. Yet, let us assume for a moment that the rth power of a d-regular graph is regular with degree d^r; then Alon–Boppana would give that the second largest eigenvalue of an r-powered graph is larger than $2\sqrt{d^r - 1} \sim 2\sqrt{d}^r$. In contrast, our Corollary 2.10 gives that the rth power of $\mathrm{ER}(n, d)$ has its second largest eigenvalue of order $\sqrt{d}^r \log(n)^{O(1)}$.

This additional logarithmic factor could suggest that powering may not give a tight generalization of Ramanujan and, thus, may not be an "optimal graph regularizer." In [32], for the r-self-avoiding-walk matrix rather than the r-powered graph, this "slack" is absorbed in the terminology of "weak" Ramanujan. However, how weak is this exactly? The above reasoning does not take into account the fact that $G^{(r)}$ is not any regular graph, but a *powered* graph (i.e., not any graph is the power of some underlying graph). So it is still possible that powering must concede more in the spectral gap that the above argument suggests; we next show that this is indeed the case. In [3], it is shown that in fact powering must concede more in the spectral gap that the above argument suggests; the statement is as follows.

Theorem 2.11 (see [3]). *Let $\{G_n\}_{n\geq 1}$ be a sequence of graphs such that $\mathrm{diam}(G_n) = \omega(1)$, and let $\{r_n\}_{n\geq 1}$ be a sequence of positive integers such that $r_n = \varepsilon \cdot \mathrm{diam}(G_n)$. Then,*

$$\lambda_2(G_n^{(r_n)}) \geq (1 - o_\varepsilon(1))(r_n + 1)\hat{d}_{r_n}^{r_n/2}(G_n), \tag{2.2}$$

where

$$\hat{d}_r(G) = \left(\frac{1}{r+1} \sum_{i=0}^{r} \sqrt{\delta^{(i)}(G)\delta^{(r-i)}(G)} \right)^{2/r}, \tag{2.3}$$

$$\delta^{(i)}(G) = \min_{(x,y)\in E(G)} |\{v \in V(G) : d_G(x, v) = i, d_G(y, v) \geq i\}|. \tag{2.4}$$

We will use this result to prove a lower bound on the second-largest eigenvalue of random d-regular graphs after powering.

Theorem 2.12. *Let G be a random d-regular graph (where $d = o(n^{1/10})$, and let $r = \varepsilon \log(n)$, where $\varepsilon \log d < 1/5$. Then, with high probability,*

$$\lambda_2(G^{(r)}) \geq (1 - o_d(1))\varepsilon\sqrt{d}^r \log(n). \tag{2.5}$$

Theorem 2.12 says that even for a random d-regular graph (with large degrees), we could not hope to get a better spectral gap for its $\varepsilon \log(n)$-power than that of an ER random graph (Corollary 2.10), except for the exponent on the logarithmic factor that we do not investigate.

2.6. Derivation of graph powering as a clique-robust linearization of BP. An important contribution of the paper is to provide a derivation of graph powering starting from a Bayes optimal estimator and connecting back to spectral operators. This requires the approximation of posterior distributions using BP and the nonbacktracking linearization. While the main ideas for this development already appear in papers such as [28], we provide in section SM1 of the supplementary materials a slightly different and more detailed account, with a formal description of the symmetry breaking and the influence of nonedges. Having reached the nonbacktracking spectral operator, we describe its weakness on the GBM, which we introduce in this paper as a simple test model for testing robustness to cliques. We then provide both theoretical and experimental evidence for the failure of the spectral nonbacktracking algorithm on the GBM due to cliques. We then proceed to correcting the linearization of BP in order to afford such robustness with the graph powering operator.

We also discuss the alternatives of the Laplacians and normalized Laplacians, with theoretical and experimental evidence for their failures on the SBM. We conclude in section SM3 of the supplementary materials with a spectral meta-algorithm that mixes powering and normalized Laplacians to improve the efficiency, and we also give connections to graph sparsification.

3. Comparisons of algorithms.

In this section, we give numerical results for some of the main algorithms for community detection on the SBM and GBM. For the GBM, see Figure 6. For the SBM, see Figure 7. In both cases, graph powering outperforms the other methods.

We also introduce an HBM, which superposes an SBM and a GBM. The HBM has the advantage of simultaneously having a short diameter, having abundant tangles, and being sparse. In Figure 8, we use the HBM to illustrate the fact that graph powering is robust to the superposition of both types of edges (geometric and abstract edges), while other algorithms suffer on either of the other types of edges.

Definition 3.1. *Suppose we are given a positive integer n and $a, b, s, t, h \geq 0$; then we define* $\mathrm{HBM}(n, a, b, s, t, h_1, h_2)$ *as the probability distribution over ordered pairs (X, G) as follows. Let $(X, G_1) \sim \mathrm{SBM}(n, a, b)$, and let G_2 be drawn from $\mathrm{GBM}(n, s, t)$ with X for the community labels, such that G_1, G_2 are independent conditionally on X. For each pair of vertices u, v, independently keep the edge from G_1 with probability h_1 and the edge from G_2 with probability h_2, and merge the edges if both are kept. Call the resulting graph G.*

4. Proof of weak recovery on SBM.

4.1. Proof of Theorem 2.7. Our proof of Theorem 2.7 uses the following theorem of Massoulié [32], which is analogous to Theorems 2.7 and 2.6, but for the matrix $A^{\{r\}}$ counting self-avoiding walks (i.e., paths) of length r. While this matrix does not offer the desired robustness to cliques, the following result will be used for graph powering in the case of SBMs.

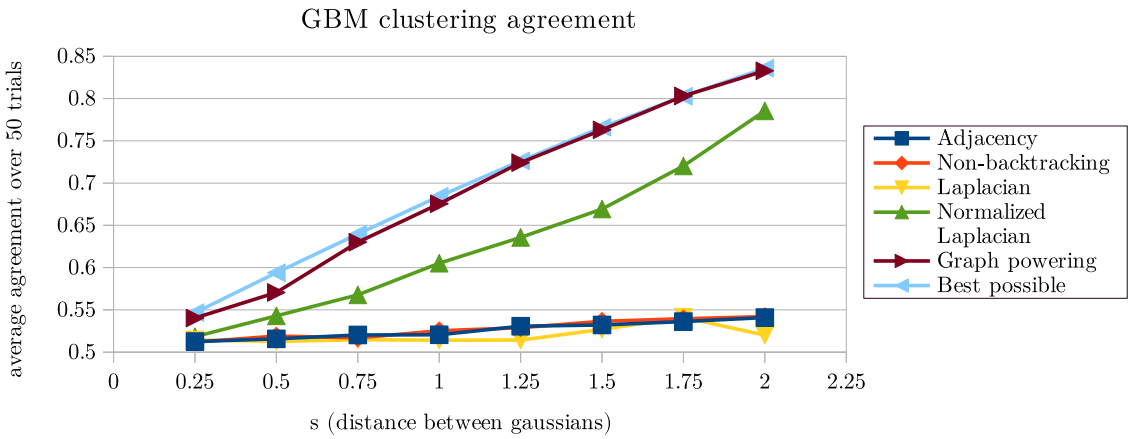


Figure 6. *The truth-agreement values of clusters calculated by different methods on* $G \sim \mathrm{GBM}(n, s, t)$ *with* $n = 1000$, $t = 10$. *Each entry is the average agreement over* 50 *runs. The spectral clustering on the powered adjacency matrix has powering parameter* $r = 0.3d$, *where* d *is the diameter. "Best possible" indicates the information-theoretic upper bound for agreement if the locations of the vertices in the plane were known to the algorithm. Notice that graph powering outperforms all other methods and achieves close-to-optimal community recovery.*

