Topics in Graph Clustering

Yali Wan

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2017

Reading Committee:

Marina Meila, Chair

Elena A. Erosheva

Maryam Fazel

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Topics in Graph Clustering

Yali Wan

Chair of the Supervisory Committee:
Professor Marina Meila
Department of Statistics

In this thesis, two problems in social networks will be studied.

In the first part of the thesis, we focus on community recovery problems for social networks. There have been many recent theoretical advances in the model-based community recovery for network data. In the center of it are the Stochastic Block Model (SBM) and its extension, Degree Corrected Stochastic Block Model (DC-SBM). Under assumptions on the balance and separation of clusters, theoretical guarantees have been provided to ensure the recovery of the true clusters with high probability.We firstly benchmark the current recovery theorems on DC-SBM through experimental approaches. The experiments suggest that there are still lots of cases that are recoverable but not predicted by the current reovery theorems. We then introduce a wider class of network models called Preference Frame Model. We show that under weaker assumptions, the communities or clusters can be recovered by spectral clustering algorithm with essentially the same guarantees. The model-based results, despite their importance, are limited by a strong and difficult-to-verify assumption that the observed data are generated from the model. We present the model-free community recovery, where we do not make assumptions about the data generating process and provide theoretical guarantees for the performance of the model based clustering algorithms in this framework.

In the second part of the thesis, we propose a perturbation framework to measure the robustness of graph properties. Although there are already perturbation methods proposed to tackle this prob-

lem, they are limited by the fact that the strength of the perturbation cannot be well controlled. We firstly provide a perturbation framework on graphs by introducing weights on the nodes, of which the magnitude of perturbation can be easily controlled through the variance of the weights. Meanwhile, the topology of the graphs are also preserved to avoid uncontrollable strength in the perturbation. We then extend the measure of robustness in the robust statistics literature to the graph properties.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Above all, I would like to thank my advisor Marina Meila, who is a truly kind person and the best advisor one can imagine. I have learnt much more than doing research from her. I would also like to thank my undergraduate advisor Mu Zhu from university of Waterloo, who enlightened me and directed me to the field of machine learning.

I must also thank my friends for their friendship. Lina Lin, my office mate and English writing consultant. Hongyu Xu, my best friend since high school. Chenglin Wei, my friend and greatest support. I am more than lucky to have you in my life.

Lastly, this is dedicated to my parents, who have never doubted that I could reach this point.

DEDICATION

To My Parents

Chapter 1

INTRODUCTION

1.1 Overview

This thesis is about the study of social networks. Throughout the thesis, we treat social networks as graphs and focus on studying their graph properties. The study of social networks has aroused lots of attention lately. One of the major interest is on finding the guarantees for the recovery of community structure and evaluating the stability of the graph properties in social networks. Current work has been focused on model-based community recovery. Particular interests are in Stochastic Block Model (SBM) and its extension, Degree Corrected Stochastic Block Model (DC-SBM). In this work, we firstly benchmark the current community recovery theorems empirically. We then present a new network model called Preference Frame Model (PFM), which includes the current block models. We provide the recovery guarantees under this model framework. We then talk about the model-free community recovery, in which we provide theoretical guarantees for the results of model-based clustering algorithm without making assumptions on data generation process. At last, we propose a perturbation framework to evaluate the robustness of graph properties, including clustering, the number of weakly connected components, etc.

Chapter 2: We define thresholds as the conditions in community recovery theorems that are sufficient and necessary. While one is close to establishing thresholds for recovery in SBM [1, 34, 37], the recovery thresholds are not known yet for the more general DC-SBM. Recent results for the recovery under DC-SBM consist of *sufficient but not necessary* conditions, therefore one would expect there exist cases that can be recoverable but do not satisfy the assumptions in the recovery theorems.

In this chapter, we generate benchmark networks which are empirically verified to be recoverable with zero or low error by spectral clustering algorithm. The difficulties of these cases are controlled in ways including: balance of degree distribution, cluster separation, graph density and etc.. We further verify whether the recovery of these cases are predicted by the theorems. If it is hard to find examples not predicted by theory, we may conclude that the theory is strong. Oth-

erwise, it suggests that there is much more ground to cover towards deriving sharp thresholds for community recovery in DC-SBM model.

The recovery theorems we compare include all the theorems published up to 2015, including[52, 22, 57, 33, 11, 18, 56, 62]. The benchmarking experiments suggest that current results are not close to the yet unknown thresholds for recovery and our understanding of the problem is not complete. We create a software tool for researchers, which is written with accessibility and modularity in mind. It allows them to create test cases of controlled difficulty and can be easily extended to test and compare new recovery theorems.

Chapter 3: Under the "block-model" assumptions, most of the advances in recovery of communities are done by spectral clustering. In this part of our work, we introduce a wider class of network models, called Preference Frame Models (PFM), which subsumes many current models including SBM and DC-SBM. We show that under PFM model assumption, the communities can be recovered by spectral clustering algorithm with the same guarantees.

PFM models have merits in several ways. Firstly, under weaker assumptions, the communities are proved to be recoverable with the same guarantees by spectral clustering algorithm. Secondly, it subsumes the current models including SBM and DC-SBM, meanwhile being more flexible by including more degrees of freedom. Thirdly, it allows for clearer and more intuitive parametrization. Meaningful parameters can be specified independently for the construction of a PFM.

Chapter 4: The results for the model-based clustering in networks, despite its fast advances, are limited by the assumption that the observed data come from a model. In this work, we propose a framework to provide *theorectical guarantees for the results of model-based clustering algorithms, without making any assumption about the data generation process*.
In the framework, we use the model's goodness of fit and prove that a very good fit implies the clustering obtained is unique up to small perturbation, and can be said to capture the data structure

in a meaningful way. We instantiate the framework by obtaining model-free guarantees for SBM and PFM models.

This work makes several contributions. Firstly, it formulates the problem of model free validation in the area of community detection in network. Secondly, all quantities in our theorems are computable from the data and the clustering $\mathcal{C}$ without undertermined constants or not available parameters. Thirdly, several model-specific stability results are obtained from simpler, more direct and more elementary techiniques in the proof. They can not only be used in the model-free recovery guarantees, but also to improve current model-based recovery theorems, where sharper thresholds can be obtained by these techniques.

Chapter 5: In this chapter, we aim at providing a perturbation framework for measuring the robustness of graph properties, such as clustering, weighted cut, the number of weakly connected components, etc. We evaluate the robustness of graph properties by measuring their sensitivity to small perturbations on the graph. In the current literature [30, 24, 14, 4], the perturbation of networks is mostly done through resampling the edges or randomly adding/deleting the nodes, where the graph topology is changed and the strength of perturbation is hard to control. Moreover, although there are already measures of robustness in the robust statistics literature, they are not applicable to the graph context since they mostly require known data distribution and i.i.d assumptions.

In our approach, we firstly propose a perturbation method, which is done through adding weights the nodes and edges. In particular, it allows us to control the strength of the perturbation smoothly and preserves the topology of the graphs. We then extend the classcial measures of robustness including Influence Function (IF) and Breakdown Point (BP) to the context of graph, and measure the robustness in graph properties.

This work has several advantages over the current literature. First, it provides a new way of

perturbation on social networks, which allows for perturbation that is easy to control and preserves the graph topology at the same time. Second, we introduce new techniques for the evaluation of perturbation. Although they have been widely used in the conventional statistics, Employing these tools in the graph properties is the first time. Third, our framework allows for deeper insight into the graph properties by evaluating the source of robustness at the node level.

1.2 Main concepts

1.2.1 Graphs, degrees, Laplacian, and clustering

Let $\mathcal{G}$ be a graph on n nodes, described by its *Similarity Matrix* A. A_{ij} describes the edge probability between node i and j. Define $d_i = \sum_{j=1}^{n} A_{ij}$ the *degree* of node i, and $D = \mathrm{diag}\{d_i\}$ the diagonal matrix of the node degrees. The *(normalized) Laplacian* of $\mathcal{G}$ is defined as[1] $L = D^{-1/2}AD^{-1/2}$. In extension, we define the *degree matrix* D and the Laplacian L associated to any matrix $A \in \mathbb{R}^{n\times n}$, with $A_{ij} = A_{ji} \geq 0$, in a similar way.

Let $\mathcal{C} = (\mathcal{C}_1, \cdots, \mathcal{C}_K)$ be a partitioning (clustering) of the nodes of $\mathcal{G}$ into K clusters. We use the shorthand notation $i \in k$ for "node i belongs to cluster k". We will represent $\mathcal{C}$ by its $n \times K$ *indicator matrix* Z, defined by

$$Z_{ik} \ = \ 1 \text{ if } i \in k, \text{ 0 otherwise, for } i = 1, \dots \ n, \ k = 1, \dots \ K. \tag{1.1}$$

Note that $Z^T Z = \mathrm{diag}\{n_k\}$ with n_k counting the number of nodes in cluster k, and $Z^T AZ = [n_{kl}]_{k,l=1}^{K}$ with n_{kl} counting the edges in $\mathcal{G}$ between clusters k and l. Moreover, for two indicator matrices Z, Z' for clusterings $\mathcal{C}, \mathcal{C}'$, $(Z^T Z')_{kk'}$ counts the number of points in the intersection of cluster k of $\mathcal{C}$ with cluster k' of $\mathcal{C}'$, and $(Z^T DZ')A_{kk'}$ computes $\sum_{i\in k\cap k'} D_i$ the volume of the same intersection.

In order to measure the distance between clusterings, we employ two measures as follows. The *Misclassification Error (ME) distance* between two clusterings $\mathcal{C}, \mathcal{C}'$ over the same set of n points

[1] Rigorously speaking, the normalized graph Laplacian is $I - L$ [20].

is

$$\mathrm{dist}(\mathcal{C}, \mathcal{C}') \;=\; 1 - \frac{1}{n} \max_{\pi \in \mathbb{S}_K} \sum_{i \in k \cap \pi(k)} 1, \tag{1.2}$$

where π ranges over all permutations of K elements $\mathbb{S}_K$, and $\pi(k)$ indexes a cluster in $\mathcal{C}'$. If the points are weighted by their degrees, a natural measure on the node set, the *Weighted ME (wME) distance* is

$$\mathrm{dist}_d(\mathcal{C}, \mathcal{C}') \;=\; 1 - \frac{1}{\sum_{i=1}^{n} d_i} \max_{\pi \in \mathbb{S}_K} \sum_{i \in k \cap \pi(k)} d_i \,. \tag{1.3}$$

While dist is more popular, we believe dist_d is more natural, especially when node degrees are dissimilar, as d can be seen as a natural measure on the set of nodes, and dist_d is equivalent to the *earth-mover's* distance. When all the nodes have the same degree, dist_d reduces to dist.

1.2.2 "Block models" for random graphs (SBM, DC-SBM)

This family of models contains Stochastic Block Models (SBM) [1, 57] and Degree-Corrected SBM (DC-SBM) [56]. They have been widely employed as canonical models to study clustering and community recovery. Under each of these model families, a graph $\hat{\mathcal{G}}$ with adjacency matrix $\hat{A}$ over n nodes is generated by sampling its edges *independently* following the law $\hat{A}_{ij} \sim Bernoulli(A_{ij})$, for all $i > j$. The two model families differ in the constraints they put on an acceptable A. Let $\mathcal{C}^*$ be a clustering. The entries of A are defined w.r.t $\mathcal{C}^*$ as follows (and we say that A is *compatible* with $\mathcal{C}^*$).

SBM : $A_{ij} = B_{kl}$ whenever $i \in k, j \in l$, with $B = [B_{kl}] \in \mathbb{R}^{K \times K}$ symmetric and non-negative.

DC-SBM : $A_{ij} = w_i w_j B_{kl}$ whenever $i \in k, j \in l$, with B as above and $w_1, \ldots, w_n$ non-negative weights associated with the graph nodes.

While perhaps not immediately obvious, the SBM is a subclass of the DC-SBM, when $w_i = 1$, $i = 1, \ldots, n$. The number of parameters in these models is respectively $\mathcal{O}(K^2), \mathcal{O}(n + K^2)$,. Another common feature of block-models, that will be significant throughout this work is that for

both, Spectral Clustering algorithms have been proved to work well for estimating $\mathcal{C}^*$ [52].

In the rest of this thesis, without further specification, we define A, L, d, D, as properties of the model class. The above quantities will be marked with a symbol $\hat{}$ when they are properties of the observed graphs.

1.2.3 Community Recovery

The main research question in community recovery, and in particular regarding the SBM/DC-SBM is this: Given a simple undirected graph $\mathcal{G}$ on n nodes, with adjacency matrix $\hat{A}$, sampled from an unknown SBM/DC-SBM, can we estimate the clustering $\mathcal{C}$ and model parameters (w, B)?[2] It is evident that the crux of the problem is finding $\mathcal{C}$. Once this is known, the parameters (w, B) can be estimated from $\hat{A}$ and $\mathcal{C}$ by the Maximum Likelihood method. Thus, my thesis as well as most results in the literature focus on the *community recovery problem*. [19] establishes a scale of definitions of "recovery"; in particular, *strong* (or *exact*) recovery denotes identifying $\mathcal{C}$ exactly; *weak* recovery denotes estimating $\mathcal{C}$ with an error err[3] of order $o(n)$; *partial* recovery (or *detection*) denotes finding $\mathcal{C}$ with $err < 1/2$. The most promising in real applications is the weak recovery; therefore in the rest of the thesis the term "recovery" will be understood to mean "weak recovery". This is also the scenario under which most results about the SBM/DC-SBM are obtained.

---

[2]It is standard to assume that K the number of clusters is known; however, several notable recovery algorithms do not require knowing K.

[3]The recovery error between the true $\mathcal{C}$ and the estimated $\hat{\mathcal{C}}$ is defined as $err = 1 - \frac{1}{n} \max_{\phi:[K]\to[K]} \sum_k |C_{\phi(k)} \cap \hat{C}_k|$.

