Community Detection in Networks using Graph Distance

Sharmodeep Bhattacharyya and Peter J. Bickel

December 20, 2013

Abstract

The study of networks has received increased attention recently not only from the social sciences and statistics but also from physicists, computer scientists and mathematicians. One of the principal problem in networks is community detection. Many algorithms have been proposed for community finding [37] [44] but most of them do not have have theoretical guarantee for sparse networks and networks close to the phase transition boundary proposed by physicists [18]. There are some exceptions but all have incomplete theoretical basis [16] [14] [29]. Here we propose an algorithm based on the graph distance of vertices in the network. We give theoretical guarantees that our method works in identifying communities for block models and can be extended for degree-corrected block models [25] and block models with the number of communities growing with number of vertices. Despite favorable simulation results, we are not yet able to conclude that our method is satisfactory for worst possible case. We illustrate on a network of political blogs, Facebook networks and some other networks.

1 Introduction

The study of networks has received increased attention recently not only from the social sciences and statistics but also from physicists, computer scientists and mathematicians. With the information boom, a huge number of network data sets have come into prominence. In biology - gene transcription networks, protein-protein interaction network, in social media - Facebook, Twitter, Linkedin networks, information networks arising in connection with text mining, technological networks such as the Internet, ecological and epidemiological networks and many others have appeared. Although the study of networks has a long history in physics, social sciences and mathematics literature and informal methods of analysis have arisen in many fields of application, statistical inference on network models as opposed to descriptive statistics, empirical modeling and some Bayesian approaches [39] [28] [23] has not been addressed extensively in

the literature. A mathematical and systematic study of statistical inference on network models has only started in recent years.

One of the fundamental questions in analysis of such data is detecting and modeling community structure within the network. A lot of algorithmic approaches to community detection have been proposed, particularly in the physics and computer science literature [41] [36] [20]. In terms of community detection, there are two different goals that researchers have tried to pursue -

- **Algorithmic Goal:** Identify the community each vertex of the network belongs to.

- **Theoretical Goal:** If the network is generated by an underlying generative model, then, what is the probability of success for the algorithm.

1.1 Algorithms

Several popular algorithms for community detection have been proposed in physics, computer science and statistics literature. Most of these algorithms show decent performance in community detection for selected real-world and simulated networks [30] and have polynomial time complexity. We shall briefly mention some of these algorithms.

1. Modularity maximizing methods [42]. One of the most popular method of community detection. The problem is NP hard but spectral relaxations of polynomial complexity exist [40].

2. Hierarchical clustering techniques [15].

3. Spectral clustering based methods [37] [16], [44] [13]. These methods are also very popular. Most of the time these methods have linear or polynomial running times. Mostly shown to work for dense graphs only.

4. Profile likelihood maximization [7]. The problem is NP hard, but heuristic algorithms have been proposed, which have good performance for dense graphs.

5. Stochastic Model based methods:

    - MCMC based likelihood maximization by Gibbs Sampling, the cavity method and belief propagation based on stochastic block model. [18]

    - Variational Likelihood Maximization based on stochastic block model [11], [6]. Polynomial running time but appears to work only for dense graphs.

- Pseudo-likelihood Maximization [14]. Fast method which works well for both dense and sparse graphs. But the method is not fully justified.
- Model-based:
  (a) Mixed Membership Block Model [2]. Iterative method and works for dense graphs. The algorithm for this model is based on variational approximation of the maximum likelihood estimation.
  (b) Degree-corrected block model [25]: Incorporates degree inhomogeneity in the model. Algorithms based on maximum likelihood and profile likelihood estimation has been developed.
  (c) Overlapping stochastic block model [31]: Stochastic block model where each vertex can lie within more than one community. The algorithm for this model is based on variational approximation of the maximum likelihood estimation.
  (d) Mixed configurations model [4]: Another extension to degree-corrected stochastic block model, where, the model is a mixture of configurations model (degree-corrected block model with one block) and each vertex can lie in more than one community. The algorithm for this model is based on the EM algorithm and maximum likelihood estimation.

6. Model based clustering [22].

1.2 Theoretical Goal

The stochastic block model (SBM) is perhaps the most commonly used and best studied model for community detection. An SBM with Q blocks states that each node belongs to a community $\boldsymbol{c} = (c_1, \ldots, c_n) \in \{1, \ldots, Q\}$ which are drawn independently from the multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_Q)$, where $\pi_i > 0$ for all i, and Q is the number of communities, assumed known. Conditional on the labels, the edge variables A_{ij} for $i < j$ are independent Bernoulli variables with

$$\mathbb{E}[A_{ij}|\boldsymbol{c}] = P_{c_i c_j}, \tag{1}$$

where $P = [P_{ab}]$ and $K = [K_{ab}]$ are $Q \times Q$ symmetric matrix. P can be considered the *connection probability* matrix, where as K is the *kernel* matrix for the connection. So, we have $P_{ab} \leq 1$ for all $a, b = 1, \ldots, Q$, $P\mathbf{1} \leq \mathbf{1}$ and $\mathbf{1}^T P \leq \mathbf{1}$ element-wise. The network is undirected, so $A_{ji} = A_{ij}$, and $A_{ii} = 0$ (no self-loops). The problem of community detection is then to infer the node labels

$\boldsymbol{c}$ from A. Thus we are not really interested in estimation or inference on parameters $\boldsymbol{\pi}$ and P, but, rather we are interested in estimating $\boldsymbol{c}$. But, it does not mean the two problems are mutually exclusive. In reality, the inferential problem and the community detection problem are quite interlinked.

The theoretical results of community detection for stochastic block models can be divided into 3 different regimes -

(a) $\frac{\mathbb{E}(\text{degree})}{\log n} \to \infty$, equivalent to, $\mathbb{P}[\text{there exists an isolated point}] \to 0$.

(b) $\mathbb{E}(\text{degree}) \to \infty$, which means existence of giant component, but also presence of isolated small components from Theorem 2.7.

(c) If $\mathbb{E}(\text{degree}) = O(1)$, phase boundaries exist, below which community identification is not possible.

Note:

(a) All of the above mentioned algorithms perform satisfactorily on regime (a).

(b) None of the above algorithms have been shown to have near perfect probability of success under either regime (b) or (c), for the full parameter space. Some algorithms like [16] [7] [13] [14] are shown to partially work in the sparse setting. Some very recent algorithms include [29] [43].

In this paper, we shall only concentrate on stochastic block models. In the future, we shall try to extend our method and results for more general models.

1.3 Contributions and Outline of the Chapter

In real life networks, most of the time we seem to see moderately sparse networks [33] [34] [35]. Most of the large or small complex networks we see seem to fall in the (b) regime of Section 1.2 we describe before, that is, $\mathbb{E}(\text{degree}) \to \infty$. We propose a simple algorithm, which performs well in practice in both regimes (b) and (c) and has some theoretical backing If degree distribution can identify block parameters then classification using our method should give reasonable result in practice.

Our algorithm is based on graph distance between vertices of the graph. We perform spectral clustering based on the graph distance matrix of the graph. By looking at the graph distance matrix instead of adjacency matrix for spectral clustering increases the performance of the community detection, as the normalized distance between cluster centers increases when we go from the adjacency matrix to the graph distance matrix. This helps in community detection even for sparse matrices. We only show theoretical results for stochastic block models. The theoretical proofs are quite intricate and involve careful coupling of

the stochastic block model with multi-type branching process to find asymptotic distribution of the typical graph distances. Then, a careful analysis of the eigenvector of the asymptotic graph distance matrix reveals the existence of separation needed for spectral clustering to succeed. This method of analysis has been used for spectral clustering analysis using the adjacency matrix also [46], but the analysis is simpler.

The rest of the paper is organized as follows. We give a summary of the preliminary results needed in Section 2. We present the algorithms in Section 3. We give an outline of proof of theoretical guarantee of performance of the method and then the details in Section 4. The numerical performance of the methods is demonstrated on a range of simulated networks and on some real world networks in Section 5. Section 6 concludes with discussion, and the Appendix contains some additional technical results.

2 Preliminaries

Let us suppose that we have a random graph G_n as the data. Let $V(G_n) = \{v_i, \ldots, v_n\}$ denote the vertices of G_n and $E(G_n) = \{e_1, \ldots, e_m\}$ denote the edges of G_n. So, the number of vertices in G_n is $|V(G_n)| = n$ and number of edges of G_n is $|E(G_n)| = m$. Let the adjacency matrix of G_n be denoted by $A_{n\times n}$. For the sake of notational simplicity, from here onwards we shall denote G_n by G having n vertices unless specifically mentioned. We consider the n vertices of G are clustered into Q different communities with each community having size n_a, $a = 1, \ldots, Q$ and $\sum_a n_a = n$. In this paper, we are interested in the problem of *vertex community identification* or *graph partitioning*. That means that we are interested in finding which of the Q different community each vertex of G belongs to. However, the problem is an *unsupervised learning* problem. So, we assume that the data is coming from an underlying model and we try to verify how good 'our' *community detection* method works for that model.

2.1 Model for Community Detection

As a model for community detection, we consider the stochastic block model. We shall define the stochastic block model shortly, but, we first we shall introduce some more general models, of which stochastic block model is a special case.

2.1.1 Bickel-Chen Model

The general non-parametric model, as described in Bickel, Chen and Levina (2011) [8], that generates the random data network G can be defined by the following equation -

$$\mathrm{P}(A_{ij} = 1|\xi_i = u, \xi_j = v) = h_n(u, v) = \rho_n w(u, v)\mathbf{1}(w \leq \rho_n^{-1}), \tag{2}$$

where, $w(u, v) \geq 0$, symmetric, $0 \leq u, v \leq 1$, $\rho_n \to 0$. For block models, the latent variable for each vertex $(\xi_1, \ldots, \xi_n)$ can be considered to be coming from a discrete and finite set. Then, each element of that set can be considered to be inducing a partition in the vertex set $V(G_n)$. Thus, we get a model for vertex partitioning, where, the set of vertices can be partitioned into finite number of disjoint classes, but however the partition to which each vertex belongs to is the latent variable in the model and thus unknown. The main goal becomes estimating this latent variable.

2.1.2 Inhomogeneous Random Graph Model

The inhomogeneous random graph model (IRGM) was introduced in Bollobás et. al. (2007) [9]. Let $\mathcal{S}$ be a separable metric space equipped with a Borel probability measure μ. For most cases $\mathcal{S} = (0, 1]$ with μ Lebesgue measure, that means a $U(0, 1)$ distribution. The "kernel" κ will be a symmetric non-negative function on $\mathcal{S} \times \mathcal{S}$. For each n we have a deterministic or random sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ of points in $\mathcal{S}$. Writing δ_x for the measure consisting of a point mass of weight 1 at x, and

$$\nu_n \equiv \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$$

for the empirical distribution of $\boldsymbol{x}$, it is assumed that ν_n converges in probability to μ as $n \to \infty$, with convergence in the usual space of probability measures on $\mathcal{S}$. One example where the convergence holds is the random case, where the x_i are independent and identically distributed on $\mathcal{S}$ with distribution μ convergence in probability holds by the law of large numbers. Of course, we do not need $(x_n)_{n\geq 1}$ to be defined for every n, but only for an infinite set of integers n. From here onwards, we shall only focus on this special case, where, $(x_1, \ldots, x_n) \overset{iid}{\sim} \mu$.