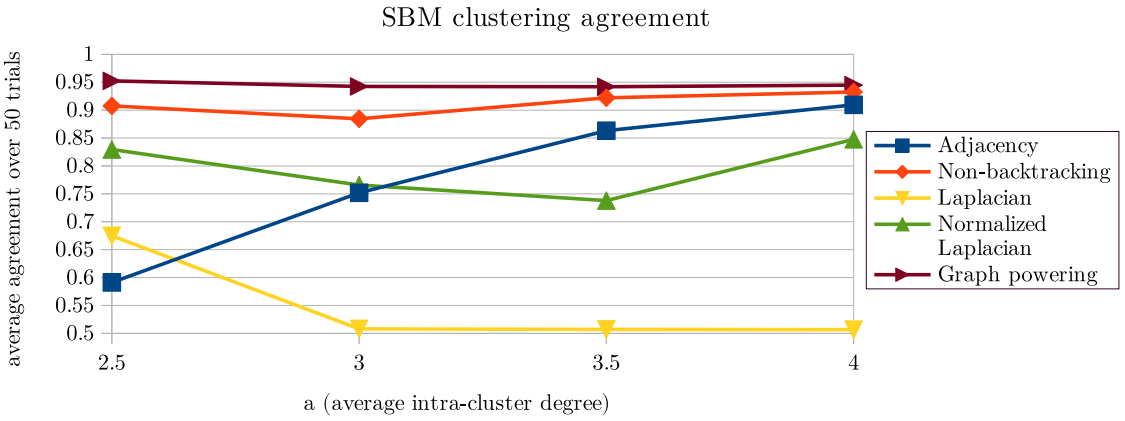


Figure 7. *The* $\mathrm{SBM}(n, a, b)$ *parameters are* $n = 4000$, a *given by the* x*-axis, and* $\mathrm{SNR} = (a-b)^2/(2(a+b))$ *chosen to be very close to* 1 *(the KS threshold for weak recovery). The graph powering parameter is* $r = 0.5d$, *where* d *is the diameter. Although the normalized Laplacian and the adjacency matrix seem to cluster the graph relatively well, when* n *is increased the performance of these methods drops dramatically. For instance, when* $n = 100000$, *the normalized Laplacian method consistently scores below* 0.51 *agreement on the parameter ranges in the above graph. Similarly, for* $n = 100000$, $\mathrm{SNR} \approx 1$ *and* $a \leq 3$, *the adjacency method scores below* 0.55 *agreement.*

Theorem 4.1 (spectral separation for the self-avoiding-walk matrix; proved in [32]). *Suppose* $a, b \geq 0$ *are such that* $(a+b)/2 \geq 1$, *and* $r = \varepsilon \log(n)$ *is such that* $\varepsilon = \Omega(1)$, $\varepsilon \log(a+b)/2 < 1/4$. *Let* $A^{\{r\}}$ *be the length-*r*-self-avoiding-walk matrix of* G ($A^{\{r\}}_{ij}$ *equals the number of self-avoiding*

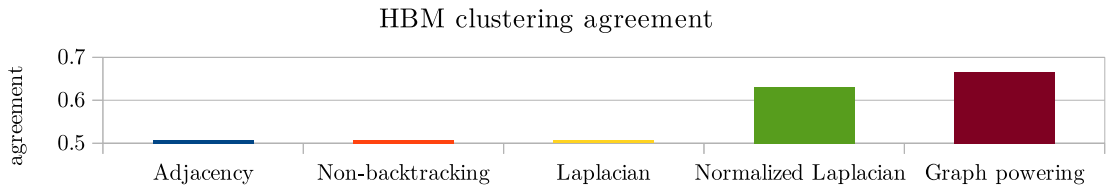


Figure 8. *Average agreement over* 50 *trials for* $\mathrm{HBM}(n, a, b, s, t, h_1, h_2)$ *with parameters* $n = 4000$, $s = 1$, $t = 10$, $h_1 = h_2 = 0.5$, $a = 2.5$, $b = 0.187$. *The graph powering parameter is* $r = 0.4d$, *where* d *is the diameter of the graph. Since the HBM combines elements of the GBM and SBM models, the adjacency, nonbacktracking, and Laplacian methods fail to recover the clusters. We conjecture that for higher* n, *the normalized Laplacian method fails as well (similarly to the SBM case).*

walks of length r *between* i *and* j*). Then* $A^{\{r\}}$ *is an* r*-spectral separator for* $\mathrm{SBM}(n, a, b)$ *in the sense of Definition* 2.5.

Remark 4.2. Theorem 4.1 does not appear in the above form in [32]. In particular, case B of the r-spectral separator definition is not addressed in [32], but it can be proved with the same techniques as case A. The polylogarithmic factors in the bounds of [32] on λ_1 and λ_2 can be removed with a more careful analysis, along the lines of the later work [9] on the spectrum of the nonbacktracking operator. Similarly, the n^ϵ factor in the bound of [32] on λ_3 can be seen to be in fact a $(\log n)^{O(1)}$ factor. Finally, the bounds in [32] are stated for $A^{\{r\}}$ alone, not for $A^{\{k\}}$ for all $k \in \{r/2, \dots, r\}$. However, the proof of [32] shows that there is a constant $C > 0$ such that for each individual $k \in \{r/2, \dots, r\}$ the bounds on the top eigenvalues of $A^{\{k\}}$ hold with probability $\geq 1 - C(\log n)^{-2}$. A union bound over $k \in \{r/2, \dots, r\}$ then gives Theorem 4.1 as stated. (This union bound is needed for the proof of Theorem 2.6 in the next section but is not needed for the proof of Theorem 2.7.)

We will only prove case A of the spectral separation property for Theorem 2.7, since the argument for case B is similar and simpler. Writing $\alpha := (a + b)/2$ and $\beta := (a - b)/2$, the proof can be broken down into two steps:

1. We show that the distance-r indicator matrix $A^{[r]}$ (defined as $A^{[r]}_{ij} = 1$ if and only if $d_G(i, j) = r$) is a small perturbation of the length-r-self-avoiding-walk matrix. In particular, we show that with high probability, the difference $B = B(r) := A^{[r]} - A^{\{r\}}$ has small spectral norm $\|B\|_2 = O(\alpha^{r/2} \log^3 n)$.
2. We use matrix perturbation theory to prove that the top eigenvalues and eigenvectors of $A^{[r]}$ behave like the top eigenvalues and eigenvectors of $A^{\{r\}}$. In the event that the bounds in Theorem 4.1 and step 1 hold, then we have the following:
   (a) $\|B\|_2 = o(\lambda_2(A^{\{r\}}))$, so Weyl's inequality [47] gives
   $$\lambda_1(A^{[r]}) \asymp \lambda_1(A^{\{r\}}) \asymp \alpha^r,$$
   $$\lambda_2(A^{[r]}) \asymp \lambda_2(A^{\{r\}}) \asymp \beta^r,$$
   $$|\lambda_3(A^{[r]})| \leq |\lambda_3(A^{\{r\}})| + \|B\|_2 \leq \alpha^{r/2}(\log n)^{O(1)}.$$
   (b) $\|B\|_2 = o(\max\{\lambda_1(A^{\{r\}}) - \lambda_2(A^{\{r\}}), \lambda_2(A^{\{r\}}) - \lambda_3(A^{\{r\}})\})$, so by the Davis–Kahan theorem [13], $\phi_1(A^{[r]})$ and $\phi_2(A^{[r]})$ are asymptotically aligned with

$\phi_1(A^{\{r\}})$ and $\phi_2(A^{\{r\}})$, respectively, which is enough for rounding by the median, as in [32], to achieve weak recovery. As a reminder, the Davis–Kahan theorem states the following.

