Title
Semi-Supervised Clustering of Sparse Graphs: Crossing the Information-Theoretic Threshold

Author
Sheng, Junda

Publication Date
2022

Peer reviewed|Thesis/dissertation

Semi-Supervised Clustering of Sparse Graphs: Crossing the Information-Theoretic Threshold

By

JUNDA SHENG

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

APPLIED MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Professor Thomas Strohmer, Chair

Professor Wolfgang Polonik

Professor Albert Fannjiang

Committee in Charge

2022

To my family

Contents

Semi-Supervised Clustering of Sparse Graphs:
Crossing the Information-Theoretic Threshold

Abstract

The stochastic block model, also known as the planted partition model, is considered a canonical random graph model to study clustering and community detection on network-structured data. Decades of extensive study on the problem have established many profound results, among which the phase transition for weak recovery at Kesten-Stigum threshold is particularly interesting both from a mathematical and an applied point of view. It says that no estimator can perform substantially better than chance on sparse graphs if the model parameter is below the threshold when we have access only to the network topology.

Nevertheless, if we slightly extend the horizon to the ubiquitous semi-supervised setting, such a fundamental limitation will disappear completely. We prove that with arbitrary fraction of the labels revealed, the detection problem is feasible throughout the parameter domain. Moreover, we introduce two efficient algorithms, one combinatorial and one based on optimization, to integrate label information with graph structures. Our work brings a new perspective to stochastic model of networks and semidefinite program research. The foundational change caused by semi-supervised learning demonstrates its indispensable power.

In turn, the mathematically rigorous results help us to develop powerful tools for real-world applications. We propose a variation of graph convolutional network based on our clustering algorithms, which is the first of its kind to incorporate semi-supervised approach in the design of propagation scheme. It utilizes the non-local information that is justified by the learning target. Meanwhile, it captures the essence of cluster structure instead of model statistics. Numerical experiments show it outperforms other models on challenging tasks.

Acknowledgments

My past five years at Davis have been the most precious life experience to me. The mathematical training I received and the research I conducted during my Ph.D. period shape who I am today. In this amazing journey, I have gradually developed the ability to learn and understand complicated things in a quick and systematic manner. No matter how challenging the task is in the future, there is no fear in my heart.

First, I would like to thank my advisor Professor Thomas Strohmer for his guidance and support throughout this journey. Thomas taught me how to do proper research, how to write good papers, how to give academic talks, basically everything needed to be an awesome researcher. Besides, it is always quite enjoyable to work with Thomas. I would like to thank Professor Wolfgang Polonik and Professor Albert Fannjiang for serving in my dissertation committee, Professor Naoki Saito and Professor Luis Rademacher for serving in my qualifying exam committee.

I genuinely appreciate all my research collaborators and mentors, who have greatly contributed to my academic achievement. Thanks to Roman Vershynin, March Boedihardjo, Krystle Reagan, Xiaodong Li, Shiqian Ma, Bao Wang, Robert Guy for guiding me during our research collaborations and sharing those great advice on Ph.D. life.

I am very grateful for the support and mentorship from many great industrial researchers. I want to thank Nicolas Dreyfuss, Yoann Le Calonnec at Squarepoint Capital and Akif Rahim, Yan Gao at Roku Inc. I learned a lot from their visionary insights and hands-on experiences, and my mindset on applied math and machine learning research has been reshaped ever since.

I always enjoy spending time with my colleagues and friends. Thanks to Yang Li, Joseph Pappe, Dong Min Roh, Rui Okada, Austin Tran, Fushuai Jiang, Haihan Wu, Xingmei Lou, Haolin Chen, Haotian Sun, Xiaotie Chen, Shaofeng Deng, Zhenyang Zhang, Xue Feng, Yuan Ni, Yunshen Zhou, Zhongruo Wang, Yiqun Shao for all the discussions on research and life.

Finally, I wish to thank my parents and my fiancée Yan, for their unconditional love.

CHAPTER 1

Introduction

1.1. Clustering on graphs

The basic task of *clustering* or *community detection* in its general form is, given a (possibly weighted) graph, to partition its vertices into several densely connected groups with relatively weak external connectivity. This property sometimes is also called assortativity. Clustering and community detection are central problems in machine learning and data science with various applications in scientific research and industrial development. A considerable amount of data sets can be represented in the form of a network that consists of interacting nodes, and one of the first features of interest in such situation is to understand which nodes are "similar", as an end or as preliminary step towards other learning tasks. Clustering is used to find genetically similar sub-populations [**Pad14**], to segment images [**SM00**], to study sociological behavior [**NWS02a**], to improve recommendation systems [**LSY03**], to help with natural language processing [**GB13**], etc. Since the 1970s, in different communities like social science, statistical physics and machine learning, a large diversity of algorithms have been developed such as:

- Hierarchical clustering algorithms [**Joh67**] build a hierarchy of progressive communities, by either recursive aggregation or division.
- Model-based statistical methods, including the celebrated EM clustering algorithm proposed in [**DLR77**], fit the data with cluster-exhibiting statistical models.
- Optimization approaches identify the best cluster structures in regard to carefully designed cost functions, for instance, minimizing the cut [**HS00**] and maximizing the Girvan-Newman modularity [**NG04**].

Due to the absence of labels, clustering is considered to be more difficult task than classification. The given label in the case of supervised learning serves as a clue to grouping data objects as a whole. Whereas in the case of unsupervised clustering, it is not straightforward to decide, to which cluster a pattern should belong. The challenge inspires us to study the semi-supervised approaches that are able to better extract the real groupings with a little help of the revealed labels. Intuitively, similarity is the central factor to a cluster and hence clustering process. The natural grouping of data based on some inherent similarity is to be discovered in clustering. We will see in the following that the similarity is commonly defined by a measure of distance (e.g. Euclidean distance and graph distance) among the objects. The similarity and group structure can be abstractly represented by a graph, which bridges the practical task to elegant mathematical theories.

Multiple lines of research intersect at a simple random graph model, which appears under many different names. In the machine learning and statistics literature around social networks, it is called the stochastic block model (SBM) [**HLL83**], while it is known as the planted partition model [**BCLS84**] in theoretical computer science and referred to as inhomogeneous random graph [**BJR07**] in the mathematics literature. Moreover, it can also be interpreted as a spin-glass model [**DKMZ11**], a sparse-graph code [**AS15**] or a low-rank random matrix model [**McS01**] and more.

Stochastic block model is a generative model for random graphs. The essence of SBM can be summarized as follows: Conditioned on the vertex labels, edges are generated independently and the probability only depends on which clusters the pairs of vertices belong to. This abstract and mild assumption is all we need to depict the essence of network data. On the one hand, we use the patterns of connections to distinguish different clusters. On the other hand, we are aware of that not all connections directly imply the similarity of nodes.

In general, an unweighted graph (G) consists of a collection of nodes / vertices (V) and a edges set (E) connecting different nodes. Conventionally, we call the number of nodes

$n = |V|$ and the number of edges $m = |E|$. A graph can be uniquely represented by its adjacency matrix $A \in \mathbb{R}^{n\times n}$: $A_{ij} = 1$ if $(i, j) \in E$, i.e. there exists an edge connecting node i to node j and $A_{ij} = 0$ otherwise. Throughout this dissertation we only consider undirected graphs without self-loops. So A will be a symmetric matrix and $A_{ii} = 0$, $i = 1, \dots n$. The degree of node i is defined as the number of its direct neighbours, $d(i) = \sum_j A_{ij}$. If the graph is labeled, we denote the label vector as $x \in \mathbb{R}^n$ with the corresponding indices of the nodes. A community / cluster is the collection of all nodes that share the same label. To keep the notation under control, we identify the name of a node with its index. So, without loss of generality, we assume that $V = [n]$ where $[n] \coloneqq \{1, \dots, n\}$.

DEFINITION 1.1.1 (General SBM). Let $n,\ k \in \mathbb{N}^*$, $p = (p_1, \dots, p_k)$ be a probability vector on $1, \dots, k$ (the prior distribution, $\sum_{i=1}^k p_k = 1$) and the connectivity probability $W \in \mathbb{R}^{k\times k}$ be a symmetric matrix with $W_{ij} \in [0, 1]$, $i, j \in [k]$. A random object $(x, G = ([n], E))$ obeys the SBM(n, p, W), if x is a n-dimensional random vector whose components are i.i.d. instances of p, and G is a n-vertex simple graph with adjacency matrix A where $\{A_{ij}\}$'s are independent random variables that

$$A_{ij} \sim \text{Bernoulli}(W_{x_i x_j}),\ A_{ji} = A_{ij} \quad (i < j) \tag{1.1}$$

and $A_{ii} = 0$, $i \in [n]$. The communities are defined as $S_i(x) = \{v \in V | x_v = i\}$, $i \in [k]$.

Therefore, we can explicitly put down the distribution of $(x, A) \sim \text{SBM}(n, p, W)$ for $x^* \in [k]^n$ and $A^* \in \{0, 1\}^{n\times n}$ as the following,

$$\mathrm{P}(x = x^*) = \prod_{v\in[n]} p_{x_v^*} = \prod_{i\in[k]} p_i^{|S_i(x^*)|} \tag{1.2}$$

$$\mathrm{P}(A = A^* | x = x^*) = \prod_{1\le u<v\le n} W_{x_u^* x_v^*}^{A_{uv}^*} (1 - W_{x_u^* x_v^*})^{1-A_{uv}^*} \tag{1.3}$$

$$= \prod_{1\le i\le j\le k} W_{ij}^{N_{ij}} (1 - W_{ij})^{N_{ij}^{\complement}} \tag{1.4}$$

where the numbers of edges within and across communities (N) and the numbers of non-edges within and across communities ($N^{\complement}$) are calculated as the following.

When $i \neq j$,

$$N_{ij}(x^*, A^*) = \sum_{x_u^*=i,\ x_v^*=j} A_{uv} \tag{1.5}$$

$$N_{ij}^{\complement}(x^*, A^*) = \sum_{x_u^*=i,\ x_v^*=j} (1 - A_{uv}) \tag{1.6}$$

$$= |S_i(x^*)| \cdot |S_j(x^*)| - N_{ij}(x^*, A^*). \tag{1.7}$$

When $i = j$,

$$N_{ii}(x^*, A^*) = \sum_{\substack{u<v \\ x_u^*=x_v^*=i}} A_{uv} = \frac{1}{2} \sum_{x_u^*=x_v^*=i} A_{uv} \tag{1.8}$$

$$N_{ii}^{\complement}(x^*, A^*) = \sum_{\substack{u<v \\ x_u^*=x_v^*=i}} (1 - A_{uv}) \tag{1.9}$$

$$= \frac{1}{2}|S_i(x^*)| \cdot (|S_i(x^*)| - 1) - N_{ii}(x^*, A^*). \tag{1.10}$$

The law of large numbers implies

$$\lim_{n\to\infty} \frac{|S_i|}{n} = p_i \quad a.s.. \tag{1.11}$$

Hence, instead of assuming $x \sim p$, we can also define the SBM with the label vector x drawn uniformly at random under the constraint that $|S_i| = np_i$, $i \in [k]$. For the purpose of this dissertation, these two definitions are equivalent.

Figure 1.1 demonstrates how the network datasets are described by graphs. On the left, we show the graph associated with the EMAIL-EU dataset [**YBLG17**], which is collected from an e-mail communication network of an European research institution. Nodes indicate members of the institution. Department memberships of individual researchers are considered the ground truth community. The edge between two researchers implies that

they exchanged at least one email. For clarity, we only plot a subgraph of 3 departments randomly selected from the original 42 departments. On the right, we generate one instance of $\mathrm{SBM}(25, (0.4, 0.3, 0.3)^\top, W^*)$ where

$$W^* = \begin{pmatrix} 0.75 & 0 & 0.0625 \\ 0 & 0.5 & 0.125 \\ 0.0625 & 0.125 & 0.5 \end{pmatrix}.$$

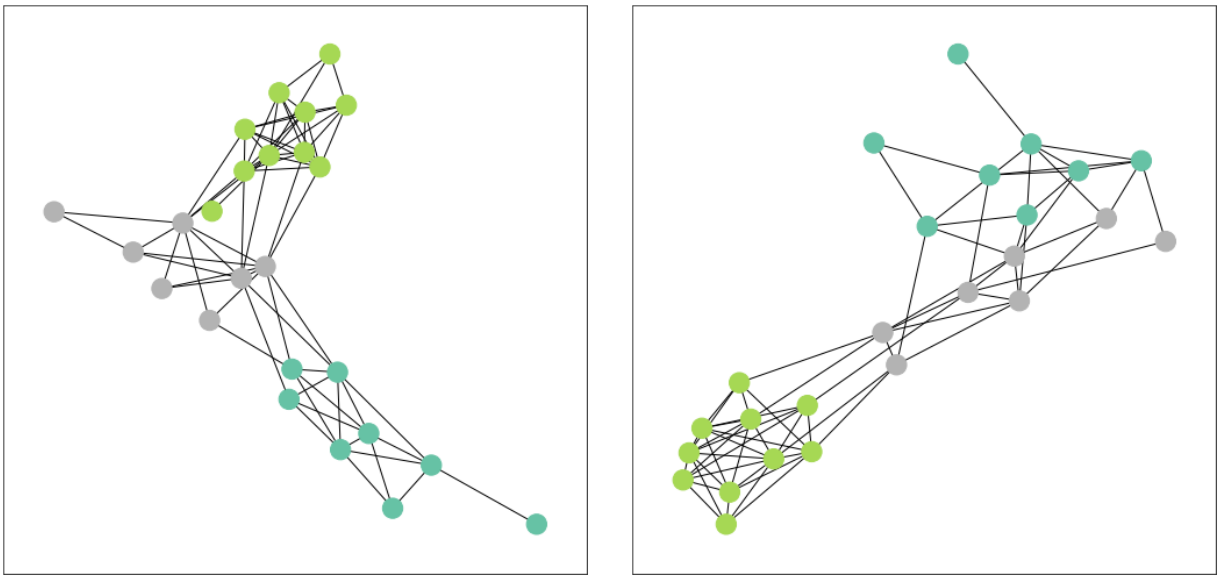

FIGURE 1.1. Community structures in the EMAIL-EU network (left) and a SBM instance (right). Nodes are color-coded according to the labels.

The symmetric SBM consisting of two blocks is also known as the planted bisection model. It takes the simplest form of the model but provides a wonderful ground for rigorous study of various clustering methods and the fundamental limits hidden in complicated problems.

DEFINITION 1.1.2 (Planted bisection model). For $n \in \mathbb{N}$ and $p, q \in (0, 1)$, let $\mathcal{G}(n, p, q)$ denote the distribution over graphs with n vertices defined as follows. The vertex set is partitioned uniformly at random into two subsets S_1, S_2 with $|S_i| = n/2$. Let E denote the

edge set. Conditional on this partition, edges are included independently with probability

$$\mathrm{P}\left((i,j)\in E|S_1,S_2\right)=\begin{cases}p & \text{if } \{i,j\}\subseteq S_1 \text{ or } \{i,j\}\subseteq S_2,\\ q & \text{if } i\in S_1, j\in S_2 \text{ or } i\in S_2, j\in S_1.\end{cases} \tag{1.12}$$

Note that if $p=q$, the planted bisection model is reduced to the so-called Erdős–Rényi random graph where all edges are generated independently with the same probability. Hence there exists no cluster structure. But if $p\gg q$, a typical graph realization will have two well-defined clusters. The scale of p and q also plays a significant role in the resulting graph, which will be discussed in detail later. They govern the amount of signal and noise in the graph generating process. As the key parameters that researchers work with, various regimes and thresholds are described by them.

The SBMs generate labels for vertices before the graph. The ground truth allows us to formally discuss the presence of community structures and measure the performance of algorithms in a meaningful way. It also supplies a natural basis to rigorously define the semi-supervised clustering problem. But as a parametrized statistical model, one can only hope that it serves as a good fit for the real data. Although not necessarily a realistic model, SBM provides us an insightful abstraction and captures some of the key phenomena [**MNS15**, **CX16**, **BMNN16**, **ABH16**, **AS18**].

Given a single realization of the graph G, our goal is to recover the labels x, up to certain level of accuracy. Formally, the ground truth of the underlying community structure is encoded using the vector $x\in\{+1,-1\}^n$, with $x_i=+1$ if $i\in S_1$, and $x_i=-1$ if $i\in S_2$. An estimator is a map $\hat{x}: G_n\to\{+1,-1\}^n$ where G_n is the space of graphs over n vertices. We define the *overlap* between an estimator and the ground truth as

$$\mathrm{Overlap}(x,\hat{x}(G))=\frac{1}{n}|\langle x,\hat{x}(G)\rangle|. \tag{1.13}$$

Overlap induces a measure on the same probability space as the model, which represents how well an (unsupervised) estimator performs on the recovery task. To intuitively interpret the result, we put requirements on its asymptotic behavior, which takes place with high probability as $n \to \infty$.

DEFINITION 1.1.3. Let $G \sim \mathcal{G}(n, p, q)$. The following recovery requirements are solved if there exists an algorithm that takes G as an input and outputs $\hat{x} = \hat{x}(G)$ such that

- Exact recovery: $\mathrm{P}\{\mathrm{Overlap}(x, \hat{x}(G)) = 1\} = 1 - o(1)$
- Weak recovery: $\mathrm{P}\{\mathrm{Overlap}(x, \hat{x}(G)) \geq \Omega(1)\} = 1 - o(1)$

In other words, exact recovery requires the entire partition to be correctly identified. Weak recovery only asks for substantially better performance than chance. In some literature, exact recovery is simply called recovery. Weak recovery is also called detection since as long as one can weakly recover the ground truth, there must exist community structure.

Note that if G is an Erdős–Rényi random graph ($p = q$) then the overlap will be $o_p(1)$ for all estimators. This can be seen by noticing that x and G are independent in this setting and then applying Markov's inequality. This has led to two additional natural questions about SBMs. On the one hand, we are interested in the distinguishability (or testing): is there a hypothesis test to distinguish random graph generated by Erdős–Rényi model (ERM) from random graph generated by SBMs that succeeds with high probability? One the other hand, we can ask about the model learnability (or parameter estimation): assuming that G is drawn from an SBM ensemble, is it possible to obtain a consistent estimator for the parameters (p, q)? Although each of these questions is of independent interest, for symmetric SBMs with two symmetric communities (planted bisection model) the following holds [**Abb18**]:

$$\text{learnability} \iff \text{weak recovery} \iff \text{distinguishability} \tag{1.14}$$

The equivalence benefits the analysis of the model in turn. For example, direct analysis about weak recovery leads to the converse of phase transition theory [**MNS15**]. While

the achievability of the phase transition threshold [**Mas14a**] is proved by counting non-backtracking walks on the graph which gives consistent estimators of parameters. The recent work [**MS16**] on the analysis of SDP approaches the problem via the hypothesis testing formulation.

SBMs demonstrate the 'fundamental limits' of clustering and community detection as some necessary and sufficient conditions for feasibility of recovery, information-theoretically or computationally. Moreover, they are usually expressed in the form of *phase transition*. Sharp transitions exist in the parameter regimes between phases where the task is resolvable or not. For example, when the average degree grows as $\log n$, if the structure is sufficiently obvious then the underlying communities can be exactly recovered [**BC09**], and the threshold at which this becomes possible has also been determined [**ABH16**]. Above this threshold, efficient algorithms exist [**ABKK17**, **AS15**, **PW17**, **DLS21**] that recover the communities exactly, labeling every vertex correctly with high probability; below this threshold, exact recovery is information-theoretically impossible.

1.2. Sparse regime and Kesten-Stigum threshold

In the sparse case where the average degree of the graph is $O(1)$, it is more difficult to find the clusters and the best we can hope for is to label the vertices with nonzero correlation or mutual information with the ground truth, i.e. weak recovery. Intuitively, we only have access to a constant amount of connections about each vertex. The intrinsic difficulty can be understood from the topological properties of the graphs in this regime. The following basic results are derived from [**ER84**]:

- For $a, b > 0$, the planted bisection model $\mathcal{G}(n, \frac{a \log n}{n}, \frac{b \log n}{n})$ is connected with high probability if and only if $\frac{a+b}{2} > 1$.
- $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ has a giant component (i.e. a component of size linear in n) with high probability if and only if $d := \frac{a+b}{2} > 1$.

The graph will only have vanishing components if the average degree is too small. Therefore, it is not possible to even weakly recover the labels. But we will see in the next section that semi-supervised approaches amazingly piece the components together with consistent labeling.

Although it is mathematically challenging to work in the sparse regime, real-world data are likely to have bounded average degrees. [**LLDM08**] and [**Str01**] studied a large collection of the benchmark data sets, including power transmission networks, web link networks and complex biological systems, which had millions of nodes with average degree of no more than 20. For instance, the LinkedIn network they studied had approximately seven million nodes, but only 30 million edges.

The phase transition for weak recovery or detection in the sparse regime was first conjectured in the paper by Decelle, Krzakala, Moore, Zdeborová [**DKMZ11**], which sparked the modern study of clustering and SBMs. Their work is based on deep but non-rigorous insights from statistical physics, derived with the cavity method (a.k.a. belief propagation). Since then, extensive excellent research has been conducted to understand this fundamental limit, see e.g. [**MNS15**,**Mas14a**,**MNS18**,**ABRS20**]. A key result is the following theorem.