Definition 2.1. *A **kernel** κ_n on a ground space $(\mathcal{S}, \mu)$ is a symmetric non-negative (Borel) measurable function on $\mathcal{S} \times \mathcal{S}$. κ is also continuous a.e. on $\mathcal{S} \times \mathcal{S}$. By a kernel on a vertex space $(\mathcal{S}, \mu, (x_n)_{n\geq 1})$ we mean a kernel on $(\mathcal{S}, \mu)$.*

Given the (random) sequence $(x_1, \ldots, x_n)$, we let $G(n, \kappa)$ be the random graph $G(n, (p_{ij}))$ with

$$p_{ij} \equiv \min\{\kappa(x_i, x_j)/n, 1\}. \tag{3}$$

In other words, $G^{\mathcal{V}}(n,\kappa)$ has n vertices $\{1,\ldots,n\}$ and, given $x_1,\ldots,x_n$, an edge ij (with $i \neq j$) exists with probability p_{ij}, independently of all other (unordered) pairs ij. Based on the graph kernel we can also define an integral operator T_κ in the following way

Definition 2.2. *The **integral operator** $T_\kappa : L^2(\mathcal{S}) \to L^2(\mathcal{S})$ corresponding to $G(n,\kappa)$, is defined as*

$$T_\kappa f(x)(\cdot) = \int_0^1 \kappa(x,y) f(y) d\mu(y),$$

where, $x \in \mathcal{S}$ and any measurable function $f \in L^1(\mathcal{S})$.

The random graph $G(n,\kappa)$ depends not only on κ but also on the choice of $x_1,\ldots,x_n$. The freedom of choice of x_i in this model gives some more flexibility than Bickel-Chen model. The asymptotic behavior of $G(n,\kappa)$ depend very much on $\mathcal{S}$ and μ. Many of these key results such as existence of giant component, typical distance, phase transition properties are proved in [9]. We shall use these results on inhomogeneous random graphs in order to prove results on graph distance for stochastic block models.

Here is further comparison of the Inhomogeneous random graph model (IRGM) with the Bickel-Chen model (BCM), to understand their similarities and dissimilarities -

(a) In BCM, $(\xi_1,\ldots,\xi_n) \overset{iid}{\sim} U(0,1)$ are the latent variables associated with the vertices $(v_1,\ldots,v_n)$ of random graph G_n. Similarly, in IRGM, $(x_1,\ldots,x_n) \sim \mu$ are the latent variables associated with the vertices $(v_1,\ldots,v_n)$ of random graph G_n. Now, if in IRGM, $(x_1,\ldots,x_n) \overset{iid}{\sim} \mu$ then the latent variable structure of the two models become equivalent.

(b) In BCM, the conditional probability of connection between two vertices given the value of their latent variables is controlled by the kernel function $h_n(u,v)$. In IRGM, the conditional probability of connection between two vertices given the value of their latent variables is controlled by the kernel function $\frac{\kappa(u,v)}{n}$.

(c) So, if $h_n(u,v) = \kappa(u,v)/n$, $\mathcal{S}[(0,1)$ and the underlying measure spaces are same and the measure μ is a uniform measure on interval $\mathcal{S} = (0,1)$, then, BCM and IRGM generates graphs from the same distribution. In fact, as noted in [7], if $\mathcal{S} = \mathbb{R}$ and μ has a positive density with respect to Lebesgue measure, then the (limiting) IRGM is equivalent to Bickel-Chen model with suitable h_n.

(d) For IRGM, let us define

$$\lambda \equiv ||T_\kappa|| \equiv \sup_{f \in L^2(\mathcal{S}), ||f||_{L^2(\mathcal{S})}=1} \int_{\mathcal{S}} \int_{\mathcal{S}} \kappa(u,v) f(u) f(v) d\mu(u) d\mu(v),$$

where, T_κ is the operator define in Definition 2.2 and $||\cdot||$ is the operator norm. In BCM,

$$\rho_n \equiv \int_0^1 \int_0^1 h_n(u,v)dudv.$$

If BCM and IRGM have same underlying measure spaces ($\mathcal{S} = (0,1), \mu = U(0,1)$) and $h_n(u,v) = \kappa(u,v)/n$ and

Case 1: $\mathbf{1}$ is the principal eigenfunction of T_κ, then

$$n\rho_n \to \lambda$$

where, λ is as defined above.

Case 2: $\mathbf{1}$ is not the principal eigenfunction of T_κ, then

$$n\rho_n \leq \lambda$$

In case of BCM $n\rho_n$ is the natural scaling parameter for the random graph, since, $\mathbb{E}[\text{Number of Edges in } G_n] = \frac{1}{2}n\rho_n$. In case of IRGM, λ is fixed. However, we shall see that the limiting behavior of the graph distance between two vertices of the network becomes dependent on the parameter λ. So, the parameter λ still remains of importance. We shall henceforth focus on IRGM, with parameter of importance being λ

2.1.3 Stochastic Block Model

The stochastic block model is perhaps the most commonly used and best studied model for community detection. We continue with IRGM framework, so the graph is sparse.

Definition 2.3. *A graph $G^Q(,(P,\boldsymbol{\pi}))$ generated from **stochastic block model** (SBM) with Q blocks and parameters $P \in (0,1)^{Q\times Q}$ and $\boldsymbol{\pi} \in (0,1)^Q$ can be defined in following way - each vertex of graph G_n from an SBM belongs to a community $\boldsymbol{c} = (c_1,\ldots,c_n) \in \{1,\ldots,Q\}$ which are drawn independently from the multinomial distribution with parameter $\boldsymbol{\pi} = (\pi_1,\ldots,\pi_Q)$, where $\pi_i > 0$ for all i. Conditional on the labels, the edge variables A_{ij} for $i < j$ are independent Bernoulli variables with*

$$\mathbb{E}[A_{ij}|\boldsymbol{c}] = P_{c_ic_j} = \min\{\frac{K_{c_ic_j}}{n}, 1\}, \tag{4}$$

*where $P = [P_{ab}]$ and $K = [K_{ab}]$ are $Q\times Q$ symmetric matrices. P is known as the **connection probability** matrix and K as the **kernel** matrix for the connection. So, we have $P_{ab} \leq 1$ for all $a,b = 1,\ldots,Q$, $P\mathbf{1} \leq \mathbf{1}$ and $\mathbf{1}^T P \leq \mathbf{1}$ element-wise.*

The network is undirected, so $A_{ji} = A_{ij}$, and $A_{ii} = 0$ (no self-loops). The problem of community detection is then to infer the node labels $\boldsymbol{c}$ from A. Thus we are not really interested in estimation or inference on parameters $\boldsymbol{\pi}$ and P, but, rather we are interested in estimating $\boldsymbol{c}$. But, it does not mean the two problems are mutually exclusive, in reality, the inferential problem and the community detection problem are quite interlinked.

We can see that SBM is a special case of both Bickel-Chen model and IRGM. In IRGM, if we consider $\mathcal{S}$ to be a finite set, $(x_1, \ldots, x_n) \in [Q]^n$ ($[Q] = \{1, \ldots, Q\}$) with $x_i \overset{iid}{\sim} Mult(n, \boldsymbol{\pi})$ and kernel $\kappa : [Q] \to [Q]$ as $\kappa(a, b) = K_{ab}$ $(a, b = 1, \ldots, Q)$, then the resulting IRGM graph follows stochastic block model. So, for SBM we can define an *integral operator* on $[Q]$ with measure $\{\pi_1, \ldots, \pi_Q\}$.

Definition 2.4. *The* ***integral operator*** $T_K : \ell^1(\mathcal{S}) \to \ell^1(\mathcal{S})$ *corresponding to* $G^Q(n, (P, \boldsymbol{\pi}))$*, is defined as*

$$(T_K(x))_a = \sum_{b=1}^{Q} K_{ab}\pi_b x_b, \text{ for } a = 1, \ldots, Q$$

where, $x \in \mathbb{R}^Q$.

The stochastic block model has deep connections with Multi-type branching process, just as, Erodös-Rényi random graph model (ERRGM) has connections with the branching process. Let us introduce branching process first.

2.2 Multi-type Branching Process

We shall try to link network formed by SBM with the tree network generated by multi-type Galton-Watson branching process. In our case, the Multi-type branching process (MTBP) has type space $S = \{1, \ldots, Q\}$, where a particle of type $a \in S$ is replaced in the next generation by a set of particles distributed as a Poisson process on S with intensity $(K_{ab}\pi_b)_{b=1}^{Q}$. We denote this branching process, started with a single particle of type a, by $\mathcal{B}_{K,\pi}(a)$. We write $\mathcal{B}_{K,\pi}$ for the same process with the type of the initial particle random, distributed according to $\boldsymbol{\pi}$.

Definition 2.5. *(a) Define* $\rho_k(K, \pi; a)$ *as the probability that the branching process*
$\mathcal{B}_{K,\pi}(a)$ *has a total population of exactly* k *particles.*

(b) Define $\rho_{\geq k}(K, \pi; a)$ *as the probability that the total population is at least* k*.*

(c) Define $\rho(K, \pi; a)$ *as the probability that the branching process survives for eternity.*

(d) Define,

$$\rho_k(K,\pi) \equiv \sum_{a=1}^{Q} \rho_k(K,\pi;a)\pi_a, \quad \rho \equiv \rho(K,\pi) \equiv \sum_{a=1}^{Q} \rho(K,\pi;a)\pi_a \tag{5}$$

*and define $\rho_{\geq k}(K)$ analogously. Thus, $\rho(K,\pi)$ is the **survival probability** of the branching process $\mathcal{B}_{K,\pi}$ given that its initial distribution is $\boldsymbol{\pi}$*

If the probability that a particle has infinitely many children is 0, then $\rho(K,\pi;a)$ is equal to $\rho_\infty(a)$, the probability that the total population is infinite. As we shall see later, the branching process $\mathcal{B}_{K,\pi}(a)$ arises naturally when exploring a component of G_n starting at a vertex of type a; this is directly analogous to the use of the single-type Poisson branching process in the analysis of the Erdös-Rényi graph $G(n, c/n)$.

2.3 Known Results for Stochastic Block Model

The performance of community detection algorithms depends on the parameters $\boldsymbol{\pi}$ and P. We refer to Definition 2.3 for definition of stochastic block models. An important condition that we usually put on parameter P is *irreducibility.*

Definition 2.6. *A connection matrix P on a $\mathcal{S} = \{1,\ldots,Q\}$ is **reducible** if there exists $A \subset \mathcal{S}$ with $0 < |A| < Q$ such that $P = 0$ a.e. on $A \times (\mathcal{S} - A)$; otherwise P is **irreducible**. Thus P is **irreducible** if $A \subseteq \mathcal{S}$ and $P = 0$ a.e. on $A \times (\mathcal{S} - A)$ implies $|A| = 0$ or $|A| = Q$.*

So, the results on existence of giant components in [9] also apply for SBM. The following theorem describes the result on existence of giant components.

