Feature Extraction and Link Prediction via Regular Partition

Shizhou Xu, Yuan Ni, Quanyan Zhu

October 2020

Abstract

A fundamental problem in the analysis of relational data, such as graphs and networks, is to extract a common structure that underlies the relationships between individual entities. In this paper, we emphasize on the probabilistic connectivity pattern between communities and shed the light on the potential of applying Szemeredi's regularity lemma to community detection, feature extraction, and link prediction on large graphs. In particular, we first give theoretical guarantee of community detection via ϵ-regular partitions for sample graph generated by stochastic block models under mild assumptions. Thereby, we show that regularity based clustering effectively captures the probabilistic connectivity pattern. Furthermore, under the assumption of latent feature induces connectivity, we show that the principal components of the covariance matrix generated by element-wise mean in the regular partition are close to for latent features when vertex data is polluted by i.i.d. noise. Therefore, we propose: (1) a latent feature extraction algorithm which finds the principal components of the elements in a regular partition, and (2) a link prediction method that uses the maximum-likelihood method to find an optimal sample graph embedding on the data subspace spanned by the extracted latent features.

1 Introduction

A fundamental problem in the analysis of relational data is to recognize the a common structure that contributes to the formation of connections between individual entities. In particular, pattern recognition and community detection for graphs have been studied in various areas such as social behavior (Fortunato, 2010; Hoffman, 2018), finance (Bartesaghi, 2019), and image segmentation (Shi et al., 2002). Here, graph is a mathematical description of the relationships. For instance, a Facebook networked data regards every account as a vertex attached with a high-dimensional feature vector, and each edge as the friendship connection between two accounts. For relational data structure recognition, the essential importance is to understand the mechanism between the feature vectors attached to vertices and the observed relationships in between. In short, the main challenge is to develop a method that extracts the relationship between the vertex data and the edge data. One major step is to cluster vertices that share similar connectivity characteristics, and thereby make analysis of the relationship available. Various works have been done in this direction: community detection (Abbe, 2015), and recovery of stochastic block model (Abbe et al., 2016; Ling et al., 2020). The main method used in the above works is spectral clustering. For example, Abbe and Bandeira used spectral method to study the recovery of the two-community stochastic block model $G(n, p, q)$ with $p > q$, which is also known as the planted bisection model, by reducing the community detection problem to a min-cut problem. Results for multi-community detection with inner-connection p and inter-connection q is also given in (Agarwal et al., 2017). Despite the successes above, the spectral method is not efficient in community detection for general SBMs with complex inner and inter connection densities. Because more complex probability matrix in SBMs hinders the problem reduction from community detection to min-cut or bottleneck problems, for which the method is designed.

In contrast, Szemeredi's regularity lemma (Szemeredi, 1978), which states that every large enough dense graph could be well-approximated by a random graph, is designed to check the probabilistic pattern on graphs and give us a different solution to connectivity pattern based clustering problem. In the original paper, Szemeredi uses regularity between clusters to check to whether enough randomness pattern exists between clusters, generates finer partition on the irregular clusters, and terminates the process until the partitioned graph is close to the original graph, where the closeness is quantified by the index of partition. Although the lemma has become a fundamental tool in graph theory, theoretical computer science, and combinatorics, limited works have applied the lemma to data analysis, not to mention community detection and feature extraction. One reason is the worst case upper bound of partition cardinality grows as an exponential tower with height equal to some power of ϵ^{-1} (Gowers, 1997). Therefore, exact ϵ-regular partition is not practical in reality. More recently, (Sperotto, 2007; Peillilo, 2017, Fionacci, 2020) showed approximate ϵ-partition is efficient in graph pattern recognition and image segmentation purpose.

In this paper, we apply the Szemeredi ϵ-regular partition as a black box to solve community detection, relative features extraction, and edge prediction. In particular, we show ϵ-regular partition is efficient in clustering vertices for community detection purpose. By using graphs generated by symmetric stochastic block models (SSBM) as a benchmark for theoretical analysis and experiments, we first show a theoretical guarantee of community detection and recovery of SSBM via regular partitions on sample graphs under mild assumptions, then use the elements of the regular partition to extract relative features to the random graph pattern, and finally apply the extracted feature to embed the graph onto the feature data space for link prediction purpose. In short, this paper first provides theoretical guarantee for community detection via ϵ-regular partition, then use the partition to extract features, and finally use the extracted features to generate a graph embedding on the feature vector space for link prediction.

The rest of the paper is organized as follows: section 2 shows that a regular partition gives a good summery SSBM for any sample graph by proving the closeness between any graph generated by the SSBM and the original graph on the graphon space. Section 3 gives a theoretical performance guarantee of community detection for ϵ-regular partitions by applying graphs generated by SSBM as a benchmark. Moreover, the recovery of SSBM is shown to be closely connected to estimation of graphons. Section 4 proposes two algorithms for feature extraction via regular partitions and graph embedding on the extracted feature data space respectively. Finally, section 5 gives numerical experiments on the community detection, feature extraction, and link prediction.

2 An Encoder for Dense Graphs

In this section, we give an estimation result on the pattern recognition accuracy for ϵ-regular partitions. For any unknown graph G, let G^s be a sample graph generated under uniform sampling. We prove that, with high probability, the random graph model W^s generated by an ϵ-regular partition on G^s gives an accurate pattern summary for not only G^s but also G. We need the estimation result because, if the connectivity pattern summarized by W^s differs significantly or with high probability from the one on G, then there is no ground for further study in community detection, feature extraction, and link prediction based on W^s.

In the rest of the section, we first state two versions of regularity lemma, then follow the procedure:

$$G \xrightarrow{\text{sampling}} G^s \xrightarrow{\text{regular partition}} W^s \xrightarrow{\text{blow-up/lifting}} \tilde{G}.$$

Here, sampling and regular partition is analogous to encoding, W^s is analogous to feature, and blow-up/lifting to decoding.

The aim is to show $\tilde{G}$, which is the graph recovered from the ϵ-partition generated random graph model W^s, is close to G, the original graph, under mild assumptions. The closeness is quantified by the cut distance on the graphon space $(\mathcal{W}, \delta_\Box)$, where $\mathcal{W} := \{w : [0,1]^2 \to [0,1] |$ be borel-measurable$\}$ and the cut distance $\delta_\Box$ is defined as the following: for any $w_1, w_2 \in \mathcal{W}$,

$$\delta_\Box(w_1, w_2) := \inf_{\phi \in \mathcal{S}([0,1])} ||w_1 - w_2^\phi||_\Box,$$

where $\mathcal{S}([0,1])$ denotes all the measure preserving maps from $[0,1]$ to $[0,1]$, $w_2^\phi(x,y) := w_2(\phi(x), \phi(y))$, and

$$||w_1 - w_2||_\Box := \sup_{S,T \subset [0,1]} |\int_{S \times T} w_1(x,y) - w_2(x,y) dxdy|.$$

Intuitively, $\delta_\Box(G_1, G_2) := \delta_\Box(w_{G_1}, w_{G_2})$, where w_{G_i} denotes the normalized adjacency matrix to $[0,1] \times [0,1]$.

Cut distance is a natural quantification of the difference between dense unlabeled graphs, we refer interested readers to (Lovasz, 2016) for detailed explanation and justification.

To prove the closeness between G and $\tilde{G}$ on the graphon space, we need to first state the two versions of regularity lemma.

