Clustering in the Geometric Block Model

PACM Certificate Independent Work

Enric Boix
advised by Prof. Emmanuel Abbe

Contents

1 Introduction

The problem of graph clustering is to detect hidden "communities" in large networks: typically this implies partitioning the vertices of the network into separate groups, such that each group has many edges within it, and not too many edges to the other groups. Graph clustering is a staple in the analysis of large graphs, such as social networks and the Internet, and also has applications in recommendation systems and in medical diagnosis [2].

One way to formalize the problem of graph clustering is to define a random graph model and to postulate that it is representative of real-world graphs. The task then becomes to create clustering algorithms that provably work well

for this model, and to study the ranges of model parameters for which good clustering can, or cannot, be achieved.

1.1 The Stochastic Block Model (SSBM)

The simplest and best-known example of such a random graph model is the Stochastic Block Model (SBM), which was first proposed by [6]. In this paper, we will restrict ourselves to the symmetric, 2-community version of the Stochastic Block Model, although there exist generalizations:

Definition 1 (Stochastic Block Model)**.** *We sample an n-vertex graph G from the Stochastic Block Model distribution $\mathsf{SSBM}(n, p, q)$ as follows: Choose vertex labels X^n uniformly from $\{-1, +1\}^n$. Given X^n, let every edge (i, j) with $i < j$ be in G independently of the other edges, with probability given by*

$$\mathbb{P}[(i, j) \in E(G) \mid X_i = X_j] = p,$$

$$\mathbb{P}[(i, j) \in E(G) \mid X_i \neq X_j] = q.$$

More informally, the SSBM has two hidden communities of roughly equal size (vertices with label $+1$, versus vertices of label -1). Within each community, each pair of vertices is connected with probability p, and between communities, each pair of vertices is connected with probability q. The goal of a clustering algorithm for the SSBM is to partition the vertices of $G \sim \mathsf{SSBM}(n, a, b)$ into two groups that align with the two communities $\{v : X_v = +1\}$ and $\{v : X_v = -1\}$.

Notice that if $p = q$, then we recover the Erdös-Rényi distribution $G(n, p)$, and hence it is statistically impossible to recover the hidden clusters, because the graph G becomes independent of the vertex labels X^n. And if $p = 1$ and $q = 0$, then recovery becomes simple: G will have two connected components corresponding to the two hidden clusters.

1.2 The Geometric Block Model (GBM)

In this paper we will also consider a new random graph model, proposed by my advisor Emmanuel Abbe, which we will call the Geometric Block Model. This model has several parameters that can be tuned as desired, but for clarity we will present and analyze a simple definition:

Definition 2 (Geometric Block Model)**.** *Let $\gamma_+, \gamma_- : \mathbb{R}^d \to [0, 1]$ be distributions of points on $\mathbb{R}^d$, for some dimension d. We sample an n-vertex graph G from the Geometric Block Model distribution $\mathsf{GBM}(n, \gamma_+, \gamma_-, T)$ as follows: Choose vertex labels X^n uniformly from $\{-1, +1\}^n$. For each $i \in [n]$, sample a point p_i from γ_{X_i}. Then, for each pair $i < j \in [n]$, let $(i, j) \in E(G)$ if and only if $||p_i - p_j||_2 \leq T$.*

In other words, the geometric block model is constructed by assigning a binary label $X_i \in \{-1, +1\}$ to each vertex i, sampling a point $p_i \in \mathbb{R}^d$ from

a distribution corresponding to that label, and putting an edge between every pair of vertices i, j whose corresponding points p_i, p_j are close enough.

The reason we consider the GBM is because the SSBM has certain properties that can make it somewhat unrealistic in many settings. For instance, in the "sparse" regime of the SSBM, when the expected degree is constant, there will only be a constant number of triangles u, v, w in expectation. Indeed, most small neighborhoods of vertices in the SSBM will be trees. This lack of "transitivity" is a property that is not shared by many real networks, such as social networks. In a social network: if u knows v, and v knows w, then there is a good chance that u also knows v.

In contrast to the SSBM, the GBM does have this "transitivity" property: if u and v are adjacent and v and w are adjacent, then p_u and p_w are close by the triangle inequality, and therefore u and w have a good chance of being adjacent.

Variants and Hybrid Models There are several possible variants of Geomtric Block Model defined above. For example, we could (1) have the connection probability between vertices i and j be a function $f(d, x, y)$ of their distance i and of their labels x, y, and/or (2) have k communities instead of just 2.

Notice that if we extend the model as in possibility (1), then by letting the connection probability

$$f(d, x, y) = \begin{cases} p, & x = y \\ q, & x \neq y \end{cases}$$

we can recover the SSBM. Hence, we can generalize the Geometric Block Model to be a quite versatile, and hopefully a quite realistic, random graph model that has some of the properties of the pure SSBM and some of the properties of the pure GBM.

For the purpose of this paper, however, we will concentrate on the following two special cases of the GBM. Some of the tools used in our analysis of these models can be readily extended to more general settings.

Definition 3 (Gaussian Geometric Block Model)**.** *The n-vertex, distance-D, threshold-T Gaussian Geometric Block Model is a random variable* $\mathsf{GGBM}(n, D, T)$ *with distribution given by* $\mathsf{GBM}(n, \gamma_+, \gamma_-, T)$*, where $\gamma_\pm$ is the distribution of the normal random variable $\mathcal{N}((\pm\frac{D}{2}, 0), I_2)$ on $\mathbb{R}^2$.*

Definition 4 (Square Geometric Block Model)**.** *The n-vertex, distance-D, threshold-T Square Geometric Block Model is a random variable* $\mathsf{SGBM}(n, D, T)$ *with distribution given by* $\mathsf{GBM}(n, \gamma_+, \gamma_-, T)$*, where $\gamma_\pm$ is the uniform distribution on the square in $\mathbb{R}^2$ with length-1 sides parallel to the x- and y-axes, and which is centered at $(\pm\frac{D}{2}, 0)$.*

In the GGBM and in the SGBM, the points with label ± 1 will be clustered around the point $(\pm\frac{D}{2}, 0)$, so two clusters will naturally arise when edges are added between points at distance $\leq T$. As D increases, these clusters of points move farther apart, and detecting the clusters in the GGBM and SGBM will become easier.