Theorem 2.7 ([9])**.** *Let us define operator T_K as in definition 2.4,*

(i) If $||T_K|| \leq 1$ ($||\cdot||$ refer to operator norm), then the size of largest component is $o_P(n)$, while if $||T_K|| > 1$, then the size of largest component is $\Theta_P(n)$ whp.

(ii) If P is irreducible, then $\frac{1}{n}$(Size of largest component) $\to \boldsymbol{\pi}^T\rho$, where, $\rho \in [0,1]^Q$ is the survival probability as defined in (5).

The theoretical results on community detection depend on the 3 different regime on which the generative model is based on -

(a) $\frac{\mathbb{E}(\text{degree})}{\log n} \to \infty$, equivalent to, $\mathbb{P}[\text{there exists an isolated point}] \to 0$. In this setting, there are several algorithms, such as those described in Section 1, can identify correct community with high probability under quite relaxed conditions on parameters P and $\boldsymbol{\pi}$. See [13] (Theorem 2 and 3), [44] (Theorem 3.1), [16] (Theorem 1).

(b) $\mathbb{E}(\text{degree}) \to \infty$, which means existence of giant component, but also presence of isolated small components from Theorem 2.7. In this setting, algorithms proposed in [16], [14] is proved to identify community labels that are highly correlated with original community labels with high probability.

(c) If $\mathbb{E}(\text{degree}) = O(1)$, phase boundaries exist, below which community identification is not possible. These results and rigorous proof are given in [38]. The results can be summarized for 2-block model with parameters $P_{11} = a, P_{12} = b, P_{22} = a$ as

Theorem 2.8 ([38]). *(i) If $(a-b)^2 < 2(a+b)$ then probability model of SBM and ERRGM with $p = \frac{a+b}{2n}$ are mutually contiguous. Moreover, if $(a-b)^2 < 2(a+b)$, there exists no consistent estimators of a and b.*

(ii) If $(a-b)^2 > 2(a+b)$ then probability model of SBM and ERRGM with $p = \frac{a+b}{2n}$ are asymptotically orthogonal.

So, in the range $(a-b)^2 > 2(a+b)$, there should exists an algorithm which identifies highly correct clustering with high probability at least within the giant components.

3 Algorithm

The algorithm we propose depend on the graph distance or geodesic distance between vertices in a graph.

Definition 3.1. ***Graph distance*** *or* ***Geodesic distance*** *between two vertices i and j of graph G is given by the length of the shortest path between the vertices i and j, if they are connected. Otherwise, the distance is infinite.*

So, for any two vertices $u, v \in V(G)$, graph distance, d_g is defined by

$$d_g(u,v) = \begin{cases} |V(e)|, & \text{if } e \text{ is the shortest path connecting } u \text{ and } v \\ \infty, & u \text{ and } v \text{ are not connected} \end{cases}$$

For sake of numerical convenience, we shall replace ∞ by a large number for value of $d_g(u,v)$, when, u and v are not connected. The main steps of the algorithm can be described as follows

1. Find the graph distance matrix $D = [d_g(v_i, v_j)]_{i,j=1}^n$ for a given network but with distance upper bounded by $k \log n$. Assign non-connected vertices an arbitrary high value B.

2. Perform hierarchical clustering to identify the giant component G^C of graph G. Let $n_C = |V(G^C)|$.

3. Normalize the graph distance matrix on G^C, D^C by

$$\bar{D}^C = -\left(I - \frac{1}{n_C}\mathbf{1}\mathbf{1}^T\right)(D^C)^2\left(I - \frac{1}{n_C}\mathbf{1}\mathbf{1}^T\right)$$

4. Perform eigenvalue decomposition on $\bar{D}^C$.

5. Consider the top Q eigenvectors of normalized distance matrix $\bar{D}^C$ and $\tilde{\mathbf{W}}$ be the $n \times Q$ matrix formed by arranging the Q eigenvectors as columns in $\tilde{\mathbf{W}}$. Perform Q-means clustering on the rows $\tilde{\mathbf{W}}$, that means, find an $n \times Q$ matrix $\mathbf{C}$, which has Q distinct rows and minimizes $||\mathbf{C} - \tilde{\mathbf{W}}||_F$.

6. (Alternative to 5.) Perform Gaussian mixture model based clustering on the rows of $\tilde{\mathbf{W}}$, when there is an indication of highly-varying average degree between the communities.

7. Let $\hat{\xi} : V \mapsto [Q]$ be the block assignment function according to the clustering of the rows of $\tilde{\mathbf{W}}$ performed in either Step 5 or 6.

Here are some important observations about the algorithm -

(a) There are standard algorithms for graph distance finding in the algorithmic graph theory literature. In algorithmic graph theory literature the problem is known as the **all pairs shortest path** problem. The two most popular algorithms are Floyd-Warshall [19] [48] and Johnson's algorithm [24]. The time complexity of the Floyd-Warshall algorithm is $O(n^3)$, where as, the time complexity of Johnson's algorithm is $O(n^2 \log n + ne)$ [32] ($n = |V(G_n)|$ and $e = |E(G_n)|$). So, for sparse graphs, Johnson's algorithm is faster than Floyd-Warshall. Memory storage is also another issue for this algorithm, since the algorithm involves a matrix multiplication step of complexity $\Omega(n^2)$. Recently, there also has been some progress on parallel implementation of all-pairs shortest path problem [45] [10] [21], which addresses both memory and computation aspects of the algorithm and lets us scale the algorithm for large graphs, both dense and sparse.

(b) The Step 3 of the algorithm is nothing but the classical multi-dimensional scaling (MDS) of the graph distance matrix. In MDS, we try to find vectors $(x_1, \ldots, x_n)$, where, $x_i \in \mathbb{R}^Q$, such that,

$$\sum_{i,j=1}^{n} \left(||x_i - x_j||_2 - (D^C)_{ij}\right)^2$$

is minimized. The minimizer is attained by the rows of the matrix formed by the top Q eigenvectors of $\bar{D}^C$ as columns. So, performing spectral clustering on $\bar{D}^C$ is the same as performing Q-means clustering on the multi-dimensional scaled space.

Instead of $\bar{D}^C$, we could also use the matrix $(D^C)^2$, but then, the topmost eigenvector does not carry any information about the clustering. Similarly, we can also use the matrix D^C directly for spectral clustering, but, in that case, D^C is not a positive semi-definite matrix and as a result we have to consider the eigenvectors corresponding to largest absolute eigenvalues (since eigenvalues can be negative).

(c) In the Step 5 of the algorithm Q-means clustering if the expected degree of the blocks are equal. However, if the expected degree of the blocks are different, it leads to multi scale behavior in the eigenvectors of the normalized distance matrix. So, we perform Gaussian Mixture Model (GMM) based clustering instead of Q-means to take into account the multi scale behavior.

4 Theory

Let us consider that we have a random graph G_n as the data. Let $V(G_n) = \{v_i, \ldots, v_n\}$ denote the vertices of G_n and $E(G_n) = \{e_1, \ldots, e_m\}$ denote the edges of G_n. So, the number of vertices in G_n is $|V(G_n)| = n$ and number of edges of G_n is $|E(G_n)| = m$. Let the adjacency matrix of G_n be denoted by $A_{n\times n}$. For sake of notational simplicity, from here onwards we shall denote G_n by G having n vertices unless specifically mentioned. There are Q communities for the vertices and each community has $(n_a)_{a=1}^Q$ number of vertices. In this paper, we are interested in the problem of *vertex community identification* or *graph partitioning*. However, the problem is an *unsupervised learning* problem. So, we assume that the data is coming from an underlying model and we try to verify how good 'our' *community detection* method works for that model.

The theoretical analysis of the algorithm has two main parts -

I. Finding the limiting distribution of graph distance between two typical vertices of type a and type b (where, $a, b = 1, \ldots, Q$). This part of the analysis is highly dependent on results from multi-type branching processes and their relation with stochastic block models. The proof techniques and results are borrowed from [9], [5] and [3].

II. Finding the behavior of top Q eigenvectors of the graph distance matrix D using the limiting distribution of the typical graph distances. This part of analysis is highly dependent on perturbation theory of linear operators. The proof techniques and results are borrowed from [26], [12] and [46].

4.1 Results of Part I

We shall give limiting results for *typical distance* between vertices in G_n. If u and $v \in V(G_n)$ are two vertices in G_n, which has been selected uniformly at random from type a and type b respectively, where, $a, b = 1, \ldots, Q$ are the different communities. Then, the graph distance between u and v is $d_G(u, v)$. Now, the operator that controls the process is T_K as defined in Definition 2.4. T_K is another representation of the matrix $\tilde{K}_{Q\times Q}$, which is defined as

$$\tilde{K}_{ab} \equiv \pi_a K_{ab} \pi_b, \text{ for } a, b = 1, \ldots, Q \tag{6}$$

The matrix $\tilde{K}$ defines the quadratic form for $T_K : \ell^1(\mathcal{S}, \pi) \to \ell^1(\mathcal{S}, \pi)$. So, we have that

$$\lambda \equiv ||T_K|| = \lambda_{max}(\tilde{K}). \tag{7}$$

The relation between λ and $\mathbb{E}$[number of Edges in G_n] is given Section 2.1.2. Here, we use λ as the scaling operator, not either average, minimum or maximum degree of vertices as used in [46] and [44]. But, we already know that, if the graph is *homogeneous*, then, $\mathbb{E}$[number of Edges in G_n] $= \frac{1}{2}\lambda$ and otherwise $\mathbb{E}$[number of Edges in G_n] $\leq \lambda$.

Let us also denote, $\nu \in \mathbb{R}^Q$ as the eigenvector of $\tilde{K}$ corresponding to λ. We at first, try to find an asymptotic bound on the graph distance $d_G(u, v)$ for vertices $u, v \in V(G)$.

Theorem 4.1. *Let $\lambda > 1$ (defined in Eq. (7)), then, the graph distance $d_G(u, v)$ between two uniformly chosen vertices of type a and b respectively, conditioned on being connected, satisfies the following asymptotic relation -*

(i)

$$\mathrm{P}\left[d_G(u, v) < (1 - \varepsilon)\frac{\log n}{\log|\lambda|/\log(\nu_a \nu_b)}\right] = o(1) \tag{8}$$

(ii)

$$\mathrm{P}\left[d_G(u, v) > (1 + \varepsilon)\frac{\log n}{\log|\lambda|/\log(\nu_a \nu_b)}\right] = o(1) \tag{9}$$

Now, let us consider the limiting operator $\mathbb{D}$ defined as

Definition 4.2. *The **normalized limiting matrix** is an $n\times n$ matrix, $\mathbb{D}$, which in limit as $n \to \infty$ becomes an operator on l_2 (space of convergent sequences), is defined as $\mathbb{D} = [\mathbb{D}_{ij}]_{i,j=1}^n$, where,*

$$\mathbb{D}_{ij} \quad = \quad \begin{cases} \frac{\log(\nu_a \nu_b)}{\log|\lambda|}, & \text{if type of } v_i = a \neq b = \text{ type of } v_j \\ \frac{2\log(\nu_a)}{\log|\lambda|}, & \text{if type of } v_i = \text{ type of } v_j = a \end{cases}$$

and $\mathbb{D}_{ii} = 0$ *for all* $i = 1, \ldots, n$.
The ***graph distance matrix*** $\mathbf{D}$ *can be defined as*

$$\mathbf{D} = [d(v_i, v_j)]_{i,j=1}^{n}.$$

In Theorem 4.1 we had a point-wise result, so, we combine these point-wise results to give a matrix result -

Theorem 4.3. *Let* $\lambda = ||T_K|| > 1$*, then, within the big connected component,*

$$\mathrm{P}\left[\left|\left|\frac{\mathbf{D}}{\log n} - \mathbb{D}\right|\right|_F \leq O(n^{1-\varepsilon})\right] = 1 - o(1)$$

Thus, the above theorem gives us an idea about the limiting behavior of the normalized version of geodesic matrix $\mathbf{D}$.