Lemma 1 (Szemeredi's Regularity Lemma)**.** *For arbitrarily fixed $\epsilon > 0$ and integer $m > 0$, there exists $P(\epsilon, m)$ and $Q(\epsilon, m)$ such that: every graph $G = (V, E)$ with $n = |V| > P$ has a partition $\mathcal{P} = \{P_i\}_{i=1}^k$ on V which satisfies*

- $m \le k \le Q$
- $|P_i| \in \{\lfloor \frac{n}{k} \rfloor, \lfloor \frac{n}{k} \rfloor + 1\}$
- *All but ϵk^2 pairs of (P_i, P_j)'s are ϵ-regular*

Intuitively, the lemma states that every large enough dense graph can be well-approximate by a random graph model. Frieze and Kannan offered a more computationally friendly version below.

Lemma 2 (Frieze and Kannan's Regularity Lemma)**.** *For any graph $G = (V, E)$ with $|V| = n$ sufficiently large, and $\epsilon > 0$, we can construct in time $\tilde{O}(\epsilon^{-2})n$ a partition $\mathcal{P}$ of V which satisfies*

- $|\mathcal{P}| < k, \log k = O(\epsilon^{-2})$,
- *$\mathcal{P}$ is ϵ-sufficient.*

First, we consider each node V_i^s of the sample graph G^s as random variables from the nodes of the population graph, V. Then we can apply the multinomial sampling size analysis to estimate the proportion and the size of the induced partition $\mathcal{P}^s := \{P_i^s\}_{i=1}^k$ on V^s, where $P_i^s := \{V_k^s \in V^s : V_k^s \in P_i\}$

Lemma 3. *For an arbitrary graph G and an equitable partition $\mathcal{P}$ on V, there exists $n^s(\gamma, \alpha, \epsilon, n, k)$ such that: if $k < \alpha^{-1}$, then for all $V^s \subset V$ satisfying $|V^s| > n^s$, we have with at least probability $1 - \gamma$,*

- $|\frac{|P_i|}{n} - \frac{|P_i^s|}{N}| \leq \alpha, \forall i \in \{1, 2, ..., k\},$
- $|P_i^s| > \epsilon |P_i|, \forall i \in \{1, 2, ..., k\}.$

Proof. Since $\mathcal{P} = \{P_i\}_{i=1}^k$ is an equitable partition, by the uniformly random sampling assumption, we have for each $i \in \{1, 2, ..., N\}$, V_i^s is a multinomial random variable. Therefore, we could approximate the sample size by normal distribution: there exists $n_1(\gamma, \alpha)$ such that, for any $N > n_1$, we have

$$\mathbb{P}\{\bigcap_{i=1}^{k} |\frac{|P_i|}{n} - \frac{|P_i^s|}{N}| < \alpha\} > 1 - \gamma.$$

Now, let $n_2(\epsilon, \alpha, n, k) := \frac{\epsilon n}{1 - \alpha k}$. It follows that, if we have $N > \max\{n_1, n_2\}$, then

$$|P_i^s| > N(\frac{|P_i|}{n} - \alpha) = N(\frac{1}{k} - \alpha) > \frac{\epsilon \frac{n}{k}}{\frac{1}{k} - \alpha}(\frac{1}{k} - \alpha) = \epsilon \frac{n}{k} = \epsilon |P_i|,$$

where the two inequalities follow from $N > n_1$ and $N > n_2, k < \alpha^{-1}$ respectively. Finally, let $n^s(\gamma, \alpha, \epsilon, n, k) := \max\{n_1(\gamma, \alpha), n_2(\epsilon, \alpha, n, k)\}$, we are done. □

Since an ϵ-regular or ϵ-sufficient $\mathcal{P}$ requires $|\mathcal{P}|$ to be large when ϵ is small, the assumption $k < \alpha^{-1}$ may seem to be too strict an assumption. But it is important to notice that $n_1(\gamma, \alpha)$ is independent of $|\mathcal{P}|$, and it grows slowly as α goes to zero. That is, we can allow α to be extremely small while keeping n^s relatively small. Therefore, in the case of large dense sample networks, the assumption $k < \alpha^{-1}$ can be easily satisfied.

Next, we want to show that if a partition is ϵ-regular on G, then the induced partition on G^s also satisfies certain level of regularity.

Lemma 4. *If $\mathcal{P}$ is an ϵ-regular partition w.r.t. G satisfying $\frac{1}{\epsilon} < k < \frac{1}{\alpha}$, and G^s satisfies $N > n^s(\gamma, \alpha, \epsilon, n, k)$, where n^s is defined as in lemma 1, then $\mathcal{P}^s$ is $[4(\frac{n}{N})^2 + 2(1 + \alpha k)^2]\epsilon$-sufficient w.r.t. G^s with probability at least $1 - \gamma$.*

Proof. Let $A^s, B^s \subset V^s$ be arbitrary, $\Delta(A^s, B^s) := E(A^s, B^s) - \sum_i \sum_j w^s_{ij}|A^s_i||B^s_j|$. We first consider $A^s, B^s \subset V$, then by the relationship between ϵ-regular and ϵ-sufficient in (Frieze, Kannan), we have

$$|E(A^s, B^s) - \sum_i \sum_j w_{ij}|A^s_i||B^s_j|| < 4\epsilon n^2.$$

It remains to prove that there exists an uniform bound for $\{|w_{ij} - w^s_{ij}|\}_{1 \leq i < j \leq k}$. But by the assumption that $N > n^s$, we have from lemma 1, with probability at least $1 - \gamma$,

$$|P^s_i| > \epsilon|P_i|, \forall i \in \{1, 2, ..., k\}.$$

That, together with the assumption that $\mathcal{P}$ is an ϵ-regular partition, implies for all but ϵk^s pairs of (i, j),

$$|\frac{E(P^s_i, P^s_j)}{|P^s_i||P^s_j|} - w_{ij}| < \epsilon.$$

Since $w^s_{ij} = \frac{E(P^s_i, P^s_j)}{|P^s_i||P^s_j|}$, we have for all but ϵk^2 pairs of (i, j), $|w^s_{ij} - w_{ij}| < \epsilon$. Finally, we have

$$\begin{aligned} |\Delta_{\mathcal{P}^s}(A^s, B^s)| :=& |E(A^s, B^s) - \sum_i \sum_j w^s_{ij}|A^s_i||B^s_j|| \\ \leq& |E(A^s, B^s) - \sum_i \sum_j w_{ij}|A^s_i||B^s_j|| + \sum_i \sum_j |w^s_{ij} - w_{ij}||A^s_i||B^s_j| \\ <& 4\epsilon n^2 + 2\epsilon k^2(\frac{N}{k} + \alpha N)^2 = [4(\frac{n}{N})^2 + 2(1 + \alpha k)^2]\epsilon N^2. \end{aligned}$$

Therefore, $\mathcal{P}^s$ is $[4(\frac{n}{N})^2 + 2(1 + \alpha k)^2]\epsilon$-sufficient.