THEOREM 1.2.1. *[Kesten-Stigum threshold] Let $\mathcal{G}(n, a/n, b/n)$ be a symmetric SBM with two balanced clusters and $a, b = O(1)$. The weak recovery problem is solvable and efficiently so if and only if $(a-b)^2 > 2(a+b)$.*

In particular, if we denote the probability measures induced by the ERM $\mathcal{G}(n, \frac{a+b}{2n}, \frac{a+b}{2n})$ and the SBM $\mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$ by $P_n^{(0)}$ and $P_n^{(1)}$ correspondingly, they are mutually contiguous, that is for any sequence of events $\{E_n\}$'s, $P_n^{(0)}(E_n) \to 0$ if, and only if, $P_n^{(1)}(E_n) \to 0$.

Conventionally, the quantity $(a-b)^2/[2(a+b)]$ is called signal-to-noise ratio (SNR). It is worth noting that we only quoted the KS threshold for the two communities case ($k = 2$). For sufficiently large k, namely $k \geq 5$, there is a 'hard but detectable' area where the weak recovery is information-theoretically possible, but computationally hard [**AS18**,**BMNN16**].

This gap between the KS threshold and information-theoretic (IT) threshold only shows up in the constant degree regime, making it a fertile ground for studying the fundamental tradeoffs in community detection. We focus on the cardinal case, symmetric SBM with two balanced clusters, where two thresholds coincide and semi-supervised approach crosses them together.

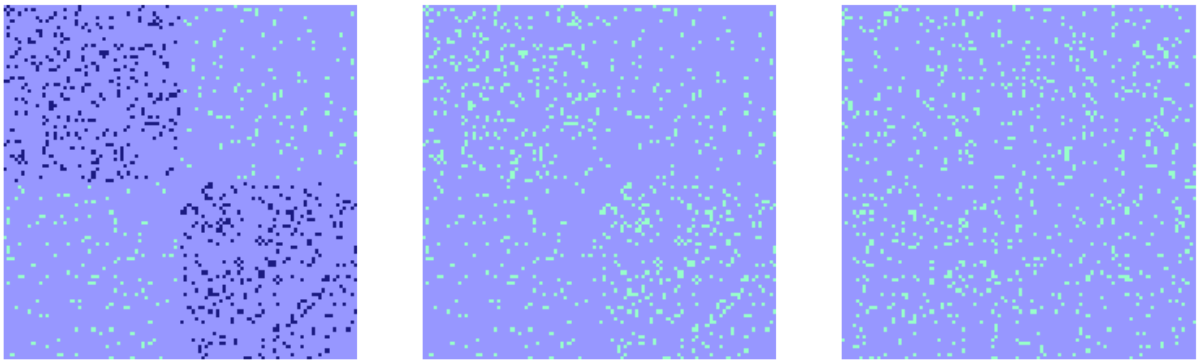

FIGURE 1.2. The left image represents the adjacency matrix of one realization of $\mathcal{G}(100, 0.12, 0.05)$, where the detection is theoretically possible. Yet the data is given non-colored (middle) and also non-ordered (right).

The terminology 'KS threshold' can be traced back to the work of Kesten and Stigum concerning reconstruction on infinite rooted trees in 1966 [**KS66**]. The problem consists in broadcasting the root label of a tree with fixed degree c down to its leaves, and trying to recover it from the leaves at large depth. We start with drawing the root label uniformly in $\{0, 1\}$. Then, in a top-down manner, we independently label every child the same as its parent with probability $1 - \epsilon$ and the opposite as its parent otherwise. Let $x^{(t)}$ denote the labels at depth t in this tree with $t = 0$ being the root. We say the reconstruction is solvable if $\lim_{t\to\infty} \mathrm{E}\,|\,\mathrm{E}(x^{(0)}|x^{(t)}) - 1/2| > 0$ or, equivalently, $\lim_{t\to\infty} I(x^{(0)}; x^{(t)}) > 0$, where I is the mutual information. Although it was shown in the original paper, reconstruction is solvable when $c(1 - 2\epsilon)^2 > 1$, the non-reconstruction was proved 30 years later, namely it is not solvable if $c(1 - 2\epsilon)^2 \leq 1$ [**BRZ95**, **EKPS00**]. Based on that finding, Mossel, Neeman, and Sly proved the converse part of Theorem 1.2.1 by coupling the local neighborhood of an SBM vertex with a Gaton-Watson tree with a Markov Process [**MNS15**]. Inspired by this

elegant approach, we propose our 'census method' to solve the semi-supervised clustering problem, and we will see in Chapter 3 how it works by amplifying revealed information with tree-like neighborhoods.

1.3. Basic algorithms

Information-theoretic bounds can provide the impossibility side of phase transitions, but we still need specific efficient algorithms for the achievability side. One straight forward approach is spectral method. Under the Definition 1.1.2, let A be the adjacency matrix of the graph $G \sim \mathcal{G}(n, a/n, b/n)$, $a > b$. Up to reordering indices, its expectation is a block matrix except for the diagonal,

$$\mathrm{E}\, A \approx \frac{1}{n} \begin{pmatrix} a & b \\ b & a \end{pmatrix} \otimes I_{n/2 \times n/2} \tag{1.15}$$

which has three eigenvalues, $(a+b)/n > (a-b)/n > 0$. 0 has multiplicity $n-2$ and the eigenvector associated with the second largest eigenvalue is $\begin{pmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{pmatrix}$ which is consistent with the ground truth of the labels. But we do not observe the expected adjacency matrix. Instead, we only have access to one realization of the model. In modern terms, community detection is a 'one-shot learning' task. But one can still hope that $A - \mathrm{E}\, A$ is small and the second eigenvector of A gives a reasonable estimator. For example, denoting the ordered eigenvalues of $\mathrm{E}\, A$ and A as $\{\lambda_i\}$'s and $\{\hat{\lambda}_i\}$'s respectively, the Courant-Fischer-Weyl min-max principle implies

$$|\hat{\lambda}_i - \lambda_i| \le \|A - \mathrm{E}\, A\|_{\mathrm{op}} \quad (\forall i \in [n]). \tag{1.16}$$

Recall that the operator norm of a symmetric matrix M, with $\xi_i(M)$ being its i-th largest eigenvalue, is $\|M\|_{\mathrm{op}} = \max(\xi_1(M), -\xi_n(M))$. If one can bound $\|A - \mathrm{E}\, A\|_{\mathrm{op}}$ by half of the least gap between the three eigenvalues mentioned above, the order will be preserved. Then the Davis-Kahan theorem guarantees the eigenvectors are correlated. Namely, if θ denotes

the angle between the second eigenvectors (spectral estimator and ground truth), we have

$$\sin\theta \le \|A - \mathrm{E}\,A\|_{\mathrm{op}} / \min\{\|\lambda_i - \lambda_2\|/2 : i \neq 2\}. \tag{1.17}$$

Thus, the key is to control the norm of the perturbation. Many deep results from random matrix theory come into play here [**Vu07**, **NN12**, **AFWZ20**].

This nice and simple approach stops working as we step into the sparse regime [**FO05**, **CO09**, **KMO09**, **DKMZ11**, **KMM+13**]. The main reason is that leading eigenvalues of A are about the order of square root of the maximum degree. High degree vertices mess up the desired order of eigenvalues. In particular, for Erdős–Rényi random graphs ($\mathcal{G}(n, d/n)$), we have $\hat{\lambda}_1 = (1+o(1))\sqrt{\log n/\log\log n}$ almost surely [**KS03**]. Furthermore, the leading eigenvectors are concentrated to these high degree 'outliers' and contain no structural information of the underlying model.

Take the star graph for example, where we assume that only the first node is connected to k neighbors. It is easy to see that the corresponding adjacency matrix has eigenvalue $\sqrt{k}$ and eigenvector $(\sqrt{k}, 1, \ldots, 1)$. Various interesting spectrum based methods are proposed to overcome this challenge [**MNS18**, **Mas14b**, **BLM15**]. The key idea is to replace adjacency with some combinatorically constructed matrices. However, they typically rely on model statistics and underlying probabilistic assumptions, which leads to the problem of adversarial robustness. For example, they are non-robust to 'helpful' perturbations. Namely, if we let an adversary to perform following changes on the graph: (1) adding edges within communities and/or (2) removing edges across communities, spectral approaches are going to fail. It is surprising since, intuitively, these changes help to emphasize community structures.

Meanwhile, semidefinite programming (SDP) sheds the light on how we may be able to overcome the limitations of spectral algorithms, which is shown to be robust when SNR is sufficiently large [**MPW16**]. In fact, it is another major line of work on clustering and community detection concerning performance of SDPs on SBMs. While a clear picture of

the unbounded degree case is figured out in [**ABH16**, **HWX16**, **AL18**, **Ban18**, **ABKK17**, **PW17**], the results for sparse networks are more complicated. [**GV14**] proved a sub-optimal condition, SNR $\geq 10^4$, using Grothendieck inequality. Then, with a Lindeberg interpolation process [**Tao11**], Montanari et al. proved that a SDP algorithm as proposed in [**MS16**] is nearly optimal for large bounded average degree by transferring the analysis of the original SDPs to analysis of SDPs of Gaussian random matrices.

THEOREM 1.3.1. *[**MS16**] Assume $G \sim G(n, a/n, b/n)$. If for some $\epsilon > 0$, SNR $\geq 1 + \epsilon$ and $d > d^*(\epsilon)$ then the SDP estimator solves the weak recovery.*

If we fix d and view ϵ as its function, the condition becomes SNR $\geq 1 + o_d(1)$. Numerical estimation and non-rigorous statistical mechanism approximation suggest that it is at most 2% sub-optimal. This result seems to be the ceiling of SDP according to the preliminary non-rigorous calculation in [**JMRT16**]. Moreover, they address the irregularity of high degree nodes by showing SDPs return similar results for Erdős–Rényi random graphs and random regular graphs, which appear to be sensitive only to the average degree. See Section 4 for more discussion on the estimation. Following their work, we propose a natural modification of SDP to incorporate revealed labels in the semi-supervised setting and show that it not only achieves, but also crosses, the KS threshold. In turn, our result brings a new perspective to study the (non-)achievability and robustness of (unsupervised) SDPs.

1.4. Semi-supervised learning

Within the field of machine learning, there are three basic approaches: supervised learning, unsupervised learning and the combination of both, semi-supervised learning. The main difference lies in the availability of labeled data. While unsupervised learning (e.g. clustering, association and dimension reduction) operates without any domain-specific guidance or preexisting knowledge, supervised learning (e.g. classification and regression) relies on all

training samples being associated with labels. However, it is often the case where existing knowledge for a problem domain doesn't fit either of these extremes.

In real-world applications, unlabeled data comes with a much lower cost not requiring expensive human annotation and laboratory experiments. For example, documents crawled from the Web, images obtained from surveillance cameras, and speech collected from broadcast are relatively more accessible comparing to their labels which are required for prediction tasks, such as sentiment orientation, intrusion detection, and phonetic transcript. Motivated by this labeling bottleneck, the semi-supervised approach utilizes both labeled and unlabeled data to perform learning tasks faster, better, and cheaper. Since the 1990s, semi-supervised learning research have enjoyed an explosion of interest with applications like natural language processing [**QCMC19**, **CX16**] and computer vision [**XHLL20**, **Lee13**].

This dissertation is closely related to the subtopic called constrained clustering, where one has some must-links (i.e. two nodes belong to the same cluster) and cannot-links (i.e. two nodes are in different clusters) as extra information. Although constrained versions of classic algorithms have been studied empirically, such as expectation–maximization [**SBHHW03**], k-means [**WCRS01**] and spectral method [**KKM03**], our results take different approaches to enhance clustering outcomes instead of hard-coding these pairwise constraints into the algorithms and provide theoretically guarantees under SBM.

A recent development in semi-supervised learning that has attracted extensive attention is called graph convolutional network (GCN) [**KW17**], which is based on an efficient variant of convolutional neural networks operating on graph structures directly. The objective functions of GCNs only involve labeled data while predictive information propagates through the graphs built in neural networks to cover unlabeled data. The benefit of integrating graph structures into deep learning approaches is twofold: (i) it efficiently embeds similarity between nodes to synchronize labeled and unlabeled samples; (ii) it significantly brings down the number of parameters by only considering the connections induced by underlying

graphs. A prototypical example of a forward model of a two-layer GCN for semi-supervised node classification on a graph is given by

$$f(Z, A) = \text{softmax}\left(\hat{A}\ \text{ReLU}\left(\hat{A}ZW^{(0)}\right)\ W^{(1)}\right) \tag{1.18}$$

where $W^{(i)},\ i \in \{0, 1\}$ are weight matrices for the hidden layer and the output layer, softmax and Relu are both vector activation functions defined as $\text{softmax}(x) = \exp(x)/\sum_i \exp(x_i)$ and $\text{ReLU}(x) = \max(0, x)$. Z stands for the features and A is the adjacency matrix. The output of each layer of GCN goes through a smoothing process defined by the *propagation model matrix* $\hat{A}$. It can be a normalized adjacency matrix, a graph Laplacian, or even the identity matrix, which reduces the model to multi-layer perception. Existing frameworks are either directly based on the adjacency matrix A [**YHC$^+$18**] or run basic clustering algorithms on A [**CLS$^+$19**] to design $\hat{A}$. But whenever GCN is applicable, some of the labels are always available. It is natural to consider making use of this label information to improve the decisive component $\hat{A}$, which can be realized directly from our semi-supervised clustering algorithms. We will discuss this interesting application further in in Chapter 5.

CHAPTER 2

Main Results

The main theoretical goal of this dissertation is to answer the long-standing open question regarding the semi-supervised learning on probabilistic graph models. We would like to quote the version from [**Abb18**]:

> "How do the fundamental limits change in a semi-supervised setting, i.e., when some of the vertex labels are revealed, exactly or probabilistically?"

2.1. Crossing the threshold

In the previous chapter, we have discussed deep research related to the clustering / community detection problem on SBMs. Establishing of the phase transition phenomenon at KS threshold is a major focal point. However, such a sharp and intrinsic limit totally disappears when an arbitrarily small fraction of the labels is revealed. This astonishing change is first observed in [**ZMZ14**] where the authors provide non-trivial conjectures based on calculations of the belief propagation approach.

The theory of semi-supervised clustering contains some fascinating and fundamental algorithmic challenges arising from both the sparse random graph model itself and the semi-supervised learning perspective. To address them rigorously, we first define the semi-supervised SBM in a way that it captures the essence of realistic semi-supervised learning scenarios and is a natural and simple generalization of unsupervised models.

Definition 2.1.1 (Semi-supervised planted bisection model). For $n \in \mathbb{N}$, $p, q \in (0,1)$ and $\rho \geq 0$, let $\mathcal{G}(n,p,q,\rho)$ denote the distribution over graphs with n vertices and n-dimensional vectors defined as follows. The vertex set is partitioned uniformly at random into two subsets

S_1, S_2 under the balance constraint $|S_1| = |S_2| = n/2$. Then conditional on the partition, two processes are undertaken independently:

- Let E denote edge set of the graph G. Edges are included independently with probability defined as follows.

$$\mathrm{P}\left((i,j) \in E | S_1, S_2\right) = \begin{cases} p & \text{if } \{i,j\} \subseteq S_1 \text{ or } \{i,j\} \subseteq S_2, \\ q & \text{if } i \in S_1, j \in S_2 \text{ or } i \in S_2, j \in S_1. \end{cases} \tag{2.1}$$

- An index set $\mathcal{R}$ of size $m \coloneqq 2\lfloor \rho \cdot \frac{n}{2} \rfloor$ is chosen uniformly at random such that $|\mathcal{R} \cap S_1| = |\mathcal{R} \cap S_2| = m/2$. The revealed labels are given as

$$\tilde{x}_i = \begin{cases} 1, & i \in \mathcal{R} \cap S_1, \\ -1, & i \in \mathcal{R} \cap S_2, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Remark 2.1.1. *The revealing process is independent from edge formation, i.e.* $G \perp \tilde{x} | S_1, S_2$. *Moreover, if we set* $\rho = 0$ *or simply ignore the revealed labels, the model is exactly the unsupervised SBM. In other words, the marginal distribution of the random graph is indeed* $\mathcal{G}(n,p,q)$ *from Definition 1.1.2.*

Remark 2.1.2. *One can also consider revealing uniformly at random over the index set independent of* $\mathcal{G}(n,p,q)$ *(instead of requiring revealed communities to have the same size), but this modification makes almost no difference in the context of this work. In practice, one can always achieve the balance requirement by either sampling a few more or dropping the uneven part.*

Remark 2.1.3. *The definition is versatile in the sense that it keeps unsupervised setting as a special case (and with it all the interesting phase transitions). On the other hand, it can be easily generalized to the multiple and/or asymmetric communities case.*

Under semi-supervised setting, the weak recovery problem is naturally defined as finding an estimator to perform substantially better than chance *on the unrevealed vertices*. And the distinguishability can be expressed as finding a test that with high probability, tells $\mathcal{G}(n, d/n, d/n, \rho)$ from $\mathcal{G}(n, a/n, b/n, \rho)$ where $d = \frac{a+b}{2}, a > b$. We will discuss these items in more detail when it comes to the corresponding section.

Based on the fact that a $\ln(n)$-neighborhood in $(G, x) \sim \mathcal{G}(n, a/n, b/n)$ asymptotically has the identical distribution as a Galton-Watson tree with Markov process, we propose our first semi-supervised clustering algorithm, called *census method*. Namely, we decide the label estimation of a certain node according to the majority of its revealed neighbors,

$$\hat{x}_v = \operatorname{sgn}\left(\sum_{i \in \{u \in \mathcal{R}:\ d(u,v)=t\}} x_i\right) \tag{2.3}$$

where $d(u, v)$ is the length of the shortest path connecting u and v. We conclude that when $\mathrm{SNR} \leq 1$, the optimal choice of t is indeed 1.

THEOREM 2.1.1. *The 1-neighbors census method solves the semi-supervised weak recovery problem with any reveal ratio $\rho > 0$ for arbitrary SNR > 0.*

Although this successfully solves the weak recovery problem, there are some limitations hindering the census method's utility in practice. Its performance depends on a sufficient amount of revealed labels, hence requiring n to be really large. Besides, without an unsupervised counterpart, it is not applicable when the revealing is not reliable.

To address these challenges, we propose our second semi-supervised clustering algorithm which performs well in practice and covers the unsupervised setting as a special case. As discussed in the previous chapter, SDPs enjoy many nice properties, among which the monotone-robustness is particularly interesting to us. In the semi-supervised setting, the revealed labels are supposed to enhance the community structure. However, the work from [**MPW16**] suggests such enhancement may not help with, but to the contrary can hurt

the performance of many other algorithms, which makes SDP an ideal starting point for us. We define the *Constrained Semidefinite Program* (CSDP) as

$$\text{CSDP}(M) = \max_{\substack{X \succeq 0 \\ X_{ii}=1,\ \forall i \in [n]}} \{\langle M, X\rangle :\ X_{ij} = x_i \cdot x_j,\ \forall i, j \in \mathcal{R}\} \tag{2.4}$$

and show that it solves the semi-supervised community detection problem.

THEOREM 2.1.2. *Let* $(G, x_{\mathcal{R}}) \sim \mathcal{G}(n, a/n, b/n, \rho)$ *and* A *be the adjacency matrix of* G. *For any* $a > b$, *there exists* $\rho_0 < 1$ *such that if* $\rho \geq \rho_0$, *the CSDP-based test* $T(G, x_{\mathcal{R}}; \Delta) = \mathbb{1}_{\{CSDP(A - \frac{d}{n}\mathbf{1}\mathbf{1}^\top) \geq n[(a-b)/2 - \Delta]\}}$ *will succeed with high probability for some* $\Delta > 0$.

2.2. Proof techniques

The technical challenges of establishing Theorem 2.1.1 are mainly due to that the advantage created by revealed labels can be easily blurred out by various approximations of the limit distribution. Instead of the central limit theorem, one needs a Berry–Esseen-type inequality to derive a more quantitative result of the convergence rate. Moreover, since the distribution of each underlying component also depends on n, the conventional sample mean formulation does not apply here. We overcome the difficulty above with a direct analysis of non-asymptotic distributions, which leads to a detailed comparison between two binomial variables with constant expectations.

It is quite surprising that this calculation can be carried out in rather elegant manner, since many other applications of this method are much more technically involved. For example to establish the independence among estimators, one may need to consider the 'leave-one-out' trick. But in our case, it comes in a very natural way.

Regarding CSDP, we first show it can be coupled to a SDP with the surrogate input matrices. Moreover, its optimal value lies in between two unsupervised SDPs associated with the same random graph model (different parameters). This means all the analytical results from SDP research can be transferred into the CSDP study. However, we notice that

it is common to make assumptions on the average degree d in the relevant literature. It is quite reasonable in the unsupervised setting since the graph topology is a strong indicator for the possibility of weak recovery, e.g. when $d \leq 1$, there will not exist a giant component that is of size linear in n.