4.1.1 Sketch of Analysis of Part I

A rough idea of the proof of part I is as follows. Fix two vertices, say 1 and 2, in the giant component. Think of a branching process starting from vertices of type 1 and 2, so that at time t, $\mathcal{B}_{P\pi}(a)(t)$ is the branching process tree from vertex of type a and includes the shortest paths to all vertices in the tree at or before time t from vertex a, $a = 1, 2$. When these two trees meet via the formation of an edge (v_1, v_2) between two vertices $v_1 \in \mathcal{B}_{P\pi}(1)(\cdot)$ and $v_2 \in \mathcal{B}_{P\pi}(2)(\cdot)$, then the shortest-length path between the two vertices 1 and 2 has been found. If $D_n(v_a)$, $a = 1, 2$, denotes the number of edges between the source a and the vertex v_a along the tree $\mathcal{B}_{P\pi}(a)$, then the graph distance $d_n(1, 2)$ is given by

$$d_n(1, 2) = D_n(v_1) + D_n(v_2) + 1 \tag{10}$$

The above idea is indeed a very rough sketch of our proof and it follows from the graph distance finding idea developed in [9]. In the paper, we embed the SBM in a multi-type branching process (MTMBP) or a single-type marked branching process (MBP), depending on whether the types of two vertices are same or not. The offspring distribution is binomial with parameters $n - 1$ and kernel P (see Section 4.4). With high probability, the vertex exploration process in the SBM can be coupled with two multi-type branching processes, bounding the vertex exploration process on SBM on both sides. Now, using the property of the two multi-type branching processes, we can bound the number of vertices explored in the vertex exploration process of a SBM graph and infer about the asymptotic limit of the graph distance.

With the above sketch of proof can be organized as follows.

1. We analyze various properties of a Galton-watson process conditioned on non-extinction, including times to grow to a particular size. In this branching process, the offspring will have a Poisson distribution.

2. We introduce multi-type branching process trees with binomially distributed offspring and make the connection between these trees and the SBM. We bound the vertices explored for an SBM graph, starting from a fixed vertex, by considering a muti-type branching process coupled with it.

3. We bound the geodesic distance using the number of vertices explored in the coupled multi-type branching processes within a certain generation. The limiting behavior of the generation give us the limiting behavior of graph distance.

4. The whole analysis is true for IRGM. So, the results are true for SBM with increasing block numbers and degree-corrected block models also.

The idea of the argument is quite simple, but making these ideas rigorous takes some technical work, particularly because we need to condition on our vertices being in the giant component.

4.2 Results of Part II

So, from Part I of the analysis, we get an idea about the point-wise asymptotic convergence of the matrix $\mathbf{D} = [d(v_i, v_j)]_{i,j=1}^{n}$ to the normalized limiting operator $\mathbb{D}$, defined in Definition 4.2.

The limiting matrix $\mathbb{D}$ can also be written in terms of limiting low-dimensional matrix, $\mathcal{D}$, which is defined as follows -

Definition 4.4. *The limiting kernel matrix, $\mathcal{D}_{Q\times Q}$ is defined as*

$$\mathcal{D}_{ab} = \begin{cases} \frac{\log(\nu_a\nu_b)}{\log|\lambda|}, & \text{if } a \neq b \\ \frac{2\log(\nu_a)}{\log|\lambda|}, & \text{if } a = b \end{cases}$$

So, we can see that if $\mathbf{J}_{n\times n} = \mathbf{1}\mathbf{1}^T$ is an $n \times n$ matrix of all ones, then, there exists a permutation of rows of $\mathbb{D}$, which is obtained by multiplying $\mathbb{D}$ with permutation matrix $\mathbf{R}$, such that,

$$\mathbb{D}\mathbf{R} = \mathcal{D} \star \mathbf{J} - Diag(\tilde{d}) \equiv [\mathcal{D}_{ab}\mathbf{J}_{ab}]_{a,b=1}^{Q} - Diag(\tilde{d}) \tag{11}$$

where, $[\mathbf{J}_{ab}]_{a,b=1}^{Q}$ is a $Q \times Q$ partition of $\mathbf{J}$ in the following way - the rows and columns are partitioned in similar fashion according to $(n_1, \ldots, n_Q)$. Note that, $(n_a)_{a=1}^{Q}$ are the number of vertices of type a in the graph G_n. So, $\mathbf{J}_{ab}$ is an $n_a \times n_b$ matrix of all ones. $\tilde{d}$ is a vector of length containing n_a elements of value $\frac{2\log(\nu_a)}{\log|\lambda|}$, $a = 1, \ldots, Q$. Note that product $\star$ can also be seen as a Khatri-Rao product of two partitioned matrices [27].

Now, we assume some conditions on the limiting low-dimensional matrix $\mathcal{D}$.

(C1) The operator T_K or the matrix $\tilde{K}$ can not have $\mathbf{1}$ as the principal eigenvector. If the principal eigenvector $\nu = \mathbf{1}$, then, $\mathcal{D}$ becomes a matrix with no difference between diagonal and off-diagonal elements and thus have no discriminatory power to do community detection.

(C2) The eigenvalues of $\mathcal{D}$, $\lambda_1(\mathcal{D}) \geq \cdots \geq \lambda_Q(\mathcal{D})$, satisfy the condition that there exists an constant α, such that, $0 < \alpha \leq \lambda_Q(\mathcal{D})$.

(C3) The eigenvectors of $\mathcal{D}$, $(v_1(\mathcal{D}), \ldots, v_Q(\mathcal{D}))$ corresponding to $\lambda_1, \ldots, \lambda_Q$, satisfy the condition that there exists a constant β, such that, rows of the $Q \times Q$ matrix $\mathbf{V} = [v_1 \cdots v_Q]$, represented as $(u_1, \ldots, u_Q)$ $(u_a \in \mathbb{R}^Q)$, satisfies the condition $0 < \beta \leq ||u_a - u_b||_2$ for all pairs of rows of $\mathbf{V}$.

(C4) The number of vertices in each type $(n_1, \ldots, n_Q)$, satisfy the condition that there exists a constant θ such that $0 < \theta < \frac{n_a}{n}$ for all $a = 1, \ldots, Q$ and all n.

Theorem 4.5. *Under the conditions (C1)-(C4), suppose that the number of blocks Q is known. Let $\hat{\xi} : V \mapsto [Q]$ be the block assignment function according to a clustering of the rows of $\tilde{\mathbf{W}}^{(n)}$ satisfying algorithm in Section 3 and $\xi : V \mapsto [Q]$ be the actual assignment. Let $\mathcal{P}_Q$ be the set of permutations on $[Q]$. With high probability and for large n it holds that*

$$\min_{\pi \in \mathcal{P}_Q} |\{u \in V : \xi(u) \neq \pi(\hat{\xi}(u))\}| = O(n^{1/2-\varepsilon}) \tag{12}$$

4.2.1 Sketch of Proof of Part II

We can consider the limiting distribution of the graph distance matrix as $\mathbf{D}$ which was proposed in Theorem 4.3, with $(\mathbf{D}_{ij}) = d_G(v_i, v_j)$, where, $v_i, v_j \in V(G)$. Our goal is to show that the eigenvectors of $\mathbf{D}$ or normalized version of it, converge to eigenvectors of $\mathcal{D}$ or $\mathbb{D}$. For that reason, we use the perturbation theory of operators, as given in Kato [26] and Davis-Kahan [17]. The steps are as follows

- We use Davis-Kahan to show convergence of eigenspace $\tilde{\mathbf{W}}$, formed by top Q eigenvectors of $\mathbf{D}/\log n$ to $\mathbf{WR}$, where, $\mathbf{W}$ is the eigenspace formed by the top Q eigenvectors of $\mathbb{D}$ and $\mathbf{R}$ is some orthogonal permutation matrix, which permutes the rows of $\mathbf{W}$.
- We show by contradiction that if the clustering assignment makes too many mistakes then the rate of convergence of $\tilde{\mathbf{W}}$ to $\mathbf{WR}$ would be violated.

4.3 Branching Process Results

The branching process $\mathcal{B}_K(a)$ is a multi-type Galton-Watson branching processes with type space $\mathcal{S} \equiv \{1, \ldots, Q\}$, a particle of type $a \in \mathcal{S}$ is replaced in the next generation by its "children", a set of particles whose types are distributed as a Poisson process on $\mathcal{S}$ with intensity $\{K_{ab}\pi_b\}_{b=1}^{Q}$. Recall the parameters $K \in \mathbb{R}^{Q\times Q}$ and $\pi \in [0,1]^Q$ with $\sum_{a=1}^{Q} \pi_a = 1$, from the definition of Stochastic block model in equation (4). The zeroth generation of $\mathcal{B}_K(a)$ consists of a single particle of type a. Also, the branching process $\mathcal{B}_K$ is just the process $\mathcal{B}_K(a)$ started with a single particle whose (random) type is distributed according to the probability measure $(\pi_1, \ldots, \pi_Q)$.

Let us recall our notation for the survival probabilities of particles in $\mathcal{B}_K(a)$. We write $\rho_k(K; a)$ for the probability that the total population consists of exactly k particles, and $\rho_{\geq k}(K; a)$ for the probability that the total population contains at least k particles. Furthermore, $\rho(K; a)$ is the probability that the branching process survives for eternity. We write $\rho_k(K), \rho_{\geq k}(K)$ and $\rho(K)$ for the corresponding probabilities for $\mathcal{B}_K$, so that, e.g., $\rho_k(K) = \sum_{a=1}^{Q} \rho_k(K; a)\pi_a$.

Now, we try to find a coupling relation between *neighborhood exploration process* of a vertex of type a in stochastic block model and multi-type Galton-Watson process, $\mathcal{B}(a)$ starting from a vertex of type a.