The SGBM is a toy model for the GGBM, which will be easier to analyze than the GGBM because of the piece-wise homogeneous density of points in the SGBM. We will only refer to it in Section 4. The rest of the GBM results in this paper will involve the GGBM.

1.3 Recovery regimes

Given a random graph model with hidden clusters, such as the SSBM or the GBM, our primary goal is to create an algorithm R that, given a graph G sampled from one of these distributions, assigns a label $\hat{X} = \hat{X}(G) = R(G)$ to the vertices of G that agrees with the hidden vertex labels $X = X(G)$.

In order to measure the quality of a vertex labelling $\hat{X}$, we have to define the agreement between two label vectors (the following definitions are adapted from [2]):

Definition 5 (Agreement between label vectors)**.** *Let $x, y \in \{+1, -1\}^n$ be two label vectors. Then their agreement $A(x, y)$ is given by*

$$A(x, y) = \max_{\pi \in S_2} \frac{1}{n} \sum_{i=1}^{n} 1(x_i = \pi(y_i)),$$

where S_2 is the group of permutations on $\{+1, -1\}$.

We can formulate the problem of recovering the hidden clusters as the problem of finding a label vector $\hat{X}$ that agrees highly with the true vertex label vector $\hat{X}$. To this end, we define the notion of "weak recovery" on a random graph model with hidden clusters.

Definition 6. *Weak recovery is solved in the* SSBM *(or* GBM*) for certain parameters, if for (X, G) drawn from the* SSBM *(or* GBM*) with those parameters, there exists $\varepsilon > 0$ and an algorithm R that takes G in as input and outputs $\hat{X}$ such that $\mathbb{P}[A(X, \hat{X}) \geq \frac{1}{2} + \varepsilon] = 1 - o(1)$.*

Essentially, weak recovery asks: given the graph and given one of the vertex labels, can we guess the other vertex labels with better than the trivial $\frac{1}{2}$ probability?

We will work with the notion of weak recovery as opposed to exact recovery, for which we would require $\mathbb{P}[A(X, \hat{X}) = 1] = 1 - o(1)$, because on the GGBM exact recovery is not possible except when the distance between the centers of the distributions $\gamma_{\pm}$ is very far apart:

Lemma 7. *Exact recovery is impossible in the* GGBM(n, D, T) *for $D < \infty$.*

Proof. Notice that the random graph G is the output of a channel on the points $p_1, \ldots, p_n$, so given the points $p_1, \ldots, p_n$ we can ignore the graph G. Even knowing these points, we cannot reconstruct the vertex labels exactly, because, given $p_i = (x_i, y_i) \in \mathbb{R}_2$, the maximum a-posteriori estimator for X is the sign of x. Since a point sampled from γ_+ has a constant probability of having negative x-coordinate, the maximum a-posteriori estimator will therefore be wrong on each vertex with constant probability. □

For the SSBM, weak recovery is also the appropriate notion of recovery in the "sparse" regime in which the expected degree of a node is constant: $\mathsf{SSBM}(n, a/n, b/n)$ for constant $a, b > 0$. This is because in sparse regime, the graph is not connected, and in particular there are many "singleton" nodes whose labels it would be impossible to guess if exact recovery were the goal. In fact, it is known that weak recovery for $\mathsf{SSBM}(n, a, b)$ is possible if and only if

$$\frac{(a-b)^2}{2(a+b)} > 1.$$

This is known as the Kesten-Stigum threshold for the SBM. The conjecture was proposed by [4], the impossibility result was proved by [9] in 2012 and the achievability result was proved concurrently by [7] and [8].

1.4 Overview of Paper

In Section 2, we report empirical results of spectral clustering algorithms on the SSBM and GGBM models.

In Section 3, we propose combining graph powering and spectral methods to get recovery for the SSBM and the GBM. We prove that a closely-related approach will work for SSBM down to the Kesten-Stigum threshold (resolving an open question of [7]). We also report favorable empirical results from graph powering on the GGBM.

In Section 4, we prove that weak recovery is possible in the SGBM when $D > 0$ and there is high enough expected *logarithmic* degree. We also give evidence that suggests that weak recovery is possible in the GGBM when $D > 0$ and there is high enough expected *constant* degree.

2 Empirical Results

A substantial part of the independent work involved implementing and testing well-known graph clustering techniques on the GGBM and on some hybrids of the SSBM and the GGBM. For simplicity and consistency with the rest of this paper, we will provide a brief overview of the empirical results for the GGBM and SSBM only.

We tested spectral methods on the following graph operators, with various levels of pre-processing "cleaning" operations:

1. The adjacency matrix A
2. the Laplacian $D - A$
3. the normalized Laplacian $I - D^{-1/2}AD^{-1/2}$
4. the non-backtracking operator B

Here, D is the diagonal degree matrix.

2.1 Empirical Results for the SSBM

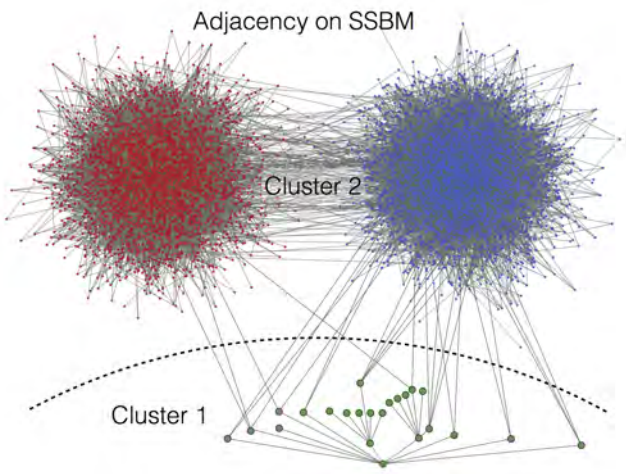
It is known ([2]) that among these techniques, only the non-backtracking operator achieves weak recovery down to the Kesten-Stigum threshold of the SSBM. We review how the other methods fare:

Adjacency Matrix This method splits the SSBM giant into (1) a small neighborhood of a large-degree vertex, and (2) the rest of the vertices. See Figure 1.