Chapter 2

BENCHMARKING RECOVERY THEOREMS FOR THE DC-SBM

There have been many recent theoretical advances in the recovery of communities from random graphs, under the assumptions of the so called "block models". For the degree corrected stochastic block model (DC-SBM), we have witnessed a series of recent results consisting of *sufficient conditions for recovery*, often by spectral algorithms. Since the conditions are not necessary, one expects that cases exist where recovery is possible, but which are not covered by the theory. In this chapter, we explore experimentally the limits of the current theories. We generate *benchmark cases*, for which recovery is possible, and with well defined parameters controlling their difficulty for recovery. Then we verify which of the existing results in the literature predict recovery. If it is hard to find examples not predicted by theory, then we may conclude that the theory is strong. This is not what our experiments show. On the contrary, the experiments suggest that there is much more ground to cover towards deriving sharp thresholds for community recovery in the DC-SBM model. The software tool we created is made publicly available as a tool for researchers. It allows them to create test cases of controlled difficulty, and can easily be extended to test and compare new recovery theorems as they are published.

2.1 Motivation

Network modeling for the purpose of community recovery has attracted intense interest in the last decade [28, 32, 29, 27, 25, 63]. More recently we have been witnessing rapid progress and a surge of novel theoretical results on recovery algorithms *with recovery guarantees*, under some modeling assumptions.

At the center of these results are two familiar models for graphs with communities, the Stochastic Block-Model (SBM) [28] and its extension Degree-Corrected SBM (DC-SBM) of [31]. For the SBM, one is close to establishing the sufficient and necessary conditions, or we call the thresholds, for recovery in various regimes, due to the pioneering work of [48, 49, 1].

For the more general DC-SBM, the recovery thresholds are not known yet. The recent progress has been in obtaining recovery guarantees under weaker and weaker conditions on the model pa-

rameters. One drawback of the present results lies in how complicated these conditions are. They are not easy to parse and have implicit dependencies on each other. This makes it hard to understand the space in which the conditions are satisfied, or to compare conditions between different papers and methods. Our work offers an empirical tool for the theoreticians: a software package that generates benchmark graphs with user controlled parameters and performs the numerical verification of the various recovery conditions on these graphs for the existing results in the literature.

We proceed as follows: we generate random graphs from the DC-SBM model, for which we verify empirically by spectral clustering that the original clustering is recoverable with low or zero error. The DC-SBM models that we generate graphs from have various difficulty in finding the clustering, which is controlled by its parametrization. Then we check if the theoretical recovery conditions from the papers under consideration hold and discuss the findings. The code we use is organized with modularity and extensibility in mind, so that it can be reused as new results are published.

While no empirical verification can be complete, we expect to get partial information about a few questions that are tedious or unresolved theoretically, such as which theorems cover more recoverable cases, and which conditions are most restrictive on the test matrices? Seen this way, our work is one of *benchmarking*. While most benchmarks are constructed for algorithms, ours is for theorems. Benchmarking and competitions are recognized as drivers of progress in other areas of research. We expect that this benchmarking exercise will also stimulate and guide useful research in community detection.

We also hope to uncover, by varying the parameters of the network models, what are the trade-offs among the various conditions of the recovery theorems. For example, how does controlling the degree imbalance gives us more freedom with cluster separability?

Third, in the case of the asymptotic results, or of those with unspecified constants, the experi-

ments allow us to estimate the graph sizes where the asymptotic regime starts. And, by calculating the asymptotic bounds, we get an idea of the unspecified universal constants. In this way, we obtain an intuitive understanding of the stringency of conditions. These findings reveal the gap between theory and actual recovery. Throughout our experiments, we focus on recovery algorithms for DC-SBM. The theorems we compare are coming from the papers [21, 57, 10, 17, 55, 62]. We also include papers on spectral clustering [52, 8] which provide recovery guarantees and are compatible with our experimental setup.

2.2 Background: community recovery problem and sparsity regimes

It is well known that the Erdös-Renyi (ER) random graph model [23] exhibits a phase transition at $p_c = (\ln n)/n$. In an ER graph, the edges between two nodes are generated independently with probability p. An ER graph with $p < p_c$ is almost surely disconnected. For the DC-SBM this has the implication that if $\max_{i,j \in C_k} w_i w_j B_{kk} < (\ln n_k)/n_k$ for some cluster C_k, the cluster will be irrecoverable (according to our definition). This also points to a separation of the recovery problems by *sparsity regimes*, with the sparser regimes being more difficult than the denser ones. As a reminder, the definition of community recovery can be found in Chapter 1.2.3. In the *dense* regime, according again to the classification of [19] the minimum node degree d_{min} is $\Omega(n)$, in the *sparse* regime $d_{min} = \Omega(\ln n)$ and the *very sparse* regime is defined by $d_{min} = o(\ln n)$ or $d_{min} = O(1)$. Currently, interest has been shifting towards obtaining recovery guarantees in the sparsest possible regime, therefore we will focus our experiments on the sparse regime, in which most results of our knowledge are proved. We note however that several more recent results break the sparsity barrier, by proving recovery is possible when only a limited number of node degress are below the $\ln n$ threshold [21, 55].

In summary, while the exact conditions for recovery in the DC-SBM are not known, there sustained interest in the problem has generated many *sufficient recovery conditions*. These control, besides the sparsity of the graph, other properties such as the distribution of node weights $w_{1:n}$ (or the degree distribution) – a more uniform distribution within a cluster promising easier recovery,

the distribution of cluster sizes – a more uniform distribution being "easier", and the "separation" of the clusters – better separated clusters begin "easier". The last property is formulated in the literature in a variety of ways, and in the Results section we will highlight a few of them. For instance, in [10] "separation" means few inter-cluster edges, while in [57] it means B is close to diagonal.

Comparing to SBM model, DC-SBM allows for more degree of freedoms, and can represent a wider class of network models. Therefore, researches on recovery theorems of DC-SBM model are well worth attention. Recent discoveries on community recovery generally put assumptions on similarity matrices and parameters, and provide upper bounds of the misclassification error. These theorems are mostly guaranteed with performance in community recovery through spectral clustering algorithms. The theorems we compare come from [Wan and Meila, 2015], [Qin and Rohe, 2013], [Rohe, Chatterjee, Yu, 2011], [Balcan, Borgs, Braverman, Chayes, 2012], [Balakrishnan, Xu, Krishnamurthy, Singh, 2011], [Ng, Jordan, Weiss, 2002], [Coja-Oglan, Lanka 2010], [Chaudhui, Chung, Tsiatas, 2012]. We include these theorems in Sections 2.6.Conditions for recovery in the DC-SBM are not known, the sustained interest in the problem has generated many *sufficient recovery conditions* [21, 57, 10, 17, 55, 62].

Weak recovery was shown to be possible [21, 57, 10, 55, 62] in the *dense* and *sparse regimes*[1], according again to the classification of [19]. These regimes are defined based on the minimum expected degree $d_{min} = \min d_{1:n}$, where $d_i = \sum_j A_{ij}$ is the expected degree of node i, and correspond respectively to $d_{min} = \Omega(n)$ and $d_{min} = \Omega(\ln n)$.

The papers cited above vary in the conditions they require to guarantee recovery, due to using different combination of parameters, and sometimes different algorithms. But their requirements lie in four main categories: (1), good separation between the communities, which can be inter-

[1]We note however that several more recent results break the sparsity barrier, by proving recovery is possible when only a limited number of node degress are below the $\ln n$ threshold [21, 55].

preted as the near block-diagonality of A; (2), the density of the graphs cannot be too low; (3), the degree distribution within clusters needs to be balanced; and (4), the cluster sizes are also required to be balanced.

Yet, the variations in the conditions make it hard to compare the stringency of the assumptions among different papers. Even the domains of applicability of these assumptions are inexplicit. For instance, it is not always explicit at what values of n some of the asymptotic conditions start holding . And more generally, it is not known how large is the gap between recoverability and the existing conditions. Therefore, this paper sets out to probe the state of the art theorems through an experimental approach.

2.3 Experiment design

Our experiment is designed as follows. First, we define a range of parameters controlling the difficulty of the benchmark problems. For each combination of parameters in this range, we

1. Generate a benchmark DC-SBM and its A matrix, according to Algorithm 1.

   Sample an adjacency matrix $\hat{A}$ from A (multiple times).

2. For each paper and for each condition in it

   Verify if the condition holds for the current model and $\hat{A}$.

By construction, all A matrices are perfectly clusterable by standard spectral clustering [62, 57, 52]. In addition, we verify for each $\hat{A}$ that the clusters can be recovered with small error.

2.3.1 Generating the benchmark matrices

All DC-SBMs we generate have $K = 5$ clusters, and number of nodes $n = 300, \ldots 30000$. Node weights in each cluster are sampled from the same Generalized Pareto ($GPareto$) distribution $GP(\mu = 0, \epsilon = 1, \sigma)$. The remaining input parameters are:

- cluster relative sizes: `balanced` (all equal) or `unbalanced` (in ratio $4:7:10:13:16$)
- the cluster level relations are parametrized by the *spectrum* $\Lambda = (\lambda_1 = 1 > \lambda_2 \geq \ldots \geq \lambda_K > 0)$ and the distribution $\rho = (\rho_1, \ldots \rho_K)$[2]
- weight distribution: `balanced` ($\sigma_l = 0$), `perturbed` ($\sigma_l = 0.01$), `unbalanced` ($\sigma_l = 0.1$) with respect to ρ (see Algorithm 1)

Input : K spectrum, Λ, ρ, cluster sizes $n_{1:K}$, σ_l

Output: A

1. Set $u = \sqrt{\rho}$, create $U = [u\, u_1\, \ldots u_{K-1}]$ orthogonal matrix, compute $B = (\operatorname{diag} u)^{-1} U \Lambda U^T (\operatorname{diag} u^{-1})$.

2.For $l = 1, \ldots K$

2.1. Sample weights in cluster $\mathcal{C}_l \sim GPareto(k = 0, \sigma = 1, \mu = 1)$.

2.2. Normalize the weights to sum to $W_l = \rho_l + s_l$ with $s_l \sim unif(-\sigma_l, \sigma_l)$.

3. Construct A using $A_{ij} = w_i w_j B_{kl}$. Normalize A by $\max_{ij} A_{ij}$.

Algorithm 1: Construction of the DC-SBM benchmark matrices.

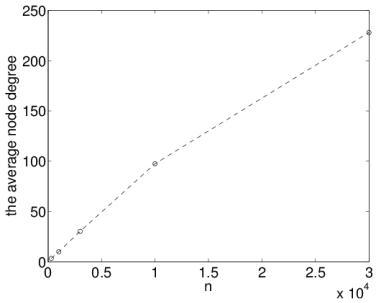
The parameters Λ control how separate the clusters are via the value λ_K, known as the *eigengap*. It was shown in [45, 57] that for A constructed as above, the *Laplacian* matrix $L = D^{-1/2} A D^{-1/2}$, which plays a central role in spectral clustering of graphs, has exactly K non-zero eigenvalues given by Λ. We do the experiments with three different sets of eigenvalues Λ, having $\lambda_K = 0.01, 0.4, 0.99$ respectively; the last value corresponding to an almost exactly block diagonal matrix A.

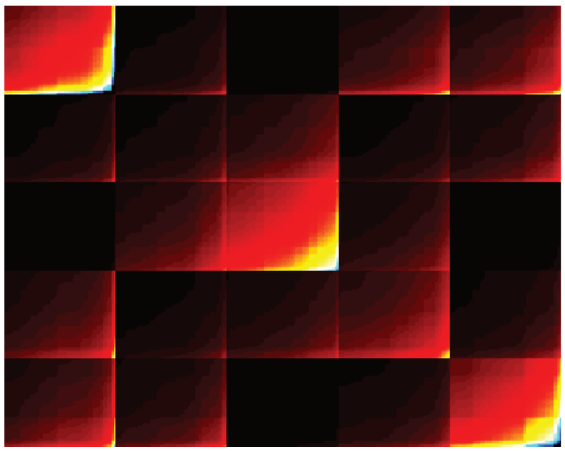
The variation from balanced to unbalanced degree distribution is further controlled by the magnitude of noise added to the cluster weight volume. In Figure 2.1 we exemplify the B and A matrices we generate. The figure also shows that in this simulation, d_i increases linearly with respect to n.

[2] In [62] the role of ρ is explained in detail. Essentially, ρ is a "default" cluster size distribution.

$$\Lambda = \begin{bmatrix} 1.00 & .75 & .50 & .26 & .01 \end{bmatrix}$$

$$B = \begin{bmatrix} 2.33 & .46 & .03 & .94 & .64 \\ .46 & 1.38 & 1.65 & .76 & 1.02 \\ .03 & 1.65 & 3.17 & .71 & .04 \\ .94 & .76 & .71 & 2.56 & .48 \\ .64 & 1.02 & .04 & .48 & 2.59 \end{bmatrix}$$





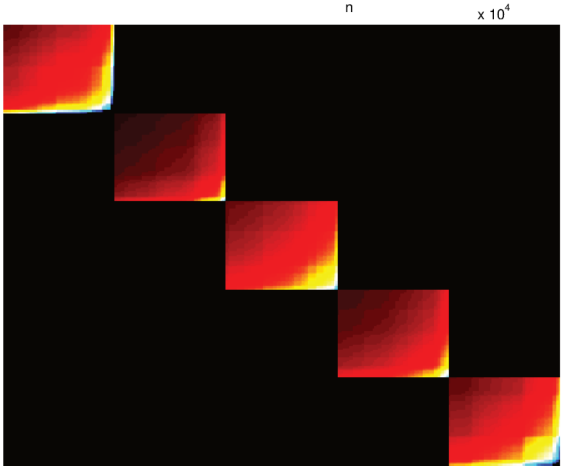


Figure 2.1: Left: example of parameters and benchmark models Λ, B, A. Note that the order of the clusters in A is not the same as in B. In particular, the first row of B corresponds to the second cluster in A, this cluster has stronger links to another cluster than within itself. Top right: the average degree versus n. Bottom left: the easiest A with $(\lambda_K = 0.99)$. Bottom right: the hardest A with $(\lambda_K = 0.01)$

2.3.2 Checking the conditions

For the main recovery theorem in each paper we check for each test case if the conditions of the theorem are satisfied. Table 2.1 describes the specific conditions we tested in each recovery theorem, for a total of about 30 conditions. For the spectral clustering papers, A is treated like the adjacency matrix of a weighted graph, in other words as a *similarity* matrix, and consequently we check the recovery conditions directly on A, without sampling.

