Graph Community Detection from Coarse Measurements: Recovery Conditions for the Coarsened Weighted Stochastic Block Model

Nafiseh Ghoroghchian
University of Toronto
Vector Institute

Gautam Dasarathy
Arizona State University

Stark C. Draper
University of Toronto

Abstract

We study the problem of community recovery from coarse measurements of a graph. In contrast to the problem of community recovery of a fully observed graph, one often encounters situations when measurements of a graph are made at low-resolution, each measurement integrating across multiple graph nodes. Such low-resolution measurements effectively induce a coarse graph with its own communities. Our objective is to develop conditions on the graph structure, the quantity, and properties of measurements, under which we can recover the community organization in this coarse graph. In this paper, we build on the stochastic block model by mathematically formalizing the coarsening process, and characterizing its impact on the community members and connections. Through this novel setup and modeling, we characterize an error bound for community recovery. The error bound yields simple and closed-form asymptotic conditions to achieve the perfect recovery of the coarse graph communities.

1 Introduction

Community detection (a.k.a. clustering) in a graph is the problem of identifying groups of nodes with similar behaviour (Fortunato and Hric, 2016; Von Luxburg, 2007; Abbe, 2017). Identifying communities is usually the first analysis tool used to draw an initial observation from data (Yang and Leskovec, 2013). A community in a graph refers to a group of nodes that are more similar to each other than to the rest of the graph. The notion of similarity most conventionally means assortativity, i.e. denser intra-community links in an unweighted graph where no weight or label is associated with the graph edges (Fortunato, 2010). However, the group similarity notion has been extended to other forms of connectivity, as well as to weighted networks (Fortunato and Hric, 2016). Cluster formation is proven to be a universal structure in real networks (Yang and Leskovec, 2015). As a result, detecting communities in networks has become a central question to a great body of prediction and inference tasks, with applications in network neuroscience (Sporns and Betzel, 2016; Bassett and Sporns, 2017; Betzel et al., 2019), social networks (Yang and Leskovec, 2013), collaboration networks (Hou et al., 2008), and biological networks (Girvan and Newman, 2002).

While existing methods for community detection have been effective in modeling, studying, and recovering communities from finely detailed, high-resolution graphs (Fortunato and Hric, 2016), there are various scenarios where a large-scale graph is not fully observable and should be coarsened due to restrictions imposed by the measuring instrument (will be exemplified shortly) (Betzel and Bassett, 2017), limitations of the storage memory, high sampling costs, computational tractability (Dabagia et al.; Serrano et al., 2009), restricted accessibility to data, and the creation of multi-scale representations for graphs (Safro et al., 2015; Loukas, 2019). Discovering the latent community structure from the coarse measured graph is a valuable objective of many graph-based tasks (Mucha et al., 2010; Betzel et al., 2019).

Although conventional community detection models can be directly applied to the coarse measured graphs (Betzel and Bassett, 2017), a fundamental understanding of the impact of coarsening on the community structure and recovery is missing. Fig. 1 illustrates how the coarse measurement process can obscure the high-resolution graph structure. The figure shows that as coarsening reduces the size of the graph, introduces

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

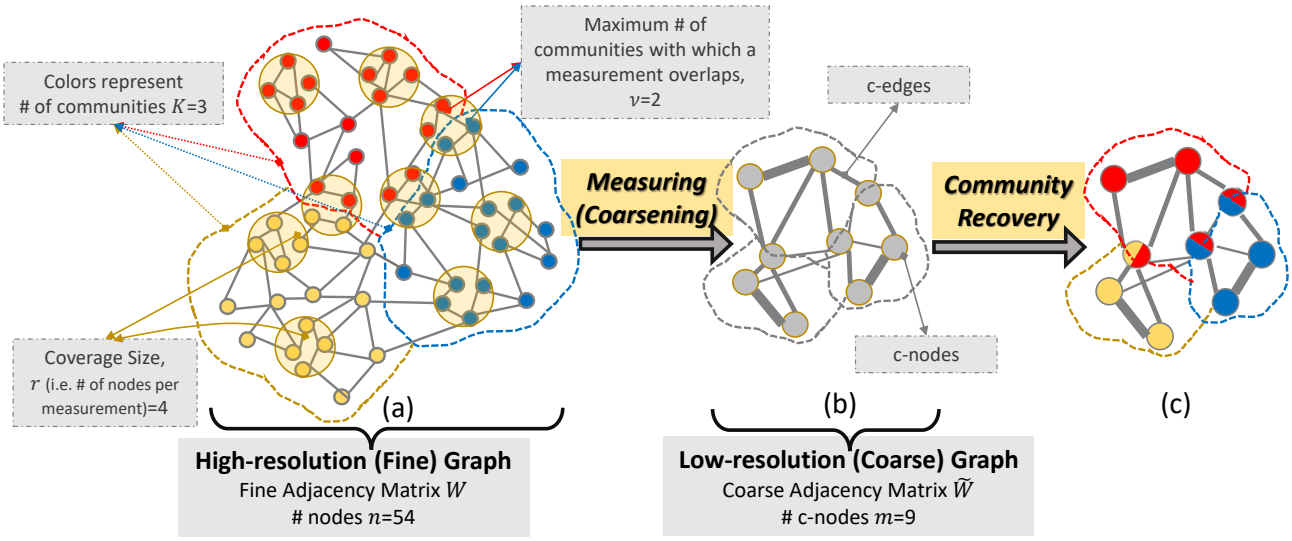


Figure 1: Visual illustration of (a) the underlying high-resolution (fine) graph, (b) the measurement (coarsening) procedure whose result is modeled as a coarse graph, and (c) the effect of the coarsening on the community structure, whose recovery is the objective of this paper. Some notations used in this paper, with their values realized for this figure, are annotated.

heterogeneity in the edge weights, which can potentially cause a drift away from the true community structure.