Laplacian Matrix This method splits the SSBM giant into (1) a small tree connected to the giant by one vertex, and (2) the rest of the vertices. See Figure 2.

Normalized Laplacian Matrix This method works on the SSBM giant until the parameters are very close to the recovery threshold, when it starts to fail. It splits the SSBM into (1) a small tree connected to the giant by one vertex, and (2) the rest of the vertices (like the Laplacian method). See Figure 3.

Figure 1: Adjacency Matrix on the SSBM



2.2 Empirical Results for the GGBM

On the GGBM, the most successful of these methods was the normalized laplacian. The other methods all failed:

Figure 2: Laplacian Matrix on the SSBM

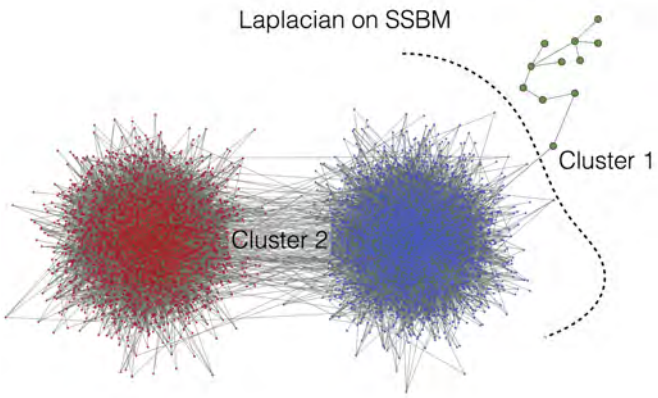
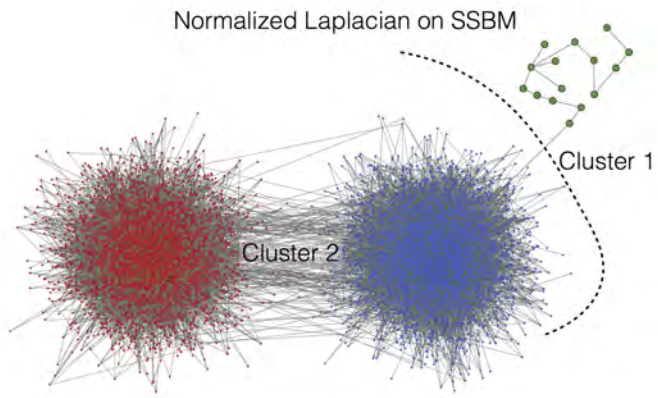


Figure 3: Normalized Laplacian Matrix on the SSBM



Adjacency Matrix The adjacency matrix fails on the GGBM giant: cluster (1) is around a clump of vertices of large degree, and cluster (2) is the rest of the vertices. See Figure 4.

Laplacian Matrix The Laplacian fails on the GGBM giant: cluster (1) is a short tail jutting out from the rest of the graph, which is assigned to cluster (2). See Figure 5.

Normalized Laplacian Matrix The normalized Laplacian works well on the GGBM giant. Indeed, Figure 6 gives an of the kind of the clustering that we optimally want on the GGBM. The partition is roughly around the line $x = 0$, which is the best that one can do even knowing the positions of the points. (It may be possible that the Normalized Laplacian method fails when the distance D between clusters is very small, but it is hard to tell based on computational simulation alone.)

Non-Backtracking Matrix The non-backtracking matrix fails on the GGBM giant: the picture is qualitatively similar to the clustering given by the adjacency matrix. See Figure 7.

Because the non-backtracking method fails, none of above the methods works well on both the GGBM and the SSBM down to the Kesten-Stigum threshold. The best compromise is probably the normalized Laplacian operator, which works well for the SSBM when the parameters are not too close to the threshold, and which also seems to work well for the GBM. However, it would be good to have a method that can provably work for both models, even in the hardest regimes. This motivates graph powering, which we will present in the next section.