Some results, such as [21, 52], depend on unspecified constants; in such cases, we calculate upper or lower bounds on these constants from the data and check if the interval obtained is non-

Paper	**Theorem**	**Conditions** (in theorems)
Balakrishnan et al [9]	Thm 1 and 2	Assumptions 1-3
Check the hierarchical structure		
Balcan et al [11]	Thm. 3.1 and 4.1	Definition 1
Check the self-determined structure. Restrictions on cluster separation		
Coja-Oglan, et al [21]	Thm. 1	C0 - C5
Restrictions on density of the graph		
Chaudhuri et al [17]	Thm. 3	Assumption 1-5
Restrictions on the balance of the node degree distribution and density of the graph		
Ng&Jordan&Weiss [52]	Thm. 2	A1-A5
Restrictions on cluster separation		
Qin&Rohe [55]	Thm. 4.4	(a-b)
Restrictions on cluster separation and density of the graph		
Rohe, Chatterjee&Yu [57]	Thm. 3.1	Equations (1-2)
Restrictions on cluster separation and density of the graph		
Wan&Meila [62]	Thm. 3	Assumption 3-6
Restrictions on the balance of the node degree and density of the graph		

Table 2.1: Theorems and conditions tested.

empty. The next section describes the conditions in more detail and summarizes our findings.

2.4 Results

Throughout our simulation, we find that most of the papers fail to cover even the easiest cases where clusterings can be recovered with no error with spectral clustering algorithm; the exceptions are [62] and [10]. The other papers approach the satisfaction of the conditions as the parameters are tuned towards their favor.

Wan&Meila [62] Since A is generated from the DC-SBM model , Assumptions 1, 2 and 5 from the theorem hold immediately, in which they require bounds in A_{ij} and d_i. From Figure 2.2, we observe that as n gets larger, Assumptions 3 and 4 hold more often, where they require the minimum node degree in both the model and the observed graph to be lower bounded by $log(n)$. This is because the node degree which grows with n is faster than $\log(n)$ in the assumptions. Another interesting fact is that these two assumptions are easier to be satisfied by harder cases where A is less block-diagonal, because as the off-diagonal entries of A increase, the magnitudes of A spread out, as a result the node degree increases above the assumption thresholds. Assumption 6 is highly associated with the balance of the degree distribution across various clusters. It can tolerate slight perturbation to the degree distribution but fails with significant unbalance. Comparing the balanced and unbalanced cluster size setting from Figure 2.2 and 2.3, unbalanced cases violate Assumption 3 and 4 more often, which may be satisified with larger n, but the computational limits prevent us from exploring larger graphs. Assumption 6 stays the same.

Balcan et al [10] The main assumption proposed by this theorem is that the clusters are self-determined communities, i.e. each node is more connected to the nodes within the same cluster than outside. Because node weights are unequal, this is hard to satisfy for low weight nodes, occurring only in the cases when λ_K is almost 1, and A is almost block-diagonal.

Coja-Oglan, et al [21] Assumptions $C0$, $C1$, $C5$ automatically hold from the DC-SBM model configuration. This model has several dependent assumptions.

$$C2\text{: } w_i \leq n^{1-\epsilon}, \forall i \in V \tag{2.1}$$

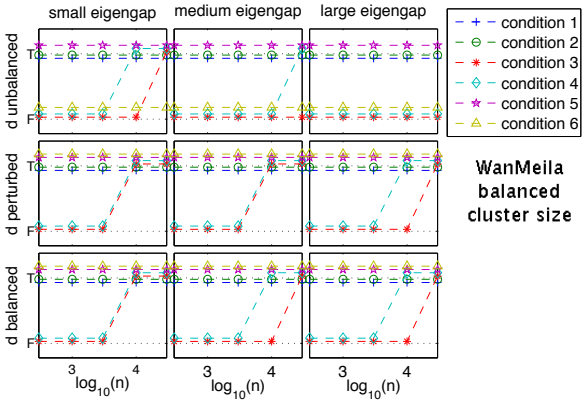


Figure 2.2: Results for Wan&Meila with balanced cluster sizes. For each of the 6 conditions, T (F) indicates whether the condition is true (false).

The range of n	$mean(\frac{min(\bar{w})}{max(\bar{w})})$
{1000, 3000, 10000, 30000}	0.08
{300, 1000, 3000, 10000}	0.07
{300, 1000, 3000, 10000, 30000}	0.03

Table 2.2: Changes in the range of D over different n ranges for Coja-Oglan, et al [21]

$$C3\text{: } w_i \geq \epsilon\bar{w}, \forall i \in V, \bar{w} = \sum w_i/n \tag{2.2}$$

$$C4\text{: } \bar{w} \geq D > 0 \tag{2.3}$$

We define $D = \bar{w}$. We use $C2$ and $C3$ to normalize the w and fix ϵ. We then record $\bar{w}$ as the maximum value of the unknown D. Since D should be independent of n, we examine its range over various sets of n. From table 2.2, we observe that $\bar{w}$ decreases significantly as the range of n increases. This suggests that asymptotically the value of D, if it exists, could be very small.

Chaudhuri et al [17] The paper assumes the extended planted partition model, which is more restricted than DC-SBM by reducing $K \times K$ parameters in B to only 2 parameters $p = B_{kk}$ and

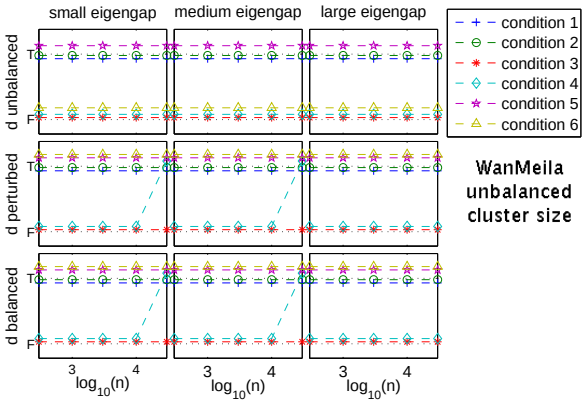


Figure 2.3: Results for Wan&Meila with unbalanced cluster sizes. For each of the 6 conditions, T (F) indicates whether the condition is true (false).

$q = B_{kl},\ k \neq l, p > q$ (we do not test for this condition). Since it assumes simpler model structure, the other assumptions for recovery should be easier to satisfy.

Assumption 1 requires the balance of cluster sizes; Assumptions 2-4 put restrictions on the degree distribution, among which Assumption 4 is the hardest to satisfy, since it requires the degree distribution having small variance in a squared degree setting normalized by the average degree. We denote as Assumption 5 the extra assumption inside Theorem 3:

$$E[d_i] \geq \frac{128}{9} \ln(6n/\delta), \text{for } i \in V \tag{2.4}$$

where $\delta \ll 1$ is associated with the probability of success. From Figure 2.4, we see that Assumptions 1-3 hold and Assumptions 4-5 fail regardless of the value of n. We further plot a critical value coming from equation (2.4)

$$1 - \min(d_i) / \left[128/9 \ln(6n/\delta)\right], \tag{2.5}$$

which should be negative for Assumption 5 to hold. This value is getting smaller when n gets larger, which indicates that Assumption 5 will hold when n is sufficiently large.

Qin&Rohe [55] Assumption (a) lower bounds the smallest eigenvalue. Assumption (b) lower bounds the expected node degree. From Figure 2.1 and Figure 2.5, we observe that, as n increases,

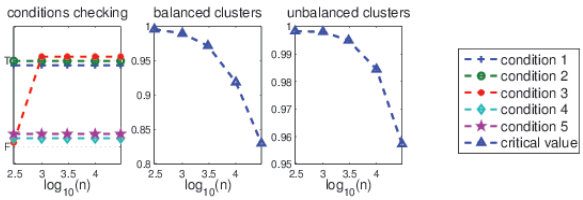


Figure 2.4: Typical results for [17]. Left: the satisfaction of the 5 conditions versus n; middle: critical value of (2.5) vs. n in balanced cluster size setting; right: the same in unbalanced cluster setting.

so does the average degree and Assumption (b) starts to hold. Assumption (a) is not met regardless of the value of n. Assumption (a) is

$$\frac{1}{8\sqrt{3}}\sqrt{\frac{K\ln(4n/\epsilon)}{\min d_i+\gamma}} \leq \lambda_K, \tag{2.6}$$

which puts a lower bound on λ_K depending on the minimum d_i. The mis-clustering rate is bounded with probability $(1-\epsilon)$ if these assumptions hold; γ is a constant. We set $\gamma = \bar{d}_i$ as suggested by the paper. Figure 2.5 displays the lower bound, which stays larger than 1 in all cases. As a result Assumption (a) fails since $\lambda_K < 1$. As n increases, lower bound in (2.6) decreases. We may anticipate the satisfaction of Assumption (b) when n is sufficiently large. Meanwhile, the balanced cluster size setting performs better than the unbalanced setting.

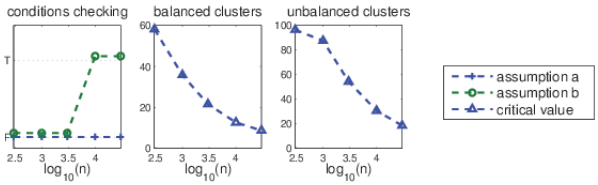


Figure 2.5: Typical results for [55]. Left: the satisfaction of the 2 conditions versus n; middle: critical value of (2.6) vs. n in balanced cluster size setting; right: the same in unbalanced cluster setting. These results are obtained with largest eigengap $\lambda_K = 0.99$ and balanced degree distribution.

Rohe, Chatterjee&Yu [57] Assumption (1) requires that the eigengap not be too small, and

Assumption (2) requires the graph to be dense enough, i.e.

$$\frac{\min d_i^2 \log n}{n^2} > 2 \tag{2.7}$$

From Figure 2.6, we observe that Assumption (1) always holds and Assumption (2) always fails. The critical value is the left hand side of the inequality (2.7). We can see that the critical value is staying far below 2 in all cases. The assumption for graph density fail for all cases.

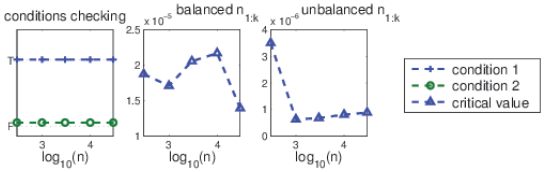


Figure 2.6: Typical results for [57], with largest eigengap $\lambda_K = 0.99$ and balanced degree distribution. Left: the satisfaction of the 2 conditions versus n; middle: critical value of (2.7) vs. n in balanced cluster size setting; right: the same in unbalanced cluster setting.

Now we discuss the two spectral clustering papers.

Ng&Jordan&Weiss [52] This paper also has dependent conditions. We eliminate the unknown parameters from Assumptions $A1 - A4$ and plug them into $A5$, then check whether it holds or not. Assumption $A5$ is defined as

$$\delta - (2 + \sqrt{2})\epsilon > 0, \tag{2.8}$$

In the above, δ is obtained from A1, and $0 < \delta < 1$; ϵ is obtained from A2 and A3, and is small as long as the similarity within the cluster is higher than that between clusters. However, in the experiments A5 always fails. Calculating ϵ from various parameter setting, we find that it is always bigger than 1. The plots in Figure 2.7 show a clear trend that as n increases, ϵ gets larger. We also observe that the balanced cluster size setting has smaller ϵ than the unbalanced setting, and is thus closer to satisfying condition A5.

Balakrishnan et al [8] The paper proposes two algorithms, one for hierachical clustering, and the other for k-way clustering with spectral method. For the hierarchical clustering, it assumes that

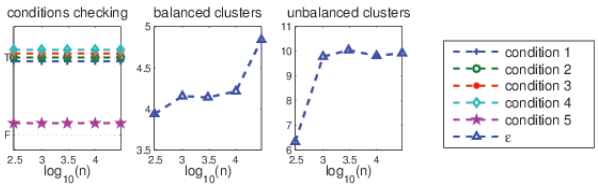


Figure 2.7: The best results for Ng, Jordan and Weiss, 2002. They are produced with balanced degree distribution and largest eigengap $\lambda_K = 0.99$. Left: the satisfaction of the 5 conditions versus n; middle: critical value of (2.8) vs. n in balanced cluster size setting; right: the same in unbalanced cluster setting.

A is constructed from the combination of a noise matrix and a hierarchical block matrix. We tested Assumptions 1–3 under all the possible hierarchical structures of the 5 clusters, and Assumption 3 is constantly violated. This is because the cluster separation is not large enough.

2.5 Discussion, conclusion and further work

In summary, because each of the eight papers we studied gives sufficient (but not necessary) guarantees for recovery, our experiments consist of generating networks for which recovery is possible and check which theory predicts it more often. We were able to generate such examples because it has been known for a long time, empirically, how to generate cases that are (likely to be) clusterable. Thus, this experiment tested the limits of the current theory. As we already mentioned, if necessary and sufficient conditions for community recovery were known, i.e *sharp recovery thresholds*, such experiments would have been unnecessary. If the current results were close to the unknown thresholds, then it would have been difficult to find clusterable examples not covered by the theory. Our results show that this is not the case.

We started with the aim of providing empirical comparisons between theoretical results in order to help readers understand their strenghs and weaknesses. We were also expecting to see a gradual degradation of the agreement between theory and reality as the test cases became harder. To our surprise, it turned out that we had to create the trivially easy $\lambda_K = 0.99$ set of test cases in order to observe that some theorems predict recoverability.