□

Now, for a partition $\mathcal{Q}^s$ on G^s, we define $\Delta_{\mathcal{Q}^s\mathcal{P}^s}(A^s, B^s) := \sum_{i=1}^q \sum_{j=1}^q Q^s_{ij}|A^s_i||B^s_j| - \sum_{i=1}^k \sum_{j=1}^k W^s_{ij}|A^s_i||B^s_j|$, where $Q^s_{ij} := \frac{E(A^s_i, B^s_j)}{|A^s_i||B^s_j|}$ and $E(A^s_i, B^s_j) := \{e(U, V) : e(U, V) = 1, U \in Q^s_i, V \in Q^s_j\}$. Therefore, let $\epsilon' := [4(\frac{n}{N})^2 + 2(1 + \alpha\frac{k}{N})^2]\epsilon$we obtain the following corollary directly from lemma 2:

Corollary 1. *For any ϵ-regular partition $\mathcal{Q}^s$ on G^s, any $A^s, B^s \subset V^s$,*

$$|\Delta_{\mathcal{Q}^s\mathcal{P}^s}(A^s, B^s)| \leq [4(\frac{n}{N})^2 + 2(1 + \alpha k)^2 + 4]\epsilon N^2.$$

Proof.

$$\begin{aligned}|\Delta_{\mathcal{Q}^s\mathcal{P}^s}(A^s, B^s)| &\leq |\Delta_{\mathcal{Q}^s}(A^s, B^s)| + |\Delta_{\mathcal{P}^s}(A^s, B^s)| \\ &\leq 4\epsilon N^2 + \epsilon' N^2.\end{aligned}$$

□

The above corollary shows the closeness between the induced partition $\mathcal{P}^s$ on G^s and any of the ϵ-sufficient partition $\mathcal{Q}^s$ on G^s. In other words, any of the ϵ-sufficient partition $\mathcal{Q}^s$ provides a "good" estimation for the induced partition on the sample graph. Now, since we have already shown above that the induced partition is proportionally close to the original partition, under the assumption of uniform random sampling, it is straightforward for one to conjecture that any of the ϵ-regular partition $\mathcal{Q}^s$ also provides accurate estimation for G.

Before the proof of theorem 1, we define a random graph $\tilde{G} := (\tilde{V}, \tilde{E})$ and a partition $\mathcal{Q}$ on $\tilde{V}$ as the following:

- $|\tilde{V}| = n$,
- $\mathbb{P}\{\tilde{V}_p \in Q_i\} = \frac{|Q_i^s|}{N}$, $\forall i \in \{1, 2, ..., n\}, i \in \{1, 2, ..., q := |\mathcal{Q}|\}$,
- $\tilde{V}_p \perp \tilde{V}_q, \forall i \neq j$,
- $\mathbb{P}\{e(\tilde{V}_p, \tilde{V}_q) = 1 | \tilde{V}_p \in Q_i, \tilde{V}_q \in Q_j\} = w_{ij}^s, \forall i \neq j \in \{1, 2, ..., q\}$,
- $(e(\tilde{V}_p, \tilde{V}_q) \perp e(\tilde{V}_a, \tilde{V}_b)) | \{\mathcal{Q}(\tilde{V}_p), \mathcal{Q}(\tilde{V}_q), \mathcal{Q}(\tilde{V}_a), \mathcal{Q}(\tilde{V}_b)\}$ where $\mathcal{Q}(\tilde{V}_p) = Q_i$ such that $\tilde{V}_p \in Q_i$.

One may notice that $\tilde{G} \sim SBM(n, \frac{\{Q^s\}_i}{N}, w_{ij}^s)$ is generated by a stochastic block model. We will explore more details in the next section. Now, we show the first main result.

Theorem 1. *Let $\tilde{G}$ be the random graph that is obtained by the procedure, where $|V^s| \geq n^s(\gamma, \alpha, \epsilon, n, k)$, then for any $A, B \subset V$, there exists $\tilde{A}, \tilde{B} \subset \tilde{V}$ such that with high probability, $d_\Box(G, \tilde{G}) < \epsilon$.*

Proof. First, denote the number of edges in $\tilde{G}$ between Q_i and Q_j by $\tilde{E}(Q_i, Q_j)$, by the construction above, we have:

$$\begin{aligned}Q_{ij} &:= \frac{\tilde{E}(Q_i, Q_j)}{|Q_i||Q_j|} \\ &= \frac{\sum_{p=1}^n \sum_{q=1}^n \mathbb{1}_{Q_i}(\tilde{V}_p)\mathbb{1}_{Q_j}(\tilde{V}_q)e(\tilde{V}_p, \tilde{V}_q)}{\sum_{p=1}^n \mathbb{1}_{Q_i}(\tilde{V}_p)\sum_{q=1}^n \mathbb{1}_{Q_j}(\tilde{V}_q)}.\end{aligned}$$

Now, $N_{ij} := \sum_{p=1}^{n} \mathbb{1}_{Q_i}(\tilde{V_p}) \sum_{q=1}^{n} \mathbb{1}_{Q_j}(\tilde{V_q})$, by independence between $\tilde{V_p}$ and $\tilde{V_q}$, $\forall p \neq q$, we have that $|Q_i|$ is the sum of n i.i.d. binomial random variables for all $i \in \{1, 2, ..., q\}$. Furthermore, for any fixed $|Q_i| = q_i, \forall i \in \{1, 2, ..., q\}$, we have $\sum_{p=1}^{n} \sum_{q=1}^{n} \mathbb{1}_{Q_i}(\tilde{V_p}) \mathbb{1}_{Q_j}(\tilde{V_q}) e(\tilde{V_p}, \tilde{V_q}) = \sum_{\tilde{V_p} \in Q_i} \sum_{\tilde{V_q} \in Q_j} e(\tilde{V_p}, \tilde{V_q})$ is also the sum of n^2 i.i.d. binomial edge random variables as in the construction above, where the independence comes from the conditional independence in construction. Therefore, for fixed $|Q_i| = q_i, \forall i$, we have from Hoeffding's inequality that $\mathbb{P}\{|Q_{ij} - Q^s_{ij}| \geq \epsilon ||Q_i| = q_i, |Q_j| = q_j\} < 2\exp{(-2q_iq_j\epsilon^2)}$.

$$\Longrightarrow \mathbb{P}\{\bigcup_{i,j}\{|Q_{ij} - Q^s_{ij}| \geq \epsilon ||Q_i| = q_i, |Q_j| = q_j\}\} < \sum_{i,j} 2\exp{(-2q_iq_j\epsilon^2)}$$
$$\Longrightarrow \mathbb{P}\{\bigcap_{i,j}\{|Q_{ij} - Q^s_{ij}| < \epsilon ||Q_i| = q_i, |Q_j| = q_j\}\} \geq 1 - \sum_{i,j} 2\exp{(-2q_iq_j\epsilon^2)}$$

Also, since $n > N > n^s(\gamma, \alpha, \epsilon, n, k) > n_1(\gamma, \alpha)$, we have

$$\mathbb{P}\{\bigcap_{i=1}^{q} |\frac{|Q_i|}{n} - \frac{|Q^s_i|}{N}| < \alpha\} >= 1 - \gamma.$$

Now, let $A, B \subset V$ be arbitrary, and assume without loss of generality that $\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij}|A_i||B_j| > \sum_{i=1}^{q} \sum_{j=1}^{q} Q_{ij}|\tilde{A}_i||\tilde{B}_j|$, it follows

$$\begin{aligned}&|\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}|A_i||B_j| - \sum_{i=1}^{q}\sum_{j=1}^{q} Q_{ij}|\tilde{A}_i||\tilde{B}_j||\\ \leq&|\Delta_{\mathcal{P}\mathcal{P}^s}(A,B)| + |\sum_{i=1}^{k}\sum_{j=1}^{k} w^s_{ij}|A_i||B_j| - \sum_{i=1}^{q}\sum_{j=1}^{q} Q^s_{ij}|\tilde{A}_i||\tilde{B}_j|| + |\Delta_{\mathcal{Q}^s\mathcal{Q}}(\tilde{A},\tilde{B})|,\end{aligned}$$

where

- $|\Delta_{\mathcal{P}\mathcal{P}^s}(A,B)| := \sum_{i=1}^{k}\sum_{j=1}^{k} |w_{ij} - w^s_{ij}|A_i||B_j|$
- $|\Delta_{\mathcal{Q}^s\mathcal{Q}}(A,B)| := \sum_{i=1}^{q}\sum_{j=1}^{q} |Q^s_{ij} - Q_{ij}|\tilde{A}_i||\tilde{B}_j|$