To establish our result without such extra assumptions, we derive a probabilistic bound on the cut norm of centered adjacency matrix and then use Grothendieck's inequality to bound the SDPs on ERMs from above. This idea follows from [**GV14**], we give a slightly different analysis to accommodate our usage. A generalized weak law of large number is also derived to address the issue that distributions of the entries change as $n \to \infty$. Then we conclude the proof with a lower bound of CSDPs on SBMs considering a witness consists of the ground truth of labels.

2.3. Outline

The rest of the dissertation is organized in the following way. In Chapter 3, we formally derive the census method and prove that it can solve the weak recovery problem throughout the entire parameter domain. In Chapter 4, we introduce the constrained SDP and the associated hypothesis test, through which we show that even under the KS threshold (also the information-theoretic threshold), the ERMs and the SBMs become distinguishable in the semi-supervised setting. In Chapter 5, we discuss the application to GCN. We end the dissertation with concluding remarks in Chapter 6.

2.4. Notation

For any $n \in \mathbb{N}$, we denote the first n integers by $[n] = \{1, 2, \ldots, n\}$. For a set S, its cardinality is denoted by $|S|$. We use lowercase letters for vectors (e.g. $v = (v_1, v_2, \ldots, v_n)$) and uppercase letters for matrices (e.g. $M = [M_{ij}]_{i,j\in[n]}$). In particular, for adjacency matrices, we omit their dependency on underlying graphs. Instead of A_G, we simply write A. $\mathbf{1}_n = (1, \ldots, 1) \in \mathbb{R}^n$ stands for the all-ones vector and I_n is the $n \times n$ identity matrix. $\mathbf{e}_i \in \mathbb{R}_n$

represents the i's standard basis vector. For two real-valued matrices A and B with the same dimensions, we define the Frobenius inner product as $\langle A, B\rangle = \sum_{i,j} A_{ij} \cdot B_{ij} = \mathrm{Tr}(A^\top B)$. Vector inner product is viewed as a special case of $n \times 1$ matrices. Let $\|v\|_p = (\sum_{i=1}^{p} \|v_i\|^p)^{1/p}$ be the ℓ_p norm of vectors with standard extension to $p = \infty$. Let $\|M\|_{p\to q} = \sup_{\|v\|_p \le 1} \|Mv\|_q$ be the ℓ_p-to-ℓ_q operator norm and $\|M\|_{\mathrm{op}} \coloneqq \|M\|_2 \coloneqq \|M\|_{2\to 2}$. Random graphs induce measures on the product space of label, edge and revealed node assignments over n vertices. For any $n \in \mathbb{N}$, it is implicitly understood that one such measure is specified with that graph size. The terminology *with high probability* means 'with probability converging to 1 as $n \to \infty$'. Also, we follow the conventional Big-Oh notation for asymptotic analysis. $o_p(1)$ stands for convergence to 0 in probability.

CHAPTER 3

Census Method

Analysis of Model 1.1.2 is a challenging task since conditioned on the graph, it is neither an Ising model, nor a Markov random field. This is mainly due to following facts: (1) The balance requirement puts a global condition on the size of each cluster; (2) Even if conditioned on sizes, there is a slight repulsion between unconnected nodes. Namely, if two nodes do not form an edge, the probability of them being in the same community is different from them being in the opposite communities.

Recent years have witnessed a series of excellent contributions on the phase transitions in the sparse regime. Our census method for semi-supervised clustering is mainly inspired by the natural connection between community detection on SBMs and reconstruction on trees, which is formally established in [**MNS15**]. Intuitively, for a vertex v in $\mathcal{G}(n, a/n, b/n)$, it is not likely that a node from its small neighborhood has an edge leading back to v. Therefore, the neighborhood looks like a random labelled tree with high probability. Furthermore, the labelling on the vertices behaves like the broadcasting a bit from the root of a tree down to its leaves (see the survey [**Mos01**] for a detailed discussion).

In this chapter, we will first look into the census method of t-neighbors, i.e., deciding the label of a node by the majority on its neighbors at depth t. We shall see that when SNR $\leq$ 1, 1-neighbors voting is optimal in terms of recovering the cluster structure via informal calculation. Then we rigorously prove that census on 1-neighbors solves the semi-supervised weak recovery for any SNR > 0 with an arbitrarily small fraction of labels revealed.

3.1. Majority of t-neighbors

Let (G, x) obey the planted bisection model $\mathcal{G}(n, a/n, b/n)$. We denote the set of all vertices by $V(G)$. For a fixed vertex v and $t \in \mathbb{N}$, let $N_t(v)$ denote the number of vertices which are t edges away from v. $\Delta_t(v)$ is defined as the difference between the numbers of t-neighbors in each community. Namely,

$$N_t(v) = |K_t(v)| \tag{3.1}$$

$$\Delta_t(v) = \sum_{u \in K_t(v)} x_u \tag{3.2}$$

where $K_t(v) \coloneqq \{u \in V(G) : \ d(u, v) = t\}$ denotes the t-neighbors of v.

If one assume that the subgraph of G induced by the vertices within t edges of v is a tree, the expected value of $N_t(v)$ is approximately $[(a + b)/2]^t$ and the expected value of $x_v \cdot \Delta_t(v)$, i.e., the expected number of these vertices in the same community as v minus the expected number of these vertices in the other community, is approximately $[(a - b)/2]^t$. So, if one can somehow independently determine which community a vertex is in with an accuracy of $1/2 + \alpha$ for some $\alpha > 0$, one will be able to predict the label of each vertex with an accuracy of roughly $1/2 + [(a - b)^2/(2(a + b))]^{t/2} \cdot \alpha$, by guessing it as the majority of v's t-neighbors. Under the unsupervised learning setting, one can get a small advantage, $\alpha \sim \Theta(1/\sqrt{n})$, by randomly initializing labels. It is guaranteed by the central limit theorem that such a fraction exists in either an agreement or disagreement form.

To amplify this lucky guess, we need t to be sufficiently large so that $[(a - b)^2/(2(a + b))]^{t/2} > \sqrt{n}$, which implies $[(a + b)/2]^t > n$. Note $d = (a + b)/2$ is the average degree. This means before the signal is strong enough for our purpose, not only our tree approximation will break down, but vertices will be exhausted. However, if we have access to some of the true labels, i.e., in the semi-supervised setting, we can leverage the tree structure for a non-vanishing advantage over random guessing.

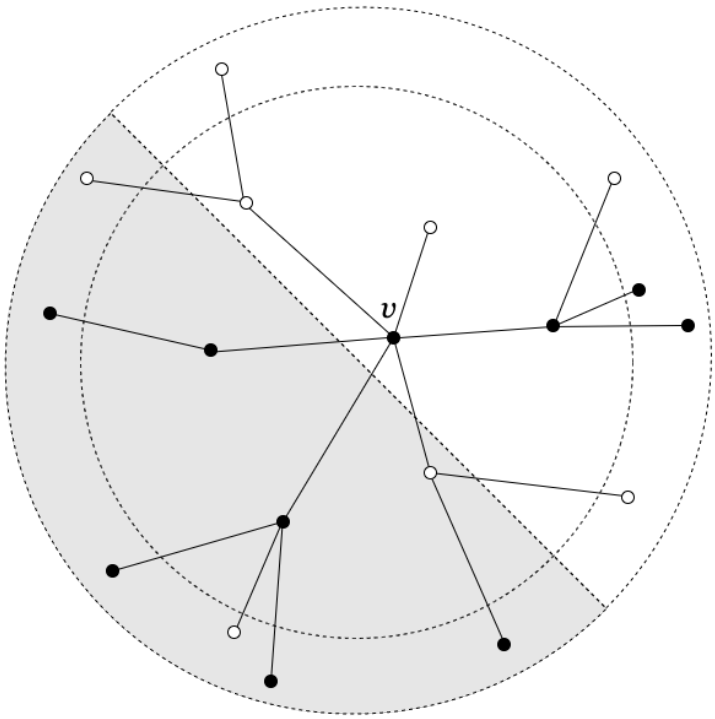

FIGURE 3.1. Neighborhood of node v with a tree structure. True clusters are coded in black and white. The shaded area indicates those nodes randomly guessed to be in the same community or the opposite community as v. The annulus represents the collection of its t-neighbors.

Let A be the adjacency matrix associated with G. Consider the random variables Y_u representing votes of directly connected neighbors,

$$Y_u = \begin{cases} x_u & \text{if } A_{uv} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

We have

$$N_1(v) = \sum_{u \in V(G)} |Y_u| \tag{3.4}$$

$$\Delta_1(v) = \sum_{u \in V(G)} Y_u \tag{3.5}$$

By definition of the planted bisection model,

$$\mathrm{P}(Y_u = 1 | x_v = 1) = \frac{\mathrm{P}(Y_u = 1, x_v = 1)}{\mathrm{P}(x_v = 1)} \approx \frac{a}{2n} \tag{3.6}$$

Similarly,

$$\mathrm{P}(Y_u = -1 | x_v = 1) \approx \frac{b}{2n} \tag{3.7}$$

It is not exact due to the balanced community constraint. But when n is large, such an effect is negligible. Furthermore, if we consider definition of the planted bisection model without balance constraint, the equation will be exact.

Without loss of generality, we only consider the case where $x_v = 1$ and omit the condition on it. We have

$$\Delta_1(v) = \sum_{u \in V(G)} Y_u \qquad \text{with} \quad Y_u = \begin{cases} 1 & \text{w.p. } \frac{a}{2n} \\ -1 & \text{w.p. } \frac{b}{2n} \\ 0 & \text{w.p. } 1 - \frac{a+b}{2n} \end{cases} \tag{3.8}$$

where the Y_u's are independent. Note that $\mathrm{E}(Y_u) = \frac{a-b}{2n}$ and $\mathrm{E}(Y_u^2) = \frac{a+b}{2n}$.

Recall that $\rho \in [0, 1]$ is the ratio of revealed labels. For the sake of simplicity, we assume the total number of revealed vertices $m = \rho n \in 2\mathbb{N}$ to be an even integer. The revealed vertices are chosen arbitrarily, denoted as $\mathcal{R} \coloneqq \{u_1, u_2, \ldots, u_{n-m}\}$. The model also provides

that the number of revealed vertices in each community is $\frac{\rho n}{2}$. Then the majority of revealed vertices among 1-neighborhood of v can be written as

$$\tilde{\Delta}_1(v) = \sum_{u \in \mathcal{R}} Y_u \tag{3.9}$$

Therefore,

$$\mathrm{E}(\tilde{\Delta}_1(v)) = \sum_{u \in \mathcal{R}} \mathrm{E}(Y_u) = \rho \frac{a-b}{2} \tag{3.10}$$

$$\mathrm{Var}(\tilde{\Delta}_1(v)) = \sum_{u \in \mathcal{R}} \mathrm{Var}(Y_u) = \rho \frac{a+b}{2} + o(1) \tag{3.11}$$

3.2. Locally tree-like structure

Proceeding to the t-neighbors, we need to understand a bit better the structure of a small neighborhood in the SBM. The neighborhoods in a sparse network locally have no loops. So they have a nice tree-like structure. Moreover, the labels also obey some random broadcasting processes on trees.

A broadcasting process transmit the information from the root of a tree to all the nodes. At each level, nodes inherit the information from its parent. But error could happen with certain amount of probability. Usually the edges are assumed to be included according to the same rule and work independently. It was firstly considered in genetics [**Cav78**] since it perfectly describes the propagation of a gene from ancestor to descendants. It can also be interpreted as a communication network that pass out the information from the root. So such processes were intensively studied in information theory and statistical physics [**Spi75**, **Hig77**, **BRZ95**]. In particular, we are interested in the following Markov process since it can be identified with the labeling process of a small neighborhood in SBM.

Definition 3.2.1 (Galton–Watson tree with Markov process). Let T be an infinite rooted tree with root v. Given a number $0 \leq \epsilon < 1$ and the offspring rate $d > 0$, we define a random

labelling $\tau \in \{1, -1\}^T$. First, draw τ_v uniformly in $\{1, -1\}$. Then recursively construct the labelling as follows.

- Generate children of each parent node according to the Poisson distribution with the expectation of d.
- Conditionally independently given τ_v, for every child u of v, set $\tau_u = \tau_v$ with probability $1 - \epsilon$ and $\tau_u = -\tau_v$ otherwise.

The following lemma shows that a $\ln(n)$-neighborhood in (G, x) looks like a Galton-Watson tree with Markov process. For any $v \in G$, let G_R be the induced subgraph on $\{u \in G : d(u, v) \leq R\}$.

LEMMA 3.2.1. *[**MNS15**] Let $R = R(n) = \frac{\ln n}{10 \ln(2(a+b))}$. There exists a coupling between (G, x) and (T, τ) such that $(G_R, x_{G_R}) = (T_R, \tau_{T_R})$ a.a.s.*

Hence, for fixed $t \in \mathbb{N}$, $t \leq R$ and any $v \notin \mathcal{R}$, we can denote the label of a vertex in v's t-neighborhood as $Y_i^{(t)} := \Pi_{k=1}^{t} {}^kY_u$, where $\{{}^kY_u\}_{k=1}^{t}$ are independent copies of Y_u. Then we have $\mathrm{E}(Y_i^{(t)}) = (\frac{a-b}{2n})^t$ and $\mathrm{E}((Y_i^{(t)})^2) = (\frac{a+b}{2n})^t$. Moreover, $\{Y_i^{(t)}\}$'s are independent. Therefore, the census of v's revealed t-neighbors can be written as

$$\tilde{\Delta}_t(v) = \sum_{i \in [\rho \cdot n^t]} Y_i^{(t)} \quad a.a.s \tag{3.12}$$

The central limit theorem suggests

$$\tilde{\Delta}_t(v) \to \mathcal{N}(\rho(\frac{a-b}{2})^t, \rho(\frac{a+b}{2})^t) \qquad \text{as } n \to \infty \tag{3.13}$$

Hence,

$$\mathrm{P}(\tilde{\Delta}_t(v) > 0 | x_v = 1) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{\rho[(a-b)/2]^t}{\sqrt{\rho[(a+b)/2]^t}\sqrt{2}}\right)\right] + o(1) \tag{3.14}$$

$$= \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\sqrt{\frac{\rho\ \mathrm{SNR}^t}{2}}\right) + o(1) \tag{3.15}$$

where $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x \exp(-t^2)\,\mathrm{d}t$ is the Gauss error function.

So one can see that once SNR is less than or equal to 1, it is not beneficial to look into t-neighbors. The optimal choice of t is 1 in this situation. Since we also know that weak recovery is solvable when SNR > 1, it makes the majority of 1-neighbors particularly interesting.

Suppose SNR ≤ 1 and include the symmetric part of $x_v = -1$, we have

$$\mathrm{P}(\operatorname{sgn}(\tilde{\Delta}_1(v)) = x_v) > \frac{1}{2} + \frac{1}{3}\sqrt{\rho\ \mathrm{SNR}} \tag{3.16}$$

Consider the estimator of unrevealed labels

$$\hat{x}_{\mathcal{R}^\complement} \coloneqq \operatorname{sgn}\left([\tilde{\Delta}_1(u_1), \tilde{\Delta}_1(u_2), \ldots, \tilde{\Delta}_1(u_{n-m})]^\top\right) \tag{3.17}$$

and the ground truth $x_{\mathcal{R}^\complement} = [x_{u_1}, x_{u_2}, \ldots, x_{u_{n-m}}]^\top$. Recall that

$$\mathrm{Overlap}(x_{\mathcal{R}^\complement}, \hat{x}_{\mathcal{R}^\complement}) = \frac{1}{n-m}|\langle x_{\mathcal{R}^\complement}, \hat{x}_{\mathcal{R}^\complement}\rangle| \tag{3.18}$$

We can conclude that

$$\mathrm{E}[\mathrm{Overlap}(x_{\mathcal{R}^\complement}, \hat{x}_{\mathcal{R}^\complement})] = \mathrm{E}\left[\frac{1}{n-m}\left|\sum_{i\in[n-m]} \mathrm{sgn}(\tilde{\Delta}_1(u_i))x_{u_i}\right|\right] \tag{3.19}$$

$$\geq \frac{1}{n-m}\left|\sum_{i\in[n-m]} \mathrm{E}\left[\mathrm{sgn}(\tilde{\Delta}_1(u_i))x_{u_i}\right]\right| \tag{3.20}$$

$$> \frac{2}{3}\sqrt{\rho\ \mathrm{SNR}} \tag{3.21}$$

The expected overlap is not vanishing which suggests the weak recovery is solvable for any SNR. But it is technically impractical to rigorously describe the limit distribution of our census estimator without blurring this edge out. From Figure 3.2, we can see that our calculation is close to the expectation. But the convergence rate depends on ρ. In particular, when both SNR and ρ are small, the asymptotic behavior of our algorithm remains unclear. Hence we go through a direct analysis to establish the desired result.

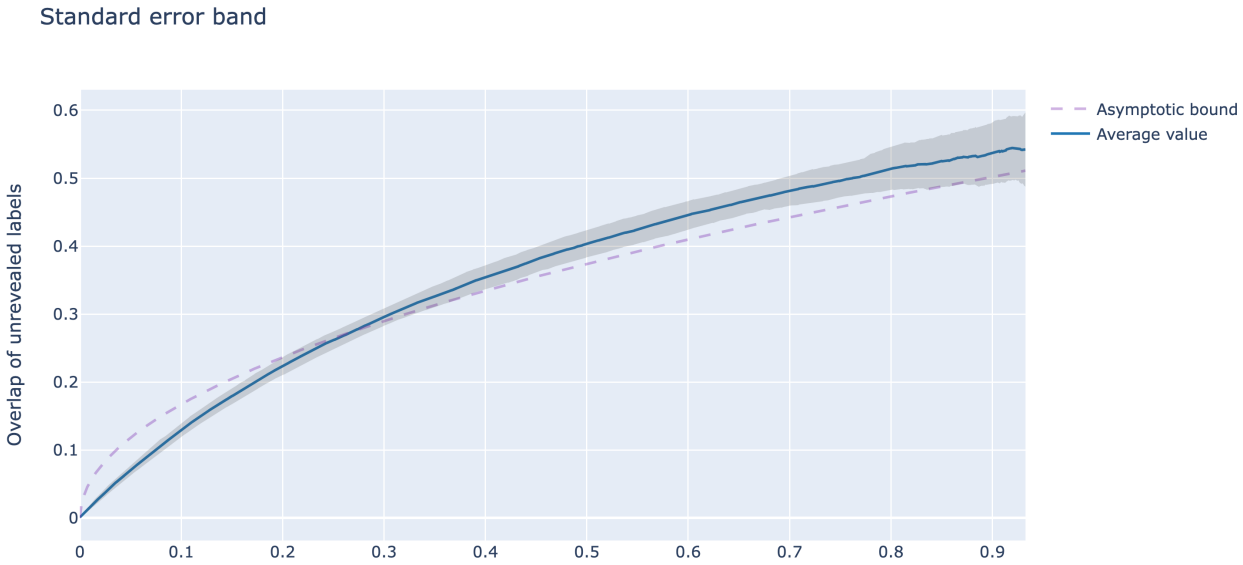

FIGURE 3.2. The simulation result of $\mathcal{G}(3000, 5/3000, 2/3000)$, SNR ≈ 0.64. Horizontal coordinate is the ratio of revealed labels. The blue curve stands for the average overlap of 60 independent realizations of the random graph with the shaded area being its standard error band. The purple dashed curve stands for the asymptotic lower bound we conclude from our calculation.

3.3. Majority of 1-neighbors

Since the algorithm is invariant under index reordering, without loss of generality, we let the adjacency matrix A be a symmetric matrix with diagonal entries $A_{ii} = 0,\ i = 1, 2, \ldots, n$. For $1 \le i < j \le n$, $\{A_{ij}\}$'s are independent,

$$A_{ij} \sim \text{Bernoulli}\left(\frac{a}{n}\right) \qquad \left(i \le \frac{n}{2} \text{ and } j \le \frac{n}{2}\right) \text{ or } \left(i \ge \frac{n}{2} \text{ and } j \ge \frac{n}{2}\right) \tag{3.22}$$

$$A_{ij} \sim \text{Bernoulli}\left(\frac{b}{n}\right) \qquad i \ge \frac{n}{2} \text{ and } j \le \frac{n}{2} \tag{3.23}$$

The true label x and revealed label $\tilde{x}$ are, respectively,

$$x_i = \begin{cases} 1, & i = 1, 2, \ldots, \frac{n}{2}, \\ -1, & i = \frac{n}{2}, \frac{n}{2}+1, \ldots, n, \end{cases} \qquad \tilde{x}_i = \begin{cases} 1, & i = 1, 2, \ldots, \frac{m}{2}, \\ -1, & i = \frac{n}{2}, \frac{n}{2}+1, \ldots, \frac{n+m}{2}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.24}$$

For a unrevealed vertex, we consider the majority of its 1-neighbors,

$$\tilde{\Delta}_1(i) = \langle A[i,:], \tilde{x} \rangle = \sum_{j:\tilde{x}_j \neq 0} A_{ij}\tilde{x}_j = \sum_{j:\tilde{x}_j \neq 0} A_{ji}\tilde{x}_j \tag{3.25}$$

Therefore, $\{\tilde{\Delta}_1(i)\}$'s are independent for all $i : \tilde{x}_i = 0$ (the unrevealed nodes) since they have no common term. Notice it is not the case for all $i \in [n]$. But we only need to predict the unrevealed labels, hence the independence is sufficient. To be more specific, if $\tilde{x}_k = 0$, then none of $\{A_{ik}\}_{i=1}^n$ will be involved in the computation of $\tilde{\Delta}_1(i)$ for any i. It brings remarkable convenience to our analysis.