The study of clustering from coarse measured graphs enables the characterization of contributing factors to their community recovery. Such characterization leads to identifying the barriers in community detection from a coarse graph, which can potentially improve the clustering by applying adjustments to the measurement and community recovery process. Such clustering characterization and recovery improvement are crucial to many fields including neuroscience. Often in the study of the brain on a large scale, the scientific measuring instruments are quite coarse and cannot directly monitor the activity of all the neurons in the brain, which is as high as 14 billion. Hence, one is restricted to collect aggregate signals from bundles of neurons (Osorio et al., 2016; Ghoroghchian et al., 2020), from which a low-resolution functional brain graph is generated (Friston, 2011; Ghoroghchian et al., 2018). The communities identified in the measured graph have been connected to brain cognitive and behavioral units, and they provide biomarkers for neurological diseases (Sporns and Betzel, 2016; Bassett and Sporns, 2017; Lynn and Bassett, 2019; Patankar et al., 2020).

Contributions: In this paper, we study the community detection from coarse measured graphs, which to the best of our knowledge is the first analysis of this problem:

- A random generative model is introduced for the coarse measured networks. A mathematical framework is defined that characterizes the measurement process, the coarse graph, as well as the relationship between the community structure of the fine and coarse graphs.
- Simple and closed-form asymptotic conditions are developed on the graph structure, the quantity and properties of the measurements, under which the community organization of the coarse graph is recovered. The recovery error is characterized, which facilitated studying the effects of various measurement- and structure-related parameters, who take part in improving or exacerbating the quality of the recovery.
- Simulations are provided to compare the derived theoretical error bound with the performance of state-of-the-art community detection methods.

Related Work: While the problem of coarsening a *known graph* has received considerable attention in the past (Karypis and Kumar, 1998; Harel and Koren, 2001; Kushnir et al., 2006; Safro et al., 2015; Loukas, 2019; Rahmani et al., 2020), to the best of our knowledge, this paper is the first to consider learning community structure from coarse summaries of an *unknown graph*.

This paper is built upon the stochastic block model (SBM), a random generative model that is widely used as a canonical model in community detection literature (Abbe, 2017). Although there are other approaches to detect communities, mainly based on modularity

maximization and statistical inference (Fortunato and Hric, 2016; Javed et al., 2018), there are advantages to SBM that fit it to our purposes. SBM provides a rich benchmark that facilitates its generalization to numerous variants (Abbe, 2017; Fortunato and Hric, 2016; Funke and Becker, 2019). Furthermore, the generative nature of SBMs allows for characterizing communities and their recovery (Abbe, 2017), which particularly serves the improvement of community detection. We start by using the vanilla symmetric SBM to model the fine scale graph, which we consider a *latent model* that underlies the observed coarse graph. Under this model, we show that the coarse graph becomes a weighted and mixed membership (or overlapping) variant of the SBM.

The mixed membership SBM (MMSBM) is another relevant paradigm to our purposes and could serve as a good model when directly applied at the measurement (coarsened) level. However in the current paper, we start with a model of the fine graph and characterize the coarse model as a function of the coarsening/measurement procedure. This is more natural given our goal is to infer community information about the underlying fine graph. Relatedly, as far as we know, most papers on MMSBM such as (Dulac et al., 2020) are algorithmic-oriented and do not contain theoretical analysis of the community recovery performance similar to our work in this paper. Few existing works that include theoretical analysis (Mao et al., 2017), do not model weighted edges and do not focus on coarsening, the two components that are crucial to our setup.

2 Model

Consider an unweighted graph $G = G(\mathcal{V}, E)$, where $\mathcal{V}$ is a set of nodes of cardinality $|\mathcal{V}| = n$, and E is a set of pairs of nodes, referred to as edges. Alternative to E and since the graph is unweighted, we can represent the edges using an adjacency matrix $W \in \{0,1\}^{n\times n}$, where each node of the graph is labeled by a unique number in the index set $[n] \triangleq \{1, 2, \cdots, n\}$, and $W_{uv} = 1$ shows the existence of an edge between nodes u and v.

We assume an underlying community structure on $\mathcal{V}$, which partitions the node set into disjointed sets $\mathcal{V} = \cup_{k=1}^{K} \mathcal{V}_k$. For all $k \in [K]$, $\mathcal{V}_k$ represents the set of the nodes that belong to community k. Each node belongs to only one of the K communities. The intra-connection among nodes in the same community is different from their connection to the rest of the graph. Let $P \in \{0,1\}^{K\times n}$ be the true community assignment matrix, where $P_{ku} = 1$ iff node u belongs to community k, i.e.

$$P_{ku} = \begin{cases} 1 & \text{if } u \in \mathcal{V}_k \\ 0 & \text{else} \end{cases} . \tag{1}$$

A graph is drawn under the Symmetric Stochastic Block Model (SSBM) characterised by p and q, where the probability of having an edge between two nodes is independently distributed according to Bernoulli(p), for two nodes in the same community, and Bernoulli(q) for nodes in different communities. Also, the nodes are assigned to communities in a uniform and independent manner. We let W be distributed according to $W \sim \text{SSBM}(n, K, p, q)$ conditional on P, i.e.,

$$W_{uv} \sim \begin{cases} \text{Bernoulli}(p) & \text{if } \exists k \in [K] : P_{ku} = 1, P_{kv} = 1 \\ \text{Bernoulli}(q) & \text{else} \end{cases} , \tag{2}$$

We assume a general scaling behaviour for p, q by defining the constants $0 < \alpha, \beta < \infty$ and a scaling factor $f(n)$, where:

$$p \triangleq \alpha f(n), \quad q \triangleq \beta f(n). \tag{3}$$

$f(n)$ tracks the changes in the graph density as a function of the graph size. As n increases, $f(n)$ may remain unchanged, or it may get smaller, i.e. the graph becomes sparser as it grows. The latter sparsity assumption has been considered in existing literature, as it fits to many real-world applications, including biological, social, and collaborative networks (Abbe, 2017; Mossel et al., 2014; Abbe et al., 2015; Abbe and Sandon, 2015a).

In real applications, G can be very large, in the order of millions or billions of nodes. In general, the population is much larger than the number of communities (e.g., there are many more citizens than cities) and so $n \gg K$. We often cannot observe the existence (or lack of existence) of all $\frac{n(n-1)}{2}$ possible connections and instead measure J *summaries* of associations.