Now, if the condition $|Q_i| = q_i, |Q_j| = q_j$ is satisfied, then with probability at least $1 - \sum_{i,j} 2\exp{(-2q_iq_j\epsilon^2)}$, $\sum_{i=1}^{q}\sum_{j=1}^{q} |Q^s_{ij} - Q_{ij}||\tilde{A}_i||\tilde{B}_j| < \epsilon n^2$. But we also have with probability at least $1 - \gamma$, $|\frac{|Q^s_i|}{N} - \frac{|Q_i|}{n}| < \alpha, \forall i \in \{1, 2, ..., q\}$. It follows with probability at least $(1 - \gamma)(1 - \sum_{i,j} 2\exp{(-2(\frac{n}{q} - \alpha n)^2\epsilon^2)})$,

$$\sum_{i=1}^{q}\sum_{j=1}^{q} |Q^s_{ij} - Q_{ij}||\tilde{A}_i||\tilde{B}_j| \leq \epsilon(\frac{n}{q} + \alpha n)^2$$

Similarly, we have:

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{k}|w_{ij}-w_{ij}^{s}||A_i||B_j| \leq&(1-\epsilon)k^2\epsilon(\frac{n}{k})^2+\epsilon k^2(\frac{n}{k})^2\\ <&2\epsilon n^2.\end{aligned}$$

It remains to bound $\sum_{i=1}^{k}\sum_{j=1}^{k}w_{ij}^{s}|A_i||B_j|-\sum_{i=1}^{q}\sum_{j=1}^{q}Q_{ij}^{s}|\tilde{A}_i||\tilde{B}_j|$:

$$\begin{aligned}&\sum_{i=1}^{k}\sum_{j=1}^{k}w_{ij}^{s}|A_i||B_j|-\sum_{i=1}^{q}\sum_{j=1}^{q}Q_{ij}^{s}|\tilde{A}_i||\tilde{B}_j|\\ \leq&\sum_{i=1}^{k}\sum_{j=1}^{k}w_{ij}^{s}(|A_i^s|\frac{n}{N}+\alpha n)(|B_j^s|\frac{n}{N}+\alpha n)-\sum_{i=1}^{q}\sum_{j=1}^{q}Q_{ij}^{s}(|A_i^s|\frac{n}{N}-\alpha n)(|B_j^s|\frac{n}{N}-\alpha n)\\ \leq&(\frac{n}{N})^2|\Delta_{\mathcal{P}^s\mathcal{Q}^s}(A_i^s,B_j^s)|+\alpha\frac{n}{N}\sum_{i=1}^{k}\sum_{j=1}^{k}(w_{ij}^{s}+Q_{ij}^{s})(|A_i^s|+|B_j^s|)+\alpha n^2(\sum_{i=1}^{k}\sum_{j=1}^{k}w_{ij}^{s}-\sum_{i=1}^{q}\sum_{j=1}^{q}Q_{ij}^{s})\end{aligned}$$

It follows from corollary 1:

- $|\Delta_{\mathcal{P}^s\mathcal{Q}^s}(A_i^s,B_j^s)|\leq[4(\frac{n}{N})^2+2(1+\alpha k)^2+4]\epsilon N^2$
- $\alpha\frac{n}{N}\sum_{i=1}^{k}\sum_{j=1}^{k}(w_{ij}^{s}+Q_{ij}^{s})(|A_i^s|+|B_j^s|)\leq\alpha\frac{n}{N}4N=4\alpha n$
- $\alpha n^2(\sum_{i=1}^{k}\sum_{j=1}^{k}w_{ij}^{s}-\sum_{i=1}^{q}\sum_{j=1}^{q}Q_{ij}^{s})\leq 2\alpha^2n^2(k+q)$

Finally, we have with probability at least $(1-\gamma)(1-2\exp{(-2(\frac{n}{q}-\alpha n)^2\epsilon^2)})$,

$$\begin{aligned}&|\sum_{i=1}^{k}\sum_{j=1}^{k}w_{ij}^{s}|A_i||B_j|-\sum_{i=1}^{q}\sum_{j=1}^{q}Q_{ij}^{s}|\tilde{A}_i||\tilde{B}_j||\\ \leq&[4(\frac{n}{N})^2+2(1+\alpha k)^2+(\frac{1}{q}+\alpha)^2+6]\epsilon n^2+4\alpha n+2\alpha^2n^2(k+q).\end{aligned}$$

□

In short, we showed that the input graph is well-approximated by the output graph when considering the regular partition as encoder, the resulting random multipartite graph as feature, and blow-up/lifting as decoder.

3 Community Detection via Regular Partitions

In section 2, we give an estimation result about the closeness between the input graph and the output graph when using a regular partition as an encoder. In this section, we show that the ϵ-regular partitions give a partial community recovery for symmetric stochastic block models. In particular, we first state the formal definition of (Symmetric) Stochastic Block Models ((S)SBM), and show that the ϵ-regular partitions solve the fundamental problem of community detection in the study of SSBM with complex probability matrix under mild assumptions.

The idea of applying ϵ-regular partition to recover stochastic block model seems straight-forward: since the lemma tells us that every large enough graph could be well-approximated by a random graph, given a graph that is generated by a SSBM, any implement of the algorithm should be able to give us a random graph that is close to the graph and therefore to the SSBM. Unfortunately, because Szemeredi's regularity lemma and its implement algorithms relies on the index of partition as the objective function, the cardinality of an ϵ-regualr partition grows exponentially. That is, ϵ-regular partition tends to over-partition the set of vertices and give more clusters than the original community number in SSBM. Therefore, we use relative connectivity statistics to relabel the partitions and thereby generate a partial recovery for the original communities, up to a permutation on the community index.

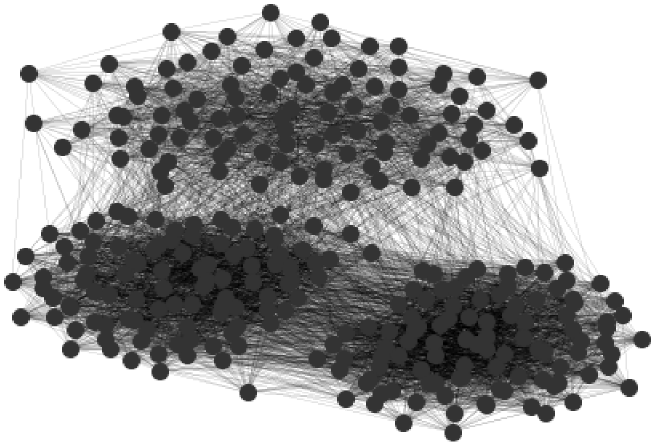


Figure 1: A sample graph that is generated from a stochastic block model with three communities.

Definition 1 (Stochastic Block Model (SBM)). *The tri-tuple (n,p,W) is called a stochatic black model, where*

- $n \in \mathbb{N}$;

- $p := \{p_i\}_{i=1}^k$, *where* $p_i \in [0,1], \forall i$, *and* $\sum_{i=1}^k p_i = 1$;
- $W \in \mathbb{R}^{k\times k}$, *where* $W_{ij} \in [0,1], \forall 0 \leq i,j \leq k$.

Definition 2 (Symmetric Stochastic Block Model (SSBM))**.** *A SBM is called symmetric if* $p_i = \frac{1}{k}, \forall i$, *and* $W_{ij} = W_{ji}, \forall 0 \leq i < j \leq k$.

Any sample graph $(X, G) \sim (n, p, W)$ can be considered as a random graph, where $X = \{x_i\}_{i=1}^n$ is a random n-dimensional vector with i.i.d. components distributed under the probability vector p, and $G := ([n], E)$ is a random simple graph with $\mathbb{P}(\{E_{ij}|x_i, x_j\}) = W_{x_i x_j}$. Here, $E_{ij} := \{(i,j),(j,i) \in E\}$.