Figure 4: Adjacency Matrix on the GGBM

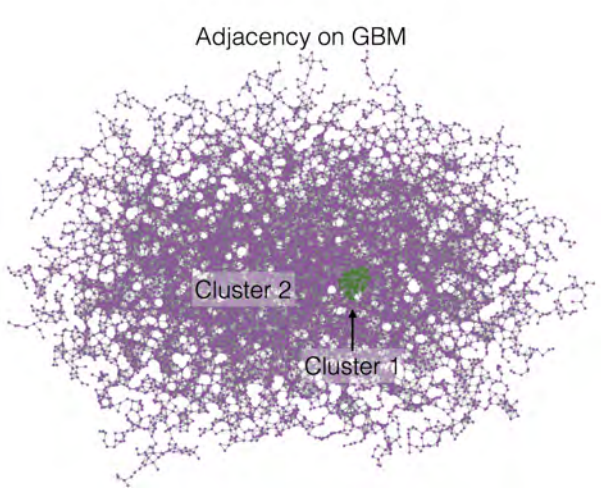


Figure 5: Laplacian Matrix on the GGBM

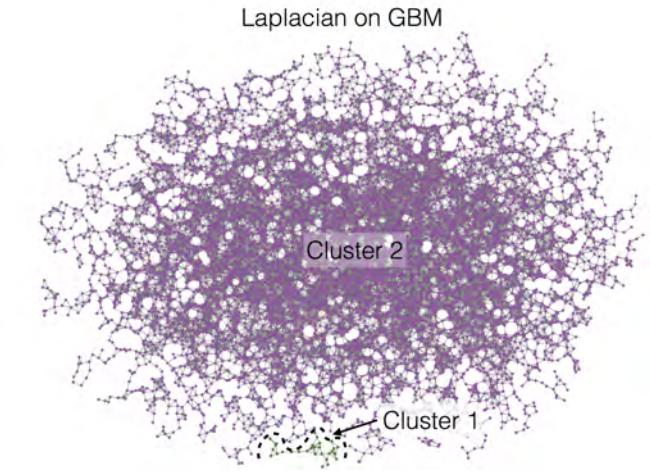
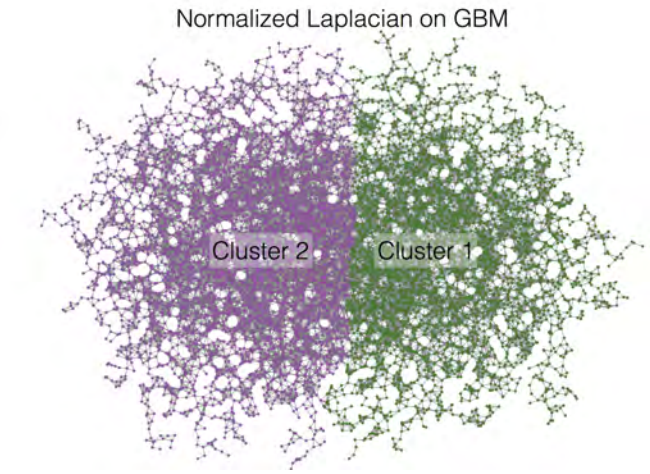


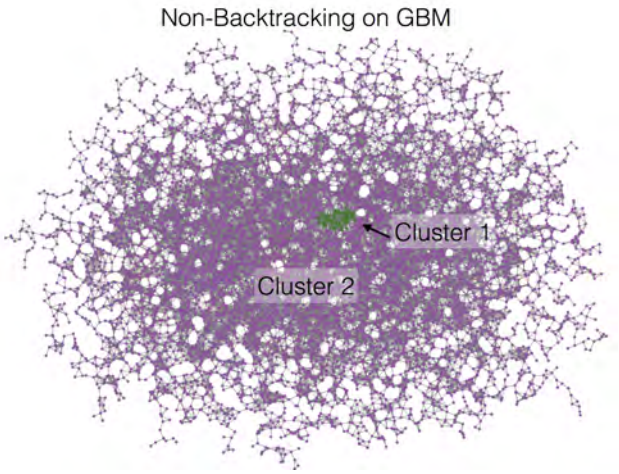
Figure 6: Normalized Laplacian Matrix on the GGBM



3 Smoothing the Graph: Graph Powering

In this section we introduce the graph powering operation. The hope is that powering a graph and then applying spectral clustering methods to it will lead to cluster recovery in both the SSBM and the GGBM. We prove in Section 3.1 that a closely-related smoothing method (the distance-l operator) works for the SSBM, all the way down to the Kesten-Stigum threshold. This gives

Figure 7: Non-Backtracking Matrix on the GGBM



strong evidence that graph powering should work on the SSBM as well. And we provide empirical evidence in Section 3.2 that the graph powering method works to cluster the GGBM.

Definition 8 (Graph power)**.** *Let G be an unweighted graph given by adjacency matrix A. Its kth power is the graph on V given by the adjacency matrix*

$$A^{(k)} := 1(I + A^k > 0).$$

There is an edge between v and w in the kth power of G if and only if there is a path of length $\leq k$ between v and w in G.

The intuition behind first powering the graph and then applying spectral methods is that the influence of (1) irregularly-high-degree vertices and (2) irregularly-low-degree vertices is reduced. The smoothing is achieved because the value A^l grows exponentially faster in l around high-degree vertices than around low-degree vertices, but, after applying the threshold function, the value of $A^{(l)}$ does not – the entries of the matrix $A^{(l)}$ all grow at roughly the same rate around low-degree vertices as around high-degree vertices. This means that spectral methods on $A^{(l)}$ no longer have the neighborhood of a high-degree vertex as the second eigenvector. Nor do they cut the graph into a tail of degree-2 vertices and the rest of the graph.

3.1 Proof: Clustering the SSBM

We will prove that spectral methods on the distance-l graph, a close relative of the lth graph power, give weak recovery for the SSBM whenever possible. This answers the open question of [7].

Definition 9 (Distance-l graph)**.** *Let G be an unweighted graph. Then the distance-l matrix $C^{(l)}$ is defined by*

$$C^{(l)}_{i,j} = 1(d_G(i,j) = l).$$

The distance-l graph is the graph with adjacency matrix $C^{(l)}$.

In other words, there is an edge between i and j in the distance-l matrix if and only if i and j are at distance l in G.

Notice that the graph power and the distance-l matrix are closely related, since for any l,

$$A^{(l)} = \sum_{k=0}^{l} C^{(k)}.$$

We will prove our clustering result by proving that the distance-l matrix and the length-l self-avoiding path matrix of [7] are very close in spectral norm:

Definition 10 ([7], Self-avoiding path matrix)**.** *Let G be an unweighted graph. The length-l self-avoiding path matrix $B^{(l)}$ is the matrix such that $B^{(l)}_{i,j}$ counts the number of self-avoiding paths of length l connecting i to j in G.*

Let G be a random graph distributed as $\mathsf{SSBM}(n, a/n, b/n)$, for constants $a, b \geq 0$, and let $B^{(l)}$, $C^{(l)}$ refer to its self-avoiding paths matrix and distance-l matrix, respectively. Let $\alpha := (a+b)/2$ be the expected degree of a vertex, and let $\beta := (a-b)/2$.

Lemma 11. *For any l, define*

$$M^{(l)} = B^{(l)} - C^{(l)}.$$

If $l = c \log n$ such that $l < \log_\alpha n^{1/4}$, then with high probability

$$\rho(M^{(l)}) = \tilde{O}(\alpha^{l/2}),$$

where $\rho(M^{(l)})$ denotes the spectral radius of $M^{(l)}$ and $\tilde{O}(f)$ denotes $O(f)$ up to poly-logarithmic factors in f.

Proof. Let E_1 be the event that no vertex has more than one cycle in its l-neighborhood. By Lemma 4.2 of [7], E_1 occurs with high probability.