Theorem 4.3 (Davis–Kahan theorem). *Suppose that* $\bar{H} = \sum_{j=1}^{n} \bar{\mu}_j \bar{u}_j \bar{u}_j^T$ *and* $H = \bar{H} + E$, *where* $\bar{\mu}_1 \geq \cdots \geq \bar{\mu}_n$, $\|\bar{u}_j\|_2 = 1$, *and* E *is symmetric. Let* u_j *be a unit eigenvector of* H *corresponding to its* j*th largest eigenvalue, and let* $\Delta_j = \min\{\bar{\mu}_{j-1} - \bar{\mu}_j, \bar{\mu}_j - \bar{\mu}_{j+1}\}$, *where we define* $\bar{\mu}_0 = +\infty$ *and* $\bar{\mu}_{n+1} = -\infty$. *We have*

$$\min_{s=\pm 1} \|s u_j - \bar{u}_j\|_2 \lesssim \frac{\|E\|_2}{\Delta_j}, \tag{4.1}$$

where $\lesssim$ *only hides an absolute constant.*

Therefore, it only remains to prove that $\|B\|_2 = O(\alpha^{r/2} \log^3 n)$ with high probability.

To understand the intuition behind our argument, it helps to imagine what would happen if the underlying graph were a tree instead of an SBM. In the tree case, there would be exactly one self-avoiding walk between every pair of vertices, and the length of this walk would be equal to the distance between the two vertices. In other words, in the tree case, the matrices $A^{\{r\}}$ and $A^{[r]}$ would be equal.

While the SBM is (with high probability) not a tree, it is with high probability locally tree-like. This means that for small r, most vertices do not have cycles in their r-neighborhoods. Therefore, most vertices' r-neighborhoods are trees, and hence $A^{\{r\}} \approx A^{[r]}$.

The observation that the SBM is locally tree-like can be formalized.

Lemma 4.4 (e.g., Lemma 4.2 of [32]). *Let* E_1 *be the event that no vertex has more than one cycle in its* r*-neighborhood. For* $r = \varepsilon \log n$ *and* $\varepsilon \log \alpha < 1/4$, E_1 *occurs with high probability.*

Conditioning on E_1, we can define the equivalence relation $\sim$ between vertices so that $v \sim w$ if and only if there is a cycle in the intersection of the r-neighborhoods of v and w. This is a well-defined equivalence relation, because every vertex has at most one cycle in its r-neighborhood. The relation $\sim$ is useful because of item (i) of the following proposition (proof postponed) connecting $\sim$ to the structure of B.

Proposition 4.5. *Condition on* E_1. *Then, for all* $i, j \in V(G)$, *the following hold:*

(i) $B_{i,j} \neq 0 \implies i \sim j$.
(ii) $B_{i,j} \neq 0 \implies$ *there are at least two length-*$(\leq r)$ *paths from* i *to* j.
(iii) $|B_{i,j}| \leq 1$.
(iv) *There are at most two length-*$(\leq r)$ *paths from* i *to* j.

We condition on E_1 in the rest of the proof, since it holds with high probability. By item (i) of Proposition 4.5, B is a block-diagonal matrix, where each block $B_{S\times S}$ corresponds to an equivalence class $S \subseteq [n]$ of $\sim$. Therefore, it suffices to separately bound the spectral norm of each block $B_{S\times S}$. To do this, we introduce the following event.

Lemma 4.6 (Theorem 2.3 of [32]). *Let* $E_2 = E_2(C)$ *be the event that for all vertices* $i \in V(G)$, *for all* $t \in [r]$, *the following holds:*

$$|\{j : d_G(i,j) \leq t\}| \leq C(\log n)^2 \alpha^t. \tag{4.2}$$

There is large enough C that $E_2(C)$ holds with high probability.

Informally, E_2 is the event that for all $t \in [r]$, each vertex's t-neighborhood is not much larger than α^t.

From now on, also condition on E_2, since it holds with high probability. Suppose S is the set of vertices in some equivalence class of $\sim$. Let $H \subseteq G$ be the cycle that is shared by the r-neighborhoods of the vertices in S. Let e be an edge of H. Then for every $i, j \in S$ such that there are two length-$(\leq r)$ self-avoiding walks from i to j, at least one of the paths must contain e. Otherwise, the cycle H is not the only cycle in the r-neighborhood of i. So by item (ii) of Proposition 4.5,

$$|\{\{i,j\} \in S \mid B_{i,j} \neq 0\}| \leq |\{(\leq r)\text{-length paths } P \subseteq G \mid e \in E(P)\}|.$$

For any $t \in [r]$,

$$|\{t\text{-length paths } P \subseteq G \mid e = (u,v) \in E(P)\}|$$

$$\leq \sum_{l=0}^{r} |\{l\text{-length paths } P \subseteq G \mid u \in V(P)\}| \cdot |\{(t-l-1)\text{-length paths } P \subseteq G \mid v \in V(P)\}|.$$

By item (iv) of Proposition 4.5 and by E_2, for any $u \in V(G)$, $l \in [r]$,

$$|\{l\text{-length paths } P \subseteq G \mid u \in V(P)\}| \leq 2C(\log n)^2 \alpha^l,$$

so

$$|\{\{i,j\} \in S \mid B_{i,j} \neq 0\}| \leq r(2C(\log n)^2)^2 \alpha^{r-1} = O(\alpha^r \log^5 n).$$

Therefore, by item (iii) of Proposition 4.5,

$$\|B_{S\times S}\|_2 \leq \|B_{S\times S}\|_F = O(\alpha^{r/2} \log^{5/2} n),$$

as desired. ($\|\cdot\|_F$ denotes the Frobenius norm.)

Proof of Proposition 4.5. Suppose $B_{i,j} \neq 0$. Since every vertex has at most one cycle in its r-neighborhood, there are at most 2 length-$(\leq r)$ paths between every pair of vertices (item (iv)), so $A^{\{r\}}_{i,j} \in \{0,1,2\}$. Also, since $A^{[r]}_{i,j} \in \{0,1\}$, the possible cases are the following:

1. $A^{[r]}_{i,j} = 0$:
   (a) $A^{\{r\}}_{i,j} = 1$. There is a path of length $< r$ between i and j, because otherwise $d_G(i,j) = r$. So there are two paths of length $\leq r$ between i and j.
   (b) $A^{\{r\}}_{i,j} = 2$. This case is impossible. There is no path of length $< r$ between i and j, because there are at most two paths of length $\leq r$ between i and j, and $A^{\{r\}}_{i,j} = 2$ tells us that there are two paths of length r between i and j. Therefore, $d_G(i,j) = r$, so $A^{[r]}_{i,j} = 1$.
2. $A^{[r]}_{i,j} = 1$:
   (a) $A^{\{r\}}_{i,j} = 0$. This case is impossible. The distance between i and j is r, so there should be a path of length r between them.
   (b) $A^{\{r\}}_{i,j} = 2$. There are two paths of length r between i and j.

So if $B_{i,j} \neq 0$, then $|B_{i,j}| = 1$, and there are exactly two ($\leq r$)-length paths between i and j. This case analysis proves items (ii) and (iii) of the claim.

The union of the two paths from i to j contains a simple cycle which is contained in the r-neighborhoods of both i and j. Therefore, $i \sim j$, proving item (i) of the claim. ∎

4.2. Proof of Theorem 2.6. Recall the definitions $A^{(r)}_{ij} = \mathbb{1}(d_G(i,j) \leq r)$ and $A^{[r]}_{ij} = \mathbb{1}(d_G(i,j) = r)$. The key to proving Theorem 2.6 is the identity

$$A^{(r)} = \sum_{k=0}^{r} A^{[k]},$$

which is just another way to write $\mathbb{1}(d_G(i,j) \leq r) = \sum_{k=0}^{r} \mathbb{1}(d_G(i,j) = k)$.