Also, we denote the set of vertices that share the same label by $V_i := \{x_i \in [n] : x_i = k\}$. it is clear from the definition that $\{V_i\}_{i=1}^k$ forms a partition of $[n]$.

Now, given a SBM, we show that, for any two communities i, j, the edge density between correctly clustered communities gives a consistent estimator for W_{ij} by the law of large number.

Proposition 1. *Given a regular partition* $\mathcal{P}$ *on* G, *if* P_i *and* P_j *are correct community recovery, then* $\frac{e(P_i,P_j)}{|P_i||P_j|}$ *is a consistent estimator for* W_{ij}.

Proof. For any fixed $\epsilon > 0$, let $Pv_iPv_j := \{|P_i| = v_i, |P_j| = v_j\}$, we have the following:

$$\begin{aligned}
&\mathbb{P}(\{|\frac{e(P_i,P_j)}{|P_i||P_j|} - W_{ij}| \geq \epsilon\}) \\
=&\sum_{v_i,v_j} \mathbb{P}(\{|\frac{\sum_{i,j} e_{ij}}{v_i v_j} - W_{ij}| \geq \epsilon\}|Pv_iPv_j)\mathbb{P}(Pv_iPv_j) \\
\leq&\frac{Var(e_{ij})}{(p_i n - \epsilon)(p_j n - \epsilon)\epsilon^2} + \mathbb{P}(\{|\frac{v_i}{n} - p_i| \geq \epsilon\}) \\
=&\frac{W_{ij}(1-W_{ij})}{(p_i n - \epsilon)(p_j n - \epsilon)\epsilon^2} + \mathbb{P}(\{|\frac{v_i}{n} - p_i| \geq \epsilon\}) \longrightarrow 0.
\end{aligned}$$

Here, the inequality follows from the conditional i.i.d. assumption of e_{ij} and the correct cluster assumption, the second equality from the conditional Bernoulli assumption of e_{ij}.

□

Notice that the assumption in the above lemma is strong, because it requires algorithms to not only assign correct label to each sample vertex but also reveal the correct number of communities. As shown below, regular partitions result in correct labeling under mild assumptions, but tend to give more clusters than communities. Worse still, since the number of clusters usually grows with the number of sample vertices, the law of large number type of argument used above cannot be used directly if one want to prove similar consistency result above.

Therefore, we have the following sufficient condition for a regular partition to give consistent estimator under the assumption of correct labeling rather than assuming the partition gives the correct community recovery.

Lemma 5. *Given a regular partition $\{P_i\}_{i=1}^{K(n)}$ on G, if $K = o(n)$ and P_i and P_j are clustered correctly, i.e. $\exists j \in [k], P_i \cap V_j = P_i$, then $\frac{e(P_i,P_j)}{|P_i||P_j|}$ is a consistent estimator for W_{ij}.*

Proof. Similar to the proof above, we have:

$$\begin{aligned}&\mathbb{P}(\{|\frac{e(P_i,P_j)}{|P_i||P_j|} - W_{ij}| \geq \epsilon\})\\ \leq&\frac{W_{ij}(1-W_{ij})}{(p_i\frac{kn}{K}-\epsilon)(p_j\frac{kn}{K}-\epsilon)\epsilon^2} + \mathbb{P}(\{|\frac{v_iK}{nk} - p_i| \geq \epsilon\}) \longrightarrow 0.\end{aligned}$$

□

Hence, we have shown that if the cardinality of a regular partition does not grow as fast as the number of vertices, any correctly labeled pair of clusters give us some information about the original SSBM probability matrix, w. The question that naturally follows is when does a regular partition give correctly labeled pair of clusters. It turns out that if a SBM is satisfies the condition called ϵ-separability, then any regular pair of clusters gives an accurate labeling.

Definition 3. *A SBM (n,p,W) is called ϵ-separable if $\forall(i,j) \in [k]^2, \exists k \neq i,j$ such that $p_k|W_{ik} - W_{jk}| > \epsilon$.*

Intuitively, ϵ-separability allows us to characterize each of the community V_i by the vector $\{p_jW_{ij}\}_j$, up to a permutation on index.

Now, we show that if a SSBM is ϵ-separable, then the regular partition with none irregular pairs gives back the community up to ϵ mixture.

Lemma 6. *Given an ϵ-regular partition $\{P_i\}_{i=1}^{K}$ with zero irregular pairs of the sample graph G that is generated by an 2ϵ-separable SSBM(n,p,w) with n large, then we have:*

- *All of the P_i consists of vertices from the same community up to ϵ-mixture: $\forall i \in [K], \exists! j \in [k]$ such that $|P_i \cap V_j| > \epsilon|P_i|$.*

- $\forall j \in [k], \exists i \in [K],$ *such that* $|P_i \cap V_j| > \epsilon |P_i|$.

Proof. Assume we have an ϵ-regular partition $\{P_i\}_{i=0}^K$ on G. We focus on the K^2 ϵ-regular pairs of $(P_i, P_j)'s$.

Claim 1: Let $A \subset V_i, B \subset V_j$ such that $|A|, |B| > \epsilon|P|$ for some $P \in \{P_i\}_{i=0}^K$, then by the 2ϵ-separability, $\exists V_k$ such that $p_k|W_{ik} - W_{jk}| > 2\epsilon$. Now, if $\exists P^* \in \{P_i\}_{i=1}^K \backslash \{P\}$ such that $|P^* \cap V_k| > \epsilon|P^*|$, let $C := P^* \cap V_k$, we have by regularity that $|\frac{e(A,C)}{|A||C|} - \frac{e(P,P^*)}{|P||P^*|}| < \epsilon, |\frac{e(B,C)}{|B||C|} - \frac{e(P,P^*)}{|P||P^*|}| < \epsilon$. This implies

$$|\frac{e(A,C)}{|A||C|} - \frac{e(B,C)}{|B||C|}| < 2\epsilon.$$

But that contradicts the fact that

$$|\frac{e(A,C)}{|A||C|} - \frac{e(B,C)}{|B||C|}| \longrightarrow |W_{ik} - W_{jk}| > \frac{2\epsilon}{p_k} \geq 2\epsilon.$$

Finally, if $\nexists P^* \subset V_k$ such that $|P^* \cap V_k| > \epsilon|V^*|$. Then,

$$|V_k| \leq \epsilon|P_i|, \forall i \in \{1, ..., K\} \implies \frac{|V_k|}{n} \longrightarrow p_k < \epsilon,$$

which contradicts $p_k > \frac{2\epsilon}{|W_ik - W_jk|} > \epsilon$.

Claim 2: We prove by contradiction. Assume $\exists i \in [k], \forall j \in [K], |V_i \cap P_j| < \epsilon|P_j|$, then it follows

$$|V_i| < \epsilon|V| \implies p_i < \epsilon \implies p_i|W_{ji} - W_{ki}| < 2\epsilon, \forall j, k$$

contradicts the 2ϵ-separable assumption. □

Therefore, combine Lemma 5 and Lemma 6, we see that a regular partition without irregular pairs on any large enough graph generated by a SSBM indeed gives us a partial recovery of the communities with high accuracy. In experiments, we found the convergence rate of irregular pairs ratio to zero is particularly fast for graphs with strong probabilistic connectivity patterns. That is, the assumption $K = o(n)$ in Lemma 5 and the non-irregular pair assumption in Lemma 6 can be easily satisfied in practice, provided that the graph indeed has clear probabilistic connectivity pattern.