We were also expecting that the theoretical results by and large improve with time, as newer results are built on previous ones. This was partially confirmed by the most recent paper, [62], whose conditions are the only ones to cover a number of instances with non-trivially separated clusters.

We now turn to examining if any particular type of condition can be held responsible for the negative results. Before we start, we need to caution the reader on drawing hasty conclusions from examining Figures 2.2–2.7. As we have already mentioned, several theorems have interdependent conditions, or conditions that depend on the same unknown value. Between theses, one can trade-off violating one condition for satisfying another.

For every one of the eight papers studied, the requirements for cluster separation failed to be met, always or occasionally. This suggest that in many cases they are too severe. Interestingly enough, [62], which fared by far the best in terms of tolerance to intercluster edges, the separation conditions involve neither the off-diagonal blocks of A, nor λ_K. Rather, they are based on the separation in the spectral mapping obtained by spectral clustering, and depend on the imbalance between the sums of the node degrees in each cluster, with respect to the distribution ρ.

Furthermore, it appears that all four types of conditions previously mentioned were violated for some of the theorems. The biggest surprise is that the requirements on graph density were also occasionally too restrictive. For instance, even though Figure 1 shows that in our examples the average degree grows *linearly* with n, and the graphs are relatively dense (average $d_i \approx 0.1n$), equation (2.6) does not hold for any n up to 30,000. Extrapolating from our graphs, we see that it may start holding if n increases by another 2–4 orders of magnitude.

What we have observed with these benchmarking experiments suggests that the current results are not close to the yet unknown thresholds for recovery. They suggest that our understanding of the problem is not complete, and that the existing conditions do not yet align with the actual combinations of parameters that make recovery challenging.

As any benchmarking work, this one, too, can be improved on. Our second contribution is the Matlab code we wrote. This is being made available as the package `ThmBench` on `github/mmp2`. The code is written with ease of use and modularity in mind. We invite researchers in the field to

construct their own test cases, which may offer new perspectives on the limits of our current understanding in this area. We also have made it easy to add new testing modules, so that as new results are published, their conditions can be benchmarked as well.

2.6 Appendix

We list all the theorems we are comparing here.

2.6.1 Balakrishnan et al [9]

Assumption 1 For all i, j, $0 < A_{ij}\beta^*$, for some constant β^*.

Assumption 2 (Balanced clusters) There is) There is a constant $eta1$ such that at every split of the hierarchy $\frac{\mathcal{C}_{max}}{\mathcal{C}_{min}}$, where $|\mathcal{C}max|$, $|\mathcal{C}_{min}|$ are the sizes of the biggest and smallest clusters respectively.

Assumption 3 (Range Restriction) For every cluster s, $min\{\alpha_{s.L}, \alpha_{s.R}\} - \beta_s > \eta(\beta_s - \alpha_s)$.

Theorem 1: Supporse that $W = A+R$ is an $n \times n$ noisy HBM where A satisifies Assumptions 1, 2, and 3. Suppose that the scale factor of R increases at $\sigma = o(min(\kappa^{*5}\sqrt{\frac{m}{logn}}, \kappa^{*4}\sqrt{\frac{m}{logn}}))$ where $\kappa^* = min(\alpha_0, \frac{\gamma^*_{Sm}}{1+\eta})$, $m > 0$ and $m = \omega(logn)$. Then for all n large enough, with probability at list $1 - 6/n$, HS, on input M, will exactly recover all clusters of size at least m.

2.6.2 Balcan et al [11]

Definition 1: Given three positive parameters θ, α, β, where $\beta < \alpha \leq 1$ and an affinity system (V, Π) we say that a subset S of V is an (θ, α, β) self-determined community with respect to (V, Π) if we have both.

- For all $i \in S$, $\phi^{\theta}_S(i) \geq \alpha|S|$.
- For all $j \notin S$, $\phi^{\theta}_S(j) \leq \beta|S|$.

Theorem 3: Given a weighted affinity system (V, A), $\theta, \alpha, \beta, \epsilon < \alpha$, and a community size t, there is an efficient procedure that constructs a non-weighted instance (V', Π) along with a mapping f from V' to V, s.t. for any (θ, α, β), community S in V there exists a $(\theta, \alpha\epsilon, \beta)$ community S in V there exisit a $(\theta, \alpha - \epsilon, \beta)$ community S' in (V', Π) with $f(S') = S$.

Theorem 4: For any $\theta, \alpha, \beta, \gamma = \alpha - \beta$, the number of weighted $(\theta, \alpha, \beta)-$self-determined communities is $B(n) = (n/\gamma)^{O(log(1/\gamma)/\alpha)}(\frac{2\theta log(1/\gamma)}{\alpha})^{O(\frac{1}{\gamma^2}\log(\frac{\theta log(1/\gamma)}{\alpha\gamma}))}$ and we can find them in time $B(n)poly(n)$.

2.6.3 Coja-Oglan, et al [21]

Theorem 1. There exist

a. a deterministic polynomial time algorithm A and

b. for any $\alpha, \epsilon, \delta > 0$, any integer $k \geq 2$, and any $k \times k$ matrix $\Phi = (\phi_{ij})_{1 \leq i,j \leq k}$ with nonnegative entries numbers $D = D(\epsilon, \delta, \Phi) > 0$ and $n_0 = n_0(\alpha, \epsilon, \delta, \Phi)$.

such that the following is true. Suppose that $n > n_0$, that $w = (w_1, ..., w_n)$ is a tuple of positive reals, and that $V = (V_1, ..., V_k)$ is a partition of $V = \{1, ..., n\}$ such that the following six conditions hold:

C0: Let $p_{uv} = \phi_{\Psi(u),\Psi(v)} \frac{w_u w_v}{\sum_{x \in V} w_x}$, $p_{uv} \leq 1$.

C1: The rows of Φ are pairwise linearly independent.

C2: For all $u \in V$, we have $w_u \leq n^{1-\epsilon}$.

C3: Let $\bar{w} = \sum_{u \in V} w_u/n$, we have $w_u \leq n^{1-\epsilon}$.

C4: $\bar{w} \geq D$.

C5: $|V_I| \geq \delta n$ for all $1 \leq i \leq k$.

The with probability at least $1-\alpha$, the algorithm applies to the random graph $\mathcal{G}=\mathcal{G}_n(\Phi,w,\mathcal{V})$ outputes a prtition $V_1',\cdots,V_k'$ such that

$$\sum_{i=1}^{k}|(V_i\setminus V_i')\cup(V_i'\setminus V_i)|=O(n/\bar{w'}^0.97), \tag{2.9}$$

where $\bar{w'}=\frac{1}{n}\sum_{(u,v)\in V}p_{uv}=\sum_{(u,v)\in V}\phi_{\Psi(u),\Psi(v)}\frac{w_uw_v}{\bar{w}n^2}$ is the expected degree.

2.6.4 Chaudhuri et al [17]

Let $G=(\mathcal{V},E)$ be a random graph drawn from an extended planted partition model which satisfies the conditions in Lemma 4. Then, there exists a constant C such that the following holds. If, for all u, $E[deg(u)]\frac{128}{9}ln(6n/\delta)$, and if for all pairs of clusters V_i and V_j,

$$(\frac{p}{Z_i}-\frac{q}{Z_j})^2\sum_{w\in V_i}d_u^2+(\frac{p}{Z_j}-\frac{q}{Z_i})^2\sum_{u\in V_j}d_u^2\geq \tag{2.10}$$

$$64(\frac{384\sqrt{ln(2n/\delta)}}{Z_i\sqrt{\tau+min_{u\in V_i}E[deg(u)]}}(\sum_{u\in V_i}\frac{d_u^2}{(E[deg(u)]+\tau)^2})^{-1/2} \tag{2.11}$$

$$+\frac{384\sqrt{ln(2n/\delta)}}{Z_j\sqrt{\tau+min_{u\in V_j}E[deg(u)]}}(\sum_{u\in V_i}\frac{d_u^2}{(E[deg(u)]+\tau)^2})^{-1/2}+min_{u\in V_i,v\in V_j}2(\lambda_u+\lambda_v))^2 \tag{2.12}$$

then, $w.p.\geq 1-6\delta$, Algorithm 1 outputs a correct clustering.

2.6.5 Ng&Jordan&Weiss [52]

Assumption A1. There exists $\delta>0$ so that, for all $i=1,\cdots,k$, $\lambda_2^{(i)}\leq 1-\delta$.

Assumption A2. There is some fixed $\epsilon_1>0$, so that for every $i_1,i_2\in\{1,\cdots,k\}$, $i_1\neq i_2$, we have that

$$\sum_{j\in S_{i_1}}\sum_{k\in S_{i_2}}\frac{A_{jk}^2}{d_jd_k}\leq\epsilon_1 \tag{2.13}$$

Assumption A3. For some fixed $\epsilon_2>0$, for every $i=1,\cdots,k$, $j\in S_i$, we have

$$\frac{sum_{k:k\notin S_i}A_{jk}}{\hat{d}_j}\leq\epsilon_2(\sum_{k,l\in S_i}\frac{A_{kl}^2}{\hat{d}_k\hat{d}_l})^{-1/2} \tag{2.14}$$

Assumption A4. There is some constant $C \geq 0$ so that for every $i = 1, \cdots, k$, $j = 1, \cdots, n_i$, we have $\hat{d}_j^{(i)} \geq (\sum_{k=1}^{n_i} \hat{d}_k^{(i)})/(Cn_i)$.

Let Assumptions $A1$, $A2$, $A3$ and $A4$ hold. Set $\epsilon = \sqrt{k(k-1)\epsilon_1 + k\epsilon_2^2}$. If $\delta \geq (2+\sqrt{2})\epsilon$, then there exist k orthogonal vectors $r_1, \cdots, r_k$ ($r_i^T r_j = 1$, if $i = j$, 0 otherwise) so that Y's row satisfy

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} \|y_j^{(i)} - r_i\|_2^2 \leq 4C(4+2\sqrt{k})^2 \frac{\epsilon^2}{(\delta - \sqrt{2}\epsilon)^2} \tag{2.15}$$

Thus, the rows of Y will form tight clusters around k well-separated points on the surface of the k-sphere according to their "true" clusters S_i.

2.6.6 Qin&Rohe [55]

Theorem 4.4. Suppose $A \in R^{N\times N}$ is san adjacency matrix of a graph $\mathcal{G}$ generated from the DC-SBM with K blocks and parameters $\{B, Z, \Theta\}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_K > 0$ be the K positive eigenvalues of L_τ. Define $\mathcal{M}$, the set of mis-clustered nodes, as in Definition 4.3. Let δ be the minimum expected degree of $\mathcal{G}$. For any $\epsilon > 0$ and sufficiently large N, assume

(a) $\sqrt{\frac{Kln(4N/\epsilon)}{\delta+\tau}} \leq \frac{1}{8\sqrt{3}}\lambda_K$

(b) $\delta + \tau > 3lnN + 3ln(4/\epsilon)$

Then with probability at least $1-\epsilon$, the mis-clustering rate of RSC with regularization constant τ is bounded

$$|\mathcal{M}|/N \leq c_1 \frac{Kln(N/\epsilon)}{Nm^2(\delta+\tau)\lambda_K^2} \tag{2.16}$$

2.6.7 Rohe, Chatterjee&Yu

Suppose $W \in R^{n\times n}$ is an adjacency matrix from the Stochastic Blockmodel with k_n blocks. Define the population graph L. Define $\|\bar{\lambda_1}\| \geq \|\bar{\lambda_2}\| \cdots \geq \|\bar{\lambda_{k_n}}\| > 0$ as the absolute values of the k_n nonzero eigenvalues of L. Define $\mathcal{M}$, the set misclustered nodes. Define $\tau_n = \min d^{(n)}/n$ and

assume there exists N such that for all $n > N$, $\tau_n^2 > 2/logn$. Define $P_n = \max(Z_n^T Z_n)_{jj}$. If $n^{-1/2}(logn)^2 = O(\lambda_{k_n})^2$, then the number of misclustered nodes is bounded

$$|M| = o(\frac{P_n(logn)^2}{\lambda_{k_n}^2 \tau_n^4 n}) \tag{2.17}$$

2.6.8 Wan&Meila [62]

The theorem is shown in Theorem 3 in Chapter 3.

Chapter 3

A CLASS OF NETWORK MODELS RECOVERABLE BY SPECTRAL CLUSTERING

3.1 Introduction

There have been many recent advances in the recovery of communities in networks, under "block-model" assumptions [57, 56, 36]. In particular, advances in recovering communities by spectral clustering algorithms. These have been extended to models including node-specific propensities. In this chapter, we argue that one can further expand the model class for which recovery by spectral clustering is possible, and describe a model that subsumes a number of existing models, which we call the PFM. We show that under the PFM model, the communities can be recovered with small error, with high probability. Our results correspond to what [19] termed the "weak recovery" regime (see Section 1.2.3), in which with high probability the fraction of nodes that are mislabeled is $o(1)$ when $n \to \infty$.

3.2 The Preference Frame Model of graphs with communities

This model embodies the assumption that interactions at the community level (which we will also call *macro* level) can be quantified by meaningful parameters. This general assumption underlies the (p, q) and the related parameterizations of the SBM as well. We define a *preference frame* to be a graph with K nodes, one for each community, that encodes the connectivity pattern at the community level by a (non-symmetric) *stochastic matrix* R. Formally, given $[K] = \{1, \dots K\}$, a $K \times K$ matrix R ($\det(R) \neq 0$) representing the transition matrix of a *reversible* Markov chain on $[K]$, the weighted graph $\mathcal{H} = ([K], R)$, with edge set $\operatorname{supp} R$ (edges correspond to entries in R not being 0) is called a *K-preference frame*. Requiring reversibility is equivalent to requiring that there is a set of *symmetric* weights on the edges from which R can be derived ([53]). We note that without the reversibility assumption, we would be modeling *directed* graphs, which we will leave for future work. We denote by ρ the left principal eigenvector of R, satisfying $\rho^T R = \rho^T$. Without loss of generality, we can assume the eigenvalue 1 or R has multiplicity 1[1] and therefore we call ρ the *stationary distribution* of R.