Informally, we know from Theorem 2.7 that the top eigenvalues of the $A^{[k]}$ matrices have the desired separation properties. So if we can prove that the top eigenvectors of the $A^{[k]}$ matrices are all roughly equal, then they will be roughly equal to the top eigenvectors of $A^{(r)}$, essentially proving Theorem 2.6.

Keeping this intuition in mind, our first step is to reduce the problem of analyzing $A^{(r)}$ to the problem of analyzing a slightly simpler matrix D:

$$D = D(r) := \sum_{k=r/2}^{r} A^{\{k\}}; \quad \text{equivalently, } A^{(r)} = D + A^{(r/2-1)} - \sum_{k=r/2}^{r} (A^{\{k\}} - A^{[k]}),$$

where $A^{\{r\}}_{ij}$ counts the number of self-avoiding walks of length r between i and j. By the proof of Theorem 2.6, we know that conditioned on $E_1 \cap E_2$ (and so with high probability),

$$\left\| \sum_{k=r/2}^{r} A^{[k]} - A^{\{k\}} \right\|_2 \leq \sum_{k=0}^{r} \|A^{[k]} - A^{\{k\}}\|_2 = O(\alpha^{r/2} (\log n)^4).$$

And since by Lemma 4.6 the neighborhoods of vertices do not grow too quickly, with high probability, $A^{(r/2-1)}$ is the adjacency matrix of a graph with maximum degree $O(\alpha^{r/2}(\log n)^2)$. Under this event, we also get the following bound:

$$\|A^{(r/2-1)}\|_2 = O(\alpha^{r/2}(\log n)^2).$$

We can conclude by the triangle inequality that $\|A^{(r)} - D\|_2 = O(\alpha^{r/2}(\log n)^4)$. Therefore, by the matrix perturbation arguments (Weyl's inequality and the Davis–Kahan inequality) used to prove Theorem 2.7, it suffices to prove that the matrix D has the spectral properties that we desire for $A^{(r)}$. Theorem 2.6 will follow. We will now show that D has a "weak Ramanujan property", similar to Theorem 2.4 of [32].

Lemma 4.7. *With high probability, D satisfies the following weak Ramanujan property:*

$$\sup_{\|u\|_2 = 1, u^T A^{\{r\}} \mathbb{1} = u^T A^{\{r\}} X = 0} \|Du\|_2 = \alpha^{r/2} (\log n)^{O(1)},$$

where $\mathbb{1}$ is the all-ones vector and $X \in \{-1, +1\}^n$ is the community label vector.

We will also need the following two lemmas. The first lemma tells us that the top eigenvectors of the $A^{\{k\}}$ matrices are pretty well aligned with the top eigenvectors of $A^{\{r\}}$.

Lemma 4.8. *There are $c_0, \delta > 0$ such that with high probability, for all $k \in \{r/2, \dots, r\}$,*

$$\left\| \frac{A^{\{r\}}1}{\|A^{\{r\}}1\|_2} - \frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} \right\|_2 \leq c_0 n^{-\delta},$$

$$\left\| \frac{A^{\{r\}}X}{\|A^{\{r\}}X\|_2} - \frac{A^{\{k\}}X}{\|A^{\{k\}}X\|_2} \right\|_2 \leq c_0 n^{-\delta}.$$

Lemma 4.9 (from [32], with minor modification as in Remark 4.2). *There are $c_1, c_2 > 0$ and $g = o(1)$ such that with high probability, for all $k \in \{r/2, \dots, r\}$, the following hold:*

(i) $c_1\alpha^k < \|A^{\{k\}}\|_2 < c_2\alpha^k$.

(ii) $A^{\{k\}}A^{\{k\}}1 = \|A^{\{k\}}A^{\{k\}}1\|_2 \left(\frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} + h_k\right)$ *for a vector h_k s.t. $\|h_k\|_2 < g = o(1)$.*

(iii) $A^{\{k\}}A^{\{k\}}X = \|A^{\{k\}}A^{\{k\}}X\|_2 \left(\frac{A^{\{k\}}X}{\|A^{\{k\}}X\|_2} + h'_k\right)$ *for a vector h'_k s.t. $\|h'_k\|_2 < g = o(1)$.*

We now show that Lemmas 4.7, 4.8, and 4.9 imply Theorem 2.6.

Following the argument of Theorem 4.1 of [32], it suffices to show that with high probability,

$$\|DA^{\{r\}}1\|_2 = \Theta(\alpha^r \|A^{\{r\}}1\|_2), \tag{4.3}$$

$$\|DA^{\{r\}}X\|_2 = \Theta(\beta^r \|A^{\{r\}}X\|_2). \tag{4.4}$$

Since $A^{\{r\}}1$ and $A^{\{r\}}X$ are asymptotically orthogonal (by Lemma 4.4 of [32]), and since D has the weak Ramanujan property of Lemma 4.7, the variational definition of eigenvalues yields that the top two eigenvectors of D will be asymptotically in the span of $A^{\{r\}}1$ and $A^{\{r\}}X$.[4] By the lower bound of (4.3) and the upper bound of (4.4), the top eigenvalue of D will be $\Theta(\alpha^r)$, with eigenvector asymptotically parallel to $A^{\{r\}}1$. Since $A^{\{r\}}X$ is asymptotically orthogonal to $A^{\{r\}}1$, the second eigenvalue of D will be $\Theta(\beta^r)$, with eigenvector asymptotically parallel to $A^{\{r\}}X$. This proves the theorem, since by Massoulié [32] $A^{\{r\}}1$ and $A^{\{r\}}X$ are in fact asymptotically parallel to the top two eigenvectors of $A^{\{r\}}$.

The inequalities in (4.3) hold because with high probability,

$$\frac{\|DA^{\{r\}}1\|_2}{\|A^{\{r\}}1\|_2} = \left\| \sum_{k=r/2}^{r} A^{\{k\}} \frac{A^{\{r\}}1}{\|A^{\{r\}}1\|_2} \right\|_2 \tag{4.5}$$

$$= \left\| \sum_{k=r/2}^{r} A^{\{k\}} \frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} \right\|_2 + O\left(\sum_{k=r/2}^{r} \|A^{\{k\}}\|_2 \left\| \frac{A^{\{r\}}1}{\|A^{\{r\}}1\|_2} - \frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} \right\|_2 \right) \tag{4.6}$$

$$= \left\| \sum_{k=r/2}^{r} A^{\{k\}} \frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} \right\|_2 + O\left(\sum_{k=0}^{r} \alpha^k n^{-\delta} \right) \tag{4.7}$$

[4] Here we assume we are in the case $\beta^2 > \alpha$ of Theorem 2.6, since the other case is similar.

$$= \left\| \sum_{k=r/2}^{r} A^{\{k\}} \frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} \right\|_2 + O(\alpha^r n^{-\delta}) \tag{4.8}$$

$$= \left\| \sum_{k=r/2}^{r} \left(\frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} + h_k \right) \frac{\|A^{\{k\}}A^{\{k\}}1\|_2}{\|A^{\{k\}}1\|_2} \right\|_2 + O(\alpha^r n^{-\delta}) \quad \text{for } \|h_k\|_2 = o(1) \tag{4.9}$$

$$\begin{aligned} &= \left\| \sum_{k=r/2}^{r} \left(\frac{A^{\{r\}}1}{\|A^{\{r\}}1\|_2} + h_k \right) \frac{\|A^{\{k\}}A^{\{k\}}1\|_2}{\|A^{\{k\}}1\|_2} \right\|_2 \\ &\quad + O\left(\sum_{k=r/2}^{r} \frac{\|A^{\{k\}}A^{\{k\}}1\|_2}{\|A^{\{k\}}1\|_2} \left\| \frac{A^{\{r\}}1}{\|A^{\{r\}}1\|_2} - \frac{A^{\{k\}}1}{\|A^{\{k\}}1\|_2} \right\|_2 \right) + O(\alpha^r n^{-\delta}) \end{aligned} \tag{4.10}$$

$$= \left\| \sum_{k=r/2}^{r} \left(\frac{A^{\{r\}}1}{\|A^{\{r\}}1\|_2} + h_k \right) \frac{\|A^{\{k\}}A^{\{k\}}1\|_2}{\|A^{\{k\}}1\|_2} \right\|_2 + O(\alpha^r n^{-\delta}) \tag{4.11}$$

$$= \left(\sum_{k=r/2}^{r} \frac{\|A^{\{k\}}A^{\{k\}}1\|_2}{\|A^{\{k\}}1\|_2} \right) (1 + o(1)) + O(\alpha^r n^{-\delta}). \tag{4.12}$$