One possible choice to collect a simplified and interpretable set of summary measurements (more explanations come shortly), is to define a set of *disjointed* measurement vectors $\{\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m\}$, all satisfying $\mathbf{b}_i \in \{0,1\}^n$ and $\mathbf{b}_i \mathbf{b}_j^\intercal = 0$ for all $i \in [m]$ different from $j \in [m]$. The latter condition means measurement vectors do not overlap, i.e. each node is measured at most one time. Each summary, denoted by s_ℓ for $\ell \in [m^2]$, is defined as:

$$\begin{aligned} s_\ell &= \sum_{u \in \text{supp}(\mathbf{b}_{\lceil \ell/m \rceil})} \sum_{v \in \text{supp}(\mathbf{b}_{\ell \bmod m})} W_{uv} \\ &= \mathbf{b}_{\lceil \ell/m \rceil} W \mathbf{b}_{\ell \bmod m}^\intercal . \end{aligned} \tag{4}$$

$\text{supp}(\mathbf{b}_i)$ denotes the support of $\mathbf{b}_i$ and $|\text{supp}(\mathbf{b}_i)|$ is the cardinality of the support. Equation (4) corresponds to the set of summary measurements one would

get if one defines an $m \times n$ matrix whose rows are $\mathbf{b}_1, \mathbf{b}_2, \cdots, \mathbf{b}_m$, and then collects m^2 non-distinct (or $\frac{m(m+1)}{2}$ distinct) measurements as in (4), forming the following matrix equality:

$$\tilde{W} = BWB^\intercal. \tag{5}$$

Such measurement model is a natural choice in existing applications. For instance, linear measurement of a high-dimensional signal appear in compressed sensing (Donoho, 2006; Draper and Malekpour, 2009) which is further applied to Electroencephalogram (EEG) signal processing (Aviyente, 2007) and image processing (Baraniuk, 2007), as well as in Covariance sketching (Dasarathy et al., 2015). For such linear measurements, the original and the measured graphs respectively model the Covariance (here, thresholded for weighted graphs) matrices of the original and the linearly measured signals. The measurement model in (5) is also a popular graph reduction method, where $\tilde{W}$ approximates W by preserving some of its spectral properties (Safro et al., 2015; Loukas, 2019; Jin et al., 2020).

Matrix $\tilde{W}$ can be thought of as the weighted adjacency matrix representation of a measured weighted graph $\tilde{G} \in \tilde{\mathcal{G}}(\tilde{\mathcal{V}}, \tilde{E})$, where $\tilde{\mathcal{V}}$ is the set of *c-nodes* [1] and $|\tilde{\mathcal{V}}| = m$. $\tilde{E}$ is the set of *c-edges*, consisting of pairs of c-nodes and a weight, i.e., $(i, j, \tilde{w})$. Note that $\tilde{W}_{ij}$'s for all $i > j$ are independent random variables if the $\mathbf{b}_i$'s are disjoint. We return to this point, and the formal statistics of $\tilde{W}$, shortly.

Definition 1. *A measurement matrix B is "r-homogeneous" if for all $i \in [m]$ there is a constant positive integer $r \leq \frac{n}{m}$ such that $|supp(\mathbf{b}_i)| = r$.*

We assume the number of measured fine nodes that represent a c-node is the same for all c-nodes. We refer to this number as the *coverage size* and denote it by r. Accordingly, the support of the rows of a homogeneous measurement matrix has cardinality equal to r. We define the *c-node profile* matrix:

$$\Phi \triangleq BP^\intercal, \tag{6}$$

whose dimension is $m \times K$ and connects the measurement matrix B to the graph of community assignment matrix P. Φ displays the impact of coarsening on the community memberships. A c-node can belong to one community or multiple communities. Each row of the c-node profile matrix, ϕ_i, is a length-K vector that counts the number of nodes in each community in G that is measured by the i-th c-node. For instance, $\Phi = \begin{bmatrix} 0 & 4 & 0 \\ 2 & 0 & 2 \\ 4 & 0 & 0 \\ 0 & 2 & 2 \end{bmatrix}$ means that, all the 4 fine nodes that map to the first (resp. the third) c-node belong to community 2 (resp. 1), while half of the 4 fine nodes mapping to the second c-node belong to community 1 and the other half belong to community 3.

[1] "c-" stands for compound or coarse.

The following Lemma derives the statistics of $\tilde{W}$.

Lemma 1. *Let $W \sim SSBM(n, K, p, q)$ from which $\tilde{W}$ in (5) is measured under the r-homogeneous measurement assumption defined in Def. 1. Then $\tilde{W}_{ij}$'s are i.i.d. random variables for all $i > j$, with distribution:*

$$\tilde{W}_{ij} \sim \quad PoissonBinomial(\{p\}^{\phi_i^\intercal \phi_j}, \{q\}^{r^2 - \phi_i^\intercal \phi_j}) \ , \tag{7}$$

where the PoissonBinomial in (22), is a compact notation for a Poisson Binomial distribution with success probabilities $\phi_i^\intercal \phi_j$ of p's and $r^2 - \phi_i^\intercal \phi_j$ of q's.

The proof is elaborated in Sec. 4.1 of the supplementary materials.

Each c-node can measure from members of one or multiple communities. We denote the maximum number of communities that overlap with a c-node, by ν, where $1 \leq \nu \leq K$. This is considered as a Community Overlap (CO) constraint, and is illustrated in the next Definition.

Definition 2. *A measurement matrix B is CO-ν with respect to a graph $G \in \mathcal{G}(\mathcal{V}, E)$ with community assignment matrix P, if the profile matrix $\Phi = BP^\intercal$ satisfies: $1 \leq |supp(\phi_i)| \leq \nu \quad \forall i \in [m]$.*

Def. 2 means that the support of each row of B corresponds to at most ν of the communities in G. The next definition is the last step to formalizing the coarse graph community structure.

Definition 3. *A measurement matrix B is "balanced" with respect to a graph $G \in \mathcal{G}(\mathcal{V}, E)$ with community assignment matrix P, if the profile matrix $\Phi = BP^\intercal$ satisfies $\Phi_{ik} = \Phi_{ik'}$ for all $i \in [m]$ and $k, k' \in supp(\phi_i)$.*

In other words, in a balanced-measured graph, an identical number of nodes are measured from each community.