We say that a deterministic weighted graph $\mathcal{G} = (\mathcal{V}, A)$ with weight matrix A (and edge set

[1] Otherwise the networks obtained would be disconnected.

$\operatorname{supp} A$) *admits a K-preference frame* $\mathcal{H} = ([K], R)$ if and only if there exists a partition $\mathcal{C}$ of the nodes $\mathcal{V}$ into K clusters $\mathcal{C} = \{\mathcal{C}_1, \dots \mathcal{C}_k\}$ of sizes $n_1, \dots, n_K$, respectively, so that the Markov chain on $\mathcal{V}$ with transition matrix P determined by A satisfies the linear constraints

$$\sum_{j \in \mathcal{C}_m} P_{ij} \;=\; R_{lm} \quad \text{for all } i \in \mathcal{C}_l, \text{ and all cluster indices } l, m \in \{1, 2, \dots k\}. \tag{3.1}$$

The matrix P is obtained from A by the standard row-normalization $P = D^{-1}A$ where $D = \operatorname{diag}\{d_{1:n}\}$, $d_i = \sum_{i=1}^{n} A_{ij}$.

A random graph family over node set $\mathcal{V}$ *admits a K-preference frame $\mathcal{H}$*, and is called a *Preference Frame Model* (PFM), if the edges i, j, $i < j$ are sampled independently from Bernoulli distributions with parameters A_{ij}. It is assumed that the edges obtained are undirected and that $A_{ij} \leq 1$ for all pairs $i \neq j$. We denote a realization from this process by $\hat{A}$. Furthermore, let $\hat{d}_i = \sum_{j \in \mathcal{V}} \hat{A}_{ij}$ and in general, throughout this chapter, we will denote computable quantities derived from the observed $\hat{A}$ with the same letter as their model counterparts, decorated with the "hat" symbol. Thus, $\hat{D} = \operatorname{diag} \hat{d}_{1:n}$, $\hat{P} = \hat{D}^{-1}\hat{A}$, and so on.

One question we will study is under what conditions the PFM model can be estimated from a given $\hat{A}$ by a standard spectral clustering algorithms. Evidently, the difficult part in this estimation problem is recovering the partition $\mathcal{C}$. If this is obtained correctly, the remaining parameters are easily estimated in a Maximum Likelihood framework.

But another question we elucidate refers to the parametrization itself. It is known that in the SBM and Degree Corrected-SBM (DC-SBM) [56], in spite of their simplicity, there are dependencies between the community level "intensive" parameters and the graph level "extensive"parameters, as we will show below. In the parametrization of the PFM , we can explicitly show which are the free parameters and which are the dependent ones.

Several network models in wide use admit a preference frame. For example, in the case of SBM(p, q), which is a special case of SBM with matrix B defined as $B_{kk} = p, B_{kl} = q$ for $k, l \in [K]$, $k \neq l$, we have

$$d_i = p(n_l - 1) + q(n - n_l) \equiv d_{\mathcal{C}_l}, \text{ for } i \in \mathcal{C}_l,$$

$$R_{lm} = \frac{qn_m}{d_{\mathcal{C}_l}} \text{ if } l \neq m,\ R_{ll} = \frac{p(n_l - 1)}{d_{\mathcal{C}_l}}, \text{ for } l, m \in \{1, 2, \ldots, k\}.$$

In the above we have introduced the notation $d_{\mathcal{C}_l} = \sum_{j \in \mathcal{C}_l} d_i$. One particular realization of the PFM is the *Homogeneous K-Preference Frame* model (HPFM). In a HPFM, each node $i \in \mathcal{V}$ is characterized by a *weight, or propensity to form ties* w_i. For each pair of communities l, m with $l \leq m$ and for each $i \in \mathcal{C}_l, j \in \mathcal{C}_m$ we sample $\hat{A}_{ij}$ with probability A_{ij} given by

$$A_{ij} = \frac{R_{ml} w_i w_j}{\rho_l}. \tag{3.2}$$

This formulation ensures detail balance in the edge expectations, i.e. $A_{ij} = A_{ji}$. The HPFM is virtually equivalent to what is known as the "degree model" [29] or "DC-SBM", up to a reparameterization[2]. Proposition 1 relates the node weights to the expected node degrees d_i. We note that the main result we prove in this paper uses independent sampling of edges only to prove the concentration of the Laplacian matrix. The HPFM model can be easily extended to other graph models with dependent edges if one could prove concentration and eigenvalue separation. For example, when R has rational entries, the subgraph induced by each block of $\hat{A}$ can be represented by a random d-regular graph with a specified degree.

Proposition 1 *In a HPFM $d_i = w_i \sum_{l=1}^{K} R_{kl} \frac{w_{\mathcal{C}_l}}{\rho_l}$ whenever $i \in \mathcal{C}_k$ and $k \in [K]$, where $w_{\mathcal{C}_l} = \sum_{j \in \mathcal{C}_l} w_j$.*

Equivalent statements that the expected degrees in each cluster are proportional to the weights exist in [21, 57] and they are instrumental in analyzing this model. This particular parametrization immediately implies in what case the degrees are globally proportional to the weights. This is, obviously, the situation when $w_{\mathcal{C}_l} \propto \rho_l$ for all $l \in [K]$.

As we see, the node degrees in a HPFM are not directly determined by the propensities w_i, but depend on those by a multiplicative constant that varies with the cluster. This type of interaction between parameters has been observed in practically all extensions of the Stochastic Block-Model

[2]Here we follow the customary definition of this model, which does not enforce $A_{ii} = 0$, even though this implies a non-zero probability of self-loops.

that we are aware of, making parameter interpretation more difficult. Our following result establishes what are the free parameters of the PFM and of their subclasses. As it will turn out, these parameters and their interactions are easily interpretable.

Proposition 2 *Let* $(n_1, \dots n_K)$ *be a partition of* n *(assumed to represent the cluster sizes of* $\mathcal{C} = \{\mathcal{C}_1, \dots \mathcal{C}_K\}$ *a partition of node set* $\mathcal{V}$*), R a non-singular $K \times K$ stochastic matrix, ρ its left principal eigenvector, and* $\pi_{\mathcal{C}_1} \in [0,1]^{n_1}, \dots \pi_{\mathcal{C}_K} \in [0,1]^{n_K}$ *probability distributions over* $\mathcal{C}_{1:K}$*. Then, there exists a PFM consistent with* $\mathcal{H} = ([K], R)$*, with clustering* $\mathcal{C}$*, and whose node degrees are given by*

$$d_i = d_{tot} \rho_k \pi_{\mathcal{C}_k i}, \tag{3.3}$$

whenever $i \in \mathcal{C}_k$*, where* $d_{tot} = \sum_{i \in \mathcal{V}} d_i$ *is a user parameter which is only restricted above by Assumption 2.*

The proof of this result is constructive, and can be found in Section 3.7

The parametrization shows to what extent one can specify independently the degree distribution of a network model, and the connectivity parameters R. Moreover, it describes the pattern of connection of a node i as a composition of a macro-level pattern, which gives the total probability of i to form connections with a cluster l, and the micro-level distribution of connections between i and the members of $\mathcal{C}_l$. These parameters are meaningful on their own and can be specified or estimated separately, as they have no hidden dependence on each other or on n, K.

The PFM enjoys a number of other interesting properties. As this chapter will show, almost all the properties that make SBM's popular and easy to understand hold also for the much more flexible PFM. In the remainder of this paper we derive recovery guarantees for the PFM. As an additional goal, we will show that in the frame we set with the PFM, the recovery conditions become clearer, more interpretable, and occasionally less restrictive than for other models.

Third, as already mentioned, the PFM includes many models that have been found useful by previous authors. Yet, the PFM class is much more flexible than those individual models, in the sense that it allows other unexplored degrees of freedom (or, in other words, achieves the same advantages as previously studied models with fewer constraints on the data). There are $O(n^2)$

parameters or degree of freedoms in PFM, while the number of parameters is less than K^2 in SBM and $O(n)$ in DC-SBM. Note that for PFM, there is an infinite number of possible random graphs $\mathcal{G}$ with the same parameters $(d_{1:n}, n_{1:k}, R)$ satisfying the constraints (3.1) and Proposition 2, yet for reliable community detection we do not need to estimate A fully, but only to look at aggregate statistics like $\sum_{j\in\mathcal{C}} \hat{A}_{ij}$.

3.3 Spectral clustering algorithm

Now, we describe the community recovery algorithms from a random graph $(\mathcal{V}, \hat{A})$ sampled from the PFM defined as above. We make the standard assumption that K is known. Our analysis is based on a very common spectral clustering algorithm used in [45] and described also in [46, 60].

Input : Graph $(\mathcal{V}, \hat{A})$ with $|\mathcal{V}| = n$ and $\hat{A} \in \{0,1\}^{n\times n}$, number of clusters K

Output: Clustering $\hat{\mathcal{C}}$

1. Compute $\hat{D} = \mathrm{diag}(\hat{d}_1, \cdots, \hat{d}_n)$ and Laplacian

$$\hat{L} = \hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2} \tag{3.4}$$

2. Calculate the K eigenvectors $\hat{Y}_1, \cdots, \hat{Y}_K$ associated with the K eigenvalues $|\hat{\lambda}_1| \geq \cdots \geq |\hat{\lambda}_K|$ of $\hat{L}$. Normalize the eigenvectors to unit length. We denote them as the first K eigenvectors in the following text;

3. Set $\hat{V}_i = \hat{D}^{-1/2}\hat{Y}_i$, $i = 1, \cdots, K$. Form matrix $\hat{V} = [\hat{V}_1 \cdots \hat{V}_K]$;

4. Treating each row of $\hat{V}$ as a point in K dimensions, cluster them by the K-means algorithm to obtain the clustering $\hat{\mathcal{C}}$.

Algorithm 2: Spectral Clustering

Note that the vectors $\hat{V}$ are the first K eigenvectors of P. The K-means algorithm is assumed to find the global optimum. For more details on good initializations for K-means in step 4 see [52].

We quantify the difference between $\hat{\mathcal{C}}$ and the true clusterings $\mathcal{C}$ by the mis-clustering rate p_{err}, which is defined as

$$p_{err} = dist(\mathcal{C}, \hat{\mathcal{C}}) \tag{3.5}$$

3.4 Main Results

Theorem 3 (Mis-clustering rate bound for HPFM and PFM) *Let the $n \times n$ matrix A admit a PFM, and $w_{1:n}, R, \rho, P, \hat{A}, d_{1:n}$ have the usual meaning. Let $\lambda_{1:n}$ be the eigenvalues of P, with $|\lambda_i| \geq |\lambda_{i+1}|$. Let $d_{min} = \min d_{1:n}$ be the minimum expected degree, $\hat{d}_{min} = \min \hat{d}_i$, and $d_{max} = \max_{ij} nA_{ij}$. Let $\gamma \geq 1$, $\epsilon > 0$ be arbitrary numbers. Assume:*

Assumption 1 *A admits a HPFM model and* (3.2) *holds.*

Assumption 2 $A_{ij} \leq 1$

Assumption 3 $\hat{d}_{min} \geq \log n$

Assumption 4 $d_{min} \geq \log n$

Assumption 5 $\exists \varkappa > 0$, $d_{max} \leq \varkappa \log n$

Assumption 6 $g_{row} > 0$, *where g_{row} is defined in Proposition 4.*

Assumption 7 *$\lambda_{1:K}$ are the eigenvalues of R, and $|\lambda_K| - |\lambda_{K+1}| = \sigma > 0$.*

We also assume that we run Algorithm 1 on A and that K-means finds the optimal solution. Then, for n sufficiently large, the following statements hold with probability at least $(1-2\exp\frac{-\epsilon^2}{2+\epsilon/\sqrt{\log n}})(1-e^{-\gamma})$.

PFM *Assumptions 2 - 7 imply*

$$p_{err} \leq \frac{Kd_{tot}}{nd_{min}g_{row}}\left[\frac{C_0\gamma^4}{\sigma^2 \log n} + \frac{4\epsilon^2}{\hat{d}_{min}}\right] \tag{3.6}$$

HPFM *Assumptions 1 - 6 imply*

$$p_{err} \leq \frac{Kd_{tot}}{nd_{min}g_{row}}\left[\frac{C_0\gamma^4}{\lambda_K^2 \log n} + \frac{4\epsilon^2}{\hat{d}_{min}}\right] \tag{3.7}$$

where C_0 is a constant.

Note that p_{err} decreases at least as $1/\log n$ when $\hat{d}_{min} = d_{min} = \log n$. This is because $\hat{d}_{min}$ and d_{min} help with the concentration of L. Using Proposition 4, the distances between rows of V, i.e, the true centers of the K-means step, are lower bounded by g_{row}/d_{tot}. After plugging in the assumptions for $d_{min}, \hat{d}_{min}, d_{max}$, we obtain

$$p_{err} \leq \frac{K\varkappa(C_0\gamma^4 + 4\epsilon^2\sigma^2)}{g_{row}\sigma^2 \log n}. \tag{3.8}$$

This shows that p_{err} decreases as $1/\log n$. Of the remaining quantities, κ controls the spread of the degrees d_i, and C_0 depends on κ and γ. Notice that λ_K and σ are eigengaps in HPFM model and PFM model respectively and depend only on the preference frame, and likewise for g_{row}. The eigengaps ensure the stability of principal spaces and the separation from the spurious eigenvalues, as shown in Proposition 6.