Equations (4.6) and (4.10) are derived by the triangle inequality. Equations (4.7) and (4.11) are consequences of Lemma 4.8. Equation (4.9) follows by Lemma 4.9. Plugging in the bound on $\frac{\|A^{\{k\}}A^{\{k\}}1\|_2}{\|A^{\{k\}}1\|_2}$ from item (i) of Lemma 4.9 gives

$$\|DA^{\{r\}}1\|_2 = \Theta(\|A^{\{r\}}1\|_2),$$

proving (4.3). A similar argument proves (4.4). We conclude by proving the auxiliary lemmas.

Proof of Lemma 4.7. Let $\delta > 0$. By the triangle inequality and the definition $D = \sum_{k=r/2}^{r} A^{\{k\}}$, it suffices to prove that for all $k \in \{r/2, \dots, r\}$,

$$\sup_{\|u\|_2=1, u^T A^{\{r\}}1 = u^T A^{\{r\}}X = 0} \|A^{\{k\}}u\|_2 \le (\log n)^{O(1)} \alpha^{r/2},$$

where the bound is uniform over k. This follows from a union bound over Theorem 2.4 in [32] (the weak Ramanujan property for $A^{\{r\}}$). ∎

Proof of Lemma 4.8. Let $\mathcal{B} \subset V(G)$ denote the set of vertices v such that there is a cycle in the r-neighborhood of v. Lemma 4.3 of [32] gives us the following bounds on entries of $A^{\{k\}}1$, $A^{\{k\}}X$ for all $k \in [r]$:

$$v \notin \mathcal{B} \implies (A^{\{k\}}1)_v = \alpha^{k-r}(A^{\{r\}}1)_v + O(\log n) + O\left(\sqrt{\log(n)\alpha^k}\right), \tag{4.13}$$

$$v \notin \mathcal{B} \implies (A^{\{k\}}X)_v = \beta^{k-r}(A^{\{r\}}X)_v + O(\log n) + O\left(\sqrt{\log(n)\alpha^k}\right), \tag{4.14}$$

$$v \in \mathcal{B} \implies \|(A^{\{k\}}1)_v\|_2 = O(\alpha^k \log(n)), \tag{4.15}$$

$$v \in \mathcal{B} \implies \|(A^{\{k\}}X)_v\|_2 = O(\beta^k \log(n)). \tag{4.16}$$

By Lemma 4.2 of [32], $|\mathcal{B}| = O(\alpha^{2r} \log^4 n)$ with high probability, so by (4.14) and (4.16),

$$\begin{aligned}\langle A^{\{r\}}X, A^{\{k\}}X\rangle &= \left(\sum_{v\notin\mathcal{B}}(A^{\{r\}}X)_v^2\beta^{k-r} + (A^{\{r\}}X)_v(O(\log n + \sqrt{\alpha^k \log n}))\right) + \sum_{v\in\mathcal{B}} O(\beta^{k+r}\log n)\\ &= \beta^{k-r}\|A^{\{r\}}X\|_2^2 + O(n\beta^r\sqrt{\alpha^k}\log^2 n) + O(\alpha^{2r}\beta^{k+r}\log^4 n).\end{aligned}$$

Noting that $\beta^2 > \alpha$, and that $\|A^{\{r\}}X\|_2^2 = \Theta(n\beta^{2r})$ up to a factor of $\log^2 n$, there is $\delta > 0$ such that

$$\langle A^{\{r\}}X, A^{\{k\}}X\rangle = \beta^{k-r}\|A^{\{r\}}X\|_2^2(1 + o(n^{-\delta})). \tag{4.17}$$

Similarly, we can show that $\|A^{\{k\}}X\|_2 = \beta^{k-r}\|A^{\{r\}}X\|_2(1 + o(n^{-\delta}))$. This implies that

$$\left\|\frac{A^{\{r\}}X}{\|A^{\{r\}}X\|_2} - \frac{A^{\{k\}}X}{\|A^{\{k\}}X\|_2}\right\|_2^2 = 2 - 2\frac{\langle A^{\{r\}}X, A^{\{k\}}X\rangle}{\|A^{\{k\}}X\|_2\|A^{\{r\}}X\|_2} = 2 - 2(1 + o(n^{-\delta'})) = o(n^{-\delta'})$$

for some $\delta' > 0$. Similar arguments, using (4.13) and (4.15), prove the analogous result for $\langle A^{\{k\}}1, A^{\{r\}}1\rangle$. ∎

Proof of Lemma 4.9. Item (i) is the statement from Theorem 4.1 that $\lambda_1(A^{\{k\}}) = \Theta(\alpha^k)$, with the additional subtlety that we can choose uniform constants in the Θ notation for $k \in \{r/2, \dots, r\}$. Items (ii) and (iii) are equivalent to stating that for all $k \in \{l/2, \dots, l\}$, $A^{\{k\}}A^{\{k\}}1$ is asymptotically in the same direction as $A^{\{k\}}1$, and $A^{\{k\}}A^{\{k\}}X$ is asymptotically in the same direction as $A^{\{k\}}X$. A union bound over Theorem 4.1 of [32] implies this is true for all $k \in \{r/2, \dots, r\}$. ∎

5. Justification for weak recovery on GBM. We now provide justification for Conjecture 2.9, that rounding the second eigenvector of $A^{(r)}$ for $r = \varepsilon \cdot \operatorname{diam}(G)$ for $\varepsilon > 0$ small enough solves weak recovery for $G \sim \mathrm{GBM}(n, s, t)$ whenever possible. Let $\kappa := \operatorname{diam}(G)$.

Justification. If graphs drawn from $\mathrm{GBM}(n, s, t)$ have no giant component with high probability, then recovering communities on $\mathrm{GBM}(n, s, t)$ with accuracy greater than $1/2$ is impossible, so the algorithm trivially recovers communities with optimal accuracy. If graphs drawn from $\mathrm{GBM}(n, s, t)$ typically have two giant components, then the two eigenvectors of $A^{(\varepsilon\kappa)}$ with the largest eigenvalues will be the ones that have positive entries for every vertex in one giant component and all other entries set to 0. So, the algorithm will assign all vertices in one giant component to one community, which attains optimal accuracy. That leaves the case where graphs drawn from $\mathrm{GBM}(n, s, t)$ have a single giant component with high probability. For the rest of this argument, assume that we are in this case. Any small component of the graph will have $o(n)$ vertices, so any eigenvector with nonzero entries corresponding to vertices in such a component must have an eigenvalue of $o(n)$. There are multiple cliques of

size $\Omega(n)$ in the giant component that do not have any edges between them, so the second largest eigenvalue of $A^{(\varepsilon\kappa)}$ is $\Omega(n)$. Therefore, the eigenvector of $A^{(\varepsilon\kappa)}$ with the second largest eigenvalue will have all of its entries corresponding to vertices outside the giant component set to 0. From now on, ignore all vertices outside of the giant component of the graph.