The objective of this paper is to recover the c-node profile matrix Φ from the measured graph $\tilde{W}$ in (5). Let a maximum a posteriori (MAP) estimator take a measured graph $\tilde{G}$ with the true c-node profile matrix Φ, and returns its estimate $\hat{\Phi}$ that assigns every c-node in $\tilde{G}$ to communities. We characterize an upper bound on the failure probability of the MAP estimator. The error refers to assigning a wrong profile to at least one c-node, up to equivalent relabelling of communities. We also study the asymptotic conditions such that this error tends to zero.

Recovering Φ in (6) from the measured matrix $\tilde{W}$, without imposing additional constraints on Φ, is gen-

erally a very hard problem. Hence, we relax the problem to achieve tractability, by putting the constraints in Def. 1, 2, and 3 on the measurement matrix (i.e. B), with respect to the community assignment matrix (i.e. P) of the graph. In many practical settings, assumptions such as homogeneity are reasonable. For instance, Electroencephalography (ECoG) signals are acquired from different brain regions using electrodes whose contact surface areas are the same. Nevertheless, a relaxation of these assumptions is of considerable interest and will serve as a compelling avenue for future exploration.

In the next section, we state and study the community recovery problem under the CO-ν constraint.

3 Recovery under the Community Overlap (CO)-ν constraint

In this section, we derive an upper bound on the MAP recovery error of the profile matrix Φ, as described at the end of Sec. 2. The recovered profile matrix $\hat{\Phi}$, estimates at most ν communities from which each c-node measures.

3.1 Main Results

We begin sketching our main result by defining

$$K^{(\nu)} = \sum_{\ell=1}^{\nu} \binom{K}{\ell}, \tag{8}$$

the profile set:

$$\begin{aligned}\Upsilon^{(\nu)} \triangleq \{\phi_i | \phi_i = \mathbf{b}_i P^{\intercal}, 1 \leq |\text{supp}(\phi_i)| \leq \nu, \\ \forall k \in \text{supp}(\phi_i) : \phi_{ik} = \tfrac{r}{|\text{supp}(\phi_i)|}\},\end{aligned} \tag{9}$$

and a one-to-one function $h : \Upsilon^{(\nu)} \to [K^{(\nu)}]$. Function h maps a c-node profile to an *extended* community indexed by $[K^{(\nu)}]$ (more explanations in Sec. 3.2). A probability matrix $U \in [0,1]^{K^{(\nu)} \times K^{(\nu)}}$ is defined, for all $\mathbf{a}, \mathbf{a}' \in \Upsilon^{(\nu)}$ and $k = h(\mathbf{a}), k' = h(\mathbf{a}')$, as:

$$U_{k,k'} \quad = \mathbb{P}(X \geq r^2(\tilde{\tau}p + (1-\tilde{\tau})q)) \ , \tag{10}$$

where $0 \leq \tilde{\tau} \leq 1$ and X is an auxiliary random variable distributed as:

$$X \sim \quad \text{PoissonBinomial}(\{p\}^{\mathbf{a}^{\intercal}\mathbf{a}'}, \{q\}^{r^2 - \mathbf{a}^{\intercal}\mathbf{a}'}) \ . \tag{11}$$

Sec. 3.2 will elaborate on the reasons behind these definitions, using the *binarization* of the coarse measured graph, i.e. mapping the c-edge weights to zero or one. Sec. 3.2 shows that the elements of matrix U in (10) essentially denote the probability of having a connection between members of the extended communities (or equivalently between c-node profiles), in the binarized coarse graph. The prior distribution on the extended communities is denoted by the probability vector $\mathbf{s}$. We define the scaled Chernoff-Hellinger (CH) divergence as:

$$\begin{aligned}D(\text{diag}(\mathbf{s})U_k, \text{diag}(\mathbf{s})U_{k'}) \triangleq \max_{0 \leq t \leq 1} \\ \sum_{k'' \in [K^{(\nu)}]} \mathbf{s}_{k''}[tU_{kk''} + (1-t)U_{k'k''} - U_{kk''}^t U_{k'k''}^{(1-t)}],\end{aligned} \tag{12}$$

where the original CH divergence is $D_+ = \frac{m}{\log m} D$. The following theorem provides an error bound for community recovery from the coarse graph.

Theorem 1. *Let $W \sim$ SSBM(n, K, p, q) from which $\tilde{W}$ in (5) is measured under the r-homogeneous, balanced, and CO-ν constraints. $\mathbf{s}$ is a length-$K^{(\nu)}$ probability vector, U_k denotes the k-th column of matrix U defined in (10), and $K^{(\nu)}$ is defined in (8). The probability that the MAP estimator fails to recover the c-node profile matrix Φ from $\tilde{W}$ (up to relabelling of Φ's columns) is upper-bounded by:*

$$\mathbb{P}(\textit{MAP failure}) \leq \sum_{\substack{k,k' \in [K^{(\nu)}] \\ k<k'}} e^{-mD(diag(\mathbf{s})U_k, diag(\mathbf{s})U_{k'})}, \tag{13}$$

where D is the scaled CH divergence in (12).

The modeling of the coarse graph under the CO-ν constraint, i.e. binarization and profile mapping to extended communities sketched before the theorem, makes the binarized coarse graph fit to the general SBM framework in (Abbe and Sandon, 2015b). In general SBM, the connection probability between members of the extended communities is no longer symmetric. Rather, this probability differs for each pair of extended communities. This way, the error bound is straightforwardly derived using equations (44) and (47) in (Abbe and Sandon, 2015b), while adjusting the notations. The rest of the detailed proof techniques for Theorem 1 is elaborated in Sec. 3.2.

Theorem 1 demonstrates that, as the connectivity probability among pairs of the extended communities become distant, the recovery error bound improves.

Remark 1. *In order to extract interpretable observations from the recovery error bound in Theorem 1, we examine the dominant term of the CH divergence in (12). For each pair of extended communities, k, k', the dominant term corresponds to an extended community k'', where the probability of its connectivity to those communities is the most distant. We derived an estimate for the dominant term in Sec. 4.2 of the supplementary materials, which demonstrates the following: the exponent of the error recovery bound (i.e.*

the CH divergence) increases as r and $|\alpha-\beta|$ increase (by fixing whichever α or β that is smaller and increasing the other one), or as ν decreases, while other parameters remain unchanged.

In the following we list the observations derived from Theorem 1 and Remark 1:

1. As we increase the measurement size (i.e. m, the number of c-nodes), the error bound decreases.

2. As the coverage size per measurement (i.e. r, the number of measured fine nodes represented by a c-node) expands, the failure error bound decreases.