3.4.1 Proof outline, techniques and main concepts

The proof of Theorem 3 (see Section 3.8) relies on three steps, which are to be found in most results dealing with spectral clustering. First, concentration bounds of the empirical Laplacian $\hat{L}$ w.r.t L are obtained. There are various conditions under which these can be obtained, and ours are most similar to the recent result of [36]. The other tools we use are Hoeffding bounds and tools from linear algebra. Second, one needs to bound the perturbation of the eigenvectors Y as a function of the perturbation in L. This is based on the pivotal results of Davis and Kahan, see e.g [56]. A crucial ingredient in these type of theorems is the size of the eigengap between the invariant subspace Y and its orthogonal complement. This is a condition that is model-dependent, and therefore we discuss the techniques we introduce for solving this problem in the PFM in the next subsection.

The third step is to bound the error of the K-means clustering algorithm. This is done by a counting argument. The crux of this step is to ensure the separation of the K distinct rows of V. This, again, is model dependent and we present our result below. The details and proof are in Section 3.7 and 3.8. All proofs are for the PFM; to specialize to the HPFM, one replaces σ with $|\lambda_K|$

3.4.2 Cluster separation and bounding the spurious eigenvalues in the PFM

Proposition 4 (Cluster separation) *Let $V, \rho, d_{1:n}$ have the usual meaning and define the* cluster volume $d_{\mathcal{C}_k} = \sum_{i \in \mathcal{C}_k} d_i$, *and c_{max}, c_{min} as $\max_k, min_k \frac{d_{\mathcal{C}_k}}{n\rho_k}$. Let $i, j \in \mathcal{V}$ be nodes belonging respectively to clusters k, m with $k \neq m$. Then,*

$$||V_{i:} - V_{j:}||^2 \geq \frac{1}{d_{tot}} \left[\frac{1}{c_{max}} \left(\frac{1}{\rho_k} + \frac{1}{\rho_m} \right) - \frac{1}{\sqrt{\rho_k \rho_m}} \left(\frac{1}{c_{min}} - \frac{1}{c_{max}} \right) \right] = \frac{g_{row}}{d_{tot}}, \tag{3.9}$$

where $g_{row} = \left[\frac{1}{c_{max}} \left(\frac{1}{\rho_k} + \frac{1}{\rho_m} \right) - \frac{1}{\sqrt{\rho_k \rho_m}} \left(\frac{1}{c_{min}} - \frac{1}{c_{max}} \right) \right]$. Moreover, if the columns of V are normalized to length 1, the above result holds by replacing $c_{max,min}$ with $\tilde{c}_{max,min} = \max, \min_k \frac{n_k}{n\rho_k}$.

In the square brackets, $c_{max,min}$ depend on the cluster-level degree distribution, while all the other quantitities depend only on the preference frame. Hence, this expression is invariant with n, and as long as it is strictly positive, we have that the cluster separation is $\Omega(1/d_{tot})$.

The next theorem is crucial in proving that L has a constant eigengap. We express the eigengap of P in terms of the preference frame $\mathcal{H}$ and the mixing inside each of the clusters $\mathcal{C}_k$. For this, we resort to *generalized stochastic matrices*, i.e. rectangular positive matrices with equal row sums, and we relate their properties to the mixing of Markov chains on bipartite graphs.

These tools are introduced here, for the sake of intuition, toghether with the main spectral result, while the rest of the proofs are in Section 3.8.

Given $\mathcal{C}$, for any vector $x \in \mathbb{R}^n$, we denote by x_k, $k = 1, \dots K$, the block of x indexed by elements of cluster k of $\mathcal{C}$. Similarly, for any square matrix $S \in \mathbb{R}^{n \times n}$, we denote by $S_{kl} = [S_{ij}]_{i \in k, j \in l}$ the block with rows indexed by $i \in k$, and columns indexed by $j \in l$.

Denote by ρ, $\lambda_{1:K}$, $\nu^{1:K} \in \mathbb{R}^K$ respectively the stationary distribution, eigenvalues[3], and eigenvectors of R.

We are interested in block stochastic matrices P for which the eigenvalues of R are the principal eigenvalues. We call $\lambda_{K+1} \dots \lambda_n$ *spurious* eigenvalues. Theorem 6 below is a sufficient condition

[3]Here too, eigenvalues will always be ordered in decreasing order of their magnitudes, with positive values preceeding negatives one of the same magnitude. Consequently, for any stochastic matrix, $\lambda_1 = 1$ always

that bounds $|\lambda_{K+1}|$ whenever each of the K^2 blocks of P is "homogeneous" in a sense that will be defined below.

When we consider the matrix $L = D^{-1/2}AD^{-1/2}$ partitioned according to $\mathcal{C}$, it will be convenient to consider the off-diagonal blocks in pairs. This is why the next result describes the properties of matrices consisting of a pair of off-diagonal blocks.

Proposition 5 (Eigenvalues for the off-diagonal blocks) *Let M be the square matrix*

$$M = \begin{bmatrix} 0 & B \\ A & 0 \end{bmatrix} \tag{3.10}$$

where $A \in \mathbb{R}^{n_2 \times n_1}$ and $B \in \mathbb{R}^{n_1 \times n_2}$, and let $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$, $x_{1,2} \in \mathbb{C}^{n_{1,2}}$ be an eigenvector of M with eigenvalue λ. Then

$$Bx_2 = \lambda x_1 \quad ABx_2 = \lambda^2 x_2 \tag{3.11}$$

$$Ax_1 = \lambda x_2 \quad BAx_1 = \lambda^2 x_1 \tag{3.12}$$

$$M^2 = \begin{bmatrix} BA & 0 \\ 0 & AB \end{bmatrix} \tag{3.13}$$

*Moreover, if M is symmetric, i.e $B = A^T$, then λ is a singular value of A, x is real, and $-\lambda$ is also an eigenvalue of M with eigenvector $[x_1^T \ -x_2^T]^T$. Assuming $n_2 \leq n_1$, and that A is full rank, one can write $A = V\Lambda U^T$ with $V \in \mathbb{R}^{n_2 \times n_2}$, $U \in \mathbb{R}^{n_1 \times n_2}$ orthogonal matrices, and Λ a diagonal matrix of non-zero singular values [**?**].*

Theorem 6 (Bounding the spurious eigenvalues of L) *Let $\mathcal{C}, L, P, D, A, R, \rho$ be defined as above, and let λ be an eigenvalue of P. Assume that (1) P is block-stochastic with respect to $\mathcal{C}$; (2) $\lambda_{1:K}$ are the eigenvalues of R, and $|\lambda_K| > 0$; (3) λ is not an eigenvalue of R; (4) denote by λ_3^{kl} (λ_2^{kk}) the third (second) largest in magnitude eigenvalue of block M_{kl} (L_{kk}) and assume that $\frac{|\lambda_3^{kl}|}{\lambda_{max}(M_{kl})} \leq c < 1$ ($\frac{|\lambda_2^{kk}|}{\lambda_{max}(L_{kk})} \leq c$). Then, the spurious eigenvalues of P are bounded by c times a constant that depends only on R.*

$$|\lambda| \leq c \max_{k=1:K} \left(r_{kk} + \sum_{l \neq k} \sqrt{r_{kl} r_{lk}} \right) \tag{3.14}$$

Remarks: The factor that multiplies c can be further bounded denoteing $a = [\sqrt{r_{kl}}]_{l=1:K}^{T}, b = [\sqrt{r_{lk}}]_{l=1:K}^{T}$

$$r_{kk} + \sum_{l \neq k} \sqrt{r_{kl} r_{lk}} \;=\; a^T b \;\leq ||a||||b|| \;=\; \sqrt{\sum_{l=1}^{K} r_{kl} \sum_{l=1}^{K} r_{lk}} \;=\; \sqrt{\sum_{l=1}^{K} r_{lk}} \tag{3.15}$$

In other words,

$$|\lambda| \leq \frac{c}{2} \max_{k=1:K} \sqrt{\sum_{l=1}^{K} r_{lk}} \tag{3.16}$$

The maximum column sum of a stochastic matrix is 1 if the matrix is doubly stochastic and larger than 1 otherwise, and can be as large as $\sqrt{K}$. However, one must remember that the interesting R matrices have "large" eigenvalues. In particular we will be interested in $\lambda_K > c$. It is expected that under these conditions, the factor depending on R to be close to 1.

The second remark is on the condition (3), that all blocks have small spurious eigenvalues. This condition is not merely a technical convenience. If a block had a large eigenvalue, near 1 or -1 (times its λ_{max}), then that block could itself be broken into two distinct clusters. In other words, the clustering $\mathcal{C}$ would not accurately capture the cluster structure of the matrix P. Hence, condition (3) amounts to requiring that no other cluster structure is present, in other words that within each block, the Markov chain induced by P *mixes well*.

3.5 Related work

Previous results we used The Laplacian concentration results use a technique introduced recently by [36], and some of the basic matrix theoretic results are based on [46] which studied the P and L matrix in the context of spectral clustering. As any of the many works we cite, we are indebted to the pioneering work on the perturbation of invariant subspaces of Davis and Kahan [58].

3.5.1 Previous related models

The configuration model for regular random graphs [16, 41] and for graphs with general fixed degrees [40, 42] is very well known. It can be shown by a simple calculation that the configuration

model also admits a K-preference frame. In the particular case when the diagonal of the R matrix is 0 and the connections between clusters are given by a bipartite configuration model with fixed degrees, K-preference frames have been studied by [51] under the name "equitable graphs"; the goal there was to provide a way to calculate the spectrum of the graph.

Since the PFM is itself an extension of the SBM, many other extensions of the latter will bear resemblance to PFM. Here we review only a subset of these, which exploit the spectral properties of the SBM and extend this to handle a large range of degree distributions [21, 57, 17]. The PFM includes each of these models as a subclass[4].

In [21] the authors study a model that coincides (up to some multiplicative constants) with the HPFM. The paper introduces an elegant algorithm that achieves partial recovery or better, which is based on the spectral properties of a random Laplacian-like matrix, and does not require knowledge of the partition size K.

The PFM also coincides with the model of [5] and [29] called the *expected degree model* w.r.t the distribution of *intra-cluster* edges, but not w.r.t the ambient edges, so the HPFM is a subclass of this model. $HPFM$ is discussed in Chapter 2.

A different approach to recovery The papers [17, 56, 36] propose regularizing the normalized Laplacian with respect to the influence of low degrees, by adding the scaled unit matrix τI to the incidence matrix $\hat{A}$, and thereby they achieve recovery for much more imbalanced degree distributions than us. Currently, we do not see an application of this interesting technique to the PFM, as the diagonal regularization destroys the separation of the intracluster and intercluster transitions, which guarantee the clustering property of the eigenvectors. Therefore, currently we cannot break the $n \log n$ limit into the ultra-sparse regime, although we recognize that this is an important current direction of research.

Recovery results like ours can be easily extended to weighted, non-random graphs, and in this sense they are relevant to the spectral clustering of these graphs, when they are assumed to be noisy versions of a $\mathcal{G}$ that admits a PFM.

---

[4] In particular, the models proposed in [21, 57, 17] are variations of the DC-SBM and thus forms of the homogeneous PFM.

3.5.2 An empirical comparison of the recovery conditions

As obtaining general results in comparing the various recovery conditions in the literature would be a tedious task, here we undertake to do a numerical comparison. While the conclusions drawn from this are not universal, they illustrate well the stringency of various conditions, as well as the gap between theory and actual recovery. For this, we construct HPFM models, and verify numerically if they satisfy the various conditions. As one may observe, our recovery theorems drawn from this chapter is model-based, i.e, one needs to assume the data is generated from a model in order for the recovery theorems to work. In Chapter 4, we will further develop model-free recovery theorems. For our model-based results, we have also clustered random graphs sampled from this model, with good results (shown in Section 3.8).

We generate A from the HPFM model with $K = 5$, $n = 5000$. Each w_i is uniformly generated from $(0.5, 1)$. $n_{1:K} = (500, 1000, 1500, 1000, 1000)$, $g_{row} > 0$, $\lambda_{1:K} = (1, 0.8, 0.6, 0.4, 0.2)$. The matrix R is given below; note its last row in which $r_{55} < \sum_{l=1}^{4} r_{5l}$.

$$R = \begin{pmatrix} .80 & .07 & .02 & .02 & .09 \\ .04 & .52 & .24 & .12 & .08 \\ .01 & .20 & .65 & .15 & .00 \\ .01 & .08 & .12 & .70 & .08 \\ .13 & .21 & .02 & .32 & .33 \end{pmatrix} \quad \rho = (.25, .44, .54, .65, .17). \tag{3.17}$$

The conditions we are verifying include besides ours, those obtained by [56], [57], [10] and [17]; since the original A is a perfect case for spectral clustering of weighted graphs, we also verify the theoretical recovery conditions for spectral clustering in [8] and [52].

Our result Theorem 3 We have $d_{min} = 77.4$, $\hat{d}_{min} = 63$, both bigger than $\log n = 8.52$. Therefore assumptions (a) and (b) hold; $d = 2500$, so (c) also holds, $g_{row} = 1.82 > 0$. After running the algorithm, the mis-clustering is rate $r = 0.0008$, which satisfies the theoretical bound. In conclusion, the dataset fits into both the assumptions and conclusion of Theorem 3.

Qin and Rohe[56] This paper has an assumption on the lower bound on λ_K, that is $\frac{1}{8\sqrt{3}}\lambda_K \geq \sqrt{\frac{K(ln(K/\epsilon)}{d_{min}}}$, so that the concentration bound holds with probability $(1 - \epsilon)$. We set $\epsilon = 0.1$ and

obtain $\lambda_K \geq 12.3$, which is impossible to hold since λ_K is upper bounded by 1[5].

Rohe, Chatterjee, Yu[57] Here, one defines $\tau_n = \frac{d_{min}}{n}$, and requires $\tau_n^2 \log n > 2$ to ensure the concentration of L. To meet this assumption, with $n = 5000$, $d_{min} \geq 2422$. While in our case $d_{min} = 77.4$. The assumption requires a very dense graph and is not satisfied in this dataset.

Balcan, Borgs Braverman, Chayes[10]Their theorem is based on self-determined community structure. It requires all the nodes to be more connected within their own cluster. However, in our graph, 1296 out of 5000 nodes have more connections to outside nodes than to nodes in their own cluster.