We assume all vertices of graph G_n generated from a stochastic block model has been assigned a community or type ξ_i (say) for vertex $v_i \in V(G_n)$. By *neighborhood exploration process* of a vertex of type a in stochastic block model, we mean that we start from a random vertex v_i of type a in the random graph G_n generated from stochastic block model. Then, we count the number of vertices of the random graph G_n are neighbors of v_i, $N(v_i)$. We repeat the neighborhood exploration process by looking at the neighbors of the vertices in $N(v_i)$. We continue until we have covered all the vertices in G_n. Since, we either consider G_n connected or only the giant component of G_n, the neighborhood exploration process will end in finite steps but the number of steps may depend on n.

Lemma 4.6. *Within the giant component, the neighborhood exploration process for a stochastic block model graph with parameters $(P, \pi) = (K/n, \pi)$, can be bounded with high probability by two multi-type branching processes with kernels $(1 - 2\epsilon)K$ and $(1 + \epsilon)K$ for some $\epsilon > 0$.*

Now, we restrict ourselves to the giant component only. So, if we condition that the exploration process does not leave the giant component, it is same as conditioning that the branching process does not die out. Under this additional condition, the branching process can be coupled with another branching process with a different kernel. The kernel of that branching process is given in following lemma.

Lemma 4.7. *If we condition a branching process, $\mathcal{B}_{K\pi}$ on survival, the new branching process has kernel* $(K_{ab}(\rho(K;a)+\rho(K;b)-\rho(K;a)\rho(K;b)))_{a,b=1}^{Q}$.

Now, we shall try to prove the limiting behavior of typical distance between vertices v and w of G_n, where, $v, w \in V(G_n)$.

Lemma 4.8. *Let us have $\lambda \equiv ||T_K|| > 1$ and let $\nu = (\nu_1, \ldots, \nu_Q)$ be the eigenvector of T_K with eigenvalue λ, then,*

$$\mathbb{E}\,|\{\{v,w\} : d_G(v,w) \leq (1-\varepsilon)\log n/\log|\lambda|\}| = O(n^{2-\varepsilon})$$

and so

$$\left|\left\{\{v,w\} : d_G(v,w) \leq (1-\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right\}\right| \leq O(n^{2-\varepsilon/2}) \text{ with high probability}$$

Now, let us try to bound the typical distance between two vertices of the any type. We shall only give an upper bound for typical distance between two vertices of any type.

Lemma 4.9. *Let us have $\lambda \equiv ||T_K|| = \lambda_{\max}(\tilde{K}) > 1$ from Eq (7) and let $\nu = (\nu_1, \ldots, \nu_Q)$ be the eigenvector of T_K with eigenvalue λ. For uniformly selected vertices $v, w \in V(G)$,*

$$\mathrm{P}\left(d_G(v,w) < (1+\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right) = 1 - exp(-\Omega(n^{2\eta}))$$

conditioned on the event that the branching process $\mathcal{B}_K$ survives.

The results on branching process and their proof techniques are mostly taken from [9]. The proof of the Lemma 4.6-4.9 are given in the Appendix.

4.4 Proof of Theorem 4.1 and Theorem 4.3

4.4.1 Proof of Theorem 4.1

We shall try to prove the limiting behavior of typical graph distance in the giant component as $n \to \infty$. The Theorem essentially follows from Lemma 4.8 and Lemma 4.9. Under the conditions mentioned in the Theorem, part (a) follows from Lemma 4.8 and part (b) follows from Lemma 4.9.

4.4.2 Proof of Theorem 4.3

From the definition 4.2, we have that $\mathbf{D}_{ij}$ = graph distance between vertices v_i and v_j, where, $v_i, v_j \in V(G_n)$.

From Lemma 4.8, we get for any vertices v and w with high probability,

$$\left|\left\{\{v,w\} : d_G(v,w) \leq (1-\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right\}\right| \leq O(n^{2-\varepsilon}).$$

Also, from Lemma 4.9, we get

$$\mathrm{P}\left(d_G(v,w) < (1+\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right) = 1 - exp(-\Omega(n^{2\eta}))$$

So, putting the two statements together, we get that with high probability,

$$\sum_{i,j=1:type(v_i)\neq type(v_j)}^{n}\left(\frac{\mathbf{D}_{ij}}{\log n} - \mathbb{D}_{ij}\right)^2 = O(n^{2-\varepsilon}) + O(n^2).\varepsilon^2 = O(n^{2-\varepsilon})$$

since, by Eq. (14) $\epsilon = O(1/\sqrt{n})$ and and $(1 - exp(-\Omega(n^{2\eta})))^{n^2} \to 1$ as $n \to \infty$. So, putting the two cases together, we get that with high probability,

$$\sum_{i,j=1}^{n}\left(\frac{\mathbf{D}_{ij}}{\log n} - \mathbb{D}_{ij}\right)^2 = O(n^{2-\varepsilon}) + O(n^2).\varepsilon^2 = O(n^{2-\varepsilon}).$$

Hence,

$$\left|\left|\frac{\mathbf{D}}{\log n} - \mathbb{D}\right|\right|_F \leq O(n^{1-\varepsilon/2}).$$

4.5 Perturbation Theory of Linear Operators

Once, we have the limiting behavior of the matrix D established in Theorem 4.3, we shall now try to see the behavior of the eigenvectors of the matrix D. Now, matrix D can be considered as a perturbation of the operator $\mathbb{D}$.

The Davis-Kahan Theorem states a bound on perturbation of eigenspace instead of eigenvector, as discussed previously. The $\sin\theta$ Theorem of Davis-Kahan [17]

Theorem 4.10 (Davis-Kahan (1970)[17])**.** *Let $\mathbf{H}, \mathbf{H}' \in \mathbb{R}^{n\times n}$ be symmetric, suppose $\mathcal{V} \subset \mathbb{R}$ is an interval, and suppose for some positive integer d that $\mathbf{W}, \mathbf{W}' \in \mathbb{R}^{n\times d}$ are such that the columns of $\mathbf{W}$ form an orthonormal basis for the sum of the eigenspaces of $\mathbf{H}$ associated with the eigenvalues of $\mathbf{H}$ in $\mathcal{V}$ and that the columns of $\mathbf{W}'$ form an orthonormal basis for the sum of the eigenspaces of $\mathbf{H}'$ associated with the eigenvalues of $\mathbf{H}'$ in $\mathcal{V}$. Let δ be the minimum distance between any eigenvalue of $\mathbf{H}$ in $\mathcal{V}$ and any eigenvalue of $\mathbf{H}$ not in $\mathcal{V}$. Then there exists an orthogonal matrix $\mathbf{R} \in \mathbb{R}^{d\times d}$ such that $||\mathbf{WR} - \mathbf{W}'||_F \leq \sqrt{2}\frac{||\mathbf{H}-\mathbf{H}'||_F}{\delta}$.*

4.6 Proof of Theorem 4.5

Now, we can try to approximate limiting operator by the graph distance matrix $\mathbf{D}$ in Frobenius norm based on Theorem 4.3 of Part I. The behavior of the eigenvalues of the limiting operator $\mathbb{D}$ can be stated as follows -

Lemma 4.11. *The eigenvalues of* $\mathbb{D}$ - $|\lambda_1(\mathbb{D})| \geq |\lambda_2(\mathbb{D})| \geq \cdots \geq |\lambda_n(\mathbb{D})|$, *can be bounded as follows -*

$$\lambda_1(\mathbb{D}) < n,\ |\lambda_K(\mathbb{D})| > Cn,\ \lambda_{Q+1}(\mathbb{D}) = -\min\{\tilde{d}_1, \ldots, \tilde{d}_Q\}, \ldots, \lambda_n = -\max\{\tilde{d}_1, \ldots, \tilde{d}_Q\} \tag{13}$$

where, $\tilde{d}$, *a vector of length* Q, *is defined in Eq.* (11) *and the smallest* $(n-Q)$ *absolute eigenvalues of* $\mathbb{D}$ *are* $-\tilde{d}$ *where* $-\tilde{d}_a$ *has multiplicity* $(n_a - 1)$ *for* $a = 1, \ldots, Q$.

Proof. The matrix $\mathbb{D}$ can be considered as a Khatri-Rao product of the matrices $\mathcal{D}$ and $\mathbf{J}$ according to equation (11). Now, there exists a constant τ such that $\log ||T_K|| > \tau > 0$, since $||T_K|| > 1$. So, we have $\lambda_1(\mathcal{D}) < \tau$. So, we have $\lambda_1(\mathcal{D}) < 1$ and since $n_a \leq n$ for all a and $\sum_a n_a = n$. So, we have $\lambda_1(\mathbb{D}) \leq n$. Now, By Assumption (C2) and (C4), $\lambda_Q(\mathcal{D}) \geq \alpha$ and $n_a \geq \gamma n$, so, $\lambda_Q(\mathbb{D}) \geq \alpha\gamma n$. Now, it is easy to see that the remaining eigenvalues of $\mathbb{D}$ is -1, since, $\mathcal{B} \star \mathbf{J}$ is a rank Q matrix and its remaining eigenvalues are zero and the eigenvalues of diagonal matrix are $\tilde{d}$ with $\tilde{d}_a$ having multiplicity (n_a) for $a = 1, \ldots, Q$. □

Corollary 4.12. *With high probability it holds that* $|\lambda_Q(\mathbf{D}/\log n)| \geq O(n)$ *and* $\lambda_{Q+1}(\mathbf{D}/\log n) \leq O(n^{1-\varepsilon})$.

Proof. By Weyl's Inequality, for all $i = 1, \ldots, n$,

$$\begin{aligned} ||\lambda_i(\mathbf{D}/\log n)| - |\lambda_i(\mathbb{D})|| &\leq \left|\left|\frac{\mathbf{D}}{\log n} - \mathbb{D}\right|\right|_F \leq O(n^{1-\varepsilon/2}) \\ &\leq O(n^{1-\varepsilon}) \end{aligned}$$

So, $|\lambda_Q(\mathbf{D}/\log n)| \geq O(n) - O(n^{1-\varepsilon}) = O(n)$ for large n and $|\lambda_{Q+1}(\mathbf{D}/\log n)| \leq -1 + O(n^{1-\varepsilon}) = O(n^{1-\varepsilon})$. □

Now, if we consider $\mathbf{W}$ is the eigenspace corresponding to top Q absolute eigenvalues of $\mathbb{D}$ and $\tilde{\mathbf{W}}$ is the eigenspace corresponding to top Q absolute eigenvalues of $\mathbf{D}$. Using Davis-Kahan

Lemma 4.13. *With high probability, there exists an orthogonal matrix* $\mathbf{R} \in \mathbb{R}^{Q\times Q}$ *such that* $||\mathbf{W}\mathbf{R} - \tilde{\mathbf{W}}||_F \leq O(n^{-\varepsilon})$

Proof. The top Q eigenvalues of both $\mathbb{D}$ and $\mathbf{D}/\log n$ lies in (Cn, ∞) for some $C > 0$. Also, the gap $\delta = O(n)$ between top Q and $Q+1$th eigenvalues of matrix $\mathbb{D}$. So, now, we can apply Davis-Kahan Theorem 4.10 and Theorem 4.3, to get that,

$$||\mathbf{W}\mathbf{R} - \tilde{\mathbf{W}}||_F \leq \sqrt{2}\frac{\left|\left|\frac{\mathbf{D}}{\log n} - \mathbb{D}\right|\right|_F}{\delta} \leq \frac{O(n^{1-\varepsilon})}{O(n)} = O(n^{-\varepsilon})$$