Finally, we propose an statistic characterization of the community label for ϵ-separated SSBM based on the observations of the connectivity between the nodes and the training data.

To state the second main result, we need the following definitions. By the results of Lemma 6, we can define a map $f : [K] \to [k]$ such that for every $i \in [K]$,

$f(i) \in [k]$ gives the unique index satisfying $|V_{f(i)} \cap P_i| > \epsilon|P_i|$. Notice the map f is unique up to a permutation on $[k]$. In practice, one could use the relative edge density among $\{P_i\}_{i=1}^K$ to find the corresponding $f(i)$. Thereby, we define $\forall i \in [k], I_i := f^{-1}(\{i\})$ being the inverse image of $\{i\}$ on $[K]$, and $\tilde{V}_i := \bigcup_{k\in I_i} P_k, \forall i \in [k]$ being the recovered communities.

Theorem 2. *Given an 2ϵ-separated SSBM, a regular partition with none irregular pairs, and an arbitrary vertex v, we have:*

$$\prod_{i=1}^{k} \mathbb{1}_{\{|\frac{|e(\{v\},\tilde{V}_i)|}{|\tilde{V}_i|} - W_{ji}| < k\epsilon\}}$$

is a consistent estimator of the label indicator $\mathbb{1}_{\{v\in V_j\}}$ for all $j \in [k]$.

Proof. Given the resulting recovered communities $\{\tilde{V}_i\}_{i=1}^k$, and any unlabeled vertex v that is from the SSBM, we have:

$$|\frac{|e(\{v\},\tilde{V}_i)|}{|\tilde{V}_i|} - W_{ji}| = |\frac{|e(\{v\},\tilde{V}_i \setminus V_i)|}{|\tilde{V}_i|} + \frac{|e(\{v\},\tilde{V}_i \cap V_i)|}{|\tilde{V}_i \cap V_i|}\frac{|\tilde{V}_i \cap V_i|}{|\tilde{V}_i|} - W_{ji}|.$$

From Lemma 6: $\forall i \in [K], \exists! j \in [k],$ such that $|P_i \cap V_j| > (1 - k\epsilon)|P_i|$, we have $\frac{|e(\{v\},\tilde{V}_i \setminus V_i)|}{|\tilde{V}_i|} \in [0, k\epsilon]$ and $\frac{|\tilde{V}_i \cap V_i|}{|\tilde{V}_i|} \in (1 - k\epsilon, 1]$ by the construction of $\{\tilde{V}_i\}_{i=1}^k$.

Now, if $v \in V_j$, then by the weak law of large number,

$$\mathbb{P}(\{|\frac{|e(\{v\},\tilde{V}_i \cap V_i)|}{|\tilde{V}_i \cap V_i|} - W_{ji}| > \delta\}) \longrightarrow 0, \forall\delta > 0, \forall i \in [k].$$

That together with the above uniform bounds further imply

$$\mathbb{P}\{|\frac{|e(\{v\},\tilde{V}_i)|}{|\tilde{V}_i|} - W_{ji}| \le k\epsilon\} \longrightarrow 1, \forall i \in [k].$$

On the other hand, if $v \in V_k$ for some $k \neq j$, then by the 2ϵ-separable assumption, $\exists i^* \in [k]$ such that $\frac{1}{k}|W_{ki^*} - W_{ji^*}| > 2\epsilon$. That together with the above uniform bounds imply

$$\mathbb{P}(\{|\frac{|e(\{v\},\tilde{V}_{i^*})|}{|\tilde{V}_{i^*}|} - W_{ji^*}| \le k\epsilon\}) \longrightarrow 0.$$

Combine the results above, we obtain,

$$\mathbb{P}(\{|\mathbb{1}_{v\in V_j} - \prod_{i=1}^{n} \mathbb{1}_{|\frac{|e(\{v\},\tilde{V}_i)|}{|\tilde{V}_i|} - W_{ji}| \le k\epsilon}| > \delta\}) \longrightarrow 0, \forall\delta > 0.$$

□

The above result gives us a reliable approach to assign community label to unlabeled nodes based on their connectivity statistics with respect to the training data. Thereby, we could predict any unobserved link given enough observations on the relative nodes' connectivity to the training data.

In conclusion, we have proved that any ϵ-regular partition on with none irregular pairs gives a partial recovery of the SSBM. Moreover, the a relabeling of the elements in an ϵ-regular partition on the training data gives a consistent estimator for the label indicator.

4 Feature Extraction and Link Prediction

Now, we start to propose algorithms for feature extraction and link prediction by exploring the proved accuracy of the ϵ-regular partitions in community detection. In particular, the regular partitions allows us to find the principal components for the covariance matrix generated by the element-wise means on the ambient feature data space. The principal components are considered major latent features with the following reason: If there exists latent features that induce the formation of such connectivity-based communities, it should explain the variance on the covariance among different connectivity patterns. But in the ϵ-regular partition case, difference in connectivity patterns are reflected in difference in the communities. Therefore the principal components of covariance matrix generated by community samples give the latent features. Finally, since we proved that each element in the ϵ-regular partition is a cluster of data points that share a similar connectivity pattern in the last section, the mean of elements can be considered as typical samples from the respective communities.

After feature extraction, we take a further step to assume there exits some parametrized kernel that give a proper graph embedding on the subspace spanned by the extracted features. Therefore, we use the maximum-likelihood estimator for link prediction purpose.

Now, we introduce Algorithm 1 below which extracts latent features from the given relational data:

For example, on a three dimensional data space $\mathbb{R}^3$, where the first dimension corresponds to age, the second corresponds to income, and the third to height. If the PCA gives us back $x_1 = (1, 0, 0)$, $x_2 = (0, 1, 0)$, and the threshold $vt = 0.95$ is satisfied. Then the outcome infer that the relationship pattern in the relational data is generated by age and income.

In the case where relational data is indeed generated by feature vectors polluted by noise, let $\mu_{P_i} := \frac{1}{|P_i|} \sum_{v \in P_i} v$ and $\tilde{V}_*$ denotes the labeled community contains P_i we have the following result to prove the closeness between element-wise

Algorithm 1: Latent Feature Extraction via Regular Partition

```
Input: (training data = $(V, E)$, regularity threshold = rt, community
labeling accuracy threshold = $\epsilon$, and variance explanation threshold =
vt);
Szemeredit Regularity Partition: $G \longrightarrow (\{P_i\}_{i=1}^K, rp)$;
if $rp < rt$ then
    print: not enough SBM pattern found with rt regularity percentage
    requirement;
else
    $\forall i \in [K]$, find the element mean $\mu_i := \frac{1}{|P_i|} \sum_{v \in P_i} v$;
    find the covariance matrix $\Sigma$ of $\{\mu_i\}_{i=1}^K$;
    obtain $p = \min_p \{p : \frac{\sum_{i=1}^p \lambda_i}{\sum_{i=1}^d \lambda_i} > vt\}$ via SVD $\Sigma = Udiag(\{\lambda_i\})U^T.$;
    find $\{u_i\}_{i=1}^p$
end
Result: Latent Features: $\{u_i\}_{i=1}^p$
```

means and feature vectors:

Lemma 7. *If distinct communities in a 2ϵ-separated SBM is generated by distinct feature vectors but polluted by i.i.d. noise with zero mean in the ambient space, then the almost sure limit* $\lim ||\mu_{P_i} - v_*|| \leq \epsilon \sum_{k \neq *} ||v_* - v_k||$.