Let w be the unit eigenvector of $A^{(\varepsilon\kappa)}$ with the largest eigenvalue. Every entry in $A^{(\varepsilon\kappa)}$ is nonnegative, so w has all nonnegative entries. Furthermore, there is a path between any two vertices of a graph drawn from $\mathrm{GBM}(n,s,t)$, so every entry in $[A^{(\varepsilon\kappa)}]^n$ is positive, which means every entry in w is positive. Now, let λ be the eigenvalue corresponding to w and let w' be a unit vector. Also, let $E^{(\varepsilon\kappa)}$ be the set of all pairs of vertices that have a path of length $\varepsilon\kappa$ or less between them. Then it must be the case that

$$\begin{aligned}\lambda &= \sum_v \lambda (w'_v)^2 \\ &= \sum_v \lambda (w'_v)^2 \sum_{v':(v,v')\in E^{(\varepsilon\kappa)}} w_{v'}/(\lambda w_v) \\ &= \sum_{(v,v')\in E^{(\varepsilon\kappa)}} (w'_v)^2 \frac{w_{v'}}{w_v}.\end{aligned}$$

This means that

$$\begin{aligned}w' \cdot A^{(\varepsilon\kappa)} w' &= \sum_{(v,v')\in E^{(\varepsilon\kappa)}} w'_v w'_{v'} \\ &= \lambda + \sum_{(v,v')\in E^{(\varepsilon\kappa)}} w'_v w'_{v'} - \frac{w_{v'}}{2w_v}(w'_v)^2 - \frac{w_v}{2w_{v'}}(w'_{v'})^2 \\ &= \lambda - \sum_{(v,v')\in E^{(\varepsilon\kappa)}} \frac{w_v w_{v'}}{2}\left(\frac{w'_v}{w_v} - \frac{w'_{v'}}{w_{v'}}\right)^2.\end{aligned}$$

If w' is the eigenvector of second greatest eigenvalue, then it must be orthogonal to w, and for any other unit vector w'' that is orthogonal to w, it must be the case that

$$\sum_{(v,v')\in E^{(\varepsilon\kappa)}} \frac{w_v w_{v'}}{2}\left(\frac{w'_v}{w_v} - \frac{w'_{v'}}{w_{v'}}\right)^2 \le \sum_{(v,v')\in E^{(\varepsilon\kappa)}} \frac{w_v w_{v'}}{2}\left(\frac{w''_v}{w_v} - \frac{w''_{v'}}{w_{v'}}\right)^2.$$

On another note, the orthogonality of w and w' implies that w' must have both positive and negative entries. Also, since $w \cdot w = w' \cdot w'$, there must exist v such that $|w'_v/w_v| \ge 1$. So, there must be a significant amount of variation in the value of $\frac{w'_v}{w_v}$, but the values of $\frac{w'_v}{w_v}$ must tend to be similar for nearby vertices. Also, vertices that are particularly close to each other geometrically will have mostly the same neighbors in terms of $E^{(\varepsilon\kappa)}$, which will result in them having similar values. So, we would expect that w'_v/w_v would be strongly positive on one side of the graph, that it would be strongly negative on the other, and that it would shift gradually between these extremes as one moved from one side to the other. Geometrically, the x-direction is the direction the giant component extends the farthest in, so we would expect that these sides would be defined in terms of x-coordinates.

Intuitively, it seems like the entries of w' would switch signs halfway between these sides, which means at the y-axis. However, we need to consider the possibility that random variation in vertices would disrupt the symmetry in a way that prevents this from happening. Given a vertex v at a given geometric location that gives it a nonvanishing probability of being in the giant component, for $r' < \varepsilon\kappa$ the expected number of vertices exactly r' edges away from v in G is $\Theta(r')$, and the expected number of vertices within $\varepsilon\kappa$ edges of v is $\Theta(n)$. Now, consider the effects of deleting a random vertex v' on the set W of vertices within $\varepsilon\kappa$ edges of v. It is possible that deleting v' disconnects v from the giant component entirely, in which case its deletion removes nearly the entire set.

Now, assume that this does not happen. Some of the edges of v' might be cut edges, but the components their removal cuts off from the giant component will typically have $O(1)$ vertices. We would expect that any two edges of v' other than cut edges would be contained in some cycle. Furthermore, due to the abundance of small cycles in the GBM, we would generally expect that there exists some m such that every such pair of edges is contained in a cycle of length at most m and $E[m^2] = O(1)$. That means that for any v'' that is still in the giant component, the length of the shortest path from v to v'' will be at most $m-4$ edges longer than it was before v' was deleted. So, only vertices that were more than $\varepsilon\kappa - (m-4)$ edges away from v are in danger of being removed from W. Furthermore, a minimum length path from v to v'' will only pass through one vertex that is r' edges away from v for each r'. So, if v' was r' edges away from v, there is only an $O(1/r')$ chance that the length of the shortest path from v to v'' is even affected by the deletion of v'. That means that given that v' is r' edges away from v, and deleting it does not remove v from the giant component, the expected number of vertices removed from W by deleting it is $O(\sqrt{n}/r')$. A random vertex is r' edges away from v with probability $\Theta(r'/n)$. So, the expected value of the square of the number of vertices removed from W by deleting v' is $O\left(\sum_{r'=1}^{\varepsilon\kappa}(r'/n)\cdot(\sqrt{n}/r')^2\right) = O(\log(n))$. By the same token, the expected value of the square of the number of vertices added to W by adding one new vertex at random is $O(\log(n))$. That means that if we add $(n-1)$ new vertices to the graph and delete all old vertices other than v, the variance in the size of W conditioned on v still being in the giant component will be $O(n\log(n))$. So, heuristically the variance of the size of W conditioned on the geometric location of v and the assumption that v is in the giant component is $O(n\log(n))$.

That means that the standard deviation of the size of this set, and the size of the subset restricted to a given geometric region, is much smaller than the expected size of the set. Furthermore, all entries of $(A^{(\varepsilon\kappa)})^{\lceil 1/\varepsilon\rceil}$ are positive, and we would expect that a power iteration method on $(A^{(\varepsilon\kappa)})$ would only need a constant number of steps to obtain a reasonable approximation of w'. So, random variation in vertex placement has little effect on the behavior of $(A^{(\varepsilon\kappa)})$ and we expect that the signs of the entries of w' have $1-o(1)$ correlation with the signs of the corresponding vertices' x-coordinates. So, this algorithm would essentially assign the vertices with positive x-coordinates to one community and the vertices with negative x-coordinates to the other. That is the best one can do to classify vertices in the GBM, so we believe this algorithm will classify vertices with optimal accuracy. ∎

6. Proof of maximality of spectral gap on ER.

Proof of Theorem 2.12. Let G be a random d-regular graph on n vertices. Let r be a

positive integer with $r < \frac{\log(n)}{5\log(d)}$, and let $x \in V$. We will first compute a probabilistic bound on the modified minimum degree $\delta^{(r)}$, as defined in the statement of Theorem 2.11.