3. By allowing measurements overlapping with fewer communities (i.e. increasing the purity of the c-nodes), the error bound drops. This intuitively makes sense due to a decrease in complexity.

4. The expansion of the gap between extra- and intra-community probabilities results in a decrease in the error bound. This is intuitively expected since communities become more distinguishable from one another.

Note that the trends listed above are true so long as the prior $\mathbf{s}$ remains unchanged, or does not change such behaviors. We also assume other parameters except for the one mentioned, remain unchanged. Otherwise, we face perturbing multiple parameters simultaneously, which might make the behavior of the error bound unpredictable and heavily depending on the parameter values.

The following corollary characterizes the asymptotic conditions such that the community recovery error, upper-bounded in Theorem 1, approaches zero.

Corollary 1. *Let $W \sim SSBM(n,K,p,q)$ from which $\tilde{W}$ in (5) is measured under the r-homogeneous, balanced, and CO-ν constraints. The probability that the MAP estimator fails to recover the c-node profile matrix Φ from $\tilde{W}$ (up to relabelling of Φ's columns), for a constant $\Delta>0$, $0<\tilde{\tau}<\frac{1}{\nu}$, tends to zero as:*

$$\begin{array}{c} r > \frac{\Delta}{\sqrt{f(n)}}, \quad \alpha \neq \beta, \\ \Delta^2(\frac{m}{n})^2 < f(n) \leq f_0, \quad n \geq n_0, \quad m,n \to \infty \end{array} . \quad (14)$$

The constant Δ is defined in equation (45) in Sec. 4.3 of the supplementary materials. The remaining parameters are assumed to remain fixed.

The condition in (14) is directly derived from the error bound in Theorem 1, by tending the exponent to infinity resulting in the error to approach zero. The complete proof is sketched in Sec. 3.2.

Corollary 1 characterizes the impact of coarsening on the community recovery. After the coarse graph is binarized, the connectivity probability between some c-edges reaches very fast to zero, and the rest to one, which facilitates the separation of communities. Moreover, the measurement coverage size r (i.e. the number of measured fine nodes combined into a c-node), and the graph binarization threshold $\tilde{\tau}$, must satisfy a lower and upper bound, respectively, to allow perfect community recovery of a coarsened graph through its binarization. The recovery conditions derived in Corollary 1, are illustrated in the last column of Table 1, and are compared with those of the classic (non-coarsened) general SBM that exist in the literature. The comparison is made in terms of various scalings of the parameters. The first column exhaustively partitions the scaling of the connection probability of the coarse graph, which can be a function of m,n,r and denoted by $\tilde{f}(m,n,r)$, for which the second column shows state-of-the-art conditions to allow or disallow exact recovery. In the third column, different scalings of the coarsening coverage size r are considered, where each scaling results in separate recovery conditions demonstrated in the last column.

3.2 Proof Techniques

The community recovery problem under the CO-ν constraint refers to the problem of estimating the c-node profile matrix Φ that corresponds to the weighted adjacency matrix $\tilde{W}$ defined in (5) and measured from $W \sim \text{SSBM}(n,K,p,q)$ under the r-homogeneous, balanced, and CO-ν constraints. This way, $\tilde{W}$ is distributed according to (7) and hence, can be thought of and modeled as a sample of a *weighted* version of the Overlapping general SBM (OSBM) random graph ensemble. The formal definition of general SBM is found in (Abbe and Sandon, 2015a). We define the weighted OSBM that models $\tilde{W}$, similar to the classic OSBM, except that the node profiles ϕ_i for all i belong to the set $\Upsilon^{(\nu)}$ defined as (9). rather than the set of any length-K binary vectors $\{0,1\}^K$. Furthermore, the weighted OSBM that models $\tilde{W}$, an edge between pairs of nodes is distributed as the Poisson Binomial distribution in (7), rather than the Bernoulli distribution in classic OSBM. Note that due to the Community Overlap (i.e. CO-ν) assumption on $\tilde{W}$, the edge distributions depend on the inner product of the pairwise profiles, which takes values between 0 and r^2, i.e. $0 \leq \phi_i^\intercal \phi_j \leq r^2$. Hence, the weighted OSBM is not symmetric.

Deriving the conditions that allow the community recovery from a weighted OSBM, except for the symmetric case (c.f. Sec. 3.3 and Sec. 1 in the supplementary materials), is an open problem (Xu et al., 2020). In the following, we exploit the properties of the special case of the weighted OSBM concerning this study, which enables its transformation to a classic

Table 1: Comparison of the recovery conditions under the CO-ν constraint, derived from Corollary 1. All scaling notations (o for strictly smaller than, and Ω for strictly greater than, disregarding constants) are defined with respect to m. $f(n)$ is the probability scaling of connections in the fine graph, Q is a constant matrix (i.e. it does not scale with other variables), Δ is a positive constant, and $\tilde{f}(m,n,r)$ represents the probability scaling of connections in the coarse graph, which is a function of the fine and coarse graphs sizes, and the coverage size (i.e. the number of measured fine nodes represented by a c-node), to allow for comparison with the classic scenario (i.e. with $n=m, r=1$).

$\tilde{f}(m,n,r)$: probability scaling of connections in coarse graph	**Classic (exact) Recovery** SBM$(m,\mathbf{s},U=Q\tilde{f}(m,m,1))$ as $m,n\to\infty$ (Abbe and Sandon, 2015a)	Scaling of coarsening coverage size, i.e. r	**Recovery, This paper** as $m,n\to\infty$
$o(\frac{\log m}{m})$	Impossible	$o(\frac{1}{\sqrt{f(n)}})$	Impossible
		$c_1\frac{1}{\sqrt{f(n)}}$	Possible if $\alpha\neq\beta, c_1>\Delta, f(n)>\Delta^2(\frac{m}{n})^2$
		$\Omega(\frac{1}{\sqrt{f(n)}})$	Possible if $\alpha\neq\beta, f(n)=\Omega((\frac{m}{n})^2)$
$c_0\frac{\log m}{m}$	Possible if $D_+>1$	$o(\sqrt{\frac{m}{\log m}})$	Impossible
		$c_1\sqrt{\frac{m}{c_0\log m}}$	Possible if $\alpha\neq\beta, c_1>\Delta, f(n)>\Delta^2(\frac{m}{n})^2$
		$\Omega(\sqrt{\frac{m}{\log m}})$	Possible if $\alpha\neq\beta, f(n)=\Omega((\frac{m}{n})^2)$
$\Omega(\frac{\log m}{m})$	Possible if $D_+>0$	$o(\frac{1}{\sqrt{f(n)}})$	Impossible
		$c_1\frac{1}{\sqrt{f(n)}}$	Possible if $\alpha\neq\beta, c_1>\Delta, f(n)>\Delta^2(\frac{m}{n})^2$
		$\Omega(\frac{1}{\sqrt{f(n)}})$	Possible if $\alpha\neq\beta, f(n)=\Omega((\frac{m}{n})^2)$