Proof. Assume that communities are generated by $\{v_i\}_{i=1}^k$, which are polluted by i.i.d. noise e with $\mathbb{E}(e) = 0$. Then by lemma 6, we have $\forall P_i, |P_i \setminus V_*| \leq (k-1)\epsilon |P_i|$ and

$$\mu_{P_i} = \frac{K}{n}\left(\sum_{v_i \in P_i \cap V_*} (v_* + e_i) + \sum_{k \neq *} \sum_{v_i \in P_i \cap V_k} (v_k + e_i) \right).$$

Finally, it follows from the strong law of large number,

$$\begin{aligned} lim_{n \to \infty} ||\mu_{P_i} - v_*|| &\leq \lim_{n \to \infty} ||\frac{K}{n} \sum_i e_i|| + \epsilon \sum_{k \neq *} ||v_* - v_i|| \\ &= \epsilon \sum_{k \neq *} ||v_* - v_i|| \end{aligned}$$

□

Before introducing the proposed algorithm for link prediction, we first show that an application of KL-divergence between the relational data and the proposed metric relationship between data leads to an optimal graph embedding to the data space.

Given two discrete probability distributions $p = \{p_i\}_{i=1}^K$ and $q = \{q_i\}_{i=1}^K$, the KL-divergence between p and q is defined as the following:

$$D_{KL}(p||q) := \sum_{i=1}^{K} p_i \log(\frac{p_i}{q_i}).$$

Intuitively, KL-divergence give a quantitative description for the difference between two probability distributions. In particular, $D_{KL}(p||q) = 0 \iff p \overset{d}{=} q$.

We will use KL-divergence to measure the closeness between the given sample connectivity and the connectivity relationship induced by a parametrized metric on the data subspace spanned by the extracted features. This method is inspired by the non-linear low-dimensional data embedding method such as t-SNE (Hinton and Roweis, 2002), which uses a the Gaussian kernel to approximate the data manifold structure by diffusion process on the ambient space.

In our case, we use a parametrized kernel function $\sigma_\theta(\mu_i, \mu_j)$ to approximate the relational manifold on the data space. Popular choices of σ_θ includes:

- Weighted l^2 norm, $||\mu_i - \mu_j||_{l^2_\theta} := \sum_{k=1}^d \theta_i(\mu_{i,k} - \mu_{j,k})^2$ with $\theta \in \mathbb{R}^d$ non-negative,
- Bi-linear form $\langle x, \theta y\rangle_{l^2}$ with $\theta \in \mathbb{R}^{d\times d}$ symmetric.

In particular, for each μ_i, we use $\{\frac{e^{-\sigma_\theta(\tilde{\mu}_i,\tilde{\mu}_j)}}{\sum_{j=1}^K e^{-\sigma_\theta(\tilde{\mu}_i,\tilde{\mu}_j)}}\}_{j=1}^K$, where $\tilde{\mu}_i := \sum_{k=1}^p \langle \mu_i, u_k\rangle_{l^2} u_k$, to approximate the probability of a connection between μ_i and μ_j on the subspace spanned by the extracted features. Our goal is to find the maximum-likelihood estimator for the empirical relation structure $\{\frac{w_{ij}}{\sum_j w_{ij}}\}_{j=1}^K$:

$$\arg\min_\theta \sum_{i=1}^{K} D_{KL}(\{\frac{w_{ij}}{\sum_j w_{ij}}\}_{j=1}^K||\{\frac{e^{-\sigma_\theta(\tilde{\mu}_i,\tilde{\mu}_j)}}{\sum_{j=1}^K e^{-\sigma_\theta(\tilde{\mu}_i,\tilde{\mu}_j)}}\}_{j=1}^K).$$

Now, we are ready to introduce algorithm 2 for link prediction.

Now, given any test data set $\{t_i\}_{i=1}^n$, we first obtain the projection $\tilde{t}_i := \sum_{k=1}^p \langle t_i, u_k\rangle_{l^2} u_k, \forall i \in [n]$, and obtain the estimated probability of edges among them by:

$$\mathbb{P}(\{e(t_i, t_j) = 1\}) = \frac{\exp\left(-\sigma_{\theta^*}(\tilde{t}_i, \tilde{t}_j)\right)}{\sum_k \exp\left(-\sigma_{\theta^*}(\tilde{t}_i, \tilde{t}_k)\right)}.$$

Therefore, one could use the generated probability model for link prediction.

Algorithm 2: Link Prediction via Graph Embedding on Metric Data Space

```
Input: (the empirical edge density between elements of a regular
partition = $\{W_{ij}\}_{i,j\in[K]}$, the extracted latent features = $\{u_i\}_{i=1}^{p}$
generated from the regular partition, the element-wise mean $\{\mu_i\}_{i=1}^{K}$, a
parametrized kernel function = $\sigma_\theta$, termination threshold = $\epsilon$);
Find $\tilde{\mu}_i = \sum_{k=1}^{p}\langle \mu_i, u_k\rangle_{l^2} u_k$, which are the projection of $\mu_i$ onto the
subspace spanned by $\{u_i\}_{i=1}^{p}$;
Let $f(\theta) := \sum_{i=1}^{K} D_{KL}(\{\frac{W_{ij}}{\sum_j W_{ij}}\}_{j=1}^{K} || \{\frac{e^{-\sigma_\theta(\tilde{\mu}_i,\tilde{\mu}_j)}}{\sum_{j=1}^{K} e^{-\sigma_\theta(\tilde{\mu}_i,\tilde{\mu}_j)}}\}_{j=1}^{K})$;
Find $\theta^* :=$ argmin $f(\theta)$;
Result: the maximum-likelihood graph embedding kernel
        $d : span(\{u_i\}_{i=1}^{p}) \longrightarrow \mathbb{R}^+ \cup \{0\}$ defined by
        $d(x,y) := \exp(\sigma_{\theta^*}(x,y)), \forall x,y \in span(\{u_i\}_{i=1}^{P})$
```

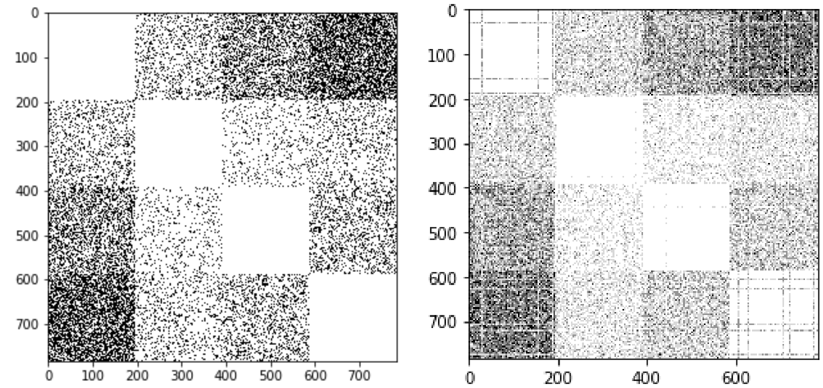


Figure 2: The left is the original graph from SBM, the right is the output weighted adjacency matrix from an ϵ-regular partition

5 Numerical Experiments

In this section, we show the numeric experiments on the application of regular partitions to solve community detection and recovery.