Suppose E_1 holds. Then we can partition the vertices of the graph so that $v \sim w$ iff v and w share a simple cycle in their l-neighborhoods. We postpone the proof of the following claim:

Claim 12. *Suppose E_1 holds. Then for all $i, j \in V$:*

(i) $|M^{(l)}_{i,j}| \leq 1$.

(ii) $M^{(l)}_{i,j} \neq 0 \implies i \sim j$.

(iii) $M^{(l)}_{i,j} \neq 0 \implies$ *there are two length-($\leq l$) paths from i to j.*

By the item (ii) of Claim 12, $M^{(l)}$ is a block-diagonal matrix, where each block corresponds to an equivalence class of $\sim$ in the vertex partition. Therefore, it suffices to bound the spectral norm of each block.

First, we will need to define the following event. Let $E_2(C)$ be the event that for all vertices $i \in V(G)$, for all $t \in \{1, \ldots, l\}$, the following holds:

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

By Theorem 2.3 of [7], we know that there is C large enough that $E_2(C)$ holds with high probability.

Claim 13. *Conditioning on $E_1 \cap E_2(C)$, the Frobenius norm of each block is upper-bounded by*

$$2Cl(\log n)^2 \alpha^{l/2} = \tilde{O}(\alpha^{l/2}).$$

This is an upper bound on the spectral norm of $M^{(l)}$, and since E_1 and $E_2(C)$ occur with high probability the theorem is proved. □

Proof (of Claim 13). Suppose by contradiction that the block corresponding $S \subseteq V$ has Frobenius norm $> 2Cl(\log n)^2 \alpha^{l/2}$, where S is the set of vertices in some equivalence class of $\sim$. Then let $H \subseteq G$ be the cycle that is shared by the l-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-($\leq l$) paths from i to j, at least one of the paths must contain e. Otherwise, the cycle H is not the only cycle in the l-neighborhood of i. So by item (iii) of Claim 12, the number of pairs $i, j \in S$ such that $M^{(l)}_{i,j} \neq 0$ is at most the number of ($\leq l$)-length paths in G that contain e.

We bound this number of such paths by

$$4C^2 l^2 (\log n)^4 \alpha^{l/2},$$

which by item (i) of Claim 12 means that the Frobenius norm of the block is at most $2Cl(\log n)^2 \alpha^{l/2}$: a contradiction.

It suffices to bound the number of length-t paths containing $e = (u, v)$ by

$$4C^2 t (\log n)^4 \alpha^{t/2}.$$

for all $t \in \{1, \ldots, l\}$. And since E_1 and $E_2(C)$ hold,

$$(\#\text{ length-}t\text{ paths containing } e = (u,v)) \leq$$

$$\sum_{r=0}^{l} (\#\text{ length-}r\text{ paths containing } u) \cdot (\#\text{ length-}(t-r-1)\text{ paths containing } v) \leq$$

$$t(2C(\log n)^2)^2 \alpha^{t-1} \leq$$

$$4C^2 t (\log n)^4 \alpha^t.$$

□

Proof (of Claim 12). Suppose $M^{(l)}_{i,j} \neq 0$. Since every vertex has at most one cycle in its l-neighborhood, there are at most 2 length-($\leq l$) self-avoiding paths between every pair of vertices. So the possible cases are:

1. $C^{(l)}_{i,j} = 0$:
   (a) $B^{(l)}_{i,j} = 1$. There is a path of length $< l$ between i and j. So there are two paths of length $\leq l$ between i and j.
   (b) $B^{(l)}_{i,j} = 2$. Impossible. There is no path of length $< l$ between i and j, because there are at most two ($\leq l$)-length paths between i and j, and there is a path of length l between i and j, so $C^{(l)}_{i,j} = 1$.
2. $C^{(l)}_{i,j} = 1$:
   (a) $B^{(l)}_{i,j} = 0$. Impossible. The distance between i and j is l, so there should be an l-length path between them.
   (b) $B^{(l)}_{i,j} = 2$. There are two paths of length l between i and j.

So if $M^{(l)}_{i,j} \neq 0$, then $|M^{(l)}_{i,j}| = 1$, and there are exactly two ($\leq l$)-length paths between i and j. This proves items (i) and (iii) of the claim.

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

We now state a version of the Davis-Kahan theorem ([3]) presented in [1]:

Theorem 14 (Davis-Kahan Theorem from [1])**.** *Suppose that $H = \sum_{j=1}^{n} \bar{\mu}_j \bar{u}_j \bar{u}_j^T$ 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 an eigenvector of H corresponding to its j-th largest eigenvalue, and $\Delta = \min\{\bar{\mu}_{j-1} - \bar{\mu}_j, \bar{\mu}_j - \bar{\mu}_{j+1}\}$, where we define $\bar{\mu}_0 = +\infty$ and $\bar{\mu}_{n+1} = -\infty$. We have*

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

In addition, if $\|E\|_2 \leq \Delta$, then

$$\min_{s=\pm 1} \|s u_j - \bar{u}_j\|_2 \lesssim \frac{\|E\bar{u}_j\|_2}{\Delta}, \tag{2}$$

where both $\lesssim$ only hide absolute constants.

Using Theorem 14, and Theorem 2.1 of [7], we can characterize the top two eigenvalues and eigenvectors of the distance-l matrix of a SSBM. The following theorem is analogous to Theorem 2.1 in [7], which characterizes the top eigenvalues and eigenvectors of the self-avoiding-paths matrix $B^{(l)}$ of a SSBM.