The estimator given by majority voting of 1-neighbors is

$$\hat{x}_i = \begin{cases} \tilde{x}_i & \text{if } \tilde{x}_i \neq 0 \\ \operatorname{sgn}^*(\tilde{\Delta}_1(i)) & \text{if } \tilde{x}_i = 0 \end{cases} \tag{3.26}$$

We toss a fair coin when $\tilde{\Delta}_1(i) = 0$ to break the tie, i.e.

$$\mathrm{P}(\operatorname{sgn}^*(\tilde{\Delta}_1(i)) = 1 | \tilde{\Delta}_1(i) = 0) = \mathrm{P}(\operatorname{sgn}^*(\tilde{\Delta}_1(i)) = -1 | \tilde{\Delta}_1(i) = 0) = \frac{1}{2} \tag{3.27}$$

Notice that it is only introduced for analysis purpose and is equivalent to the conventional sign function in practice.

3.4. Proof of weak recovery

Now we are ready to show that our semi-supervised clustering algorithm solves the community detection problem for arbitrary SNR. Suppose (G, x) is an Erdős–Rényi random graph with revealed label $\tilde{x}$, any estimator can only have vanishing correlation with the true label among the unrevealed vertices. So the semi-supervised weak recovery problem on SBM requires finding an estimator such that the correlation restricted on the unrevealed part is non-vanishing. Formally, we want to show that

$$\mathrm{P}\left(\operatorname{Overlap}(x|_{\tilde{x}_i=0}, \hat{x}|_{\tilde{x}_i=0}) \geq \Omega(1)\right) = 1 - o(1) \tag{3.28}$$

First, we prove an elementary but important lemma. It shows the difference of two Binomial distributions with the same limiting success probability does not vanish.

LEMMA 3.4.1. *Let X and Y be two independent binomial random variables with $X \sim$ Binomial$(n, \frac{a}{n})$ and $Y \sim$ Binomial$(n, \frac{b}{n})$, $a > b$. Denote $\delta = \delta(a, b) \coloneqq \frac{a-b}{2\exp(a+b)}$. Then, for*

sufficiently large n,

$$\mathrm{P}(X > Y) - \mathrm{P}(X < Y) \geq \delta \tag{3.29}$$

REMARK 3.4.1. *By symmetry, we always have* $\mathrm{P}(X > Y) - \mathrm{P}(X < Y) > 0$. *This lemma guarantees the difference will not vanish as* $n \to \infty$.

PROOF. By the law of total probability and independence, we have

$$\mathrm{P}(X > Y) = \sum_{x=1}^{n} \mathrm{P}(Y < x)\,\mathrm{P}(X = x) \tag{3.30}$$

$$= \sum_{x=1}^{n} \sum_{y=0}^{x-1} \mathrm{P}(Y = y)\,\mathrm{P}(X = x) \tag{3.31}$$

$$= \sum_{x=1}^{n} \sum_{y=0}^{x-1} \left[\binom{n}{x} \left(\frac{a}{n}\right)^{x} \left(1 - \frac{a}{n}\right)^{n-x} \binom{n}{y} \left(\frac{b}{n}\right)^{y} \left(1 - \frac{b}{n}\right)^{n-y} \right] \tag{3.32}$$

Let $\Delta \coloneqq \mathrm{P}(X > Y) - \mathrm{P}(X < Y)$, then

$$\Delta = \sum_{x=1}^{n} \binom{n}{x} \left(\frac{a}{n}\right)^{x} \left(1 - \frac{a}{n}\right)^{n-x} \left(\frac{b}{n}\right)^{x} \left(1 - \frac{b}{n}\right)^{n-x} \left\{ \sum_{y=0}^{x-1} \binom{n}{y} \left[\left(\frac{b}{n}\right)^{y-x} \left(1 - \frac{b}{n}\right)^{x-y} - \left(\frac{a}{n}\right)^{y-x} \left(1 - \frac{a}{n}\right)^{x-y} \right] \right\}$$

$$= \sum_{x=1}^{n} \binom{n}{x} \left(\frac{ab}{n}\right)^{x} \left(1 - \frac{a+b}{n} + \frac{ab}{n^2}\right)^{n-x} \left\{ \sum_{y=0}^{x-1} \binom{n}{y} \frac{1}{n^y} \left[\left(\frac{1}{b} - \frac{1}{n}\right)^{x-y} - \left(\frac{1}{a} - \frac{1}{n}\right)^{x-y} \right] \right\}$$

Let $f(x) = \alpha^x - \beta^x,\ \alpha > \beta > 0$. Since $f'(x) = \alpha^x \ln \alpha - \beta^x \ln \beta > 0$, we have $f(m) \geq f(1) = \alpha - \beta,\ \forall m \in \mathbb{N}$. So $\left(\frac{1}{b} - \frac{1}{n}\right)^{x-y} - \left(\frac{1}{a} - \frac{1}{n}\right)^{x-y} \geq \frac{a-b}{ab}$.

Also notice that $\binom{n}{m} = \prod_{i=0}^{m-1} \frac{n-i}{m-i} \geq \left(\frac{n}{m}\right)^m$, $\forall 1 \leq m \leq n$. We have,

$$\Delta \geq \sum_{x=1}^{n} \left(\frac{ab}{x}\right)^x \left(1 - \frac{a+b}{n}\right)^{n-x} \left(\sum_{y=0}^{x-1} \frac{1}{y^y} \cdot \frac{a-b}{ab}\right) \tag{3.33}$$

$$\geq (a-b)\left(1 - \frac{a+b}{n}\right)^n \tag{3.34}$$

$$\geq \frac{a-b}{2\exp(a+b)} \qquad \text{(for sufficiently large } n\text{)} \tag{3.35}$$

where we follow the convention that $0^0 = 1$. □

We resort to a classical concentration inequality to bound the overlap.

LEMMA 3.4.2 (Chernoff–Hoeffding theorem [**Che52**]). *Suppose $X_1, \ldots, X_n$ are i.i.d. random variables, taking values in $\{0, 1\}$. Let $p = \mathrm{E}(X)$ and $\epsilon > 0$. Then*

$$\mathrm{P}\left(\frac{1}{n}\sum X_i \leq p - \epsilon\right) \leq \left(\left(\frac{p}{p-\epsilon}\right)^{p-\epsilon}\left(\frac{1-p}{1-p+\epsilon}\right)^{1-p+\epsilon}\right)^n = \mathrm{e}^{-D(p-\epsilon\|p)n} \tag{3.36}$$

where $D(x\|y) = x\ln\frac{x}{y} + (1-x)\ln(\frac{1-x}{1-y})$ is the Kullback–Leibler-divergence between Bernoulli distributed random variables with parameters x and y.

We now convert the KL divergence to the total variation distance, which is easier to work with. Let P_1 and P_2 be two probability measures defined on the same sample space Ω and sigma-algebra $\mathcal{F}$. The total variation distance between them is defined as $d_{TV}(P_1, P_2) = \sup_{E\in\mathcal{F}} |P_1(E) - P_2(E)|$. Moreover, in the discrete case, we have following identity $d_{TV}(P_1, P_2) = \frac{1}{2}\|P_1 - P_2\|_1 = \sum_{\omega\in\Omega} \frac{1}{2}\|P_1(\omega) - P_2(\omega)\|$. It is related to the KL divergence through Pinsker's inequality (see, eg. [**Tsy09**], Chapter 3). For completeness, we include an elementary proof of the Bernoulli special case.

LEMMA 3.4.3. *Let P_1 and P_2 be two Bernoulli distribution, where $P_1(1) = x$ and $P_2(1) = y$. We have*

$$2(d_{TV}(P_1, P_2))^2 \leq D(x\|y) \tag{3.37}$$

PROOF. We can manipulate both sides of the inequality as

$$D(x\|y) = x \ln \frac{x}{y} + (1-x) \ln(\frac{1-x}{1-y}) \tag{3.38}$$

$$2(d_{TV}(P_1, P_2))^2 = \frac{1}{2}\|P_1 - P_2\|_1^2 = 2(x-y)^2 \tag{3.39}$$

Then we denote

$$f(x, y) = x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y} - 2(x-y)^2. \tag{3.40}$$

Therefore,

$$\frac{\partial f}{\partial y} = -\frac{x}{y} + (1-x)\frac{1}{1-y} + 4(x-y) \tag{3.41}$$

$$= \frac{-x+y}{y(1-y)} + 4(x-y) \tag{3.42}$$

$$= (x-y)(4 - \frac{1}{y(1-y)}) \tag{3.43}$$

Notice that since $0 \leq y \leq 1$, we have $y(1-y) \leq \frac{1}{4}$. So $4 - \frac{1}{y(1-y)}$ is always negative.

Thus, for fixed x, $f(x, y) \geq f(x, x) = 0, \ \forall y$. Hence,

$$D(x\|y) - 2(d_{TV}(P_1, P_2))^2 \geq 0 \tag{3.44}$$

□

Now we prove the main result for the census method.

PROOF OF THEOREM 2.1.1. Recall that for any i such that $\tilde{x}_i = 0$, our estimator is defined as $\hat{x}_i = \mathrm{sgn}^*(\tilde{\Delta}_1(i))$ and

$$\tilde{\Delta}_1(i) = \sum_{j:\tilde{x}_j \neq 0} A_{ij}\tilde{x}_j = \left(\sum_{\rho\frac{n}{2} < j \leq \frac{n}{2}} A_{ij} \right) - \left(\sum_{(1+\rho)\frac{n}{2} < j \leq n} A_{ij} \right) \tag{3.45}$$

It is indeed the difference between two independent binomial variables with parameters $(\rho n, \frac{\rho a}{\rho n})$ and $(\rho n, \frac{\rho b}{\rho n})$. By Lemma 3.4.1, we have

$$\mathrm{P}(\mathrm{sgn}(\tilde{\Delta}_1(i)) = x_i) - \mathrm{P}(\mathrm{sgn}(\tilde{\Delta}_1(i)) = -x_i) \geq \delta = \frac{\rho(a-b)}{2\mathrm{e}^{\rho(a+b)}} \tag{3.46}$$

for sufficiently large n. Also notice that

$$\mathrm{P}(\mathrm{sgn}(\tilde{\Delta}_1(i)) = -x_i) = 1 - \mathrm{P}(\mathrm{sgn}(\tilde{\Delta}_1(i)) = x_i) - \mathrm{P}(\tilde{\Delta}_1(i) = 0) \tag{3.47}$$

Therefore,

$$\mathrm{P}(\mathrm{sgn}(\tilde{\Delta}_1(i)) = x_i) \geq \frac{1+\delta}{2} - \frac{1}{2}\mathrm{P}(\tilde{\Delta}_1(i) = 0) \tag{3.48}$$

Then, by the law of total probability, we have

$$\mathrm{P}(\hat{x}_i = x_i) = \mathrm{P}(\mathrm{sgn}^*(\tilde{\Delta}_1(i)) = x_i) \tag{3.49}$$

$$= \mathrm{P}(\mathrm{sgn}(\tilde{\Delta}_1(i)) = x_i) + \frac{1}{2}\mathrm{P}(\tilde{\Delta}_1(i) = 0) \tag{3.50}$$

$$\geq \frac{1}{2} + \frac{\delta}{2} \tag{3.51}$$

Since $\{\hat{x}_i\}$'s are independent for all unrevealed vertices as $\{\tilde{\Delta}_1(i)\}$'s and $\mathrm{E}\left[\frac{\hat{x}_i x_i + 1}{2}\right] = \mathrm{P}(\hat{x}_i = x_i)$, Lemma 3.4.2 and Lemma 3.4.3 give us that

$$\mathrm{P}\left(\frac{1}{(1-\rho)n} \sum_{i:\tilde{x}_i = 0} \frac{\hat{x}_i x_i + 1}{2} \leq \frac{1}{2} + \frac{\delta}{2} - \epsilon \right) \leq \mathrm{e}^{-2\epsilon^2(1-\rho)n} \tag{3.52}$$

Taking $\epsilon = \frac{\delta}{4}$, we have

$$\mathrm{P}\left(\mathrm{Overlap}(x|_{\tilde{x}_i=0}, \hat{x}|_{\tilde{x}_i=0}) \geq \frac{\delta}{2}\right) \geq 1 - \mathrm{e}^{-\frac{\delta^2(1-\rho)n}{8}} \tag{3.53}$$

As long as $a > b$, we have $\delta > 0$, which concludes the proof. □

COROLLARY 3.4.1. *The semi-supervised SBM and ERM are not mutually contiguous for any given $a > b \geq 0$ and $\rho > 0$.*

PROOF. Let $P_n^{(0)} = \mathcal{G}(n, \frac{a+b}{2n}, \frac{a+b}{2n}, \rho)$ and $P_n^{(1)} = \mathcal{G}(n, \frac{a}{n}, \frac{b}{n}, \rho)$. Then consider the same constant $\delta > 0$ from the proof of Theorem 2.1.1 and denote the event sequence $E_n = \{\mathrm{Overlap}(x|_{\tilde{x}_i=0}, \hat{x}|_{\tilde{x}_i=0}) \geq \frac{\delta}{2}\}$ where $\hat{x}$ is our semi-supervised census estimator. We have

$$P_n^{(0)}(E_n) \to 0 \quad \text{(Law of large number)} \tag{3.54}$$

$$P_n^{(1)}(E_n) \nrightarrow 0 \quad \text{(Bounded from below)} \tag{3.55}$$

□

CHAPTER 4

Semi-Supervised SDP

We have seen that the census method solves the semi-supervised community detection problem. But the algorithm is desirable in practice only when the amount of revealed labels is sufficient to support a reasonable performance. In other words, it has no unsupervised 'fallback' built in. Meanwhile, SDPs enjoy nice properties like optimality and robustness as mentioned earlier. It is also well known that approximate information about the extremal cuts of graphs can be obtained by computing the optimizer for SDP of their adjacency matrix, see for example [**GW95**]. From both a practical and a mathematical point of view, we are interested in developing an SDP based semi-supervised clustering approach, and through which we shall be able to see the models, algorithms and phase transitions with a fresh perspective.

In this chapter, we will focus on the hypothesis testing formulation of the community detection problem. We have discussed the equivalency between it and the non-vanishing overlap formulation under the unsupervised setting. In the semi-supervised scenario it is still an interesting question to ask whether there exists a test that can distinguish SBMs from ERMs. Here we understand ERM as the special case of SBM with $a = b$. It also has ground truth of labels, which is uniformly random under the balance constraint. Given that they are originally contiguous when SNR ≤ 1, we want to show that revealed labels together with random graphs can separate them.

4.1. SDP for community detection

Under the Planted Bisection Model 1.1.2, the Maximum A Posteriori (MAP) estimator is equivalent to the Maximum Likelihood estimator, which is given by min-bisection, i.e., a

balanced partition with the least number of crossing edges. Formally, it can be written as the following optimization problem,

$$\max_{\substack{x\in\{1,-1\}^n \\ x^\top \mathbf{1}=0}} x^\top A x \tag{4.1}$$

By lifting the variable $X \coloneqq xx^\top$, we can rewrite it as

$$\hat{X}_{\mathrm{MAP}}(G) = \arg\max_{\substack{X\succeq 0 \\ X_{ii}=1,\ \forall i\in[n] \\ \mathrm{rank}(X)=1 \\ X\mathbf{1}=0}} \langle A, X\rangle \tag{4.2}$$

Although min-bisection of G is optimal (in the MAP sense) for exact recovery, finding it is NP-hard. Various relaxations have been proposed for the MAP estimator. Since the rank constraint makes the optimization difficult, we can remove it to make the problem convex. One can also get rid of the balance constraint by centralizing the adjacency matrix, $\tilde{A} \coloneqq A - \frac{d}{n}\mathbf{1}\mathbf{1}^\top$ with $d = (a+b)/2$ the average degree. This can also be justified using Lagrangian multipliers. And the resulting semidefinite relaxation is given by

$$\hat{X}_{\mathrm{SDP}}(G) = \arg\max_{\substack{X\succeq 0 \\ X_{ii}=1,\ \forall i\in[n]}} \langle \tilde{A}, X\rangle \tag{4.3}$$

The feasible region $\{X \in \mathbb{R}^{n\times n} : X \succeq 0, X_{ii} = 1\ \forall i \in [n]\}$ is indeed the space of correlation matrices, which defines a subset of the unit hypercube and is also called the *elliptope*. Although it is derived from the relaxation of MAP, one can define the SDP for general symmetric matrices as

$$\mathrm{SDP}(M_{n\times n}) = \max\{\langle M, X\rangle : X \in \mathrm{elliptope}_n\} \tag{4.4}$$

PROPOSITION 4.1.1. *For any $n \times n$ symmetric matrix M, if we denote its leading eigenvalue as λ_1, then*

$$\frac{1}{n} SDP(M) \leq \lambda_1 \tag{4.5}$$

PROOF. For any feasible $X \succeq 0$ and $X_{ii} = 1$, we have $\mathrm{Tr}(X) = n$.

$$\langle X, M = U\Lambda U^\top \rangle = \mathrm{Tr}(U^\top X U \Lambda) = \langle Y \coloneqq U^\top X U, \Lambda \rangle = \sum Y_{ii}\lambda_i \leq n\lambda_1 \tag{4.6}$$

Since $\mathrm{Tr}(Y) = n$ and $Y \succeq 0$, we have $Y_{ii} \geq 0$. So the last inequality follows. □

This proposition relates SDPs to spectra of the underlying matrices, which suffer from those high degree nodes as we mentioned in the introduction. In contrast, SDPs behave similarly on SBMs and random regular graphs. The optimal values of the SDPs for both are approximately $2n\sqrt{d}$, see [**MS16**]. Random regular graphs obey the uniform distribution over graphs with n vertices and uniform degree d, which provide a simple example to illustrate the regularity property of SDPs. We cite an intermediate result from the original proof as Lemma 4.4.3.

An important way to understand SDPs is considering the Cholesky decomposition of X, which characterizes the constraints. Since X is positive semidefinite, we always have $X = \Sigma\Sigma^\top$ with $\Sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)^\top$ and $\|\sigma_i\|_2 = 1,\ \forall i \in [n]$. Therefore, the i-th node of the graph is associated with the vector σ_i that lies on the unit sphere. $X_{ij} = \langle \sigma_i, \sigma_j \rangle$ can be interpreted as the affinity metric between node i and node j. SDP maximizes the likelihood score of this affinity matrix with respect to the given centralized adjacency matrix. The optimizer X^* is a better representation of the structure information than the vanilla adjacency matrix. Then we can identify the labels by simply running a K-means method on it or compute the eigenvector corresponding to the largest eigenvalue.

4.2. Constrained SDP

In this section, we introduce our SDP modification and prove that it solves the semi-supervised community detection problem with the hypothesis testing formulation. Let x denote labels of G(n, $\frac{a}{n}$, $\frac{b}{n}$). And m of them are revealed uniformly at random in a balanced manner. Conditioning on the ground truth of clusters, indices of revealed nodes $\mathcal{R}$ and edges are independent. So without loss of generality, we denote revealed labels $\tilde{x}$ as follows.

$$x_i = \begin{cases} 1, & i = 1, 2, \ldots, \frac{n}{2} \\ -1, & i = \frac{n}{2}, \frac{n}{2}+1, \ldots, n \end{cases} \qquad \tilde{x}_i = \begin{cases} 1, & i = 1, 2, \ldots, \frac{m}{2} \\ -1, & i = \frac{n}{2}, \frac{n}{2}+1, \ldots, \frac{n+m}{2} \\ 0, & \text{otherwise} \end{cases} \tag{4.7}$$

We have shown that the entry value of the optimizer X can be interpreted as an affinity metric among nodes. Moreover, we have $X_{ij} \in [-1, 1],\ \forall\ i, j$. It is natural to force the optimizer to have large entry values for those vertex pairs in which we have high confidence to be in the same community and vice versa. Therefore, we propose the CSDP approach to integrate the information provided by semi-supervised approach. If node i and node j are revealed to have the same label, we add the constraint $X_{ij} = 1$ to the optimization model. If they are revealed to have the opposite labels, we add $X_{ij} = -1$. Formally, the CSDP is defined as

$$\text{CSDP}(M_{n\times n}) = \max\{\langle M, X\rangle : X \in \text{elliptope}_n,\ X_{ij} = x_i \cdot x_j\ \forall i, j \in \mathcal{R}\} \tag{4.8}$$

where $\mathcal{R}$ denotes the collection of revealed nodes. After reordering the indices, we can assume it as $\{1, 2, \ldots, \frac{m}{2}\} \bigcup \{\frac{n}{2}, \frac{n}{2}+1, \ldots, \frac{n+m}{2}\}$. It is worth noting that the optimization remains a positive semidefinite programming, which can be solved efficiently, for example by interior point methods [**Ali95**].