In Figure 2, we use synthetic data generated by SBM:

$$(n = 784, p = \begin{pmatrix}0.25\\0.25\\0.25\\0.25\end{pmatrix}, W = \begin{pmatrix}0 & 0.2 & 0.4 & 0.6\\0.2 & 0 & 0.1 & 0.15\\0.4 & 0.1 & 0 & 0.25\\0.6 & 0.15 & 0.25 & 0\end{pmatrix}).$$

We have exact community recovery in this case. Since the data has been clustered into four communities that are characterized by the edge density matrix

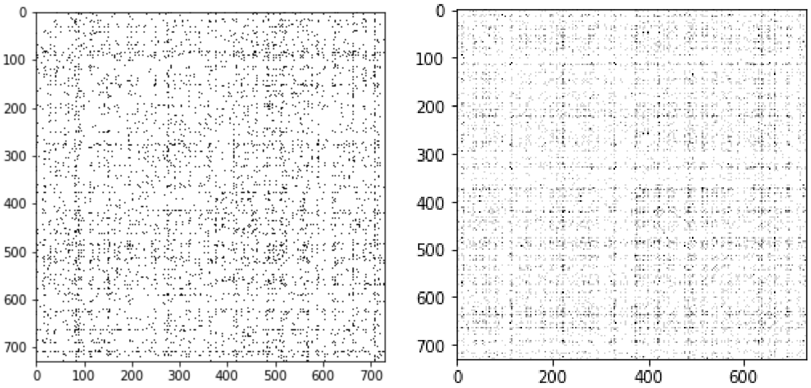
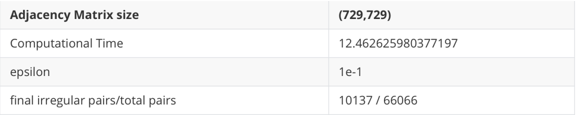


Figure 3: The left is the original graph from (Traud, 2011), the right is the output weighted adjacency matrix from an ϵ-regular partition.



Adjacency Matrix size	(729,729)
Computational Time	12.462625980377197
epsilon	1e-1
final irregular pairs/total pairs	10137 / 66066

Figure 4: An application of regular partition algorithm on the real Facebook data only gives partial community recovery.

$\{W_i\}_{i=1}^4$, the i^{th} row of W, as we proved in Theorem 2. Also, the algorithm converges after 7 iterations and gives 0 irregular percentage.

In Figure 2, the left is the original data, which consists of $e_{ij} \in \{0, 1\}$. Therefore, one cannot recover the original adjacency matrix after applying a permutation on the index. The right is the output weighted adjacency matrix from an ϵ-regular partition where $e_{ij} \in \{0, 0.2, 0.4, 0.6, 0.1, 0.15, 0.25\}$. Since each community is characterized by the W_i's, one could cluster the output weighted adjacency matrix even if a permutation is applied on the vertex index.

In Figure 3, we use real Facebook relational data from (Traud, 2011). The graph is relatively sparse. Still, an application of regular partition gives a partial recovery of the communities from the original adjacency matrix that is on the left. Figure 4 shows the details: the random-connectivity pattern on the Facebook relational data is not strong enough to generate an ϵ-regular partition, as the irregular pair rate is around $\frac{1}{6}$. In other words, the $K = o(n)$ assumption in Lemma 5 is unlikely to be satisfied.

Therefore, the numeric results on both synthetic data and real data support our theoretical guarantee results on community detection and recovery. Furthermore, the experiment result also confirms our conjecture on fast convergence rate for algorithms generating regular partitions on graph with highly structured random connectivity, despite the exponential tower-type bound of partition cardinality with respect to the accuracy ϵ.

6 Conclusion

In this paper, we shed the light on the potential of applying Szemeredi's regularity partitions as a black box to solving community detection, feature extraction, and link prediction problems.

In section 2, we show that, given an unknown large dense graph and its sample subgraph under uniform sampling, the ϵ-regular partitions give accurate pattern or information summery for the unknown graph. This estimation result suggests that the regular partitions, as a random-connectivity pattern recognition tool, can be robust to any noise that does not bring structural change. In the future works, we hope to give a theoretical guarantee on the noise-robustness of the ϵ-regular partitions.

In section 3, we proved that regular partitions give ϵ-mixture solution to community detection when using benchmark graphs that are generated by 2ϵ-separable symmetric stochastic black model. Community detection and recovery for stochastic block models with complex probability matrix is a long-standing tough problem. Our theoretical guarantee result shows that ϵ-regular partition is an efficient tool in solving the problems.

In section 4, we apply the good community detection and recovery result to propose algorithms for feature extraction and link prediction. To extract latent feature, we first use the element-wise means in the ϵ-regular partition as sample data draw randomly from the unknown feature-induced communities, then calculate the empirical covariance matrix of the element-wise means to, and finally extract the leading principal components as latent features as the subspace spanned by them tend to explain more of the variance in connectivity pattern.

In section 5, we give experiment results on the community detection and . The numerical results on community detection using synthetic data confirm our theoretical guarantee on recovering communities from SSBMs. Furthermore, the experiment result also confirms our conjecture on fast convergence rate for algorithms generating regular partitions on graph with highly structured random connectivity, despite the exponential tower-type bound of partition cardinality with respect to the accuracy ϵ.

References

S. Fortunato (2010). Community detection in graphs. *Physics Reports* **486 (3-5)**:75—174.

M. E. J. Newman, D. J. Watts, and S. H. Strogatz (2002). Random graph models of social networks. *Proc. Natl. Acad. Sci.* **99**:2566—2572.

J. Shi and J. Malik (2000). Normalized cuts and image segmentation. *IEEE Transac- tions on Pattern Analysis and Machine Intelligence* **22**:888—905.

Paolo Bartesaghi and Stefano Benati and Gian Paolo Clemente and Rosanna Grassi (2019). Multi-criteria community detection in International Trade Network. *ArXiv* abs/1911.08593.

Hoffman, Michaela et al. (2018). Detecting Clusters/Communities in Social Networks. *Multivariate behavioral research* **53,1**:57–73.

Ling, S., Strohmer, T. (2020). Certifying Global Optimality of Graph Cuts via Semidefinite Relaxation: A Performance Guarantee for Spectral Clustering. *Found Comput Math* **20**:367—421.

Agarwal N., Bandeira A.S., Koiliaris K., Kolla A. (2017). Multisection in the Stochastic Block Model Using Semidefinite Programming. *Compressed Sensing and its Applications*

Abbe, Emmanuel, Afonso S Bandeira, and Georgina Hall (2016). Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory* **62.1**:471—487.

Abbe, E. and C. Sandon (2015). Community Detection in General Stochastic Block models: Fundamental Limits and Efficient Algorithms for Recovery. *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, 670-–688.

Roweis, Sam and Hinton, Geoffrey (2002). Stochastic neighbor embedding. *Neural Information Processing Systems*, **15**:833-–840.

T. Gowers (1997). Lower bounds of tower type for Szemeredi's uniformity lemma. *Geom Func. Anal., vol. 7*, **2**:322-–337.

N. Alon, R. A. Duke, H. Lefmann, V. Rodl, and R. Yuster (1994). The algorithmic aspects of the regularity lemma. *J. Algorithms, vol. 16*, **1**:80—109.

A. Sperotto and M. Pelillo (2007). Szemeredi's regularity lemma and its applications to pairwise clustering and segmentation. *Energy Minimization Methods in Computer Vision and Pattern Recognition, 6th International Conference*:13—27.

M. Pelillo, I. Elezi, M. Fiorucci (2017). Revealing structure in large graphs: Szemerdi's regularity lemma. *Pattern Recognition Letters*, **87**:4—11.

L. Lovasz (2012). Large Networks and Graph Limits. *AMS Colloquium Publications*: Vol. 60.

E. Szemeredi (1978). Regular partitions of graphs. *Proc. Colloque Inter. CNRS*:399—401.

Traud, Amanda L., Mucha, Peter J., and Porter, Mason A. Porter (2011). Social Structure of Facebook Networks. *arXiv:1102.2166*

Traud, Amanda L., Eric D. Kelsic, Mucha, Peter J., and Mason A. Porter (2011). Comparing Community Structure to Characteristics in Online Collegiate Social Networks. *SIAM Review, in press (arXiv:0809.0960)*