(unweighted) general SBM. We propose a two-stage strategy that first binarizes $\tilde{W}$ and then represents the resultant unweighted OSBM as an unweighted classic (non-overlapping) general SBM. The binarization is motivated for two reasons. First, binarization is widely used to simplify and sparsify weighted graphs. Second, through binarization, we can leverage existing work in community detection literature to study the conditions to recover the c-node profile matrix.

3.2.1 Stage one: Binarizing $\tilde{W}$

The summation in the coarsening model (5) suggests the concentration of edge weights around a mean value. Hence, for the c-edges that corresponds to a pair of c-nodes measuring from only one community, the expectation of the weights tend to concentrate about means pr^2 or qr^2. Regarding the c-nodes measuring from multiple communities, their corresponding c-edge weights concentrate about means $p\phi_i^\intercal\phi_j+q(r^2-\phi_i^\intercal\phi_j)$. This motivates solving our weighted OSBM problem by first binarizing $\tilde{W}$. Such binarization facilitates community recovery by adopting the much more evolved tools available for unweighted graphs. We define the binarized coarse measured matrix $\tilde{W}^{(\mathrm{b})}$ as:

$$\tilde{W}^{(\mathrm{b})}_{ij}\triangleq\begin{cases}1 & \text{if } \tilde{W}_{ij}\geq r^2(\tilde{\tau}p+(1-\tilde{\tau})q)\\ 0 & \text{else}\end{cases}, \quad (15)$$

for $0\leq\tilde{\tau}\leq 1$. The chosen threshold, i.e. $r^2(\tilde{\tau}p+(1-\tilde{\tau})q)$ in (15), is a suitable choice since it is lower- and upper- bounded by qr^2, pr^2, the minimum and maximum mean values of $\tilde{W}_{ij}$ for various profile inner products. This way, we only keep the most significant edges, i.e. those whose weights are above the mean value of the intra-community connections.

3.2.2 Stage two: SBM representation of the OSBM

Through the binarization explained in Sec. 3.2.1, the coarse graph $\tilde{W}$ previously modeled as a weighted general OSBM, is converted to $\tilde{W}^{(\mathrm{b})}$, which is a classic (unweighted) general OSBM. Following the approach suggested in (Abbe, 2017), we convert the classic OSBM to an equivalent *non-overlapping* general SBM. To do so, instead of the original community set $[K]$, we use the extended community set $[K^{(\nu)}]$, where $K^{(\nu)}\triangleq|\Upsilon^{(\nu)}|$ defined in (8), where each extended community represents a possible c-node profile $\phi_i\in\Upsilon^{(\nu)}$ for all $i\in[m]$. The one-to-one function $h:\Upsilon^{(\nu)}\to[K^{(\nu)}]$ provides indexing for the extended communities, i.e. a profile vector $\phi_i\in\Upsilon^{(\nu)}$ maps to an extended community $k=h(\phi_i)$. Such conversion of profiles to extended communities, models the binarized matrix of measurements $\tilde{W}^{(\mathrm{b})}$ in (15) as a general unweighted SBM denoted by $\tilde{W}^{(\mathrm{b})}\sim\mathrm{SBM}(m,\mathbf{s},U)$, where $\mathbf{s}$ is a prior probability vector of the extended

communities. Sec. 2 in the supplementary materials provides the formal definition of the general unweighted SBM, the derivation of the matrix of community connectivity probabilities U, and the remaining of the proof techniques of Theorem 1 and Corollary 1.

3.3 Stronger recovery under the special Community Overlap (CO)-1 constraint

The results in Sec. 3.1 are applicable to coarse measured graphs under the general CO-ν constraint, for all $1 \leq \nu \leq K$. However, the CO-1 constraint is an special case, which corresponds to a weighted and *symmetric* SBM. Contrary to the general (i.e. non-symmetric) weighted SBM model which is an open problem, the community recovery from such weighted and symmetric SBM has already been addressed in the literature (Jog and Loh, 2015; Xu et al., 2020). In the following theorem, we adopt the results of (Jog and Loh, 2015) to achieve stronger recovery conditions under the CO-1 constraint, compared with those of the general CO-ν scenario in Corollary 1.

Theorem 2. *Let $W \sim SSBM(n, K, p, q)$ from which $\tilde{W}$ in (5) is measured under the r-homogeneous, and CO-1 constraints. The probability that the MAP estimator fails to recover the c-node profile matrix Φ from $\tilde{W}$ (up to relabelling of Φ's columns) from $\tilde{W}$, tends to zero as:*

$$\begin{cases} \alpha \neq \beta & \text{if} \quad m < \infty \quad r\sqrt{f(n)} \to \infty \\ \frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > \lim\limits_{m\to\infty} \left[\frac{K}{2}\frac{\log m}{r^2 f(n) m}\right] & \text{if} \quad m, n \to \infty, \end{cases} \tag{16}$$

if $f(n) \xrightarrow{n\to\infty} 0$. K is assumed to remain fixed.

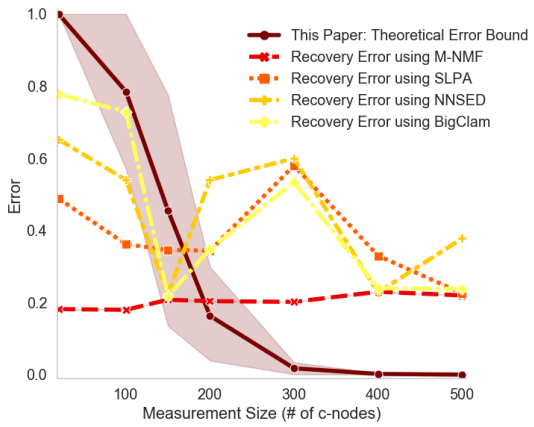
More explanations and proof details are found in Sec. 1 and Sec. 4.4 of the supplementary materials.