Then let $\mathcal{S}^{n-1} \coloneqq \{v \in \mathbb{R}^n : \|v\|_2 = 1\}$ be the unit $(n-1)$-sphere and $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n) \in (\mathcal{S}^{n-1})^n$. Consider the CSDP in the form derived from the Cholesky decomposition, $X = \Sigma\Sigma^\top$. We have the following identities,

$$\text{SDP}(M) = \max \left\{ \sum_{i,j=1}^{n} M_{ij} \langle \sigma_i, \sigma_j \rangle : \sigma_i \in \mathcal{S}^{n-1} \ \forall i \in [n] \right\} \tag{4.9}$$

$$\text{CSDP}(M) = \max_{\sigma \in (\mathcal{S}^{n-1})^n} \left\{ \sum_{i,j=1}^{n} M_{ij} \langle \sigma_i, \sigma_j \rangle : \sigma_i^\top \sigma_j = x_i x_j \ \forall i, j \in \mathcal{R} \right\} \tag{4.10}$$

$$= \max_{\sigma \in (\mathcal{S}^{n-1})^n} \left\{ \sum_{i,j \in [n] \setminus \mathcal{R}} M_{ij} \sigma_i^\top \sigma_j + \sum_{i,j \in \mathcal{R}} M_{ij} x_i x_j + 2 \sum_{i \in \mathcal{R}} \sum_{j \in [n] \setminus \mathcal{R}} x_i M_{ij} \sigma_0^\top \sigma_j \right\} \tag{4.11}$$

where $\sigma_0 \equiv x_i \sigma_i$, $\forall i \in \mathcal{R}$. Now one can consider an alternative matrix with a special margin denoting the algebraic sum of the blocks from M that are associated with $\mathcal{R}$. We define M^{agg} to be the $(n - m + 1) \times (n - m + 1)$ symmetric matrix indexed from 0 that

$$M_{00}^{\text{agg}} = \sum_{i,j \in \mathcal{R}} M_{ij} x_i x_j \tag{4.12}$$

$$M_{0j}^{\text{agg}} = \sum_{i \in \mathcal{R}} x_i M_{i,j+\frac{m}{2}}, \ \forall j \in [\frac{n}{2} - \frac{m}{2}] \tag{4.13}$$

$$M_{0j}^{\text{agg}} = \sum_{i \in \mathcal{R}} x_i M_{i,j+m}, \ \forall j \in\in [n-m] \setminus [\frac{n}{2} - \frac{m}{2}] \tag{4.14}$$

$$M_{ij}^{\text{agg}} = M_{i+\frac{m}{2}, j+\frac{m}{2}}, \ \forall i, j \in [\frac{n}{2} - \frac{m}{2}] \tag{4.15}$$

$$M_{ij}^{\text{agg}} = M_{i+m,j+m}, \ \forall i, j \in [n-m] \setminus [\frac{n}{2} - \frac{m}{2}] \tag{4.16}$$

Essentially, we aggregate the rows and columns related to revealed vertices according to their communities into the 0-th row and column. Then we reindex the matrix. It introduces spikiness to the underlying matrix.

$$M^{\text{agg}} = \left(\begin{array}{c|ccc} \sum_{i,j\in\mathcal{R}} M_{ij}x_ix_j & M_{01}^{\text{agg}} & \cdots & M_{0,n-m}^{\text{agg}} \\ \hline M_{01}^{\text{agg}} & & & \\ M_{02}^{\text{agg}} & & M_{\mathcal{R}^{\complement}} & \\ \vdots & & & \\ M_{0,n-m}^{\text{agg}} & & & \end{array}\right) \tag{4.17}$$

Although [**MS16**] takes a rather different approach to study SDPs, they also notice that the critical change comes with such built-in structures, where the authors state "we expect the phase transition in $\text{SDP}(\lambda vv^\top + W)/n$ to depend — in general — on the vector v, and in particular on how 'spiky' this is".

Combining the transformed input matrix with equation (4.11), we conclude that CSDP is indeed an SDP regarding M^{agg},

$$\text{CSDP}(M) = \max_{\substack{\sigma_i\in\mathcal{S}^{n-m}\\ i=0,1,\ldots,n-m}} \left\{ \sum_{i,j\in[n-m]} M_{ij}^{\text{agg}}\sigma_i^\top\sigma_j + M_{00}^{\text{agg}} + 2\sum_{j\in[n-m]} M_{0j}^{\text{agg}}\sigma_0^\top\sigma_j \right\} \tag{4.18}$$

$$= \text{SDP}(M^{\text{agg}}) \tag{4.19}$$

LEMMA 4.2.1. *Let $M_{\mathcal{R}^{\complement}}$ be the principle submatrix of M obtained by removing the rows and columns associated with $\mathcal{R}$. The following inequalities hold,*

$$SDP(M_{\mathcal{R}^{\complement}}) \leq CSDP(M) - M_{00}^{agg} \tag{4.20}$$

PROOF. Let X^* be the optimizer of $\mathrm{SDP}(M_{\mathcal{R}^\complement})$. Define its $(n-m+1)\times(n-m+1)$ extension $\hat{X}^*$ as

$$\hat{X}^*_{ij} = \begin{cases} 1 & i=j=0 \\ 0 & i\in[n-m],\ j=0 \\ 0 & j\in[n-m],\ i=0 \\ X^*_{ij} & \text{otherwise} \end{cases} \tag{4.21}$$

Due to the identity from above and the fact that $\hat{X}^* \in \mathrm{elliptope}_{n-m+1}$ is feasible, we can conclude that

$$\mathrm{CSDP}(M) = \mathrm{SDP}(M^{\mathrm{agg}}) \geq \langle \hat{X}^*, M^{\mathrm{agg}} \rangle = \mathrm{SDP}(M_{\mathcal{R}^\complement}) + M^{\mathrm{agg}}_{00} \tag{4.22}$$

□

So far, all the results are deterministic, M can be arbitrary symmetric matrix and $\mathcal{R}$ can be any balanced index set. Next, we will consider $M = \tilde{A} \coloneqq A - \frac{d}{n}\mathbf{1}\mathbf{1}^\top$ to study CSDPs on probabilistic models.

REMARK 4.2.1. *As shown in the Lemma 4.4.1, $\tilde{A}^{agg}_{00} \geq m\cdot\frac{a-b}{2} \geq 0$ with high probability. By definition, we have $CSDP(\tilde{A}) \leq SDP(\tilde{A})$. So, with high probability,*

$$SDP(\tilde{A}_{\mathcal{R}^\complement}) \leq CSDP(\tilde{A}) \leq SDP(\tilde{A}) \tag{4.23}$$

The CSDP always lies in between the SDPs of the original adjacency matrix and the submatrix of unrevealed vertices. Moreover, if $\tilde{A} \sim \mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$, we have $\tilde{A}_{\mathcal{R}^\complement} \sim \mathcal{G}(n-m, \frac{a(1-\rho)}{n-m}, \frac{b(1-\rho)}{n-m})$. It is worth mentioning that although $\tilde{A}_{\mathcal{R}^\complement}$ is just a submatrix of the original centered adjacency matrix, its probabilistic distribution as a random matrix is not simply changed from n nodes to $n-m$ nodes. The edge probability parameters are also changed by a factor of $(1-\rho)$. It leads to some technical challenges, which we are going to handle later.

But intuitively, from the asymptotic behavior of SDP, we can derive a rough understanding of CSDP as $n \to \infty$. Recall that the phase transition theory tells us that when SNR $\leq$ 1, SDP of SBM will not be large enough to distinguish from SDP of ERM. Therefore, the order of above quantities from inequality (4.23) *suggests that semi-supervised SDP can not help to increase the statistics associated with SBM. The best one can hope for is that it will make the statistics associated with ERM smaller by a factor depending on ρ. This turns out to be enough for community detection.*

4.3. Hypothesis test with revealed labels

Recall the community detection problem can be formalized as a binary hypothesis testing problem, whereby we want to determine, with high probability of success, whether the random graph under consideration has a community structure or not. As discussed in Section 2, we introduce semi-supervised learning to the problem by revealing a part of the labels involved in the random graph generating process. Namely, if the labels associated with a graph G over n vertices are denoted as x, we choose m of them uniformly at random denote the index set by $\mathcal{R}$, such that $\sum_{i\in\mathcal{R}} x_i = 0$.

Given a realization of the random graph G *and the revealed labels* $x_{\mathcal{R}}$, we want to decide which of the following holds,

Hypothesis 0: $(G, x_{\mathcal{R}}) \sim \mathcal{G}(n, \frac{d}{n}, \rho)$ is an Erdős–Rényi random graph with edge probability $\frac{d}{n}$, $d = \frac{a+b}{2}$ and reveal ratio ρ. We denote the corresponding distribution over graphs by P_0.

Hypothesis 1: $(G, x_{\mathcal{R}}) \sim \mathcal{G}(n, \frac{a}{n}, \frac{b}{n}, \rho)$ is a planted bisection random graph with edge probabilities $(\frac{a}{n}, \frac{b}{n})$ and reveal ratio ρ. We denote the corresponding distribution over graphs by P_1.

A statistical test T is a function defined on the graphs and revealed labels with range $\{0, 1\}$. It succeeds with high probability if

$$\mathrm{P}_0(T(G, x_{\mathcal{R}}) = 1) + \mathrm{P}_1(T(G, x_{\mathcal{R}}) = 0) \to 0 \quad (n \to \infty) \tag{4.24}$$

Notice that this is indeed a generalization of the unsupervised community detection. Simply looking into the labels, two models are indistinguishable. What characters their difference is the probabilistic law of how edges are generated, i.e., whether there is a cluster structure. The revealed labels serve as an enhancement of the graph observed. The phase transition theory says that under the unsupervised setting (the special case when $\rho = 0$), no test can succeed with high probability when SNR ≤ 1, or equivalently, $a - b \leq \sqrt{2(a + b)}$. While if SNR > 1, several polynomially computable tests are developed. SDP based test is nearly optimal, in the sense that it requires

$$\frac{a - b}{\sqrt{2(a + b)}} \geq 1 + \epsilon(d) \tag{4.25}$$

where $\epsilon(d) \to 0$ as $d \to \infty$. It is believed to be the best that SDPs can reach. As the monotone-robustness study suggests [**MPW16**], this gap may be necessary, since SDP is indeed solving a harder problem where no algorithm can approach the threshold. However, we are going to see that when ρ is sufficiently large, SDPs can not only reach but cross the threshold.

4.4. Semi-supervised detectability

With the problem and algorithm defined clearly, we are ready to prove that SBMs and ERMs can be consistently distinguished in the semi-supervised setting. We take a 'divide and conquer' approach to establish an upper bound of CSDPs on ERMs, while we bound the CSDPs on SBMs from below with a witness that consists of the ground truth of labels, $X = xx^\top$.

LEMMA 4.4.1. *Let (A, x) obey the planted bisection model $G(n, \frac{a}{n}, \frac{b}{n})$ and denote $\langle xx^\top, \tilde{A}\rangle$ as Y. Then for any $\epsilon > 0$, we have $Y/n \in [\frac{a-b}{2} - \epsilon, \frac{a-b}{2} + \epsilon]$ with probability converging to one as $n \to \infty$.*

PROOF.

$$Y = \langle xx^\top, \tilde{A}\rangle = \langle xx^\top, A\rangle - \frac{d}{n}\langle xx^\top, \mathbf{1}\mathbf{1}^\top\rangle \tag{4.26}$$

$$\stackrel{d}{=} 2 \cdot \left[\mathrm{Bin}\left(\left(\frac{n}{2}\right)^2 - \frac{n}{2},\ \frac{a}{n}\right) - \mathrm{Bin}\left(\left(\frac{n}{2}\right)^2,\ \frac{b}{n}\right)\right] \tag{4.27}$$

We have $\mathrm{E}\, Y = \frac{n}{2}(a-b) - a$ and

$$\mathrm{Var}\, Y = 4\left(a\left(\frac{n}{4} - \frac{1}{2}\right)\left(1 - \frac{a}{n}\right) + b\ \frac{n}{4}\left(1 - \frac{b}{n}\right)\right) \leq n(a+b) \tag{4.28}$$

Then Chebyshev's inequality implies that for any $\delta\ in(0,1)$

$$\mathrm{P}\left(|Y - \frac{n}{2}(a-b) + a| \geq \sqrt{n(a+b)} \cdot n^{(1-\delta)/2}\right) \leq \frac{1}{n^{1-\delta}} \tag{4.29}$$

$$\implies \quad \mathrm{P}\left(|\frac{Y}{n} - \frac{a-b}{2} + \frac{a}{n}| \geq \frac{\sqrt{a+b}}{n^{\delta/2}}\right) \leq \frac{1}{n^{1-\delta}} \tag{4.30}$$

(4.31)

Hence, for sufficiently large n, we have

$$\mathrm{P}\left(\frac{Y}{n} \geq \frac{a-b}{2} + \epsilon\right) + \mathrm{P}\left(\frac{Y}{n} \leq \frac{a-b}{2} - \epsilon\right) \leq \frac{1}{n^{1-\delta}} \tag{4.32}$$

Therefore,

$$\mathrm{P}\left(\frac{Y}{n} \in [\frac{a-b}{2} - \epsilon, \frac{a-b}{2} + \epsilon]\right) \geq 1 - \frac{1}{n^{1-\delta}} \tag{4.33}$$

□

Besides bounding the outcomes on SBMs from below, this lemma can also be applied to the 'all revealed blocks' to estimate $\tilde{A}_{00}^{\text{agg}}$, which is used several times throughout our proofs.

LEMMA 4.4.2. *Let $G \sim \mathcal{G}(n, \frac{a}{n}, \frac{b}{n})$, $d = \frac{a+b}{2}$ and $\tilde{A} = A - \frac{d}{n}\mathbf{1}\mathbf{1}^\top$ be its centered adjacency matrix. Then for any $\epsilon > 0$ and $\gamma > 0$, with probability at least $1 - \frac{1}{n^{1-\gamma}}$, for all $n \geq n_0(a, b, \epsilon, \gamma)$, we have*

$$\mathit{CSDP}(\tilde{A}) \geq n \left(\frac{a-b}{2} - \epsilon\right) \tag{4.34}$$

PROOF. We prove the lower bound by considering a witness of the constrained optimization problem. Notice that $xx^\top$ is feasible for both SDP and CSDP, where x is the label vector associated with G. Therefore,

$$\text{CSDP}(\tilde{A}) \geq \langle xx^\top, \tilde{A} \rangle \tag{4.35}$$

Then, we can apply Lemma 4.4.1 to get the result. □

This result holds for any SNR > 0 and suggests the following test for the semi-supervised community detection problem:

$$T(G, x_{\mathcal{R}}; \Delta) = \begin{cases} 1 & \text{if CSDP}(\tilde{A}) \geq n[(a-b)/2 - \Delta] \\ 0 & \text{otherwise} \end{cases} \tag{4.36}$$

The following lemmas bound the CSDP of ERM from above. Intuitively, the contribution from the blocks of adjacency matrix, where columns or rows are associated with revealed nodes, concentrates well around zero. So the 'effective dimension' of the SDP is reduced, which leads to a smaller optimal value. However, it is not directly equivalent to a model with a smaller n since the connectivity probability depends on the original dimension. There are some technical issues we need to deal with.

LEMMA 4.4.3 (Theorem 1, [**MS16**]. Reformulated.). *Let $G \sim \mathcal{G}(n, \frac{d}{n})$ and $\tilde{A} = A - \frac{d}{n}\mathbf{1}\mathbf{1}^\top$ be its centered adjacency matrix. There exists absolute constants C and $d_0 > 1$ such that if $d \geq d_0$, then with high probability,*

$$\frac{1}{n\sqrt{d}} SDP(\tilde{A}) \leq 2 + \frac{C \log d}{d^{1/10}} \tag{4.37}$$

This result is rigorously derived with profound insights from mathematical physics. However, there is an implicit condition on the average degree d in the proof. In fact, it is common to assume at least $d > 1$ in the literature concerning unsupervised clustering because otherwise the graph has no giant component, not to mention reconstruction, as discussed in Section 1.2. However, our approach leads to a subgraph with possibly small effective average degree. Moreover, we do not want to be limited by the topology structure, although which is indeed a fundamental limit in the unsupervised setting. Theorem 2.1.2 shows that semi-supervised SDPs are able to integrate those sublinear components. To achieve that we resort to Grothendieck's inequality and carry out the analysis without assumption on d.

THEOREM 4.4.1 (Grothendieck's inequality [**Gro52**]). *Let M be a $n \times n$ real matrix. If for any $s, t \in \{-1, 1\}^n$,*

$$\big| \sum_{i,j} M_{ij} s_i t_j \big| \leq 1 \tag{4.38}$$

Then for all vectors $X_i, Y_i \in \{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$, $i = 1, 2, \ldots, n$, we have

$$\big| \sum_{i,j} M_{ij} \langle X_i, Y_i \rangle \big| \leq K_G \tag{4.39}$$

Here K_{G} is an absolute constant called Grothendieck's constant. We consider the upper bound derived in [**BMMN11**],

$$K_{\mathrm{G}} < \frac{\pi}{2 \log(1 + \sqrt{2})} \leq 1.78 \tag{4.40}$$

Notice that if we restrict the vectors X_i's and Y_i's to the unit sphere $\mathcal{S}^{n-1}$, the inequality still holds. Since s, t are arbitrary, the left hand side of equation is the $\ell_\infty \to \ell_1$ norm of matrix M, which is

$$\|M\|_{\infty\to 1} = \max_{\|x\|_\infty \le 1} \|Mx\|_1 = \max_{s,t\in\{-1,1\}^n} s^\top M t = \max_{s,t\in\{-1,1\}^n} \Big|\sum_{i,j} M_{ij} s_i t_j\Big| \tag{4.41}$$

This norm is also known as the cut norm, whose importance in algorithmic problems is well understood in theoretical computer science community. Now we can rewrite the theorem in the matrix form and combine it with the elliptope definition of SDP from equation (4.4).

LEMMA 4.4.4. *For arbitrary matrix $M \in \mathbb{R}^{n\times n}$, we have*

$$SDP(M) \le \max_{X\in elliptope_n} \big|\langle M, X\rangle\big| \le K_G \|M\|_{\infty\to 1} \tag{4.42}$$

Next, we use Bernstein's inequality to establish a probabilistic bound on the cut norm of $A - \mathrm{E}\, A$ where A is the adjacency matrix of $\mathcal{G}(n, \frac{d}{n})$.

THEOREM 4.4.2 (Bernstein's inequality [**Pia38**]). *Let $\{X_i\}_{i=1}^n$ be independent random variables such that $\mathrm{E}\, X_i = 0$ and $|X_i| \le M$ for any $i \in [n]$. Denote the average variance as $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}(X_i)$. Then for any $t \ge 0$,*

$$\mathrm{P}\left(\frac{1}{n}\sum_{i=1}^n X_i > t\right) \le \exp\left(-\frac{nt^2/2}{\sigma^2 + \frac{Mt}{3}}\right) \tag{4.43}$$

LEMMA 4.4.5. *Let A be the adjacency matrix of an ERM, $\mathcal{G}(n, \frac{d}{n})$. Then, with probability at least $1 - 5^{-n+2}$,*

$$\|A - \mathrm{E}\, A\|_{\infty\to 1} \le 6(1+d)n \tag{4.44}$$

PROOF. According to the identity from equation (4.38), we want to bound

$$\|A - \mathrm{E}\, A\|_{\infty\to 1} = \max_{s,t\in\{-1,1\}^n} \sum_{i,j} (A - \mathrm{E}\, A)_{ij}\ s_i t_j \tag{4.45}$$

$$= \max_{s,t\in\{-1,1\}^n} \sum_{i<j} (A - \mathrm{E}\, A)_{ij}\ (s_i t_j + s_j t_i) \tag{4.46}$$

For fixed $s, t \in \{-1, 1\}^n$, denote

$$X_{ij} = (A - \mathrm{E}\, A)_{ij}\ (s_i t_j + s_j t_i) \quad (1 \le i < j \le n) \tag{4.47}$$

Then we have $\mathrm{E}\, X_{ij} = 0$, $|X_{ij}| \le 2$ and $\mathrm{Var}(X_{ij}) \le 4\frac{d}{n}$ for any $i < j$. There are totally $n(n-1)/2$ of $\{X_{ij}\}$'s. And they are independent by the definition of ERM. So Bernstein's inequality implies

$$\mathrm{P}\left(\frac{2}{n(n-1)} \sum_{i<j} X_{ij} > t\right) \le \exp\left(-\frac{n(n-1)t^2/4}{\frac{4d}{n} + \frac{2t}{3}}\right) \tag{4.48}$$

Let $t = 12(1+d)/n$, which guarantees $4d/n + 2t/3 < t$. Hence,

$$\mathrm{P}\left(\sum_{i<j} X_{ij} > 6(1+d)n\right) \le \exp\big(-3(n-1)\big) \tag{4.49}$$

Apply the union bound to all 2^{2n} possible (s, t), we have

$$\mathrm{P}\left(\max_{s,t\in\{-1,1\}^n} \sum_{i<j} (A - \mathrm{E}\, A)_{ij}\ (s_i t_j + s_j t_i) > 6(1+d)n\right) \le 2^{2n} \cdot e^{-3(n-1)} \tag{4.50}$$

We conclude the proof with the identity of $\ell_\infty \to \ell_1$ norm and the fact that right hand side of the above inequality is less than 5^{-n+2}. □

Since the distribution of each entry in the matrix changes as $n \to \infty$, we now develop a slightly generalized version of the weak law of large numbers to accommodate our purpose.