Theorem 15. *Let $G = \mathsf{SSBM}(n, a/n, b/n)$ for two constants $a, b \geq 0$ such that the expected degree $\alpha := (a+b)/2 > 1$, and such that $\beta^2 > \alpha$, where $\beta := (a-b)/2$. Let $X \in \{-1, +1\}^n$ be the hidden label vector of G. Then, for $l = c \log n$ such that $l < \log_\alpha n^{1/4}$, we have the following with high probability:*

1. *The first eigenvalue of $C^{(l)}$ is $\Theta(\alpha^l)$ up to logarithmic factors. The corresponding eigenvector is asymptotically parallel to $B^{(l)}\vec{1}$.*
2. *The second eigenvalue of $C^{(l)}$ is $\Omega(\beta^l)$ up to logarithmic factors. The corresponding eigenvector is asymptotically parallel to $B^{(l)}X$.*
3. *For any $\epsilon > 0$, all other eigenvalues are $O(n^\epsilon \sqrt{\alpha^l})$.*

Proof. Write $C^{(l)} = B^{(l)} - M^{(l)}$. By Lemma 11, $\rho(M^{(l)}) = O(n^\epsilon \alpha^{l/2})$, for all $\epsilon > 0$. Therefore, if $C^{(l)}$ has three eigenvectors of eigenvalue $\omega(n^\epsilon \alpha^{l/2})$ for some $\epsilon > 0$, there are three orthogonal unit vectors v_1, v_2, v_3 such that $\|C^{(l)} v_i\|_2 = \omega(n^\epsilon \alpha^{l/2})$, and hence $\|B^{(l)} v_i\|_2 = \omega(n^\epsilon \alpha^{l/2})$ by triangle inequality. This contradicts Theorem 2.1 of [7], which states that with high probability $B^{(l)}$ only two vectors with eigenvalue $\omega(n^\varepsilon \alpha^{l/2})$, and hence item 3 is true.

For items 2 and 3, notice that we can apply the Davis-Kahan inequality because $\rho(M^{(l)}) = O(n^\epsilon \alpha^{l/2})$ for all $\epsilon > 0$. And in both cases, for all $\epsilon > 0$, $\Delta = \omega(n^\epsilon \alpha^{l/2})$. Hence the first eigenvector of $C^{(l)}$ asymptotically aligns with the first eigenvector of $B^{(l)}$, and the second eigenvector of $C^{(l)}$ asymptotically aligns with the second eigenvector of $B^{(l)}$. Since the spectral norm is bounded, we get that the first and second eigenvalues of $C^{(l)}$ match with the first and second eigenvalues of $B^{(l)}$, which are $\Theta(\alpha^l)$ and $\Omega(\beta^l)$, respectively. □

Incidentally, Lemma 4.4 of [7] states that $B^{(l)}1$ is asymptotically aligned with $C^{(l)}1$ and $B^{(l)}X$ is asymptotically aligned with $C^{(l)}X$. But as stated, Theorem 15 already allows us to weakly recover X from the second eigenvector of $C^{(l)}$, by using the same procedure that [7] uses to recover X from the second eigenvector of $B^{(l)}$.

Graph Powering on the SSBM

This characterization of the top two eigenvectors of the distance-l matrix indicates that the top two eigenvectors of the adjacency matrix of an l-powered graph should also be aligned with $B^{(l)}1$ and $B^{(l)}X$, respectively. Indeed, writing

$$A^{(l)} = \sum_{t=0}^{l} C^{(t)} = \sum_{t=0}^{l/2} C^{(t)} + \sum_{t=l/2+1}^{l} C^{(t)},$$

we notice that by Theorem 15 and triangle inequality, $\rho\left(\sum_{t=0}^{l/2} C^{(t)}\right) = O(l\alpha^{l/2})$ up to logarithmic factors, so the corresponding terms $C^{(t)}$ for $t \leq l/2$ in the sum are negligible by the Davis-Kahan theorem. And moreover one can see that for $t > l/2$, the top two eigenvectors of $C^{(t)}$ are asymptotically well aligned with the top two eigenvectors of $C^{(l)}$ by combining Theorem 15 above, and

Theorem 2.3 in [7]. Therefore, heuristically, the top two eigenvectors of $A^{(l)}$ should asymptotically align to the top two eigenvectors of $C^{(l)}$, giving weak recovery via the graph powering operator.

3.2 Empirical Evidence: Clustering the GGBM

Powering the GGBM and then clustering based on the top eigenvectors of the powered adjacency matrix appears to give weak recovery. The clusters are similar to those obtained via the normalized Laplacian, without powering.

4 Weak recovery for the GBM

In this section, we prove that weak recovery is possible in the SGBM when the expected degree is $\geq C \log n$ for high enough constant C. We also conjecture that weak recovery is possible in the GGBM for large enough constant expected degree, and we provide some justification for our conjecture.

4.1 Proof: Expected logarithmic degree (SGBM)

Our first goal will be to prove that in the regime of expected logarithmic degree with high leading constant, the graph distances will closely resemble the straight-line distances in the plane.

Lemma 16. *Let $G \sim \mathsf{SGBM}(n, D, T)$, where $T > 4\sqrt{\frac{\log n}{n}}$. Then with high probability G is connected and for all $u, v \in V(G)$, $d_G(u, v) \leq 4d(p_u, p_v)/T$.*

Proof. Choose $r, c \in \mathbb{Z}_{>0}$ and divide $[-(D+1)/2, +(D+1)/2] \times [-1/2, 1/2]$ into an $r \times c$ array of bins $(A_{i,j})_{(i,j)\in\{0,\ldots,r-1\}\times\{0,\ldots,c-1\}}$ such that

$$A_{i,j} = (-\frac{(D+1)}{2} + \frac{i(D+1)}{r}, -\frac{1}{2} + \frac{j}{c}) + [0, \frac{D+1}{r}] \times [0, \frac{1}{c}].$$

Let $S = \{p_v \mid v \in V(G)\}$ be the set of points associated with the vertices of G, and let $E_{i,j}$ be the event that $A_{i,j} \cap S \neq \emptyset$.

$$\mathbb{P}[E_{i,j}] = 1 - \mathbb{P}[p_k \notin A_{i,j} \forall k \in [n]] \geq 1 - \left(1 - \frac{(D+1)}{2rc}\right)^n.$$

So by union bound

$$\mathbb{P}[E_{i,j} \forall (i,j)] \geq 1 - rc\left(1 - \frac{(D+1)}{2rc}\right)^n \geq 1 - rc\exp\left(-\frac{n(D+1)}{2rc}\right).$$

So choosing r, c so that $rc \leq \frac{(D+1)}{2}\frac{n}{\log n}$, we have

$$\mathbb{P}[E_{i,j} \forall (i,j)] \geq 1 - \frac{(D+1)}{2}\frac{1}{\log n} \to 1.$$

Hence, let $r = \lfloor (D+1)c \rfloor$, and $c = \lfloor \frac{1}{\sqrt{2}} \sqrt{\frac{n}{\log n}} \rfloor$ from now on. And notice that for any $p \in A_{i,j}$, $q \in A_{i\pm1,j} \cup A_{i,j\pm1}$, $d(p,q) \leq 4\sqrt{\frac{\log n}{n}}$.