4 Numerical Results

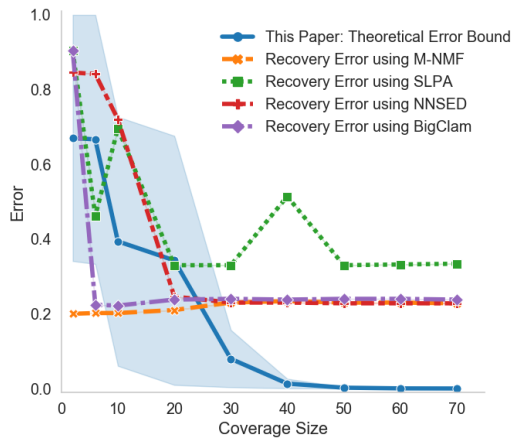
In this section, we evaluate the error behavior of the community recovery from synthetically generated coarse measured graphs. We compare the theoretical error bounds derived in Sec. 3, with state-of-the-art community detection methods from existing works that are applied to the generated coarse graphs. [2] It should be noted these algorithmic methods only output the index of the nodes estimated to be assigned. This translates into the recovery of a binarized version of the community assignment matrix $\mathbf{\Phi}$. Refer to Sec. 3 in supplementary materials for the detailed methodology used in this section.

[2] The Python code to reproduce the results of this paper is available at: `https://github.com/NaGho/Community-Detection-From-Coarse-Measured-Graphs`.

In Fig. 2, the theoretical error bound (solid line), as well as the community recovery error for multiple state-of-the-art overlapping community detection methods (Rossetti et al., 2019) are plotted[3]. The methods include Modularized non-negative matrix factorization (M-NMF) (Wang et al., 2017), Speaker-listener Label Propagation Algorithm (SLPA) (Xie et al., 2011, 2013), Non-Negative Symmetric Encoder-Decoder (NNSED) (Sun et al., 2017), and Cluster Affiliation Model for Big Networks (BigClam) (Yang and Leskovec, 2013) (dashed lines). Note that we have evaluated these methods for various hyper-parameters and plotted their best performance.



(a) w.r.t. m for $r = 50$.



(b) w.r.t. r for $m = 400$.

Figure 2: Community recovery error for $n = 30000$ fine nodes, $\nu = 2$ community overlap (CO), $K = 5$ communities, and $\alpha = 500, \beta = 50$ intra- and extra- community constants for the probability of connectivity.

From Fig. 2a, we observe that as we increase the mea-

[3] The results in this section are computed assuming p, q, K, r are known. However, using model selection methods, heuristics can be developed to estimate these parameters when they are not known apriori.

surement size (i.e. m, the number of c-nodes), the theoretical error bound drops monotonically. Similarly, Fig. 2b plots the community recovery error with respect to the coverage size r (i.e. the number of measured fine nodes combined into a c-node), demonstrating that increasing the coverage size monotonically improves the theoretical community recovery error. These observations confirm the expectations made subsequent to Theorem 1. Although the simulated methods, both in Fig. 2a and Fig. 2b, do not perform as predictable as the theoretical error bound, most of them show an overall decrease in their recovery error when respectively, the number of measurements and the coverage size increase. Note that the light shade in Fig. 2 around the theoretical bound represents the ambiguity in the calculation of the bound (c.f. Sec. 3 of the supplementary materials).

Note that the theoretical bound is the *upper bound* for the MAP estimator. Fig. 2 shows that the upper bound seem to be loose in certain regimes (e.g. for small m, r), in which existing methods perform better. However, as the measurement- and the coverage sizes increase, the theoretical error bound becomes tight and outperforms existing community detection methods with an increasing gap.

5 Conclusion and Future Work

We introduced a mathematical framework based on the stochastic block model, that characterizes community recovery from coarse measured graphs. We developed theoretical conditions, on the quantity and properties of the measurements with respect to the community structure of the high-resolution graph, to achieve perfect recovery. The assumptions of homogeneous and balanced measurements were essential to this work. We leave to future work the relaxation of these assumptions. Moreover, community recovery in a coarse measured graph, in which communities modeled using the weighted and overlapping stochastic block model, utilized edge weight binarization. Future work can look into community recovery without binarization, in which one would use full graph weight distribution for recovery.

Finally, a significant gap was observed between the performance of state-of-the-art community detection algorithms, with the theoretical error bounds derived in this paper, in certain regimes. This gap motivates future work to improve existing clustering algorithms to achieve its theoretical potential. An algorithmic investigation into recovery performance, e.g. similar to the variational inference approaches used in (Aicher et al., 2015; Dulac et al., 2020), is a promising direction to future work and would complement our theoretical analyses.

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada through a Discovery Research Grant; the National Science Foundation Grants CCF-2029044 and CCF-2048223; the National Institutes of Health Grant 1R01GM140468-01; Connaught International Scholarship for Doctoral Students; and the Vector Postgraduate Affiliate Award by Vector Institute for AI.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 670–688. IEEE, 2015a.

Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. 2015b.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1): 471–487, 2015.

Christopher Aicher, Abigail Z Jacobs, and Aaron Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2015.

Selin Aviyente. Compressed sensing framework for eeg compression. In *2007 IEEE/SP 14th workshop on Statistical Signal Processing*, pages 181–184. IEEE, 2007.

Richard G Baraniuk. Compressive sensing [lecture notes]. *IEEE Signal Processing Magazine*, 24(4): 118–121, 2007.

Danielle S Bassett and Olaf Sporns. Network neuroscience. *Nature Neuroscience*, 20(3):353–364, 2017.

Richard F Betzel and Danielle S Bassett. Multi-scale brain networks. *Neuroimage*, 160:73–83, 2017.

Richard F Betzel, Maxwell A Bertolero, Evan M Gordon, Caterina Gratton, Nico UF Dosenbach, and Danielle S Bassett. The community structure of functional brain networks exhibits scale-specific patterns of inter-and intra-subject variability. *Neuroimage*, 202:115990, 2019.