□

Now, the relationship between the rows of W can be specified based on Assumption (C3) as follows -

Lemma 4.14. *For any two rows i, j of $\mathbf{W}_{n\times Q}$ matrix, $||u_i - u_j||_2 \geq O(1/\sqrt{n})$, if type of $v_i \neq$ type of v_j.*

Proof. The matrix $\mathbb{D}$ can be considered as a Khatri-Rao product of the matrices $\mathcal{D}$ and $\mathbf{J}$ according to equation (11). Now, by Assumption (C3), we have a constant difference between the rows of matrix $\mathcal{D}$. So, rows of $\mathbb{D}$ as well as the projection of $\mathbb{D}$ into into its top Q eigenspace has difference of order $O(n^{-1/2})$ between rows of matrix. □

Now, if we consider Q-means criterion as the clustering criterion on $\tilde{\mathbf{W}}$, then, for the Q-means minimizer centroid matrix $\mathbf{C}$ is an $n \times Q$ matrix with Q distinct rows corresponding to the Q centroids of Q-means algorithm. By property of Q-means objective function and Lemma 4.13, with high probability,

$$\begin{aligned} ||\mathbf{C} - \tilde{\mathbf{W}}||_F &\leq ||\mathbf{WR} - \tilde{\mathbf{W}}||_F \\ ||\mathbf{C} - \mathbf{WR}||_F &\leq ||\mathbf{C} - \tilde{\mathbf{W}}||_F + ||\mathbf{WR} - \tilde{\mathbf{W}}||_F \\ &\leq 2||\mathbf{WR} - \tilde{\mathbf{W}}||_F \\ &\leq O(n^{-\varepsilon}) \end{aligned}$$

By Lemma 4.14, for large n, we can get constant C, such that, Q balls, $B_1, \ldots, B_Q$, of radius $r = Cn^{-1/2}$ around Q distinct rows of $\mathbf{W}$ are disjoint.

Now note that with high probability the number of rows i such that $||\mathbf{C}_i - (\mathbf{WR})_i|| > r$ is at most $O(n^{1/2-\varepsilon})$. If the statement does not hold then,

$$\begin{aligned} ||\mathbf{C} - \mathbf{WR}||_F &> r.O(n^{1/2-\varepsilon}) \\ &\geq Cn^{-1/2}.O(n^{1/2-\varepsilon}) = O(n^{-\varepsilon}) \end{aligned}$$

So, we get a contradiction, since $||\mathbf{C} - \mathbf{WR}||_F \leq O(n^{-\varepsilon})$. Thus, the number of mistakes should be at most of order $O(n^{1/2-\varepsilon})$.

So, for each $v_i \in V(G_n)$, if ξ_i is the type of v_i and $\hat{\xi}_i$ is the type of v_i as estimated from applying Q-means on top Q eigenspace of geodesic matrix $\mathbf{D}$, we get that with high probability, for some small $0 < \varepsilon$,

$$\min_{\pi \in \mathcal{P}_Q} |\{u \in V : \xi(u) \neq \pi(\hat{\xi}(u))\}| = O(n^{1/2-\varepsilon})$$

5 Application

We investigate the empirical performance of the algorithm in several different setup. At first, we use simulated networks from stochastic block model to find

Figure 1: The LHS is the performance of graph distance based method and RHS is the performance of Pseudolikelihood method on same generative SBM.

the empirical performance of the algorithm. Then, we apply our method to find communities in several real world networks.

5.1 Simulation

We simulate networks from stochastic block models with $Q = 3$ blocks. Let w correspond to a Q-block model defined by parameters $\theta = (\boldsymbol{\pi}, \rho_n, S)$, where π_a is the probability of a node being assigned to block a as before, and

$$\mathbf{F}_{ab} = \mathrm{P}(A_{ij} = 1 | i \in a, j \in b) = \rho_n S_{ab}, \quad 1 \leq a, b \leq K.$$

and the probability of node i to be assigned to block a to be π_a $(a = 1, \ldots, K)$.

5.1.1 Equal Density Clusters

We consider a stochastic block model with $Q = 3$. We consider the parameter matrix $\mathbf{F} = 0.012(1+0.1\nu)(\tilde{\lambda}F^{(1)}+(1-\tilde{\lambda})F^{(2)})$, where, $F^{(1)}_{3\times3} = \mathrm{Diag}(0.9, 0.9, 0.9)$ and $F^{(2)}_{3\times3} = 0.1\mathbf{J}_2$, where, $\mathbf{J}_2$ is a 2×2 matrix of all 1's and ν varies from 1 to 15 to give networks of different density. So, we get $\rho_n = \boldsymbol{\pi}^T\mathbf{F}\boldsymbol{\pi}$. We now, vary $\tilde{\lambda}$ to get different combinations of $\mathbf{F}$ as well as ρ_n.

In the following figures, we try to see the behavior of mean and variances of the count statistics, as we vary λ_n as we vary ν.

5.1.2 Unequal Density Clusters

We consider a stochastic block model with $Q = 3$. We consider the parameter matrix $\mathbf{F} = 0.012(1+0.1\nu)(\tilde{\lambda}F^{(1)}+(1-\tilde{\lambda})F^{(2)})$, where, $F^{(1)}_{3\times3} = \mathrm{Diag}(0.1, 0.5, 0.9)$ and $F^{(2)}_{3\times3} = 0.1\mathbf{J}_2$, where, $\mathbf{J}_2$ is a 2×2 matrix of all 1's and ν varies from 1 to 15 to give networks of different density. So, we get $\rho_n = \boldsymbol{\pi}^T\mathbf{F}\boldsymbol{\pi}$. We now, vary $\tilde{\lambda}$ to get different combinations of $\mathbf{F}$ as well as ρ_n.

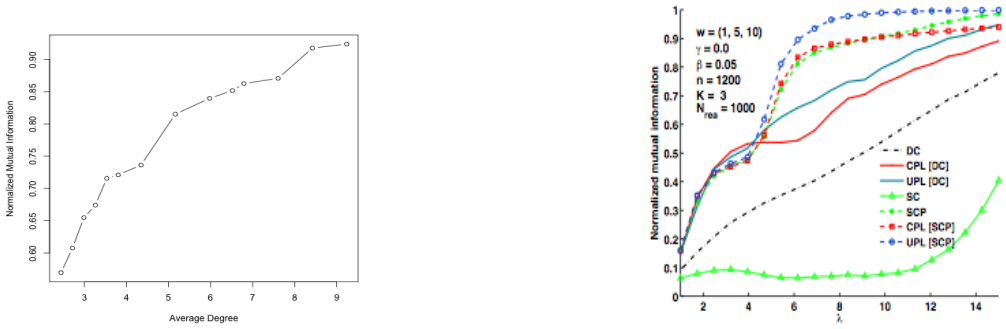

Figure 2: The LHS is the performance of graph distance based method and RHS is the performance of Pseudolikelihood method on same generative SBM.

Figure 3: The LHS is community allocation and RHS is the one estimated by graph distance for Facebook Caltech network with 3 dorms.

In the following figures, we try to see the behavior of mean and variances of the count statistics, as we vary λ_n as we vary ν.

5.2 Application to Real Network Data

5.2.1 Facebook Collegiate Network

In this application, we try to find communities for Facebook collegiate networks. The networks were presented in the paper by Traud et.al. (2011) [47]. The network is formed by Facebook users acting as nodes and if two Facebook users are "friends" there is an edge between the corresponding nodes. Along with the network structure, we also have the data on covariates of the nodes. Each node has covariates: gender, class year, and data fields that represent (using anonymous numerical identifiers) high school, major, and dormitory residence. We consider the network of a specific college (Caltech). We compare the communities found with the dormitory affiliation of the nodes.

Figure 4: The LHS is community allocation and RHS is the one estimated by graph distance for Political Web blogs Network.

5.2.2 Political Web Blogs Network

This dataset on political blogs was compiled by [1] soon after the 2004 U.S. presidential election. The nodes are blogs focused on US politics and the edges are hyperlinks between these blogs. Each blog was manually labeled as liberal or conservative by [1], and we treat these as true community labels. We ignore directions of the hyperlinks and analyze the largest connected component of this network, which has 1222 nodes and the average degree of 27. The distribution of degrees is highly skewed to the right (the median degree is 13, and the maximum is 351). This is a network where the degree distribution is heavy-tailed and the graph is inhomogeneous.

6 Conclusion

The proposed graph distance based community detection algorithm gives a very general way for community detection for graphs over a large range of densities - from very sparse graphs to very dense graphs. We theoretically prove the efficacy of the method under the model that the graph is generated from stochastic block model with fixed number of blocks. We prove that the proportion of mislabeled communities goes to zero as the number of vertices $n \to \infty$. This result is true for graphs coming from stochastic block model under certain conditions on the stochastic block model parameters. These conditions are satisfied above the threshold of block identification for two blocks as given in [38]. The condition (C1) of $\mathbf{1}$ not being the eigenvector of $\tilde{K}$ for our community identification result to hold, seems to be an artificial one, as simulation suggests that our method is able to identify communities, even when $\mathbf{1}$ is an eigenvector of $\tilde{K}$.

We demonstrate the empirical performance of the method by using both simulated and real world networks. We compare with the pseudo-likelihood method and show that they have similar empirical performances. We demonstrate the

empirical performance by applying the method for community detection in several real world networks too.

The method also works when number of blocks in the stochastic block model brows with n (number of vertices) and for degree-corrected block model [25]. We conjecture that under these models too the method will have the theoretical guarantee of correct community detection. The proof can be obtained by using similar techniques that we have used in this paper.

References

[1] Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

[2] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, 9:1981–2014, 2008.

[3] Krishna B Athreya and Peter E Ney. *Branching processes*, volume 28. Springer-Verlag Berlin, 1972.

[4] Brian Ball, Brian Karrer, and MEJ Newman. Efficient and principled method for detecting communities in networks. *Physical Review E*, 84(3):036103, 2011.

[5] Shankar Bhamidi, Remco Van der Hofstad, and Gerard Hooghiemstra. First passage percolation on the erds-renyi random graph. *Combinatorics, Probability & Computing*, 20(5):683–707, 2011.

[6] Peter Bickel, David Choi, Xiangyu Chang, and Hai Zhang. Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *arXiv preprint arXiv:1207.0865*, 2012.

[7] Peter J Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106(50):21068–21073, 2009.

[8] Peter J Bickel, Aiyou Chen, and Elizaveta Levina. The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301, 2011.

[9] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[10] Aydın Buluç, John R Gilbert, and Ceren Budak. Solving path problems on the gpu. *Parallel Computing*, 36(5):241–253, 2010.

[11] Alain Celisse, J-J Daudin, and Laurent Pierre. Consistency of maximum-likelihood and variational estimators in the stochastic block model. *arXiv preprint arXiv:1105.3288*, 2011.

[12] Françoise Chatelin. *Spectral Approximation of Linear Operators*. SIAM, 1983.