LEMMA 4.4.6. *For any n, let $\{X_i^{(n)}\}_{i=1}^n$ be a collection of independent random variables. Assume there exist universal constants μ and σ, such that $\mathrm{E}\, X_i^{(n)} \leq \mu < \infty$ and $\mathrm{Var}(X_i^{(n)}) \leq \sigma^2 < \infty$ for any $n \in \mathbb{N}$ and $i \leq n$. If we denote the sample mean as*

$$\bar{X}^{(n)} = \frac{X_1^{(n)} + X_2^{(n)} + \cdots + X_n^{(n)}}{n} \tag{4.51}$$

then for any $\epsilon > 0$,

$$\mathrm{P}\left(\bar{X}^{(n)} \geq \mu + \epsilon\right) \to 0 \quad as \quad n \to \infty \tag{4.52}$$

PROOF. For any $n \in \mathbb{N}$, we have

$$\mathrm{Var}(\bar{X}^{(n)}) = \frac{1}{n^2}\mathrm{Var}(X_1^{(n)} + X_2^{(n)} + \cdots + X_n^{(n)}) \tag{4.53}$$

$$= \frac{\sum_{i=1}^n \mathrm{Var}(X_i^{(n)})}{n^2} \qquad \text{(by independence)} \tag{4.54}$$

$$\leq \sigma^2/n \qquad \text{(by uniform boundedness)} \tag{4.55}$$

Then Chebyshev's inequality ensures

$$\mathrm{P}\left(|\bar{X}^{(n)} - \mathrm{E}\,\bar{X}^{(n)}| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \tag{4.56}$$

$$\implies \mathrm{P}\left(\bar{X}^{(n)} \geq \frac{1}{n}\sum_{i=1}^n \mathrm{E}\, X_i^{(n)} + \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \tag{4.57}$$

$$\implies \mathrm{P}\left(\bar{X}^{(n)} \geq \mu + \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \tag{4.58}$$

□

REMARK 4.4.1. *This result does not require the random variables to be identically distributed. In fact, the distributions may depend on n. And random variables associated with different n are not necessary to be independent.*

LEMMA 4.4.7. *Let $G \sim \mathcal{G}(n, \frac{d}{n})$, x be the labels, $\mathcal{R}$ be the revealed indices and $\tilde{A} = A - \frac{d}{n}\mathbf{1}\mathbf{1}^\top$ be its centered adjacency matrix. Define*

$$B_{ij} = \begin{cases} \sum_{i,j\in\mathcal{R}} \tilde{A}_{ij} x_i x_j & i = j = 0 \\ \sum_{k\in\mathcal{R}} x_k \tilde{A}_{kj}, & i = 0,\ j \in [n] \setminus \mathcal{R} \\ \sum_{k\in\mathcal{R}} x_k \tilde{A}_{ik}, & j = 0,\ i \in [n] \setminus \mathcal{R} \\ 0 & \textit{otherwise} \end{cases} \tag{4.59}$$

Then for any $\epsilon > 0$, with high probability,

$$\mathit{SDP}(B) \le 2dm(1 - \frac{m}{n}) + (2n - m)\epsilon \tag{4.60}$$

PROOF. Notice that for any feasible X of above optimization problem, we have $X \succeq 0, X_{ii} = 1\ \forall i \in [n+1]$. So, for any $i,\ j \in [n+1]$,

$$(\mathbf{e}_i \pm \mathbf{e}_j)^\top X (\mathbf{e}_i \pm \mathbf{e}_j) = 2 \pm 2X_{ij} \ge 0 \quad \Longrightarrow \quad |X_{ij}| \le 1 \tag{4.61}$$

Therefore,

$$\mathrm{SDP}(B) = \max\{\langle B, X\rangle : X \in \mathrm{elliptope}_{n+1}\} \tag{4.62}$$

$$= B_{00} + 2\max\left\{\sum_{j\in[n]\setminus\mathcal{R}} B_{0j}X_{0j} : X \in \mathrm{elliptope}_{n+1}\right\} \tag{4.63}$$

$$\le B_{00} + 2\sum_{j\in[n]\setminus\mathcal{R}} |B_{0j}| \tag{4.64}$$

Note that $\{B_{0j} : j \in [n]\setminus\mathcal{R}\}$'s are independent random variables. Moreover, if we let $B_1,\ B_2$ be two independent binomial random variables with the same parameter $(\frac{m}{2}, \frac{d}{n})$ and denote their difference as $Z \coloneqq B_1 - B_2$, we have $B_{0j} \overset{\mathrm{d}}{=} Z$ for any $j \in [n] \setminus \mathcal{R}$ with $\mathrm{E}\, Z = 0$ and $\mathrm{Var}\, Z \le d\frac{m}{n}$.

Since $Z^2 \geq |Z|$, we have

$$\mathrm{E}\,|Z| \leq \mathrm{E}(Z^2) = \mathrm{Var}\,Z \leq d\frac{m}{n} \tag{4.65}$$

$$\mathrm{Var}\,|Z| = E(Z^2) - (\mathrm{E}\,|Z|)^2 \leq \mathrm{Var}\,Z \leq d\frac{m}{n} \tag{4.66}$$

Then Lemma 4.4.6 can be applied to

$$\bar{X}^{(n)} := \frac{\sum_{j\in[n]\setminus\mathcal{R}} |B_{0j}|}{n-m} \tag{4.67}$$

So, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathrm{P}\left(\frac{1}{n-m}\sum_{j\in[n]\setminus\mathcal{R}} |B_{0j}| > d\frac{m}{n} + \epsilon\right) = 0 \tag{4.68}$$

Hence, $\sum_{j\in[n]\setminus\mathcal{R}} |B_{0j}| \leq (n-m)(d\frac{m}{n} + \epsilon)$ with high probability.

Lemma 4.4.1 implies, with high probability,

$$B_{00} \leq \epsilon m \tag{4.69}$$

Combining the above results with the union bound completes the proof. □

Returning to the semi-supervised SDP, based on the notions from Section 4.2, we consider the following decomposition of the transformed input matrix M^{agg} with the unrevealed part and revealed part as

$$M^{\mathrm{agg}} = M^{(\mathcal{R}^{\complement})} + M^{(\mathcal{R})} \tag{4.70}$$

where we define

$$M_{ij}^{(\mathcal{R})} = \begin{cases} M_{ij}^{\mathrm{agg}} & i = 0 \text{ or } j = 0 \\ 0 & \text{otherwise} \end{cases} \tag{4.71}$$

To prove the main result of semi-supervised SDP, we first control the $M^{(\mathcal{R}^\complement)}$ part by Grothendieck's inequity and then bound the contribution of $M^{(\mathcal{R})}$ with the generalized law of large numbers shown above.

Proof of Theorem 2.1.2. Notice that Lemma 4.4.2 guarantees the test to succeed under the SBM. We only need to show, under ERM, $\mathrm{CSDP}(\tilde{A}) < n[(a-b)/2 - \Delta]$ w.h.p.. According to the identity from equation (4.18), we have

$$\mathrm{CSDP}(\tilde{A}) = \mathrm{SDP}(\tilde{A}^{\mathrm{agg}}) \tag{4.72}$$

$$= \max\{\langle \tilde{A}^{\mathrm{agg}}, X\rangle : X \in \mathrm{elliptope}_n\} \tag{4.73}$$

$$= \max\{\langle \tilde{A}^{(\mathcal{R}^\complement)} + \tilde{A}^{(\mathcal{R})}, X\rangle : X \in \mathrm{elliptope}_n\} \tag{4.74}$$

$$\leq \mathrm{SDP}(\tilde{A}^{(\mathcal{R}^\complement)}) + \mathrm{SDP}(\tilde{A}^{(\mathcal{R})}) \tag{4.75}$$

Recall that $\tilde{A}_{\mathcal{R}^\complement}$ is the principal submatrix of $\tilde{A}$ obtained by removing the rows and columns associated with $\mathcal{R}$. By definition, we have $\mathrm{SDP}(\tilde{A}_{\mathcal{R}^\complement}) = \mathrm{SDP}(\tilde{A}^{(\mathcal{R}^\complement)})$. Under the null hypothesis, $\tilde{A}_{\mathcal{R}^\complement}$ has the same distribution as the centered adjacency matrix associated with $\mathcal{G}(n-m, \frac{(1-\rho)d}{n-m})$. Also,

$$\mathrm{SDP}\left(\tilde{A}^{(\mathcal{R}^\complement)}\right) = \mathrm{SDP}\left(A_{\mathcal{R}^\complement} - \mathrm{E}\, A_{\mathcal{R}^\complement} - \frac{(1-\rho)d}{n-m} I_{n-m}\right) \tag{4.76}$$

$$= \mathrm{SDP}\left(A_{\mathcal{R}^\complement} - \mathrm{E}\, A_{\mathcal{R}^\complement}\right) - (1-\rho)d \tag{4.77}$$

According to the Grothendieck's inequality and Lemma 4.4.5, we conclude that, with probability at least $1 - 5^{-(1-\rho)n+2}$,

$$\mathrm{SDP}\left(\tilde{A}^{(\mathcal{R}^\complement)}\right) \leq 6K_{\mathrm{G}}[1 + (1-\rho)d](n-m) \tag{4.78}$$

$$< 12(1+d)(1-\rho)n \tag{4.79}$$

Combining the result from Lemma 4.4.7 with $\epsilon = d(1-\rho)^2$, we have

$$\frac{1}{n}\text{CSDP}(\tilde{A}) \leq 14(1-\rho)(1+d) \quad \text{w.h.p.} \tag{4.80}$$

Taking $\Delta = (a-b)/40$ and $\rho_0 = 1 - \frac{a-b}{30(1+d)}$, we conclude, if $\frac{m}{n} \geq \rho_0$,

$$\mathrm{P}_0(T(G, x_{\mathcal{R}}) = 1) \to 0 \qquad (n \to \infty) \tag{4.81}$$

□

4.5. Numerical simulation

We include some simulation results below. $\rho \in [0,1]$ is the ratio of revealed labels. Results associated with unsupervised SDPs are identified as $\rho = 0$. As discussed in Section 3, to make the comparison fair and keep the problem meaningful, all overlaps are restricted to the unrevealed labels.

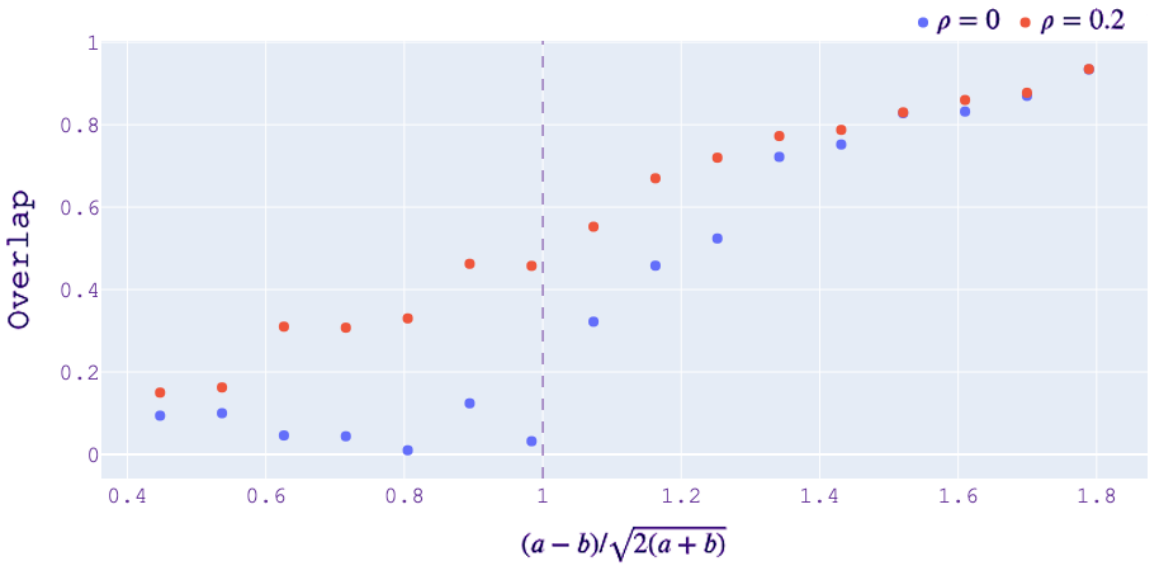

FIGURE 4.1. Disappearance of the phase transition.

Each point in Figure 4.1 represents one realization of a SBM with $n = 1000$. The dashed line stands for the KS and information-theoretic threshold. The graphs are shared by both the unsupervised and the semi-supervised SDPs. Overlaps of the unsupervised algorithm essentially drop down to zero on the left-hand side. While, with 20% of the labels revealed, the outcome of our constraint SDP algorithm goes down gradually as the SNR decreases and remains substantially greater than zero even when SNR ≤ 1.

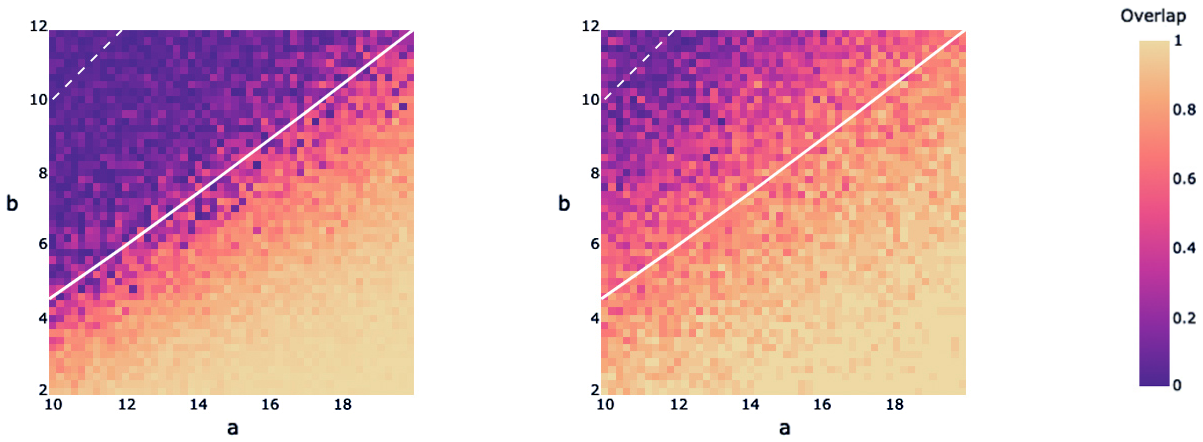

FIGURE 4.2. Overlap heatmaps of the unsupervised (left) and the semi-supervised (right) SDPs. The coordinates correspond to the model parameters a and b. The solid line represents the KS and information-theoretic threshold. The dash line corresponds to $a = b$.

The phase transition theory 1.2.1 guarantees that the upper left half of the left image will be totally dark as $n \to \infty$. But we see semi-supervised SDPs successfully 'light up' the entire area between the two reference lines, see Figure 4.2. Moreover, when n is sufficiently large, there will be no pixel with value equals to 0.

Figure 4.3 shows color-coded entry values of optimizer X^* in different settings and suggests that representing of the underlying community structure is significantly enhanced by the semi-supervised approach, while no such structure is introduced in ERM setting.

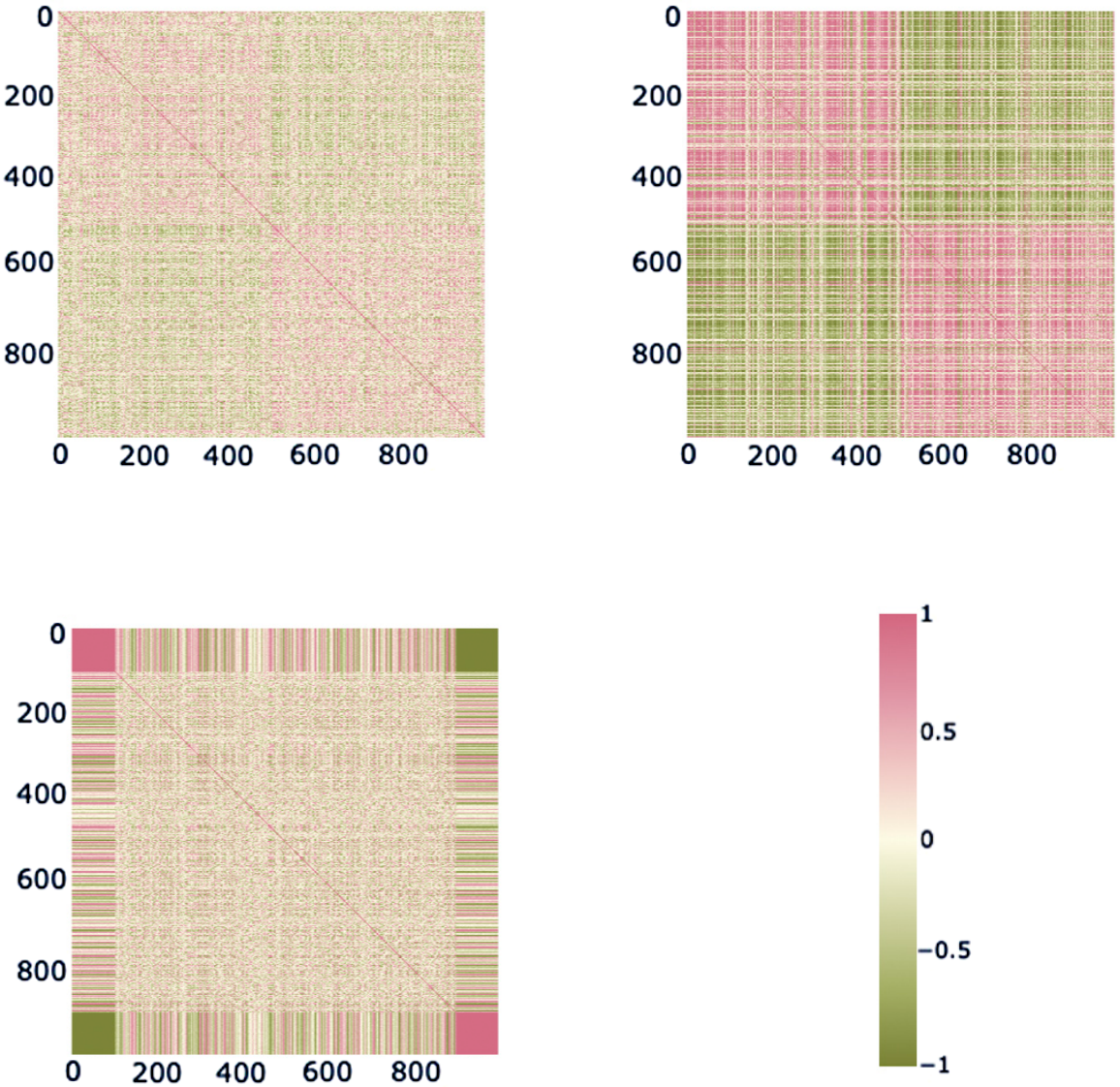

FIGURE 4.3. Visualization of the optimizer X^*. The upper row is concerned with one realization of the SBM $\mathcal{G}(1000, 12/1000, 5/1000)$, where the left image shows the value of optimizer for the unsupervised SDP and the right image is associated with the semi-supervised SDP with $\rho = 0.2$. The lower left image is optimizer for one realization of the ERM of the same size with the associated average degree $d = 8.5$, indices of which are reordered such that the entries related to revealed labels are gathered in four corners. It could be understood as the situation of null hypothesis we defined in Section 4.2.