Hence, setting the threshold $T = 4\sqrt{\frac{\log n}{n}}$ means that every bin of side-length $\frac{\sqrt{2}}{4}T$ contains a point, which is connected to all points in adjacent bins. Since all the points in S are in some bin, with high probability the graph G is connected, and, for all $u, v \in V(G)$,

$$d_G(u,v) \leq 4d(p_u, p_v)/T.$$

□

Lemma 17. *Let $G \sim \mathsf{SGBM}(n, D, T)$ as above, for $T > 4\sqrt{\frac{\log n}{n}}$. Then for all $\varepsilon > 0$, with high probability*

$$d_G(u,v) = d(p_u, p_v)(1 + o(1))/T$$

for all $u, v \in V(G)$ such that $d(p_u, p_v) \geq n^{-1/2+\varepsilon}$.

Proof. It suffices to prove that $d_G(u,v) \leq d(p_u, p_v)(1 + o(1))/T$ because the other direction is clear.

Cut $[-(D+1)/2, +(D+1)/2] \times [-1/2, +1/2]$ into squares of side length h. Then, for every pair of squares A_i, A_j, build a sequence of "stepping stone" bins $D_{i,j} = (D_{i,j}^{(1)}, \ldots, D_{i,j}^{(m_{i,j})})$ as follows:

1. Rotate the plane around the center of A_i so that the centers of A_i and A_j are separated by a horizontal line segment of distance d, and A_j is to the right of A_i.

2. Construct the sequence $(D_{i,j}^{(1)}, \ldots, D_{i,j}^{(m_{i,j})})$ of squares whose sides are parallel to the current x- and y-axes such that

   (a) The side length of each square is $s \leq h\sqrt{2}$.

   (b) $A_k \subset D_{i,j}^{(1)}$.

   (c) Space the squares so that their centers are distance $T - 3s$ apart, and square $D_{i,j}^{(l+1)}$ is to the right of square $D_{i,j}^{(l)}$.

   (d) $m_{i,j} = \lceil d/(T - 3s) \rceil$.

3. Rotate the plane back to its original orientation.

We say that a stepping stone sequence $D_{i,j}$ is "good" if, writing $S = \{p_v \mid v \in V(G)\}$,

$$\sum_{l=1}^{m_{i,j}} 1(S \cap D_{i,j}^{(l)} = \emptyset) \leq 2m_{i,j}/(\log n).$$

Notice that if $p_v \in A_i$, $p_u \in A_j$, and $D_{i,j}$ is good, then by Lemma 16 and triangle inequality, with high probability

$$d_G(u,v) \leq \frac{d(p_v,p_u)}{T-3s}(1+o(1)) + O(1),$$

because thinking of each bin $D_{i,j}^{(l)}$ as a stepping stone, we can follow the stepping stone sequence for all but a $o(1)$ fraction of the stepping stones, and whenever a stepping stone is missing we can use Lemma 16 to skip it with a slowdown of only about $3d(p_u,p_v)/(T-3s)$.

For the last part of the proof, we couple S with a homogeneous Poisson point process $\mathcal{P}$ with density $\lambda = \frac{n}{5}$ on $[-(D+1)/2, +(D+1)/2] \times [-1/2, +1/2]$ such that with high probability, $\mathcal{P} \subset S$, because conditioned on $|\mathcal{P}|$, the distribution of points of $\mathcal{P}$ is uniform, and $|\mathcal{P}| < n$ with high probability. Therefore, we may condition on $\mathcal{P} \subset S$. (For more details on this kind of coupling, a good reference is Chapter 1 of [10].)

Set $h = c_1\sqrt{\frac{\log\log n}{n}}$ for a constant c_1 to be determined later. We prove that for all $\varepsilon > 0$, with high probability $D_{i,j}$ is good for all (i,j) such that the centers of A_i and A_j are distance $\geq n^{-1/2+3\varepsilon}$ apart.

Pick such a pair (i,j). Then $m_{i,j} \geq n^{2\varepsilon}/\sqrt{\log n}$. Notice that we can assume that all the bins $D_{i,j}^{(1)}, \ldots, D_{i,j}^{(m_{i,j})}$ are disjoint, since $h = o(T)$. The Let E_l be the event that $D_{i,j}^{(l)} \cap \mathcal{P} = \emptyset$. For appropriate c_1,

$$\mathbb{P}[E_l] = \exp(-\frac{nh^2}{2}) = 1/(\log n),$$

and since the bins are disjoint, all the events $E_1, \ldots, E_{m_{i,j}}$ are independent. Since the expected number of empty bins is $\sum_{l=1}^{m_{i,j}} \mathbb{P}[E_l] = m_{i,j}/(\log n) \geq n^{2\varepsilon}/(\log n)^{3/2} \geq n^{\varepsilon}$, by Chernoff bounds, with probability $\geq 1 - \exp(-c_2 n^{\varepsilon})$ at most $2m_{i,j}/(\log n)$ of the events E_l do not hold. If E_l holds, then $\emptyset \neq D_{i,j}^{(l)} \cap \mathcal{P} \subset D_{i,j}^{(l)} \cap S$. Hence, $D_{i,j}$ is good with probability $1 - \exp(-c_2 n^{\varepsilon})$.

Since there are $O(n^2)$ pairs of bins (i,j), by union bound all $D_{i,j}$ for which the centers of A_i and A_j are farther than $n^{-1/2+3\varepsilon}$ are good with high probability. □

Solution to Weak Recovery Given Lemmas 16 and 17, we can solve weak recovery on the SGBM when $T > 4\sqrt{\frac{\log n}{n}}$.