[13] Kamalika Chaudhuri, Fan Chung Graham, and Alexander Tsiatas. Spectral clustering of graphs with general degrees in the extended planted partition model. *Journal of Machine Learning Research-Proceedings Track*, 23:35–1, 2012.

[14] Aiyou Chen, Arash A Amini, Peter J Bickel, and Elizaveta Levina. Fitting community models to large sparse networks. *arXiv preprint arXiv:1207.2340*, 2012.

[15] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[16] Amin Coja-Oghlan and André Lanka. Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics*, 23(4):1682–1714, 2009.

[17] Chandler Davis and William Morton Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[18] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[19] Robert W Floyd. Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345, 1962.

[20] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[21] Mayiz B Habbal, Haris N Koutsopoulos, and Steven R Lerman. A decomposition algorithm for the all-pairs shortest path problem on massively parallel computer architectures. *Transportation Science*, 28(4):292–308, 1994.

[22] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[23] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[24] Donald B Johnson. Efficient algorithms for shortest paths in sparse networks. *Journal of the ACM (JACM)*, 24(1):1–13, 1977.

[25] Brian Karrer and Mark EJ Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[26] Tosio Katō. *Perturbation theory for linear operators*, volume 132. springer, 1995.

[27] CG Khatri and C Radhakrishna Rao. Solutions to some functional equations and their applications to characterization of probability distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 167–180, 1968.

[28] Eric D Kolaczyk. *Statistical analysis of network data*. Springer, 2009.

[29] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption: clustering sparse networks. *arXiv preprint arXiv:1306.5550*, 2013.

[30] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

[31] Pierre Latouche, Etienne Birmelé, and Christophe Ambroise. Overlapping stochastic block models with application to the french political blogosphere. *The Annals of Applied Statistics*, 5(1):309–336, 2011.

[32] Charles E Leiserson, Ronald L Rivest, Clifford Stein, and Thomas H Cormen. *Introduction to algorithms*. The MIT press, 2001.

[33] Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):2, 2007.

[34] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th international conference on World Wide Web*, pages 695–704. ACM, 2008.

[35] Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the ab-

sence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

[36] Jure Leskovec, Kevin J Lang, and Michael Mahoney. Empirical comparison of algorithms for network community detection. In *Proceedings of the 19th international conference on World wide web*, pages 631–640. ACM, 2010.

[37] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[38] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

[39] Mark Newman. *Networks: an introduction*. OUP Oxford, 2009.

[40] Mark EJ Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[41] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[42] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[43] MEJ Newman. Spectral community detection in sparse networks. *arXiv preprint arXiv:1308.6494*, 2013.

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

[45] Edgar Solomonik, Aydın Buluç, and James Demmel. Minimizing communication in all-pairs shortest paths. *University of California at Berkeley, Berkeley, US*, 2012.

[46] Daniel L Sussman, Minh Tang, Donniell E Fishkind, and Carey E Priebe. A consistent adjacency spectral embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107(499):1119–1128, 2012.

[47] Amanda L Traud, Eric D Kelsic, Peter J Mucha, and Mason A Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM review*, 53(3):526–543, 2011.

[48] Stephen Warshall. A theorem on boolean matrices. *Journal of the ACM (JACM)*, 9(1):11–12, 1962.

Appendix: Branching Process Results

A1. Proof of Lemma 4.6

We have n_a vertices of type a ,$a = 1, \dots, Q$, and that $n_a/n \overset{a.s.}{\to} \pi_a$. From now on we condition on $n_1, \dots, n_Q$; we may thus assume that $n_1, \dots, n_Q$ are deterministic with $n_a/n \to \pi_a$. Let $\omega(n)$ be any function such that $\omega(n) \to \infty$ and $\omega(n)/n \to 0$. We call a component of $G_n \equiv G(n, P) = G(n, K/n)$ big if it has at least $\omega(n)$ vertices. Let B be the union of the big components, so $|B| = N_{\geq \omega(n)}(G_n)$. Fix $\epsilon > 0$.We may assume that n is so large that $\omega(n)/n < \epsilon$ π_i and $|n_a/n - \pi_a| < \epsilon$ π_a for every a; thus $(1-\epsilon)\pi_a n < n_a < (1+\epsilon)\pi_a n$. We may also assume that $n > \max K$, as K is a function on the finite set $\mathcal{S} \times \mathcal{S}$. Since, n_a/n is a $\sqrt{n}$-consistent estimator of π_a, we get that

$$\epsilon = O(n^{-1/2}). \tag{14}$$

Select a vertex and explore its component in the usual way, that means looking at its neighbors, one vertex at a time. We first reveal all edges from the initial vertex, and put all neighbors that we find in a list of unexplored vertices; we then choose one of these and reveal its entire neighborhood, and so on. Stop when we have found at least $\omega(n)$ vertices (so $x \in B$), or when there are no unexplored vertices left (so we have found the entire component and $x \notin B$).

Consider one step in this exploration, and assume that we are about to reveal the neighborhood of a vertex x of type a. Let us write n'_b for the number of unused vertices of type b remaining. Note that $n_b \geq n'_b \geq n_b - \omega(n)$, so

$$(1-2\epsilon)\pi_b < n'_b/n < (1+\epsilon)\pi_b \tag{15}$$

The number of new neighbors of x of type b has a binomial $Bin(n'_b, K_{ab}/n)$ distribution, and the numbers for different b are independent. The total variation distance between a binomial $Bin(n, p)$ distribution and the Poisson distribution with the same mean is at most p. Hence the total variation distance between the binomial distribution above and the Poisson distribution $Poi(K_{ab}n'_b/n)$ is at most $K_{ab}/n = O(1/n)$. Also, by (15),

$$(1-2\epsilon)K_{ab}\pi_b \leq K_{ab}n'_b/n \leq (1+\epsilon)K_{ab}\pi_b. \tag{16}$$

Since we perform at most $\omega(n)$ steps in the exploration, we may, with an error probability of $O(\omega(n)/n) = o(1)$, couple the exploration with two multi-type branching processes $\mathcal{B}((1-2\epsilon)K)$ and $\mathcal{B}((1+\epsilon)K)$ such that the first process always finds at most as many new vertices of each type as the exploration, and the second process finds at least as many. Consequently, for a vertex x of type a,

$$\rho_{\geq\omega(n)}((1-2\epsilon)K; a) + o(1) \leq \mathrm{P}(x \in B) \leq \rho_{\geq\omega(n)}((1+\epsilon)K; a) + o(1). \tag{17}$$

Since $\omega(n) \to \infty$, by Lemma 9.5 of [9], we have $\rho_{\geq \omega(n)}(K;a) \to \rho(K;a)$ for every matrix or finitary kernel K, which parametrizes the offspring distribution of the branching process in the sense that the number of offsprings of type b coming from a parent of type a follows $Poi(K_{ab}\pi_b)$ distribution. So we can rewrite (17) as

$$\rho((1-2\epsilon)K;a) + o(1) \leq P(x \in B) \leq \rho((1+\epsilon)K;a) + o(1). \tag{18}$$

A2. proof of Lemma 4.7

We need to consider certain branching process expectations $\sigma(K)$ and $\sigma_{\geq k}(K)$ in place of $\rho(K)$ and $\rho_{\geq k}(K)$. In preparation for the proof, we shall relate $\zeta(K)$ to the branching process $\mathcal{B}_K$ via $\sigma(K)$. As before, we assume that K is a kernel on $(\mathcal{S}, \pi)$ with $K \in L^1$.

Let A be a Poisson process on $\mathcal{S}$, with intensity given by a finite measure λ, so that A is a random multi-set on $\mathcal{S}$. If g is a bounded measurable function on multi-sets on $\mathcal{S}$, it is easy to see that

$$\mathbb{E}(|A|g(A)) = \sum_{i \in \mathcal{S}} \mathbb{E}g(A \cup \{i\})\lambda_i \tag{19}$$

For details see Proposition 10.4 of [9].

Let $B(x)$ denote the first generation of the branching process $\mathcal{B}_K(x)$. Thus $B(x)$ is given by a Poisson process on $\mathcal{S}$ with intensity $K(x,y)\pi_x$. Suppose that $\sum_b K_{ab}\pi_b < \infty$ for every $a = 1, \ldots, Q$, so $B(x)$ is finite. Let $\sigma(K;x)$ denote the expectation of $|B(x)|\mathbf{1}[|\mathcal{B}_K(x)| = \infty]$, recalling that under the assumption $\sum_b K_{ab}\pi_b < \infty$ for every a, the branching process $\mathcal{B}_K(x)$ dies out if and only if $|\mathcal{B}_K(x)| < \infty$. Then

$$\begin{aligned}
\sum_{b=1}^{Q} K_{xb}\pi_b - \sigma(K;x) &= \mathbb{E}\left[|B(x)|\mathbf{1}(\mathcal{B}_K(x) < \infty)\right] \\
&= \mathbb{E}\left(|B(x)| \prod_{z \in B(x)} \rho(K;z)\right) \\
&= \sum_{b=1}^{Q} K_{xb}(1-\rho(K;b))\mathbb{E}\left(\prod_{z \in B(x)} \rho(K;z)\right)\pi_b \\
&= \sum_{b=1}^{Q} K_{xb}(1-\rho(K;b))(1-\rho(K;x))\pi_b
\end{aligned}$$

Here the penultimate step is from (19); the last step uses the fact that the branching process dies out if and only if none of the children of the initial particle survives. Writing B for the first generation of $\mathcal{B}_K$ conditioned on survival

becomes

$$\sigma(K) \equiv \mathbb{E}|B|\mathbf{1}[|\mathcal{B}_K| = \infty] = \sum_{x=1}^{Q} \sigma(K;x)\pi_x$$

Then, integrating over x and subtracting from $\sum_{a,b} K_{ab}\pi_a\pi_b$, we get,

$$\sigma(K) = \sum_{a,b} K_{ab}\left(1 - (1-\rho(K;a))(1-\rho(K;b))\right)\pi_a\pi_b \tag{20}$$

So, the kernel for the conditioned branching process becomes

$$K_{ab}\left(\rho(K;a) + \rho(K;b) - \rho(K;a)\rho(K;b)\right) \tag{21}$$

A3. proof of Lemma 4.8

We have $\mathcal{S}$ is finite, say $\mathcal{S} = \{1, 2, \ldots, Q\}$. Let $\Gamma_d(v) \equiv \Gamma_d(v, G_n)$ denote the d-distance set of v in G_n, i.e., the set of vertices of G_n at graph distance exactly d from v, and let $\Gamma_{\leq d}(v) \equiv \Gamma_{\leq d}(v, G_n)$ denote the d-neighborhood $\cup_{d' \leq d}\Gamma_{d'}(v)$ of v.

Let $0 < \varepsilon < 1/10$ be arbitrary. The proof of (18) involved first showing that, for n large enough, the neighborhood exploration process starting at a given vertex v of G_n with type a (chosen without inspecting G_n) could be coupled with the branching process $\mathcal{B}_{(1+\varepsilon)K'}(i)$, where the K' is defined by equation (21), so that the branching process is conditioned to survive. However, henceforth we shall abuse notation and denote K' as K.