Ng, Jordan, Weiss[52] require $\lambda_2 < 1 - \delta$, where $\delta > (2 + 2\sqrt{2})\epsilon$, $\epsilon = \sqrt{K(K-1)\epsilon_1 + K\epsilon_2^2}$, $\epsilon_1 \geq \max_{i_1, i_2 \in \{1,\cdots,K\}} \sum_{j \in C_{i_1}} \sum_{k \in C_{i_2}} \frac{\hat{A}_{jk}^2}{\hat{d}_j \hat{d}_k}$, $\epsilon_2 \geq \max_{i \in \{1,\cdots,K\}} \frac{\sum_{k:k \in A_i}}{\hat{d}_j} (\sum_{k,l \in A_i} \frac{\hat{A}_{kl}^2}{\hat{d}_k \hat{d}_l})^{1/2}$. On the given data, we find that $\epsilon \geq 36.69$, and $\delta \geq 125.28$, which is impossible to hold since δ needs to be smaller than 1.

Chaudhuri, Chung, Tsiatas[17] The recovery theorem of this paper requires $d_i \geq \frac{128}{9} \ln(6n/\delta)$, so that when all the assumptions hold, it recovers the clustering correctly with probability at least $1 - 6\delta$. We set $\delta = 0.01$, and obtain that $d_i = 77.40$, $\frac{128}{9} \ln(6n/\delta) = 212.11$. Therefore the assumption fails as well.

For our method, the hardest condition to satisfy, and the most different from the others, was Assumption 6. We repeated this experiment with the other weights distributions for which this Assumption fails. The assumptions in the related papers continued to be violated. In [Qin and Rohe], we obtain $\lambda_K \geq 17.32$. In [Rohe, Chatterjee, Yu], we still needs $d_{min} \geq 2422$. In [Balcan, Borgs Braverman, Chayes], we get 1609 points more connected to the outside nodes of its cluster. In [Balakrishnan, Xu, Krishnamurthy, Singh], we get $\sigma = 0.172$ and needs to satisfy $\sigma = o(0.3292)$. In [Ng, Jordan, Weiss], we obtain $\delta \geq 175.35$. Therefore, the assumptions in these papers are all violated as well.

[5]To make $\lambda \leq 1$ possible, one needs $d_{min} \geq 11718$.

3.6 Conclusion

In this paper, we have introduced the Preference Frame Model, which is more flexible and subsumes many current models including SBM and DC-SBM. It produces state-of-the art recovery rates comparable to existing models. To accomplish this, we used a parametrization that is clearer and more intuitive. The theoretical results are based on the new geometric techniques which control the eigengaps of the matrices with piecewise constant eigenvectors.

We note that the main result Theorem 3 uses independent sampling of edges only to prove the concentration of the laplacian matrix. The KPFM model can be easily extended to other graph models with dependent edges if one could prove concentration and eigenvalue separation. For example, when R has rational entries, the subgraph induced by each block of $\hat{A}$ can be represented by a random d-regular graph with a specified degree.

3.7 Matrix theoretical results

This proposition below collected various basic facts moved elsewhere in the references[45, 60].

Proposition 7 ([45, 60]) *1. The matrices P and L have the same eigenvalues, denoted $\lambda_{1:n}$ (slightly abusively).*

2. *Every eigenvalue of R is also an eigenvalue of P.*

3. *If v is an eigenvector of P with eigenvalue λ, then $u = D^{1/2}v$ is is an eigenvector of L with the same eigenvalue.*

4. *In particular, $P\mathbf{1} = \mathbf{1}$ is the Frobenius vector of P, and therefore $Ls = s$, with $s_i = \sqrt{d_i}$, for $i = 1, \ldots n$, is the Frobenius eigenvector of L.*

5. *The stationary distribution of P is π, with $\pi_i \propto d_i$, $i = 1, \ldots n$.*

6. *R is diagonalizable, implying that it has K independent eigenvectors. (This follows from the fact that the Markov chain defined by P is reversible, implying that R also defines a*

reversible Markov chain.)

7. *Let λ be an eigenvalue of R, ν its eigenvector, and v the eigenvector of P corresponding to λ. Then $v_i = \nu_l$ whenever $i \in l$. In other words, P has K eigenvectors that are "telescoped" versions of the eigenvectors of R.*

Let $A, w_{1:n}, K, R, \rho, d_{1:n}, P$, etc have the usual meaning. Let $B = \mathrm{diag}(\rho)^{-1/2} R \,\mathrm{diag}(\rho)^{-1/2}$. Denote

$$X \in \mathbb{R}^{K\times K} \quad \text{the eigenvector matrix of } B\text{, orthonormal} \tag{3.18}$$

$$U \in \mathbb{R}^{K\times K} \quad \text{the eigenvector matrix of } R\text{, with } \mathrm{diag}(\rho)^{1/2} U = X \tag{3.19}$$

$$Y \in \mathbb{R}^{n\times K} \quad \text{principal eigenvectors of } L\text{, orthonormal, } Y_{il} \propto \frac{\sqrt{d_i} X_{kl}}{\sqrt{\rho_k}} \text{ if } i \in \mathcal{C}_k \tag{3.20}$$

$$y_{1:K} \quad \text{normalization constants for the columns of } Y\text{,} \tag{3.21}$$

$$V \in \mathbb{R}^{n\times K} \quad \text{principal eigenvectors of } P\text{, with, } V_{ik} = \frac{1}{\sqrt{d_i}} Y_{ik} \tag{3.22}$$

Denote also $d_{tot} = \sum_{i=1}^n d_i$, $d_k = \sum_{i\in\mathcal{C}_k} d_i, \pi_k = \frac{d_k}{d_{tot}}$, and $\max_k, \min_k \frac{\pi_k}{\rho_k} = c_{max,min}$.

Proof of Proposition 2 We construct a distribution π over $\mathcal{V}$ by $\pi' = [\pi'_1 \dots \pi'_K]$ with $\pi_k \in [0,1]^{n_k}$ the elements of π indexed by cluster $\mathcal{C}_k$. Let $\pi_k = \pi_{\mathcal{C}_k}\rho_k$ for all $k = 1, \dots K$.

We will verify that π is the stationary distribution of P. Fix $i \in \mathcal{C}_l$. We calculate $\pi' P$ at node i.

$$(\pi' P)_i \;=\; \sum_{k=1}^K \pi'_k P_{kl,:i} \;=\; \sum_{k=1}^K \rho_k \pi'_{\mathcal{C}_k} \tilde{P}_{kl,:i} r_{kl} \;=\; \sum_{k=1}^K \rho_k \pi_{\mathcal{C}_l,i} r_{kl} \;=\; \pi_{\mathcal{C}_l,i} \sum_{k=1}^K \rho_k r_{kl} \;=\; \pi_{\mathcal{C}_l,i}\rho_l \tag{3.23}$$

Above, we slightly abused notation by using i both as an index in $\mathcal{V}$ and in $\mathcal{C}_l$.

It remains to show that $\pi\mathbf{1} = 1$ which is straightforward and left to the reader. Now, from Proposition 9 $\pi_i \propto d_i$ for $i \in \mathcal{V}$; the normalization constant for the r.h.s. of this expression is $\sum_{i\in\mathcal{V}} d_i = d_{tot}$.

Proof of Proposition 4 Define

$$y_l^2 \;=\; \sum_i \frac{d_i X_{kl}^2}{\rho_k} \;=\; \sum_{k=1}^K \frac{d_k X_{kl}^2}{\rho_k}. \tag{3.24}$$

Hence,

$$c_{min}d_{tot} \leq y_l^2 \leq c_{max}d_{tot}\sum_k X_{kl}^2 = c_{max}d_{tot}. \tag{3.25}$$

. Now, we can derive bounds on $||V_{i:}||^2$, the length of row i of V.

$$||V_{i:}||^2 = \sum_{l=1}^{K}\frac{1}{y_l^2}\frac{X_{kl}^2}{\rho_k} = \frac{1}{\rho_k}\sum_{l=1}^{K}\frac{1}{y_l^2}X_{kl}^2 \geq \frac{1}{\rho_k}\frac{1}{d_{tot}c_{max}} \quad \text{when } i \in \mathcal{C}_k. \tag{3.26}$$

We now need to bound the cross-terms $V_{i:}^T V_{j:}$. Denote by $l^+ = \{l \in [K],\, X_{kl}X_{ml} \geq 0\}$ and $l^- = [K] \setminus l^+$. Then, we have

$$\sum_{l=1}^{K}\frac{1}{y_l^2}X_{kl}X_{ml} = \sum_{l^+}\frac{1}{y_l^2}X_{kl}X_{ml} - \sum_{l^-}\frac{1}{y_l^2}|X_{kl}X_{ml}| \tag{3.27}$$

$$\leq \frac{1}{c_{min}d_{tot}}X^+ - \frac{1}{c_{max}d_{tot}}X^+ = \left(\frac{1}{c_{min}d_{tot}} - \frac{1}{c_{max}d_{tot}}\right)X^+ \tag{3.28}$$

where $X^+ = \sum_{l^+} X_{kl}X_{ml} = \sum_{l^-}|X_{kl}X_{ml}| = \frac{1}{2}\sum_{l=1}^{K}|X_{kl}X_{ml}| \leq \frac{1}{2}$, because $X_{m:} \perp X_{k:}$.

Now, putting these together we obtain the desired result.

$$||V_{i:} - V_{j:}||^2 = ||V_{i:}||^2 + ||V_{j:}||^2 - 2V_{i:}^T V_{j:} \tag{3.29}$$

$$\geq \frac{1}{d_{tot}c_{max}}\left(\frac{1}{\rho_k} + \frac{1}{\rho_m}\right) - \frac{2}{\sqrt{\rho_k\rho_m}}\frac{1}{2}\left(\frac{1}{c_{min}d_{tot}} - \frac{1}{c_{max}d_{tot}}\right) \tag{3.30}$$

$$= \frac{1}{d_{tot}}\left[\frac{1}{c_{max}}\left(\frac{1}{\rho_k} + \frac{1}{\rho_m}\right) - \frac{1}{\sqrt{\rho_k\rho_m}}\left(\frac{1}{c_{min}} - \frac{1}{c_{max}}\right)\right] \tag{3.31}$$

In the case when the columns of V are normalized, we have that

$$V_{il}^{norm} = \frac{1}{v_l}\frac{X_{kl}}{\sqrt{\rho_k}} \quad \text{whenever } i \in k \tag{3.32}$$

and $v_l^2 = \sum_{k=1}^{K} n_k\frac{X_{kl}^2}{\rho_k}$. Hence $\tilde{c}_{min} = \min_k \frac{n_k}{n\rho_k} \leq v_l^2 \leq \tilde{c}_{max} = \max_k \frac{n_k}{n\rho_k}$. From this the second result follows.

In the square brackets, $c_{max,min}$ depend on $\pi_{1:k}$ the degree distribution, while all the other quantities depend only of the preference frame. Hence, this expression is invariant with n, and as long as it is strictly positive, we have that the cluster separation is $\Theta(n^{-2})$.

In particular, when all ρ_k are equal, $c_{max} \leq 2c_{min}$ suffices. These bounds are not tight since $2X^+$ is never 1.

We firstly establish the main facts needed for the proof. These are properties of the eigenvalues for the off-diagonal blocks of L.

Proposition 8 (Maximum eigenvalue of a block of L) *1) $L_{kk}s_k = r_{kk}s_k$ and therefore $\lambda_{max}(L_{kk}) = r_{kk}$ for $k = 1, \dots K$. 2) $L_{kl}s_l = r_{kl}s_k$ and $\lambda_{max}(M_{kl}) = \sqrt{r_{kl}r_{lk}}$ for all $k, l = 1, \dots K$, with $k \neq l$ and*

$$M_{kl} \;=\; \begin{bmatrix} 0 & L_{kl} \\ L_{lk} & 0 \end{bmatrix} \tag{3.33}$$

Proof For part 1, note that $L_{kk} = \operatorname{diag}(s_k)^{-1} A_{kk} \operatorname{diag}(s_k)^{-1}$ and that $\frac{1}{r_{kk}} P_{kk}$ is a stochastic matrix. Then, by Proposition 9 $\lambda_{max}(L_{kk}) = \lambda_{max}(P_{kk}) = r_{kk}$, and the corresponding eigenvector of L_{kk} is s_k.

For part 2, we show that the vector $x = [s_k'/\sqrt{\rho_k}\ s_l'/\sqrt{\rho_l}]'$ is an eigenvector of M_{kl}.

$$M_{kl}x \;=\; \begin{bmatrix} L_{kl}s_l/\sqrt{\rho_l} \\ L_{lk}s_k/\sqrt{\rho_k} \end{bmatrix} \;=\; \begin{bmatrix} r_{kl}s_k/\sqrt{\rho_l} \\ r_{lk}s_l/\sqrt{\rho_k} \end{bmatrix} \;=\; \begin{bmatrix} \sqrt{r_{kl}r_{lk}}s_k\sqrt{\frac{r_{kl}}{r_{lk}\rho_l}} \\ \sqrt{r_{kl}r_{lk}}s_l\sqrt{\frac{r_{lk}}{r_{kl}\rho_k}} \end{bmatrix} \;=\; \sqrt{r_{kl}r_{lk}} \begin{bmatrix} s_k/\sqrt{\rho_k} \\ s_l/\sqrt{\rho_l} \end{bmatrix} \tag{3.34}$$

The last equality holds because

$$\frac{r_{kl}}{r_{lk}\rho_l} \;=\; \frac{r_{kl}\rho_k}{r_{lk}\rho_l}\frac{1}{\rho_k} \;=\; \frac{1}{\rho_k} \tag{3.35}$$

using the detailed balance of the reversible Markov chain defined by R. Since the eigenvector x is positive, it must correspond to the largest eigenvalue of M_{kl}.