We want to bound the probability of the event E that there exists a vertex x for which $B_r(x)$ contains two different cycles. A cycle in $B_r(x)$ occurs when two vertices in $S_i(x)$ (for $1 \le i \le r$) are adjacent or when a vertex in $S_i(x)$ (for $2 \le i \le r$) has two neighbors in S_{i-1}. If $B_r(x)$ contains two cycles, taking the described edge in $B_r(x)$ and the shortest paths from those vertices to x, we find those cycles are part a subgraph of $B_r(x)$ of one of the following descriptions: (i) a c_1-cycle and a c_2-cycle joined by an l-path, where $c_i \le 2r+1$ and $l + \lfloor c_1/2 \rfloor + \lfloor c_2/2 \rfloor \le 2r$, or (ii) a c_1-cycle and a c_2-cycle sharing a path of l vertices, where $c_i \le 2r+1$ and $l \le r$. Observe that the total number of such subgraphs up to isomorphism is $\Theta(r^3)$ and that every such subgraph $H = (V_H, E_H)$ has $|E_H| \le 4r+2$ and $|V_H| = |E_H| - 1$. A standard result (see [26, Corollary 2.2]) of random regular graphs tells us that the expected number of instances of any such graph H in a random regular graph is $E[\#H] = \Theta(d^{|E_H|}/n)$; it follows from taking a union bound that $P(E) \le \Theta(d^{4r+2}r^3/n)$. Because $r < \frac{\log n}{5\log d}$, $P(E) =\le \Theta(d^2 r^3/n^{1/5}) = o_n(1)$, so with high probability no ball $B_r(x)$ around any vertex x contains two cycles.

In the high probability case $\overline{E}$, we can tightly bound $\delta^{(r)}$. The assumption that $B_r(x)$ contains at most one cycle means that it closely resembles a tree; this makes the following computation of $\delta^{(r)}$ straightforward.

Let $x \in V$. There are $d(d-1)^{r-1}$ nonbacktracking walks of length r starting at x. For any $y \sim x$, exactly $(d-1)^{r-1}$ of those walks start $x, y, \ldots$. Because a vertex at distance r from x and $r-1$ from y must be the endpoint of such a walk, there are at most $(d-1)^{r-1}$ of those vertices. A cycle can occur in $B_r(x)$ in one of two ways.

First, an even cycle means that there is some unique vertex z in the cycle that maximizes $d_G(x,z)$. There are exactly two nonbacktracking walks of length $d_G(x,z)$ from x to z, and these walks have different second-last vertices. Let $i = d_G(x,z)$; there are $(d-2)(d-1)^{r-i-1}$ ways to extend those paths to length r so that the endpoint is in $S_r(x)$. For any of the $d(d-1)^{i-1}-2$ other nonbacktracking walks of length i, there are $(d-1)^{r-i}$ ways to extend them to distance r, each corresponding to a unique vertex in $S_r(x)$. In total we find $|S_r(x)| = d(d-1)^{r-1} - d(d-1)^{r-i-1}$. Because $i \ge 2$, $|S_r(x)| \ge d(d-1)^{r-1} - d(d-1)^{r-3}$ if there is an even cycle.

Second, an odd cycle means that there is a unique pair of two adjacent vertices z_1, z_2 with $j = d_G(x,z_1) = d_G(x,z_2)$. Each of those vertices can be extended to a nonbacktracking walk ending in a unique vertex of $S_r(x)$ in $(d-2)(d-1)^{r-j-1}$ different ways. For any of the $d(d-1)^{j-1} - 2$ other nonbacktracking walks of length j, there are $(d-1)^{r-j}$ ways to extend them to distance r, each corresponding to a unique vertex in $S_r(x)$. In total we have $|S_r(x)| = d(d-1)^{r-1} - 2(d-1)^{r-j-1}$. Because $j \ge 1$, it follows that $|S_r(x)| \ge d(d-1)^{r-1} - 2(d-1)^{r-2}$.

As there is at most one cycle which is either odd or even, we find the overall bound $|S_r(x)| \ge d(d-1)^{r-1} - 2(d-1)^{r-2}$. Because there are at most $(d-1)^{r-1}$ vertices in $S_r(x)$ at distance $r-1$ from y, we are left with $\delta^{(r)} \ge (d-1)^r - 2(d-1)^{r-2}$. It is clearly the case that $\delta^{(r)} \le (d-1)^r$, so with high probability $\delta^{(r)} = (1 + o_d(1))\,(d-1)^r$.

Now, we need to complete the proof by citing Theorem 2.11 with this bound for $\delta^{(r)}$. Observe that in the high probability case, for all $0 \le i \le r$, $\delta^{(i)}\delta^{(r-i)} = (1+o_d(1))(d-1)^r$.

The statement of Theorem 2.11 is that in that case,

$$\lambda_2(G^{(r)}) \geq (1 - o_d(1))(r+1)\sqrt{d}^r. \tag{6.1}$$

The final result follows from the choice of $r = \varepsilon \log(n)$. ∎

7. Powering weighted graphs. In this paper, we assumed that our original graph is unweighted. It is a natural question to investigate how to generalize graph powering to weighted graphs. We make here a few remarks.

To use our algorithms to recover communities on a weighted graph, we would first have to consider what the weights mean. The simplest possibility would be that the degree of evidence that an edge provides that its vertices are in the same community is proportional to its weight. Assume for simplicity that the weights are positive and have this property. Powering would need to assign weights to the edges it is adding based on the weights of the edges in the path. To the degree that our goal in powering is to account for indirect evidence between communities while avoiding doublecounting and feedback, we may want to set the weight of the edge powering puts between two vertices equal to the weight of the strongest path between them. Generally, the degree of evidence a path provides that the vertices on either end are in the same community should be proportional to the product of the weights of its edges. As such, we may consider a path as having a weight equal to the product of the weights of its edges times some function of the path's length. The obvious choice there would be a function that is exponential in the length of the path. If it grows too slowly with length, we would only assign significant weights to short paths, in which case powering the graph would have little effect on it, and the problems that it was added to fix might prevent the algorithm from working. If the function grows too quickly with length, then the algorithm would essentially ignore short paths, which would be suboptimal, although it could still provide meaningful reconstructions. We refer the reader to [3] for further discussions on this.

REFERENCES

[1] E. Abbe, *Community detection and stochastic block models*, Found. Trends Commun. Inf. Theory, 14 (2018), pp. 1–162, https://doi.org/10.1561/0100000067.

[2] E. Abbe, F. Baccelli, and A. Sankararaman, *Community Detection on Euclidean Random Graphs*, preprint, https://arxiv.org/abs/1706.09942, 2017.

[3] E. Abbe, E. Boix, P. Ralli, and C. Sandon, *Graph Powering and Spectral Robustness*, preprint version of this paper, https://arxiv.org/pdf/1809.04818.pdf, 2018.

[4] E. Abbe, E. Boix, and C. Sandon, *Graph Powering and Spectral Gap Extraction*, manuscript; results partly presented at the Simons Institute and partly available in E. Boix PACM Thesis, Princeton University, Princeton, NJ, 2017.

[5] E. Abbe, L. Massoulié, A. Montanari, A. Sly, and N. Srivastava, *Group Synchronization on Grids*, preprint, https://arxiv.org/abs/1706.08561, 2017.

[6] E. Abbe and C. Sandon, *Detection in the Stochastic Block Model with Multiple Clusters: Proof of the Achievability Conjectures, Acyclic BP, and the Information-Computation Gap*, preprint, https://arxiv.org/abs/1512.09080, 2015.

[7] E. Abbe and C. Sandon, *Proof of the achievability conjectures for the general stochastic block model*, Comm. Pure Appl. Math., 71 (2017), pp. 1334–1406, https://doi.org/10.1002/cpa.21719.

[8] S. Balakrishnan, M. Xu, A. Krishnamurthy, and A. Singh, *Noise thresholds for spectral clustering*, in Advances in Neural Information Processing Systems, Vol. 24, J. Shawe-Taylor, R. S. Zemel, P. L.

Bartlett, F. Pereira, and K. Q. Weinberger, eds., Curran Associates, Inc., Red Hook, NY, 2011, pp. 954–962.

[9] C. Bordenave, M. Lelarge, and L. Massoulié, *Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs*, in Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science (FOCS '15), Washington, D.C., 2015, pp. 1347–1357, https://doi.org/10.1109/FOCS.2015.86.