SDP optimal value, when SNR is above KS/IT

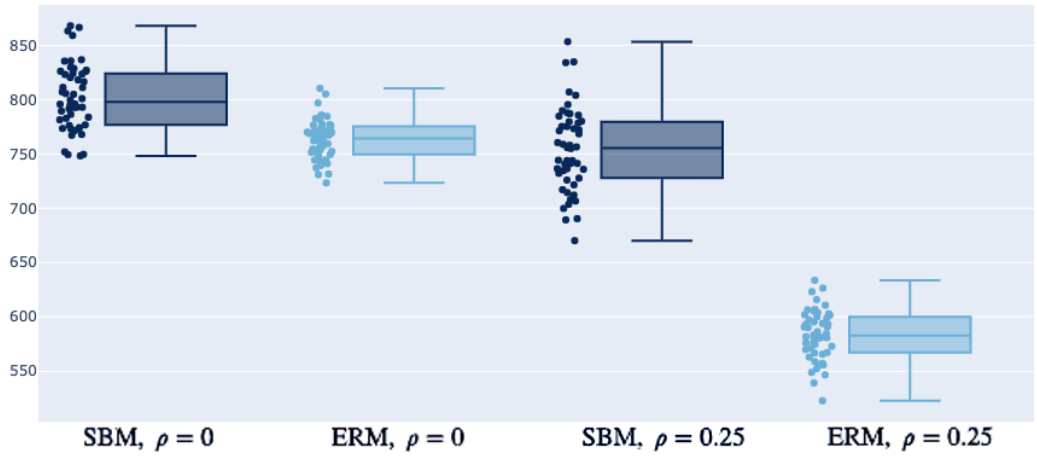

FIGURE 4.4. $a = 9,\ b = 2 \quad (d = 5.5,\ \text{SNR} \approx 2.23)$

SDP optimal value, when SNR is below KS/IT

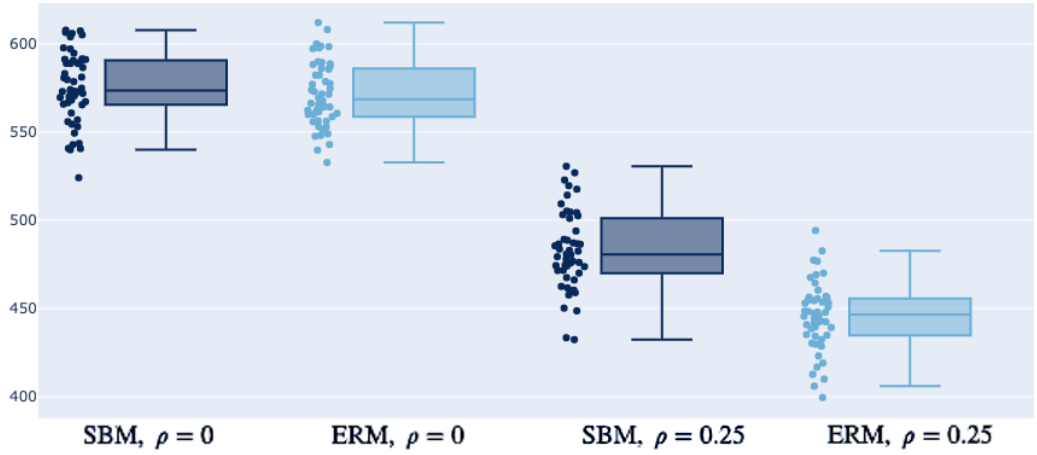

FIGURE 4.5. $a = 5,\ b = 2 \quad (d = 3.5,\ \text{SNR} \approx 0.64)$

To see how such a better representation leads to a successful test that is originally impossible, we consider the following simulations. We generate 50 independent realizations of underlying random graphs ($n = 200$) and compute their SDP values with and without the semi-supervised constraints ($\rho = 0.25$). Particularly, the parameters in Figure 4.4 are chosen to have SNR > 1. The left two boxes imply that we can tell the difference between SBM and the ERM with the same average degree $d = (a + b)/2$. However, as in Figure 4.5, the vanilla SDPs give essentially the same result since the two models become contiguous if SNR ≤ 1. As we have proved in Theorem 2.1.2, our semi-supervised SDP algorithm still manages to distinguish them by bringing down the optimal value of ERM more significantly comparing to its effect on SBM, which is confirmed by the right two boxes.

CHAPTER 5

Application to GCN

In this chapter, we present our first of its kind semi-supervised design of the propagation model for graph-based deep learning. To avoid repetition, we refer to the theoretical results in previous chapters directly and focus on addressing the intuition behind our model. We include the experimental results on both the real-world and the synthetic datasets.

5.1. Deep learning on graphs

Deep neural networks have achieved significant success on a large amount of high-dimensional datasets, where carefully designed architectures are used to exploit the intrinsic properties of data. Famous examples include images [**KSH12**], text [**Bis95**] and biomedical research [**Web18**]. The differentiable blocks in these applications are devoted to the particular types of data. It remains challenging to replicate their success to other types of data sets. As one of the largest class of data in the real world, *graphs* (or arbitrary data sets with graph structure) are ubiquitous from quantitative finance [**EK10**] to biomedical application [**GSR$^+$17**], social network [**NWS02b**], computer program [**ABK18**], recommender system [**BKW17**], etc.

In scenarios of deep learning, a graph consists of a set of nodes and a collection of edges connecting them. Each node stands for one element of a data set and is possibly associated with features and label, which is usually the target of underlying learning task. An edge could be directed or undirected and represents relation between the two end nodes, e.g., similarity, social connection or citation.

Recent years have witnessed rapid developments of new deep learning methods that are capable of learning on graph-structured data. The most prominent advancement is known

as Graph Convolutional Networks [**HYL17**], [**KW17**], [**MBB17**]. It has become a standard approach for node classification and an efficient building block of graph processing.

The key idea of GCNs is to learn the features using both the content information and the graph structure. Namely, the learning outcome of a single convolution layer is the aggregated result of a node's 1-hop neighborhood on the graph. In contrast to the pure content-based deep learning models (e.g., convolutional neural networks, recurrent neural networks), GCNs take into consideration the connectivity between vertices and draw conclusions from the entire neighborhood instead of the input features associated with a single node. GCN-based methods have attracted a vast amount of research interest as they have set a new bar on countless recommender system benchmarks (see [**HYL17**] for a survey). A core question of GCN research is how to design more effective communication protocols that better leverage the structural information to improve task performance of the GCN [**LMBB19**, **BGLA21**, **GWG19**, **KBG19**, **LHW18**].

Recall that in GCN, the communication protocol among different nodes (or the propagation model) is specified by $\hat{A}$ as in equation (1.18). Essentially, the input of activation function in each intermediate layer is a linear combination of the node representations from its previous layer.

For example, the original GCN uses the adjacency matrix with self-loops, $I + A$. Hence, the activation function takes the sum of the learned features over the 1-neighborhood of each node as input. As the prevalent approach in the field of graph-based deep learning, this formulation is also referred to as Message Passing Neural Networks (MPNNs) [**GSR$^+$17**]. From our analysis, we can see that they only allow the messages to be passed between neighboring nodes in each layer. Although MPNNs can leverage higher-order neighborhoods with a deeper structure, it is unreasonable to limit the messages to 1-neighbors. It turns out that by increasing the number of layers, the performance is not improved [**KW17**]. So it is quite natural to consider introducing some global information to the propagation model.

The MPNN-like models which only rely on the immediate neighborhood information, are often categorized as spatial methods. And another major type of propagation model is based on the spectral decomposition of $\hat{A}$. The spectral-based methods capture and make use of more complex graph properties [**DBV16**]. However, as we have discussed in Section 1.3, these methods rely heavily on the model statistics. Since the graphs from real-world applications are often complicated and noisy, these spectral properties become quite delicate. Therefore, spectral-based methods are routinely outperformed by MPNNs on benchmark tasks, see e.g. [**KW17**, **VCC$^+$18**, **XHLJ19**]. It is desirable to have a model that enjoys the robustness of MPNNs while utilizes the global information like spectral-based methods.

5.2. Semi-supervised propagation model

Although GCN is originally proposed to solve semi-supervised learning problems, the information hidden in the revealed labels is not fully used. A general deep learning classifier takes features associated with each sample as input and predicts the corresponding label. Through objective function and backpropagation, the model 'learns' from the true labels. So it produces better prediction using the features. As shown in Figure 5.1, GCN reconciles the node features and the additional graph structure to predict the labels. But the graph is either used as given or pre-processed in an unsupervised manner. To our knowledge, no existing method makes use of the label information for a better understanding of the graph, i.e., the dashed line in the following diagram is missing.

We propose our semi-supervised propagation model matrix for GCN as

$$(\hat{A}_{ij}) = (\mathbf{1}_{\{X^*_{ij} \geq \theta\}}) \tag{5.1}$$

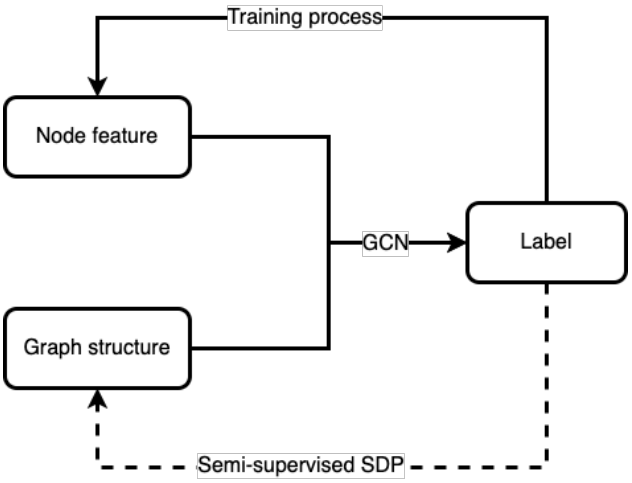

FIGURE 5.1. Information flow in graph-based deep learning models.

with $\theta < 1$ and X^* being the optimizer of the following semidefinite program:

$$\begin{aligned} \max_{X} \quad & \langle A - \frac{d}{n}\mathbf{1}\mathbf{1}^\top, X\rangle \\ \text{s.t.} \quad & X_{ii} = 1 \qquad \forall\ i \in [n] \\ & X \succeq 0 \\ & X_{ij} < \gamma \qquad \text{if } i, j \in Tr \text{ and } y_i \neq y_j \end{aligned} \tag{5.2}$$

where d is the average degree of the graph, A is the adjacency matrix, n is the total number of nodes. And Tr denotes the collection of all indices in the training set, y denotes the label vector. $\gamma \in [-1, 1]$ is a hyperparameter that controls the maximum correlation between the representations of different classes.

The main results established by this dissertation show that semi-supervised SDP captures the cluster structure in a graph even when it is extremely challenging to do so. Our method then uses such structural information to benefit the learning process. Furthermore, it can be trivially combined with any existing graph-based algorithm in a plug-and-play manner, i.e., without changing the model or affecting the computational complexity.

Intuition. Without the last constraint, the problem reduces to an SDP discussed in (4.4) and (4.9). Hence, the Cholesky decomposition argument is still valid. Each node is embedded as a vector on the unit sphere in $\mathbb{R}^n$. X^* is the correlation matrix of these embeddings. The larger the value of X^*_{ij} is, the more similar node i and node j are. Therefore, more message should be passed between them in the deep neural network, which makes X^* a good basis for our propagation model matrix. Furthermore, the similarity is based on the connectivity pattern of each node. So representation is spatially localized. And by restricting the embeddings to the unit sphere, we introduce the global consideration to the optimization problem. Due to the adversarial robustness of SDP, the use of global information is reliable and faithful to the minimum assumption on graph structure, i.e., node clusters are densely connected internally and relatively weakly connected externally.

Limitation. Although SDP-based propagation seems to have ticked all the boxes, it has a drawback that limits its real-world application. The unsupervised SDPs rely on the homophily assumption, i.e., "birds of a feather flock together" [**VCC$^+$18**]. It is a very common assumption that is shared by many methods. However, in practice, this is usually violated to some degree and it seems non-straightforward to overcome. Take the commonly considered benchmark citation datasets (Cora, Citeseer, PubMed, etc.) for instance. It is common for a subfield to be more closely connected to a subfield of another field than some other topics within its own field. The machine learning publications in the field of statistics are very likely to cite the publications categorized as machine learning in the field of computer science. On the other hand, the connections between machine learning of computer science and programming language of computer science are much sparser. But these relations are not reflected by the labels. Those 'unexpected' structures could be amplified by SDP hence affect the performance. Therefore, we need to find a way to justify the graph information according to the learning target.

Semi-supervised approach. To integrate the structural information with the prediction task, we resort to the semi-supervised SDP. Namely, the last constraint from formula (5.2) effectively separates the embeddings apart as long as they have different labels. This can been seen through the similarity score interpretation. It is worth noting that $X^*_{ij} = 1$ implies $X^*_{ik} = X^*_{jk}$ for all $k \in [n]$, i.e., the training samples from the same class now share essentially the same embedding. So we do not hard code the nodes from the same class to have similarity score of 1 as in equation (4.8). Otherwise, it will lead to a severe overfitting issue. On the other hand, we also relax the constraints for those nodes from different classes. Instead of -1 similarity score, we only put an upper bound γ on the corresponding entries. So the problem remains feasible in the multiclass setting. (E.g., if we force the similarity among three nodes from three classes to all be -1, we immediately have a contradiction.) In the citation example mentioned earlier, the subfields of CS-ML and Stats-ML will be embedded not so close due to the semi-supervised constraint. Then the deep learning component can do better in putting them into the correct classes.

Sparsification. The rounding step in equation (5.1) helps to reduce noise and improve the generalization. Essentially, we only keep those similarity relations of high confidence. So if the two class embeddings are far away, they remains separated after the rounding process. If two embeddings are close, by rounding, we get rid of the noisy patterns in the embeddings which are mistakenly learnt by deep learning component due to overfitting. The remaining difference is intrinsic and generalize well to the unseen data. This step also leads to a sparse $\hat{A}$, which brings additional benefit to the model implementation.

5.3. Experimental results

In this section, we focus on the semi-supervised node classification task, which is considered a practical and representative test for various graph-based deep learning models. We consider the following two datasets.

- **Cora [SNB+08].** The dataset consists of publications focusing on the field of machine learning research. These papers are divided into classes. Totally, the dataset has 2708 document nodes, 5429 citation edges. The feature of each paper is a 0/1-valued word vector indicating the absence/presence of the corresponding word from a dictionary of 1433 unique words.
- **Synthetic SBM.** We first generate a realization of stochastic block model for the given parameters. Then node features are sampled from standard Gaussian distributions such that the centers of clusters are located on vertices of a hypercube.

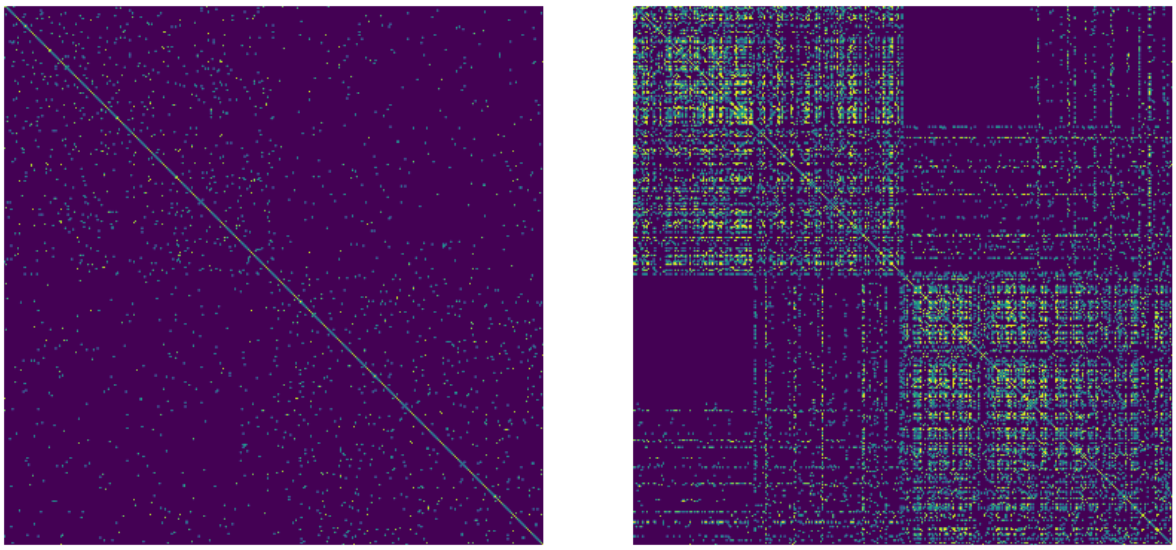

FIGURE 5.2. The propagation matrices of MPNN (left) and our method (right) for the same SBM instance. Semi-supervised SDP allows the information to be passed globally with the justification of labels.

All experiments share the same early stopping criteria. The hyperparameters (θ and γ) are chosen by the same grid search on the validation set. We use the test set only once for generating test results. For Cora, we compute the average test accuracy over 20 random splits. We realized 100 independent instances for each parameter setting of SBM and report the average performance on the test set.

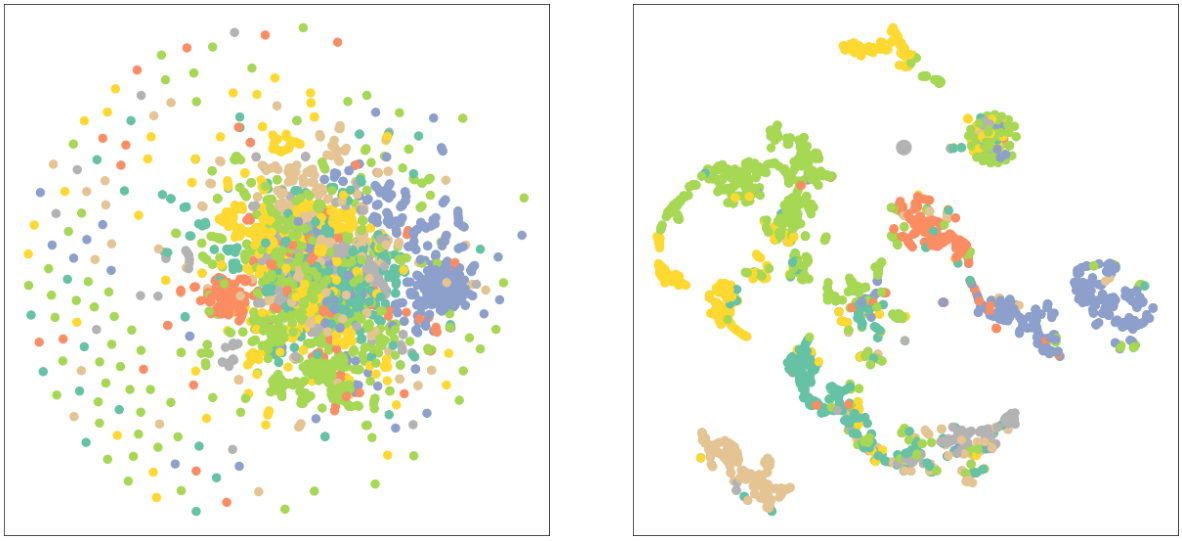

FIGURE 5.3. t-SNE [**vdMH08**] visualization of embeddings associated with MPNN (left) and CSDP (right) from the Cora dataset with 10% of the labels revealed. The ground truth of the classes is coded in color.

Since the focus of our work is on the benefit of structural information instead of the design of downstream deep learning components, we fix a relatively simple deep learning framework for all the propagation models considered. The feature matrix is fed into graph convolutional layer with a ReLU activation. It is followed by a dropout layer with dropout probability of 0.5. Then another graph convolutional layer is used to generate the scoring vector for prediction. We use cross-entropy for the objective function and Adam with learning rate of 0.01 as the optimizer. And the hidden dimension is set as 16. Although it is not a complicated model, the message passing approach performs at its best with this model [**KW17**].

To better isolate the effect of graph information on the learning task, we consider the featureless Cora, i.e., the original feature matrix is replaced by an identity matrix of the same dimension as the total number of nodes on the graph. It becomes a much harder task to classify the publications only based the connectivity. As shown in Figure 5.4, when there is only 5 training samples per class, MPNN performs not so well. In contrast, the

FIGURE 5.4. Node classification accuracy on featureless Cora. The test accuracy improves as the number of training samples increases. But unsupervised SDP is less efficient in making use of this additional information.

performance of SDP and CSDP approaches is decent and very close to each other. This is because when the labels are limited, the extra semi-supervised constraint are not going to make much difference. As we have access to more training data, this constraint kicks in and the accuracy of CSDP-GCN goes up consistently. Although the MPNN also enjoys an increase in accuracy, there is still a significant margin between it and CSDP.

For the experiments on synthetic SBM dataset, we specify the model to have 500 nodes which can be divided into two symmetric blocks and fix the dimension of feature vector to be 2000. The SNR represents how clear this block structure is reflected in the graph. We report the results for SNR $= 0.125$ (Figure 5.5) and SNR $= 1.5$ (Figure 5.6) where the height of the bar stands for the average accuracy and the interval on top of each bar represents

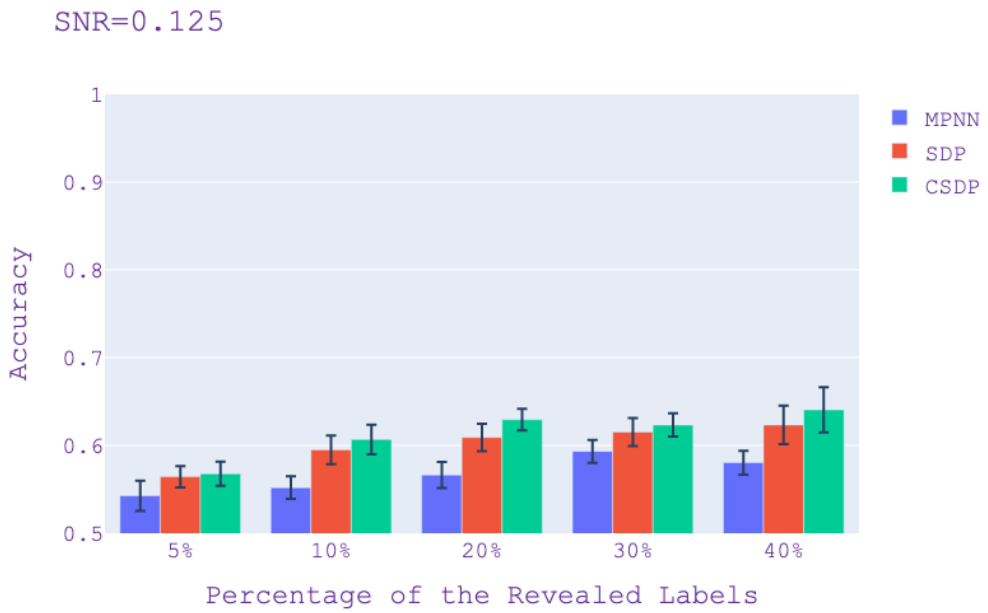

FIGURE 5.5. In a challenging setting (i.e., edges are sparse and unreliable; features are not strong indicators of the classes), CSDP-based approch outperforms others.