The neighborhood exploration process and multi-type branching process can be coupled so that for every d, $|\Gamma_d(v)|$ is at most the number N_d of particles in generation d of $\mathcal{B}_{(1+2\varepsilon)K}(i)$. The number of vertices at generation d of type c of branching process $\mathcal{B}_{(1+2\varepsilon)K}(a)$, denoted by $N^a_{d,c}$ and the number of vertices of type c at distance d from v for the neighborhood exploration process of G_n is denoted by $|\Gamma^a_{d,c}(v)|$f, where, $c = 1, \ldots, Q$.

Elementary properties of the branching process imply that $\mathbb{E}N_d = O\left(||T_{(1+2\varepsilon)K}||^d\right) = O(((1+2\varepsilon)\lambda)^d)$, where $\lambda = ||T_K|| > 1$.

Let $N^a_t(c)$ be the number of particles of type c in the t-th generation of $\mathcal{B}_K(a)$, then, N^a_t is the vector $(N^a_t(1), \ldots, N^a_t(Q))$. Also, let $\nu = (\nu_1, \ldots, \nu_Q)$ be the eigenvector of T_K with eigenvalue λ (unique, up to normalization, as P is irreducible). From standard branching process results, we have

$$N^a_t/\lambda^t \overset{a.s.}{\to} X\nu, \tag{22}$$

where $X \geq 0$ is a real-valued random variable, X is continuous except that it has some mass at 0, and $X = 0$ if and only if the branching process eventually

dies out and lastly,

$$\mathbb{E}X = \nu_a.$$

under the conditions given in Theorem V.6.1 and Theorem V.6.2 of [3].

Set $D = (1-10\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda$. Then $D < (1-\varepsilon)\log(n/\nu_a\nu_b)/\log((1+2\varepsilon)\lambda)$ if ε is small enough, which we shall assume. Thus,

$$\mathbb{E}|\Gamma_{\le D}(v)| \le \mathbb{E}\sum_{d=0}^{D} N_d = O(((1+2\varepsilon)\lambda)^D) = O(n^{1-\varepsilon})$$

So, summing over v, we have

$$\sum_{v\in V(G_n)} |\Gamma_{\le D}(v)| = |\{\{v,w\} : d_G(v,w) \le (1-\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda\}|$$

and its expected value to be

$$\mathbb{E}\,|\{\{v,w\} : d_G(v,w) \le (1-\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda\}| = \mathbb{E}\sum_{v\in V(G_n)} |\Gamma_{\le D}(v)| = O(n^{2-\varepsilon})$$

The above statement is equivalent to

$$\mathbb{E}\left|\left\{\{v,w\} : d_G(v,w) \le (1-\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right\}\right| = \mathbb{E}\sum_{v\in V(G_n)} |\Gamma_{\le D}(v)| = O(n^{2-\varepsilon})$$

So, by Markov's Theorem, we have,

$$\mathrm{P}\left[\left|\left\{\{v,w\} : d_G(v,w) \le (1-\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right\}\right| \le O(n^{2-\varepsilon/2})\right] = o(1)$$

for any fixed $\varepsilon > 0$.

A4. proof of Lemma 4.9

We consider the multi-type branching process with probability kernel $P_{ab} = \frac{K_{ab}}{n}$ $\forall a,b = 1,\dots,Q$ and the corresponding random graph G_n generated from stochastic block model has in total n nodes. We condition that branching process $\mathcal{B}_K$ survives.

Note that an upper bound 1 is obvious, since we are bounding a probability, so it suffices to prove a corresponding lower bound. We may and shall assume that $K_{ab} > 0$ for some a, b.

Fix $0 < \eta < 1/10$. We shall assume that η is small enough that $(1-2\eta)\lambda > 1$. In the argument leading to (18) in proof of Lemma 4.6, we showed that, given $\omega(n)$ with $\omega(n) = o(n)$ and a vertex v of type a, the neighborhood exploration process of v in G_n could be coupled with the branching process $\mathcal{B}_{(1-2\eta)K}(a)$ so that whp the former dominates until it reaches size $\omega(n)$. More precisely, writing $N_{d,c}$ for

the number of particles of type c in generation d of $\mathcal{B}_{(1-2\eta)K}(a)$, and $\Gamma_{d,c}(v)$ for the set of type c vertices at graph distance d from v, whp

$$|\Gamma_{d,c}(v)| \geq N_{d,c},\ c = 1, \ldots, Q, \quad \text{for all } d \text{ s.t. } |\Gamma_{\leq d}(v)| < \omega(n). \tag{23}$$

This relation between the number of vertices at generation d of type c of branching process $\mathcal{B}_{(1-2\eta)K}(a)$, denoted by $N_{d,c}$ and the number of vertices of type c at distance d from v for the neighborhood exploration process of G_n, denoted by $|\Gamma_{d,c}(v)|$ becomes highly important later on in this proof, where, $c = 1, \ldots, Q$. Note that the relation only holds when $|\Gamma_{\leq d}(v)| < \omega(n)$ for some $\omega(n)$ such that $\omega(n)/n \to 0$ as $n \to \infty$.

Let $N_t^a(c)$ be the number of particles of type c in the t-th generation of $\mathcal{B}_K(a)$, then, N_t^a is the vector $(N_t^a(1), \ldots, N_t^a(Q))$. Also, let $\nu = (\nu_1, \ldots, \nu_Q)$ be the eigenvector of T_K with eigenvalue λ (unique, up to normalization, as P is irreducible). From standard branching process results, we have

$$N_t^a/\lambda^t \overset{a.s.}{\to} X\nu, \tag{24}$$

where $X \geq 0$ is a real-valued random variable, X is continuous except that it has some mass at 0, and $X = 0$ if and only if the branching process eventually dies out and lastly,

$$\mathbb{E}X = \nu_a$$

under the conditions given in Theorem V.6.1 and Theorem V.6.2 of [3].

Let D be the integer part of $\log((n/\nu_a\nu_b)^{1/2+2\eta})/\log((1-2\eta)\lambda)$. From (24), conditioned on survival of branching process $\mathcal{B}_K(a)$, whp either $N_D^a = 0$, or $N_{D,c}^a \geq n^{1/2+\eta}$ for each c (note that $N_{D,c}^a$ comes from branching process $\mathcal{B}_{(1-2\eta)K}(a)$ not branching process $\mathcal{B}_K(a)$). Furthermore, as $\lim_{d\to\infty} \mathrm{P}(N_d^a \neq 0) = \rho((1-2\eta)K)$ and $D \to \infty$, we have $\mathrm{P}(N_D^a \neq 0) \to \rho((1-2\eta)K)$. Thus, if n is large enough,

$$\mathrm{P}\left(\forall c : N_{D,c}^a \geq n^{1/2+\eta}\right) \geq \rho((1-2\eta)K) - \eta.$$

Now, we have conditioned that the branching process with kernel K is conditioned to survive. The right-hand side tends to $\rho(K) = 1$ as $\eta \to 0$. Hence, given any fixed $\gamma > 0$, if we choose $\eta > 0$ small enough we have

$$\mathrm{P}\left(\forall c : N_{D,c}^a \geq n^{1/2+\eta}\right) \geq 1 - \gamma$$

for n large enough.

Now, the neighborhood exploration process and branching process can be coupled so that for every d, $|\Gamma_d(v)|$ is at most the number M_d of particles in generation d of $\mathcal{B}_{(1+2\varepsilon)K}(a)$ from Lemma 4.6 and Eq (16). So, we have,

$$\mathbb{E}|\Gamma_{\leq D}(v)| \leq \mathbb{E}\sum_{d=0}^{D} M_d = O(((1+2\varepsilon)\lambda)^D) = o(n^{2/3})$$

if η is small enough, since D be the integer part of $\log(n^{1/2+2\eta})/\log((1-2\eta)\lambda)$. Note that the power $2/3$ here is arbitrary, we could have any power in the range $(1/2, 1)$. Hence,

$$|\Gamma_{\leq D}(v)| \leq n^{2/3} \quad whp,$$

and whp the coupling described in (23) extends at least to the D-neighborhood. So, now, we are in a position to apply Eq (23), as we have $|\Gamma_{\leq D}(v)| \leq n_a^{2/3} < \omega(n)$, with $\omega(n)/n \to 0$.

Now let v and w be two fixed vertices of $G(n, P)$, of types a and b respectively. We explore both their neighborhoods at the same time, stopping either when we reach distance D in both neighborhoods, or we find an edge from one to the other, in which case v and w are within graph distance $2D+1$. We consider two independent branching processes $\mathcal{B}_{(1-2\eta)K}(a)$, $\mathcal{B}'_{(1-2\eta)K}(b)$, with $N^a_{d,c}$ and $N^b_{d,c}$ vertices of type c in generation d respectively. By previous equation, whp we encounter $o(n)$ vertices in the explorations so, by the argument leading to (23), whp either the explorations meet, or $|\Gamma^a_{D,c}(v)| \geq N^a_{D,c}$ and $|\Gamma^b_{D,c}(w)| \geq N^b_{D,c}$, $c = 1, \ldots, Q$, with the explorations not meeting. Using bound on $N^a_{d,c}$ and the independence of the branching processes, it follows that

$$\mathrm{P}\left(d(v,w) \leq 2D+1 \text{ or } \forall c : |\Gamma^a_{D,c}(v)|, |\Gamma^b_{D,c}(w)| \geq n^{1/2+\eta}\right) \geq (\rho(K)-\gamma)^2 - o(1).$$

Note that the two events in the above probability statement are not disjoint. We shall try to find the probability that the second event in the above equation holds but not the first. We have not examined any edges from $\Gamma_D(v)$ to $\Gamma_D(w)$, so these edges are present independently with their original unconditioned probabilities. For any c_1, c_2, the expected number of these edges is at least $|\Gamma^a_{D,c_1}(v)||\Gamma^b_{D,c_2}(w)|K_{c_1c_2}/n$. Choosing c_1, c_2 such that $K_{c_1c_2} > 0$, this expectation is $\Omega((n^{1/2+\eta})^2/n) = \Omega(n^{2\eta})$. It follows that at least one edge is present with probability $1 - \exp(-\Omega(n^{2\eta})) = 1 - o(1)$. If such an edge is present, then $d(v,w) \leq 2D+1$. So, the probability that the second event in the above equation holds but not the first is $o(1)$. Thus, the last equation implies that

$$\mathrm{P}(d(v,w) \leq 2D+1) \geq (1-\gamma)^2 - o(1) \geq 1 - 2\gamma - o(1).$$

Choosing η small enough, we have $2D+1 \leq (1+\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda$. As γ is arbitrary, we have

$$\mathrm{P}(d(v,w) \leq (1+\varepsilon)\log(n/\nu_a\nu_b)/\log\lambda) \geq 1 - \exp(-\Omega(n^{2\eta})).$$

The above statement is equivalent to

$$\mathrm{P}\left(d(v,w) \leq (1+\varepsilon)\frac{\log n}{\log\lambda/\log(\nu_a\nu_b)}\right) \geq 1 - \exp(-\Omega(n^{2\eta})).$$

and the lemma follows.