[10] P. Chin, A. Rao, and V. Vu, *Stochastic Block Model and Community Detection in the Sparse Graphs: A Spectral Algorithm with Optimal Rate of Recovery*, preprint, https://arxiv.org/abs/1501.05021, 2015.

[11] K. W. Church and P. Hanks, *Word association norms, mutual information, and lexicography*, Comput. Linguist., 16 (1990), pp. 22–29.

[12] A. Coja-Oghlan, *Graph partitioning via adaptive spectral techniques*, Combin. Probab. Comput., 19 (2010), pp. 227–284, https://doi.org/10.1017/S0963548309990514.

[13] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46, https://doi.org/10.1137/0707001.

[14] I. Diakonikolas, G. Kamath, D. M. Kane, J. Z. Li, A. Moitra, and A. Stewart, *Robust estimators in high dimensions without the computational intractability*, in Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS '16), pp. 655–664.

[15] J. Friedman, *A proof of Alon's second eigenvalue conjecture*, in Proceedings of the 35th Annual ACM Symposium on Theory of Computing (STOC '03), New York, NY, 2003, pp. 720–724.

[16] S. Galhotra, A. Mazumdar, S. Pal, and B. Saha, *The Geometric Block Model*, preprint, https://arxiv.org/abs/1709.05510, 2017.

[17] O. Guédon and R. Vershynin, *Community detection in sparse networks via Grothendieck's inequality*, Probab. Theory Related Fields, 165 (2016), pp. 1025–1049, https://doi.org/10.1007/s00440-015-0659-z.

[18] S. Hoory, N. Linial, and A. Wigderson, *Expander graphs and their applications*, Bull. Amer. Math. Soc. (N.S.), 43 (2006), pp. 439–561, https://doi.org/10.1090/S0273-0979-06-01126-8.

[19] S. B. Hopkins and J. Li, *Mixture models, robustness, and sum of squares proofs*, in Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC '18), New York, NY, 2018, pp. 1021–1034.

[20] S. B. Hopkins and D. Steurer, *Bayesian Estimation from Few Samples: Community Detection and Related Problems*, preprint, https://arxiv.org/abs/1710.00264, 2017.

[21] M. D. P. Horton, H. M. Stark, and A. A. Terras, *What are zeta functions of graphs and what are they good for?*, in Quantum Graphs and Their Applications, Contemp. Math. 415, G. Berkolaiko, R. Carlson, S. A. Fulling, and P. Kuchment, eds., AMS, Providence, RI, 2006, pp. 173–189.

[22] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, *Performance of a Community Detection Algorithm Based on Semidefinite Programming*, preprint, https://arxiv.org/abs/1603.09045, 2016.

[23] A. Joseph and B. Yu, *Impact of regularization on spectral clustering*, Ann. Statist., 44 (2016), pp. 1765–1791.

[24] B. Karrer, E. Levina, and M. E. J. Newman, *Robustness of community structure in networks*, Phys. Rev. E (3), 77 (2008), 046119, https://doi.org/10.1103/PhysRevE.77.046119.

[25] T. Kawamoto and Y. Kabashima, *Limitations in the spectral method for graph partitioning: Detectability threshold and localization of eigenvectors*, Phys. Rev. E (3), 91 (2015), 062803.

[26] J. H. Kim, B. Sudakov, and V. Vu, *Small subgraphs of random regular graphs*, Discrete Math., 307 (2007), pp. 1961–1967, https://doi.org/10.1016/j.disc.2006.09.032.

[27] P. K. Kothari, J. Steinhardt, and D. Steurer, *Robust moment estimation and improved clustering via sum of squares*, in Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing (STOC '18), New York, NY, 2018, pp. 1035–1046, https://doi.org/10.1145/3188745.3188970.

[28] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborova, and P. Zhang, *Spectral redemption in clustering sparse networks*, Proc. Natl. Acad. Sci. USA, 110 (2013), pp. 20935–20940, https://doi.org/10.1073/pnas.1312486110,

[29] C. M. Le, E. Levina, and R. Vershynin, *Sparse Random Graphs: Regularization and Concentration of the Laplacian*, preprint, https://arxiv.org/abs/1502.03049, 2015.

[30] Z. Li, J. Liu, S. Chen, and X. Tang, *Noise robust spectral clustering*, in Proceedings of the 11th IEEE International Conference on Computer Vision, 2007, pp. 1–8, https://doi.org/10.1109/ICCV.2007.4409061.

[31] A. Lubostzky, R. Phillips, and P. Sarnak, *Ramanujan graphs*, Combinatorica, 8 (1988), pp. 261–277.
[32] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, in Proceedings of the 46th Annual Symposium on Theory of Computing (STOC '14), New York, NY, 2014, pp. 1–10, https://hal.archives-ouvertes.fr/hal-00969235.
[33] A. Moitra, W. Perry, and A. S. Wein, *How robust are reconstruction thresholds for community detection?*, in Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC '16), New York, NY, 2016, pp. 828–841.
[34] M. Mondelli and A. Montanari, *Fundamental Limits of Weak Recovery with Applications to Phase Retrieval*, preprint, https://arxiv.org/abs/1708.05932, 2017.
[35] A. Montanari and S. Sen, *Semidefinite programs on sparse random graphs and their application to community detection*, in Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (STOC '16), New York, NY, 2016, pp. 814–827, https://doi.org/10.1145/2897518.2897548,
[36] E. Mossel, J. Neeman, and A. Sly, *A proof of the block model threshold conjecture*, Combinatorica, 38 (2018), pp. 665–708.
[37] J. Mu, S. Bhat, and P. Viswanath, *All-but-the-Top: Simple and Effective Postprocessing for Word Representations*, preprint, https://arxiv.org/abs/1702.01417, 2017.
[38] A. Nilli, *On the second eigenvalue of a graph*, Discrete Math., 91 (1991), pp. 207–210.
[39] L. Orecchia, S. Sachdeva, and N. K. Vishnoi, *Approximating the exponential, the Lanczos method and an $\tilde{O}(m)$-time spectral algorithm for balanced separator*, in Proceedings of the 44th Annual ACM Symposium on Theory of Computing (STOC '12), New York, NY, 2012, pp. 1141–1160, https://doi.org/10.1145/2213977.2214080.
[40] M. Penrose, *Random Geometric Graphs*, Oxford Scholarship Online, Oxford University Press, Oxford, UK, 2003.
[41] S. Sachdeva and N. K. Vishnoi, *Faster algorithms via approximation theory*, Found. Trends Theor. Comput. Sci., 9 (2014), pp. 125–210, https://doi.org/10.1561/0400000065.
[42] D. A. Spielman and N. Srivastava, *Graph sparsification by effective resistances*, SIAM J. Comput., 40 (2011), pp. 1913–1926, https://doi.org/10.1137/080734029.
[43] L. Stephan and L. Massoulié, *Robustness of Spectral Methods for Community Detection*, preprint, https://arxiv.org/abs/1811.05808, 2018.
[44] P. D. Turney and P. Pantel, *From frequency to meaning: Vector space models of semantics*, J. Artificial Intelligence Res., 37 (2010), pp. 141–188.
[45] U. von Luxburg, *A tutorial on spectral clustering*, Stat. Comput., 17 (2007), pp. 395–416, https://doi.org/10.1007/s11222-007-9033-z.
[46] V. Vu, *A simple SVD algorithm for finding hidden partitions*, Combin. Probab. Comput., 27 (2018), pp. 124–140.
[47] H. Weyl, *Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung)*, Math. Ann., 71 (1912), pp. 441–479.
[48] C.-L. Yao, G. Chen, and T.-D. Guo, *Large deviations for the graph distance in supercritical continuum percolation*, J. Appl. Probab., 48 (2011), pp. 154–172.