Max Dabagia, Konrad P Kording, and Eva L Dyer. Comparing high-dimensional neural recordings by aligning their low-dimensional latent representations.

Gautam Dasarathy, Parikshit Shah, Badri Narayan Bhaskar, and Robert D Nowak. Sketching sparse matrices, covariances, and graphs via tensor products. *IEEE Transactions on Information Theory*, 61 (3):1373–1388, 2015.

David L Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

Stark C Draper and Sheida Malekpour. Compressed sensing over finite fields. In *2009 IEEE International Symposium on Information Theory*, pages 669–673. IEEE, 2009.

Adrien Dulac, Eric Gaussier, and Christine Largeron. Mixed-membership stochastic block models for weighted networks. In *Conference on Uncertainty in Artificial Intelligence*, pages 679–688. PMLR, 2020.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

Santo Fortunato and Darko Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

Karl J Friston. Functional and effective connectivity: a review. *Brain Connectivity*, 1(1):13–36, 2011.

Thorben Funke and Till Becker. Stochastic block models: A comparison of variants and inference methods. *PloS One*, 14(4):e0215296, 2019.

Nafiseh Ghoroghchian, Stark C Draper, and Roman Genov. A hierarchical graph signal processing approach to inference from spatiotemporal signals. In *2018 29th Biennial Symposium on Communications (BSC)*, pages 1–5. IEEE, 2018.

Nafiseh Ghoroghchian, David M Groppe, Roman Genov, Taufik A Valiante, and Stark C Draper. Node-centric graph learning from data for brain state identification. *IEEE Transactions on Signal and Information Processing over Networks*, 6:120–132, 2020.

Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12): 7821–7826, 2002.

David Harel and Yehuda Koren. On clustering using random walks. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 18–41. Springer, 2001.

Haiyan Hou, Hildrun Kretschmer, and Zeyuan Liu. The structure of scientific collaboration networks in scientometrics. *Scientometrics*, 75(2):189–202, 2008.

Muhammad Aqib Javed, Muhammad Shahzad Younis, Siddique Latif, Junaid Qadir, and Adeel Baig. Community detection in networks: A multidisciplinary review. *Journal of Network and Computer Applications*, 108:87–111, 2018.

Yu Jin, Andreas Loukas, and Joseph JaJa. Graph coarsening with preserved spectral properties. In *International Conference on Artificial Intelligence and Statistics*, pages 4452–4462. PMLR, 2020.

Varun Jog and Po-Ling Loh. Information-theoretic bounds for exact recovery in weighted stochastic block models using the renyi divergence. *arXiv preprint arXiv:1509.06418*, 2015.

George Karypis and Vipin Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on scientific Computing*, 20 (1):359–392, 1998.

Dan Kushnir, Meirav Galun, and Achi Brandt. Fast multiscale clustering and manifold identification. *Pattern Recognition*, 39(10):1876–1891, 2006.

Andreas Loukas. Graph reduction with spectral and cut guarantees. *Journal of Machine Learning Research*, 20(116):1–42, 2019.

Christopher W Lynn and Danielle S Bassett. The physics of brain network structure, function and control. *Nature Reviews Physics*, 1(5):318, 2019.

Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. On mixed memberships and symmetric nonnegative matrix factorizations. In *International Conference on Machine Learning*, pages 2324–2333. PMLR, 2017.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. *arXiv preprint arXiv:1407.1591*, 3(5), 2014.

Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.

Ivan Osorio, Hitten P Zaveri, Mark G Frei, and Susan Arthurs. *Epilepsy: the intersection of neurosciences, biology, mathematics, engineering, and physics*. CRC Press, 2016.

Shubhankar P Patankar, Jason Z Kim, Fabio Pasqualetti, and Danielle S Bassett. Path-dependent connectivity, not modularity, consistently predicts controllability of structural brain networks. *Network Neuroscience*, pages 1–31, 2020.

Mostafa Rahmani, Andre Beckus, Adel Karimian, and George K Atia. Scalable and robust community detection with randomized sketching. *IEEE Transactions on Signal Processing*, 68:962–977, 2020.

Giulio Rossetti, Luca Pappalardo, and Salvatore Rinzivillo. A novel approach to evaluate community detection algorithms on ground truth. In *Complex Networks VII*, pages 133–144. Springer, 2016.

Giulio Rossetti, Letizia Milli, and Rémy Cazabet. Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1):52, 2019.

Ilya Safro, Peter Sanders, and Christian Schulz. Advanced coarsening schemes for graph partitioning. *Journal of Experimental Algorithmics (JEA)*, 19:1–24, 2015.

M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the National Academy of Sciences*, 106(16):6483–6488, 2009.

Olaf Sporns and Richard F Betzel. Modular brain networks. *Annual Review of Psychology*, 67:613–640, 2016.

Bing-Jie Sun, Huawei Shen, Jinhua Gao, Wentao Ouyang, and Xueqi Cheng. A non-negative symmetric encoder-decoder approach for community detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 597–606, 2017.

Wenpin Tang and Fengmin Tang. The poisson binomial distribution–old & new. *arXiv preprint arXiv:1908.10024*, 2019.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

Xiao Wang, Peng Cui, Jing Wang, Jian Pei, Wenwu Zhu, and Shiqiang Yang. Community preserving network embedding. In *AAAI*, volume 17, pages 203–209, 2017.

Jierui Xie, Boleslaw K Szymanski, and Xiaoming Liu. Slpa: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 344–349. IEEE, 2011.

Jierui Xie, Stephen Kelley, and Boleslaw K Szymanski. Overlapping community detection in networks: The state-of-the-art and comparative study. *ACM Computing Surveys (CSUR)*, 45(4):1–35, 2013.

Min Xu, Varun Jog, Po-Ling Loh, et al. Optimal rates for community estimation in the weighted stochastic block model. *The Annals of Statistics*, 48(1):183–204, 2020.

Jaewon Yang and Jure Leskovec. Overlapping community detection at scale: a nonnegative matrix factorization approach. In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, pages 587–596. ACM, 2013.

Jaewon Yang and Jure Leskovec. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems*, 42(1):181–213, 2015.

Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.