The algorithm we propose is: let u, v be the vertices that are farthest away from each other on G. Now let $N_u(l)$ be a neighborhood of depth l of u. Assign label $\hat{X}_w = +1$ to all vertices $w \in N_u(D/(2T))$, and assign an independent $\hat{X}_w \sim \text{Rad}(1/2)$ to each of the other vertices.

By the proof of Lemma 16, we know that the $S = \{p_w \mid w \in V(G)\}$ will contain points within distance $O(\sqrt{\frac{\log n}{n}})$ of the corners of the box

$$B \equiv [-(D+1)/2, +(D+1)/2] \times [-1/2, +1/2].$$

Therefore, letting $r \equiv \sqrt{(D+1)^2+1}$ denote the distance between the two corners of the box, there will be vertices in G that have graph distance at least r/T. Moreover, by Lemma 17, the maximum graph distance between a pair of vertices will be $r(1+o(1))/T$.

Let u, v be the two points that are furthest away from each other in G. By Lemma 17, $d(p_u, p_v) = r(1-o(1))$. Hence, p_u is distance $o(1)$ from a corner of B. For every vertex $w \in N_u(D/(2T))$, $d(p_u, p_w) \leq \frac{D}{2}(1+o(1))$. So by triangle inequality all $w \in N_u(D/(2T))$ have the same label as u, because they are all in a region in which only one of the distributions γ_+ or γ_- has support. Moreover, $N_u(D/(2T))$ contains all the vertices w for which $d(p_u, p_w) \leq \frac{D}{2}(1-o(1))$, for some $o(1)$. Therefore, with high probability $N_u(D/(2T))$ contains at least αn vertices, for some $\alpha(D) > 0$.

Hence, with high probability, the algorithm outputs a guess $\hat{X}$ of the hidden labels X that has agreement $A(X, \hat{X}) = \frac{1-\alpha}{2} + \alpha + o(1) = \frac{1}{2} + \frac{\alpha}{2} + o(1) > \frac{1}{2}$, solving weak recovery.

4.2 Conjecture: Expected constant degree (GGBM)

We conjecture that weak recovery is possible in the GGBM for any $D > 0$ when the expected degree is a large enough constant c. The conjecture is heuristically justified by the following algorithm, that succeeds with probability $q > 0$, and that outperforms random guessing when it succeeds:

Proposed Clustering Algorithm

1. Pick a random vertex v, and let $N_v(an)$ be the set of an vertices that are closest to it in the graph (breaking ties arbitrarily).
2. Give label $+1$ to all vertices in N_v. Give independent Rad(1/2) labels to all other vertices.

a is a small constant that we set based on c and D.

Non-rigorous justification The idea behind the algorithm is that if the expected degree c is high enough, then there is $\epsilon(a) > 0$ such that $\epsilon \to 0$ as $a \to 0$, and such that for a constant fraction of vertices v, we have

$$\{p_u \mid u \in N_v(an)\} \subset D_\epsilon(p_v),$$

where $D_\epsilon(p_v)$ is the disk of radius ϵ centered at v. If a point is sampled from $D_\varepsilon(p_v)$, then it has probability $\frac{1}{2} + \delta_{p_v}$ of being in one community, and probability $\frac{1}{2} - \delta_{p_v}$ of being in the other community. So, roughly speaking, the algorithm achieves agreement $\frac{1}{2} + a\delta_{p_v} + o(1)$, given that it chooses v as its random vertex. Letting c large enough, we can ensure that there is uniform $\delta > 0$ such that a constant fraction f' of vertices have the property above and also have $\delta_v > \delta$. So, heuristically, with probability $\approx f'$ the algorithm will achieve agreement $\frac{1}{2} + a\delta > \frac{1}{2}$, as desired.

It seems that one can show that $\{p_u \mid u \in N_v(an)\} \subset D_\epsilon(p_v)$ with constant probability by (1) using arguments from [10] relating finite random geometric graphs to infinite geometric graphs in which the vertices are points drawn from a Poisson point process (continuum percolation), and by then (2) using arguments on the concentration of the ratio of graph-distance to straight-line-distance in homogeneous continuum percolation ([5], [11]).

5 Acknowledgements

I would like to thank Prof. Emmanuel Abbe for being an incredible mentor – both in this project and in my senior thesis. I would also like to thank Princeton's math department for funding me over the summer of 2017, during which I completed most of this work.

References

[1] Emmanuel Abbe. "Community Detection and Stochastic Block Models". In: *In preparation* (2018).

[2] Emmanuel Abbe. "Community detection and stochastic block models: recent developments". In: *arXiv preprint arXiv:1703.10146* (2017).

[3] Chandler Davis and William Morton Kahan. "The rotation of eigenvectors by a perturbation. III". In: *SIAM Journal on Numerical Analysis* 7.1 (1970), pp. 1–46.

[4] Aurelien Decelle et al. "Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications". In: *Physical Review E* 84.6 (2011), p. 066106.

[5] Olivier Garet and Régine Marchand. "Large deviations for the chemical distance in supercritical Bernoulli percolation". In: *The Annals of Probability* (2007), pp. 833–866.

[6] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. "Stochastic blockmodels: First steps". In: *Social networks* 5.2 (1983), pp. 109–137.

[7] Laurent Massoulié. "Community detection thresholds and the weak Ramanujan property". In: *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*. ACM. 2014, pp. 694–703.

[8] Elchanan Mossel, Joe Neeman, and Allan Sly. "A proof of the block model threshold conjecture". In: *Combinatorica* (), pp. 1–44.

[9] Elchanan Mossel, Joe Neeman, and Allan Sly. "Reconstruction and estimation in the planted partition model". In: *Probability Theory and Related Fields* 162.3-4 (2015), pp. 431–461.

[10] Mathew Penrose. *Random geometric graphs*. 5. Oxford university press, 2003.

[11] Chang-Long Yao, Ge Chen, and Tian-De Guo. “Large deviations for the graph distance in supercritical continuum percolation”. In: *Journal of Applied Probability* 48.1 (2011), pp. 154–172.