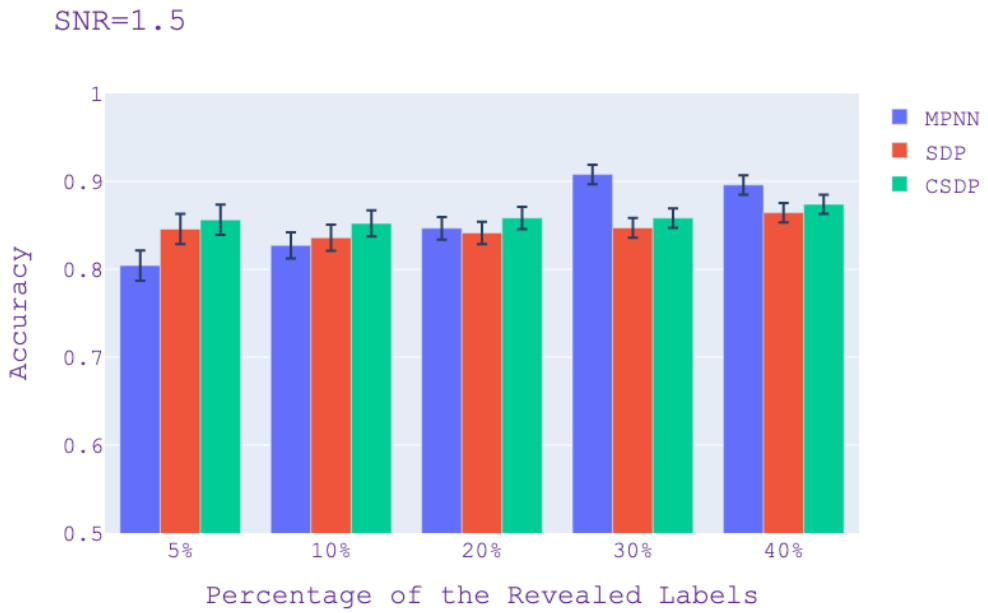

FIGURE 5.6. When the edges strongly imply the class similarity, MPNN can learn those local patterns with enough training data.

the corresponding standard error. In the previous setting, we simulate the situation where interclass edges commonly exist and the features are not dependable due to large random noise. In this case, CSDP makes the best use of the structural information and significantly increases the performance over the neighborhood-based message passing scheme. And when SNR is large, the edge become a reliable indicator of class similarity. So MPNN which learns the small neighborhoods separately and provides a better accuracy when the percentage of training samples is high. In both situations, we observe that the semi-supervised approach consistently improves the accuracy over unsupervised SDP across various sizes of the training set. This suggests that bounding those interclass correlations successfully reconciles SDP's dependence on the homophily assumption with the label ground truth.

CHAPTER 6

Conclusion

The census method comes from the combinatorial perspective, while the CSDP is inspired by convex optimization research. Both algorithms are computationally efficient. The former has no requirement on the reveal ratio. The latter one is more practical and backward compatible to the unsupervised setting. By carefully integrating the revealed information with the observed graph structure, we can not only improve the performance of clustering algorithms but resolve initially unsolvable problems. The fundamental changes brought by semi-supervised approach let us cross KS threshold, information-theoretical threshold and even the topological limitation.

Our work provides a different angle to study stochastic models of network and semidefinite programs. In real-world situations, it is almost always the case that we will have a certain amount of samples being understood fairly well. So an abstract model should be able to capture the existence of such knowledge instead of being blindly restricted to unsupervised setting. Combining the universality of 'revealed' information and the insight derived from our census method, it is arguable that the phase transitions, although very mathematically beautiful, will never be an issue in practice. Our results on CSDPs, in turn, could be used to study the performance of SDPs, e.g. prove or disprove it can reach the phase transition threshold or the monotone-robustness threshold by a limiting process of $\rho \to 0$.

Besides the mathematical curiosity, a major reason we study these foundational problems, e.g. clustering on random graph, is to develop better tools for realistic applications via theoretical guidance. Inspired by our theoretical results, we propose the CSDP-based

propagation model, which can be easily adapted to various graph-based deep learning architectures. In particular, it naturally coincides with the key idea behind GCN, i.e., making similar nodes share the activation. CSDP model will provide a learning objective justified graph representation, which not only contains more information of the underlying class structure but also auto-calibrates to the specific learning task. We conduct rigorous and representative experiments that show our method outperforms the widely adopted MPNN model, especially when it comes to the challenging learning tasks.

Bibliography

[Abb18] E. Abbe, *Community detection and stochastic block models: Recent developments*, Journal of Machine Learning Research **18** (2018), no. 177, 1–86.

[ABH16] E. Abbe, A. S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, IEEE Transactions on Information Theory **62** (2016), 471–487.

[ABK18] M. Allamanis, M. Brockschmidt, and M. Khademi, *Learning to represent programs with graphs*, International Conference on Learning Representations, 2018.

[ABKK17] N. Agarwal, A. S. Bandeira, K. Koiliaris, and A. Kolla, *Multisection in the stochastic block model using semidefinite programming*, pp. 125–162, Springer International Publishing, Cham, 2017.

[ABRS20] E. Abbe, E. Boix, P. Ralli, and C. Sandon, *Graph powering and spectral robustness*, SIAM J. Math. Data Sci. **2** (2020), 132–157.

[AFWZ20] E. Abbe, J. Fan, K. Wang, and Y. Zhong, *Entrywise eigenvector analysis of random matrices with low expected rank.*, Annals of statistics **48 3** (2020), 1452–1474.

[AL18] A. A. Amini and E. Levina, *On semidefinite relaxations for the block model*, The Annals of Statistics **46** (2018), no. 1, 149 – 179.

[Ali95] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization*, SIAM J. Optim. **5** (1995), 13–51.

[AS15] E. Abbe and C. Sandon, *Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery*, 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, 2015, pp. 670–688.

[AS18] E. Abbe and C. Sandon, *Proof of the achievability conjectures for the general stochastic block model*, Communications on Pure and Applied Mathematics **71** (2018), no. 7, 1334–1406, https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.21719.

[Ban18] A. S. Bandeira, *Random Laplacian matrices and convex relaxations*, Foundations of Computational Mathematics **18** (2018), 345–379.

[BC09] P. J. Bickel and A. Chen, *A nonparametric view of network models and Newman-Girvan and other modularities*, Proceedings of the National Academy of Sciences **106** (2009), no. 50, 21068–21073, https://www.pnas.org/doi/pdf/10.1073/pnas.0907096106.

[BCLS84] T. Bui, S. Chaudhuri, T. Leighton, and M. Sipser, *Graph bisection algorithins with good average case behavior*, 25th Annual Symposium on Foundations of Computer Science, 1984., 1984, pp. 181–192.

[BGLA21] F. M. Bianchi, D. Grattarola, L. Livi, and C. Alippi, *Graph neural networks with convolutional arma filters*, IEEE Transactions on Pattern Analysis and Machine Intelligence (2021), 1–1.

[Bis95] C. M. Bishop, *Neural networks for pattern recognition*, Oxford University Press, Inc., USA, 1995.

[BJR07] B. Bollobás, S. Janson, and O. Riordan, *The phase transition in inhomogeneous random graphs*, Random Struct. Algorithms **31** (2007), no. 1, 3–122.

[BKW17] R. v. d. Berg, T. N. Kipf, and M. Welling, *Graph convolutional matrix completion*, 2017.

[BLM15] C. Bordenave, M. Lelarge, and L. Massoulié, *Non-backtracking spectrum of random graphs: Community detection and non-regular Ramanujan graphs*, 2015 IEEE 56th Annual Symposium on Foundations of Computer Science (2015), 1347–1357.

[BMMN11] M. Braverman, K. Makarychev, Y. Makarychev, and A. Naor, *The Grothendieck constant is strictly smaller than Krivine's bound*, 2011 IEEE 52nd Annual Symposium on Foundations of Computer Science (2011), 453–462.

[BMNN16] J. Banks, C. Moore, J. Neeman, and P. Netrapalli, *Information-theoretic thresholds for community detection in sparse networks*, 29th Annual Conference on Learning Theory (Columbia University, New York, New York, USA) (V. Feldman, A. Rakhlin, and O. Shamir, eds.), Proceedings of Machine Learning Research, vol. 49, PMLR, 23–26 Jun 2016, pp. 383–416.

[BRZ95] P. Bleher, J. Ruiz, and V. A. Zagrebnov, *On the purity of the limiting gibbs state for the Ising model on the Bethe lattice*, Journal of Statistical Physics **79** (1995), 473–482.

[Cav78] J. A. Cavender, *Taxonomy with confidence*, Mathematical Biosciences **40** (1978), no. 3, 271–280.

[Che52] H. Chernoff, *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Annals of Mathematical Statistics **23** (1952), 493–507.

[CLS+19] W.-L. Chiang, X. Liu, S. Si, Y. Li, S. Bengio, and C.-J. Hsieh, *Cluster-GCN: An efficient algorithm for training deep and large graph convolutional networks*, Proceedings of the 25th

ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (New York, NY, USA), KDD '19, Association for Computing Machinery, 2019, p. 257–266.

[CO09] A. Coja-Oghlan, *Graph partitioning via adaptive spectral techniques*, Combinatorics, Probability and Computing **19** (2009), 227 – 284.

[CX16] Y. Chen and J. Xu, *Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices*, J. Mach. Learn. Res. **17** (2016), 27:1–27:57.

[DBV16] M. Defferrard, X. Bresson, and P. Vandergheynst, *Convolutional neural networks on graphs with fast localized spectral filtering*, Proceedings of the 30th International Conference on Neural Information Processing Systems (Red Hook, NY, USA), NIPS'16, Curran Associates Inc., 2016, p. 3844–3852.

[DKMZ11] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Physical review. E, Statistical, nonlinear, and soft matter physics **84 6 Pt 2** (2011), 066106.

[DLR77] A. P. Dempster, N. M. Laird, and D. B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the Royal Statistical Society: Series B (Methodological) **39** (1977), no. 1, 1–22, https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x.

[DLS21] S. Deng, S. Ling, and T. Strohmer, *Strong consistency, graph Laplacians, and the stochastic block model*, J. Mach. Learn. Res. **22** (2021), 117:1–117:44.

[EK10] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*, Cambridge University Press, 2010.

[EKPS00] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman, *Broadcasting on trees and the Ising model*, The Annals of Applied Probability **10** (2000), no. 2, 410 – 433.

[ER84] P. L. Erdos and A. Rényi, *On the evolution of random graphs*, Transactions of the American Mathematical Society **286** (1984), 257–257.

[FO05] U. Feige and E. O. Ofek, *Spectral techniques applied to sparse random graphs*, Random Struct. Algorithms **27** (2005), 251–275.

[GB13] P. K. Gopalan and D. M. Blei, *Efficient discovery of overlapping communities in massive networks*, Proceedings of the National Academy of Sciences **110** (2013), no. 36, 14534–14539, https://www.pnas.org/doi/pdf/10.1073/pnas.1221839110.

[Gro52] A. Grothendieck, *Résumé des résultats essentiels dans la théorie des produits tensoriels topologiques et des espaces nucléaires*, Annales de l'Institut Fourier **4** (1952), 73–112 (fr).

[GSR+17] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, *Neural message passing for quantum chemistry*, Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, JMLR.org, 2017, p. 1263–1272.

[GV14] O. Guédon and R. Vershynin, *Community detection in sparse networks via Grothendieck's inequality*, Probability Theory and Related Fields **165** (2014), 1025–1049.

[GW95] M. X. Goemans and D. P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM **42** (1995), no. 6, 1115–1145.

[GWG19] J. Gasteiger, S. Weißenberger, and S. Günnemann, *Diffusion improves graph learning*, Conference on Neural Information Processing Systems (NeurIPS), 2019.

[Hig77] Y. Higuchi, *Remarks on the limiting gibbs states on a (d+1)-tree*, Publications of The Research Institute for Mathematical Sciences **13** (1977), 335–348.

[HLL83] P. Holland, K. B. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Social Networks **5** (1983), 109–137.

[HS00] E. Hartuv and R. Shamir, *A clustering algorithm based on graph connectivity*, Information Processing Letters **76** (2000), no. 4, 175–181.

[HWX16] B. E. Hajek, Y. Wu, and J. Xu, *Achieving exact cluster recovery threshold via semidefinite programming*, IEEE Transactions on Information Theory **62** (2016), 2788–2797.

[HYL17] W. L. Hamilton, R. Ying, and J. Leskovec, *Representation learning on graphs: Methods and applications*, IEEE Data Eng. Bull. **40** (2017), no. 3, 52–74.

[JMRT16] A. Javanmard, A. Montanari, and F. Ricci-Tersenghi, *Phase transitions in semidefinite relaxations*, Proceedings of the National Academy of Sciences **113** (2016), E2218 – E2223.

[Joh67] S. C. Johnson, *Hierarchical clustering schemes*, Psychometrika **32** (1967), no. 3, 241–254.

[KBG19] J. Klicpera, A. Bojchevski, and S. Günnemann, *Predict then propagate: Graph neural networks meet personalized pagerank*, ICLR, 2019.

[KKM03] S. D. Kamvar, D. Klein, and C. D. Manning, *Spectral learning*, IJCAI, 2003.

[KMM+13] F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang, *Spectral redemption in clustering sparse networks*, Proceedings of the National Academy of Sciences **110** (2013), 20935 – 20940.

[KMO09] R. H. Keshavan, A. Montanari, and S. Oh, *Matrix completion from noisy entries*, J. Mach. Learn. Res., 2009.

[KS66] H. Kesten and B. P. Stigum, *A Limit Theorem for Multidimensional Galton-Watson Processes*, The Annals of Mathematical Statistics **37** (1966), no. 5, 1211 – 1223.

[KS03] M. Krivelevich and B. Sudakov, *The largest eigenvalue of sparse random graphs*, Combinatorics, Probability and Computing **12** (2003), 61 – 72.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, Communications of the ACM **60** (2012), 84 – 90.

[KW17] T. N. Kipf and M. Welling, *Semi-supervised classification with graph convolutional networks*, 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings, OpenReview.net, 2017.

[Lee13] D.-H. Lee, *Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks*, Workshop on challenges in representation learning, International Conference on Machine Learning, ICML '13, vol. 3, 2013, p. 896.

[LHW18] Q. Li, Z. Han, and X.-M. Wu, *Deeper insights into graph convolutional networks for semi-supervised learning*, AAAI, 2018.

[LLDM08] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, *Statistical properties of community structure in large social and information networks*, Proceedings of the 17th International Conference on World Wide Web (New York, NY, USA), WWW '08, Association for Computing Machinery, 2008, p. 695–704.

[LMBB19] R. Levie, F. Monti, X. Bresson, and M. M. Bronstein, *Cayleynets: Graph convolutional neural networks with complex rational spectral filters*, IEEE Transactions on Signal Processing **67** (2019), 97–109.

[LSY03] G. Linden, B. Smith, and J. York, *Amazon.com recommendations: item-to-item collaborative filtering*, IEEE Internet Computing **7** (2003), no. 1, 76–80.

[Mas14a] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '14, Association for Computing Machinery, 2014, p. 694–703.

[Mas14b] ———, *Community detection thresholds and the weak ramanujan property*, Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '14, Association for Computing Machinery, 2014, p. 694–703.

[MBB17] F. Monti, M. M. Bronstein, and X. Bresson, *Geometric matrix completion with recurrent multi-graph neural networks*, Proceedings of the 31st International Conference on Neural Information Processing Systems (Red Hook, NY, USA), NIPS'17, Curran Associates Inc., 2017, p. 3700–3710.

[McS01] F. McSherry, *Spectral partitioning of random graphs*, Proceedings 42nd IEEE Symposium on Foundations of Computer Science, 2001, pp. 529–537.

[MNS15] E. Mossel, J. Neeman, and A. Sly, *Reconstruction and estimation in the planted partition model*, Probability Theory and Related Fields **162** (2015), no. 3, 431–461.

[MNS18] E. Mossel, J. Neeman, and A. Sly, *A proof of the block model threshold conjecture*, Combinatorica **38** (2018), 665–708.

[Mos01] E. Mossel, *Survey: Information flow on trees*, Graphs, Morphisms and Statistical Physics, 2001.

[MPW16] A. Moitra, W. Perry, and A. S. Wein, *How robust are reconstruction thresholds for community detection?*, Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '16, Association for Computing Machinery, 2016, p. 828–841.

[MS16] A. Montanari and S. Sen, *Semidefinite programs on sparse random graphs and their application to community detection*, Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing (New York, NY, USA), STOC '16, Association for Computing Machinery, 2016, p. 814–827.

[NG04] M. E. J. Newman and M. Girvan, *Finding and evaluating community structure in networks.*, Physical review. E, Statistical, nonlinear, and soft matter physics **69 2 Pt 2** (2004), 026113.

[NN12] R. R. Nadakuditi and M. E. J. Newman, *Graph spectra and the detectability of community structure in networks*, Physical review letters **108 18** (2012), 188701.

[NWS02a] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, *Random graph models of social networks*, Proceedings of the National Academy of Sciences **99** (2002), no. suppl_1, 2566–2572, https://www.pnas.org/doi/pdf/10.1073/pnas.012582999.

[NWS02b] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, *Random graph models of social networks*, Proceedings of the National Academy of Sciences of the United States of America **99** (2002), 2566 – 2572.

[Pad14] B. Padhukasahasram, *Inferring ancestry from population genomic data and its applications*, Frontiers in Genetics **5** (2014), 204.

[Pia38] H. T. H. Piaggio, *Introduction to mathematical probability. by j. v. uspensky. pp. ix, 411. 30s. 1937. (mcgraw-hill)*, The Mathematical Gazette **22** (1938), no. 249, 202–204.

[PW17] W. Perry and A. S. Wein, *A semidefinite program for unbalanced multisection in the stochastic block model*, 2017 International Conference on Sampling Theory and Applications (SampTA) (2017), 64–67.

[QCMC19] Z. Qiu, E. Cho, X. Ma, and W. M. Campbell, *Graph-based semi-supervised learning for natural language understanding*, EMNLP, 2019.

[SBHHW03] N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall, *Computing gaussian mixture models with EM using equivalence constraints*, NIPS, 2003.

[SM00] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000), no. 8, 888–905.

[SNB+08] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad, *Collective classification in network data*, AI Magazine **29** (2008), no. 3, 93.

[Spi75] F. Spitzer, *Markov random fields on an infinite tree*, The Annals of Probability **3** (1975), no. 3, 387–398.

[Str01] S. H. Strogatz, *Exploring complex networks*, Nature **410** (2001), no. 6825, 268–276.

[Tao11] T. Tao, *Topics in random matrix theory*, American Physical Society, 2011.

[Tsy09] A. B. Tsybakov, *Introduction to nonparametric estimation*, Springer series in statistics, 2009.

[VCC+18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, *Graph attention networks*, International Conference on Learning Representations, 2018.

[vdMH08] L. van der Maaten and G. Hinton, *Visualizing data using t-sne*, Journal of Machine Learning Research **9** (2008), no. 86, 2579–2605.

[Vu07] V. H. Vu, *Spectral norm of random matrices*, Combinatorica **27** (2007), 721–736.

[WCRS01] K. L. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, *Constrained k-means clustering with background knowledge*, ICML, 2001.

[Web18] S. Webb, *Deep learning for biology*, Nature **554** (2018), 555–557.

[XHLJ19] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, *How powerful are graph neural networks?*, International Conference on Learning Representations, 2019.

[XHLL20] Q. Xie, E. H. Hovy, M.-T. Luong, and Q. V. Le, *Self-training with noisy student improves ImageNet classification*, 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020), 10684–10695.

[YBLG17] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, *Local higher-order graph clustering*, Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (New York, NY, USA), KDD '17, Association for Computing Machinery, 2017, p. 555–564.

[YHC+18] R. Ying, R. He, K. Chen, P. Eksombatchai, W. L. Hamilton, and J. Leskovec, *Graph convolutional neural networks for web-scale recommender systems*, Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018 (Y. Guo and F. Farooq, eds.), ACM, 2018, pp. 974–983.

[ZMZ14] P. Zhang, C. Moore, and L. Zdeborová, *Phase transitions in semisupervised clustering of sparse networks*, Phys. Rev. E **90** (2014), 052802.