Proposition 9 *Let x be a* spurious eigenvector *of L, associated to spurious eigenvalue λ, and denote by $s_k, x_k \in \mathbb{R}^{n_k}$ the restrictions of s, x to cluster k. Then $x_k \perp s_k$ for all $k = 1, \dots K$.*

Proof Let u be a principal eigenvector of L, and u_k its k-th block. From Proposition 9 it follows that $u_k = s_k\nu_k$, where ν is an eigenvector of R. Because L is symmetric, we know that $x \perp u$, which can be written equivalently as

$$\sum_{k=1}^{K} x_k' s_k \nu_k \;=\; 0 \tag{3.36}$$

Let $\xi_k = x'_k s_k$. If we write (3.36) for all K eigenvectors of R, we obtain the linear system

$$[\nu_1\ \nu_2\ \ldots \nu_K]\xi_k\ =\ 0 \tag{3.37}$$

Since the matrix $[\nu_1\ \nu_2\ \ldots \nu_K]$ is non-singular, the system admits only the trivial solution $\xi_k = 0$. □

Proof of Theorem 6 Let $x \in \mathbb{R}^n$ be a vector orthogonal to the K principal eigenvectors of L. Hence, by Proposition 9 each block x_k of x is orthogonal to s_k. In addition, for any pair k, l with $k \neq l$, $[x'_k \quad x'_l] \begin{bmatrix} s'_k/\sqrt{\rho_k} \\ s_l/\sqrt{\rho_l}] \end{bmatrix} = 0$, so x is orthogonal to the Frobenius eigenvector of all the off-diagonal blocks M_{kl}. We assume w.l.o.g. that $||x|| = 1$ and calculate

$$\begin{aligned}
|x'Lx| &= |\sum_{k=1}^{K} x'_k L_{kk} x_k + \sum_{k<l} [x'_k x'_l] M_{kl} \begin{bmatrix} x_k \\ x_l \end{bmatrix} && (3.38)\\
&\leq \sum_{k=1}^{K} r_{kk} |\lambda_2^{kk}| ||x_k||^2 + \sum_{k<l} \sqrt{r_{kl} r_{lk}} |\lambda_2^{kl}| (||x_k||^2 + ||x_l||^2) && (3.39)\\
&\leq c \sum_{k=1}^{K} ||x_k||^2 \left[r_{kk} + \sum_{k \neq l} \sqrt{r_{kl} r_{lk}} \right] && (3.40)\\
&\leq c \max_k \left[r_{kk} + \sum_{k<l} \sqrt{r_{kl} r_{lk}} \right] && (3.41)
\end{aligned}$$

From this the result follows. □

3.8 Matrix concentration results and the proof of theorem 3

Proposition 10 (Modified theorem from Le and Vershynin [36]) *(Concentration of the regularized Graph Laplacian) Let $\mathcal{G}, L, \hat{L}$ have the usual meaning and let d_{min} be the minimum expected degree of $\mathcal{G}$, that is $d_{min} = \min d_{1,\cdots,n}$, and denote $\hat{d}_{min}$ the analogous quantity for the observed degree of $\mathcal{G}$. Denote $d_{max} = \max_{ij} nA_{ij}$, $\gamma \geq 1$. $||.||$ denotes the spectral norm. If (a) $\hat{d}_{min} \geq \log(n)$, (b) $d_{min} \geq \log(n)$, (c) $\exists \varkappa > 0$, $d \leq \varkappa \log n$, then with probability at least $1 - e^{-r}$,*

$$||\hat{L} - L|| \leq \frac{\Psi \gamma^2}{\sqrt{\log n}} \tag{3.42}$$

where Ψ is a constant.

Proof of Proposition 10 The proof mainly follows from [36]. In the original theorem, they add equal weights to all entries of the similarity matrix to ensure the concentration of L. In this modified theorem, instead of adding weights we add the assumptions that $\hat{d}_{min}$ and d_{min} are bounded from below and prove that it will come to similar conclusion.

Denote $\tau = \log n$. In the step 1 of their proof, we modify E as $E := \hat{L} - L$, and get $E = M + T$, with $M = \hat{D}^{-1/2}(\hat{A} - A)\hat{D}^{-1/2}$, and $T = \hat{D}^{-1/2}A\hat{D}^{-1/2} - D^{-1/2}AD^{-1/2}$.

In the step 2 they give bound on $||M||$. We modify their Δ to be $\Delta_{ii} = 1$ if $\hat{d}_i \leq 8\gamma d_{max}$ and $\Delta_{ii} = \hat{d}_i/\gamma$ otherwise. The rest of the proof in step 2 still hold to this modification. We therefore obtain the bound for M as $||M|| \leq \frac{C_2\gamma^2}{\tau}(\sqrt{d_{max}} + \sqrt{\tau})$ with probability at least $1 - 2n^{-\gamma}$. C_2 is a constant.

We then follow the step 3 of their proof and bound the spectral norm with the Hilbert-Schmidt norm. We get $||T|| \leq ||T||_{HS} = \sum_{i,j=1}^{n} T_{ij}^2$, where $T_{ij} = A_{ij}[1/\sqrt{\hat{\delta}_{ij}} - 1/\sqrt{\delta_{ij}}]$ and $\hat{\delta}_{ij} = \hat{d}_i\hat{d}_j$ and $\delta_{ij} = d_i d_j$. The rest of the proof in step 3 can then be easily adapted to our modification and thus we obtain similar bound for T as $||T||^2 \leq \frac{C_6\gamma^4 d^5}{\tau^6}$ with probability $1 - e^{-2\gamma}$, C_6 a constant. Combining the results from step 2 and step 3 into the inequality D$||E|| \leq ||M|| + ||T||$, we obtain $||E|| \leq \frac{C_7\gamma^2}{\sqrt{\tau}}[(d_{max}/\tau)^{5/2} + (d_{max}/r)^{1/2} + 1] \leq \frac{\Psi\gamma^2}{\sqrt{\log n}}$ with high probability.

Proposition 11 *Let $L, \hat{L}, Y, \hat{Y}$ have the usual meaning. Let $P_{\hat{L}}$ denote the projection onto the span of $\hat{L}$'s first K left singular vectors, $\hat{\Lambda}$ and Λ denote the diagonal matrices of the first K eigenvalues of $\hat{L}$ and L accordingly. Then $P_{\hat{L}}\hat{L} = \hat{Y}\hat{\Lambda}\hat{Y}^T$, and*

$$||P_{\hat{L}}\hat{L} - L||_F^2 = ||\hat{Y}\hat{\Lambda}\hat{Y}^T - Y\Lambda Y^T||_F^2 \leq 8K||\hat{L} - L||^2 \tag{3.43}$$

Proposition 12 (Davis-Kahan theorem) *This is Davis-Kahan theorem[58] perturbation result. It puts results that relates the perturbation of L to Y.*

Let $S_0 \subset \mathbb{R}$ be an interval. Denote Y_{S_0} as an orthonormal matrix whose column space is equal to the eigenspace of L corresponding to the eigenvalues in $\lambda_{S_0}(L)$, where

$$\lambda_{S_0}(L) = \{\{\lambda_1, \cdots, \lambda_n\} \cap S_0\} \tag{3.44}$$

Denote by $\hat{Y}_{S_0}$ the analogous quantity for $P_{\hat{L}}\hat{L}$. Define the distance between S_0 and the spectrum of L outside of S_0 as

$$\Delta = \min\{|\lambda - s|; \lambda \textit{ eigenvalue of } L,\ \lambda \notin S_0, s \in S_0\} \tag{3.45}$$

If Y_{S_0} and $\hat{Y}_{S_0}$ are of the same dimension, then there is an orthogonal matrix O that depends on Y_{S_0} and $\hat{Y}_{S_0}$, such that

$$||Y_{S_0} - \hat{Y}_{S_0}||_F^2 \leq \frac{2||P_{\hat{L}} - L||_F^2}{\Delta^2} \tag{3.46}$$

Lemma 1 *Assume $x_1, x_2, \cdots, x_n \in R^n$ form an orthonormal basis. then $||x_i - x_j||_2 = \sqrt{2}$, $\forall i, j \in \{1, \cdots, n\}$*

Proof of Lemma 1 define $X = (x_1, x_2, \cdots, x_n)$ Then $X^T X = I$. $\sqrt{2} = ||X^T x_1 - X^T x_2||_2 = (x_1 - x_2)^T X X^T (x_1 - x_2) = ||x_1 - x_2||_2$

Lemma 2 *Assumption the HPFM model holds, the eigenvalues $\lambda_1, \cdots, \lambda_K$ of R are the K eigenvalues of P that have the largest absolute values. The ith eigenvector of P associated with these eigenvalues can be represented by the eigenvectors of R as*

$$(\underbrace{u_{i1}, \cdots, u_{i1}}_{n_1}, \underbrace{u_{i2}, \cdots, u_{i2}}_{n_2} \cdots, \underbrace{u_{iK}, \cdots, u_{iK}}_{u_K})$$

$i \in \{1, \cdots, K\}$

Proof of Lemma 2 Since P is block stochastic and $R_{lk} = \sum_{j \in \mathcal{C}_k} P_{ij}$, $i \in \mathcal{C}_l, j \in \mathcal{C}_k$. The first K eigenvectors of P are piecewise constant with respect to the clusters. Assume $x \in R^n$ is one of the first K eigenvectors of P associated with λ.

$$Px = \lambda x \tag{3.47}$$

Since x is piecewise constant, we define a vector $u \in R^K$, $u_k = x_j$ if node $j \in \mathcal{C}_k$.
Assume node $i \in \mathcal{C}_l$.

$$\lambda x_i = \sum_{j=1}^{n} [P_{ij} x_j] = \sum_{k=1}^{K} [\sum_{j \in \mathcal{C}_k} P_{ij} x_j] \tag{3.48}$$

$$\text{that is, } \lambda u_l = \sum_{k=1}^{K} [R_{lk} u_k] \tag{3.49}$$

$i = 1, \cdots, n$.

Therefore u is an eigenvector of R associated with the eigenvalue λ.

Lemma 3 *Assume* $d_i = \sum_{j=1}^{n} Bernoulli(A_{ij})$, $A_{ij} \leq 1$, *then*

$$P(|\sqrt{\hat{d}_i} - \sqrt{d_i}| \leq \epsilon) \geq 1 - 2\exp[-\frac{\epsilon^2}{2 + \epsilon/\sqrt{d_i}}] \tag{3.50}$$

Proof of Lemma 3 Using Chernoff bound, we can get that,

$$P[|\hat{d}_i - d_i| < \delta d_i] \geq 1 - 2e^{-\frac{\delta^2 d_i}{2+\delta}}$$

$$P[|\sqrt{\hat{d}_i} - \sqrt{d_i}| < \delta\sqrt{d_i}] \geq P[|\sqrt{\hat{d}_i} - \sqrt{d_i}| < \frac{\delta d_i}{\sqrt{\hat{d}_i} + \sqrt{d_i}}] \geq 1 - 2e^{-\frac{\delta^2 d_i}{2+\delta}}$$

Denote $\epsilon = \delta\sqrt{d_i}$, we can get

$$P(|\sqrt{\hat{d}_i} - \sqrt{d_i}| \leq \epsilon) \geq 1 - 2\exp[-\frac{\epsilon^2}{2 + \epsilon/\sqrt{d_i}}]$$

Proof of theorem 3 Let B, R, X, U, Y, V have the usual meaning. $\lambda_1, \cdots, \lambda_K$ are the eigenvalues of R.

In the HPFM model case, since $\lambda_{K+1} = \cdots = \lambda_n = 0$, the eigengap for L between the first K leading eigenvalues and the rest of the eigenvalues is λ_K. Denote $S_0 = (\frac{\lambda_K}{2}, 2) \subset \mathbb{R}$.

Using proposition 10, when n is sufficiently large, we obtain

$$|\lambda_K - \hat{\lambda}_K| \leq ||L - \hat{L}|| \leq \frac{\Psi\gamma^2}{\sqrt{\log n}} \leq \lambda_K/10 \tag{3.51}$$

we therefore have $\lambda_{S_0}(\hat{L}) = \{\hat{\lambda}_1, \cdots, \hat{\lambda}_K\}$.

Therefore Y and $\hat{Y}$ are of the same dimension. Using proposition 12, we obtain,

$$\frac{1}{2}||\hat{Y} - Y\mathcal{O}||_F^2 \leq \frac{||P_{\hat{L}}\hat{L} - L||_F^2}{\Delta^2} \tag{3.52}$$

where $\mathcal{O}$ is an orthonormal matrix. $\Delta = \lambda_K/2$. Further apply proposition 11, we have the following inequality holds,

$$\frac{1}{2}||\hat{Y} - Y\mathcal{O}||_F^2 \leq \frac{4||P_{\hat{L}}\hat{L} - L||_F^2}{\lambda_K^2} \leq \frac{32K||\hat{L} - L||^2}{\lambda_K^2} \tag{3.53}$$

Now we take a closer look at $\frac{1}{2}||\hat{Y} - Y\mathcal{O}||_F^2$ on the left hand side.

$$\frac{1}{2}||\hat{Y} - Y\mathcal{O}||_F^2 = \frac{1}{2}||D^{1/2}V - \hat{D}^{1/2}\hat{V}||_F^2 = \frac{1}{2}||D^{1/2}(\hat{V} - V) + (\hat{D}^{1/2} - D^{1/2})\hat{V}||_F^2 \tag{3.54}$$

Using (3.53), we obtain,

$$\underbrace{||D^{1/2}(\hat{V} - V)||_F^2}_{(a)} \leq \underbrace{||(\hat{D}^{1/2} - D^{1/2})\hat{V}||_F^2}_{(b)} + \frac{64K||\hat{L} - L||^2}{\lambda_K^2} \tag{3.55}$$

$$(b) \leq \max_i (d_i^{1/2} - \hat{d}_i^{1/2})^2 ||\hat{V}||_F^2 \tag{3.56}$$

Using lemma 3, we can get,

$$P(|\sqrt{\hat{d}_i} - \sqrt{d_i}| \leq \varepsilon) \geq 1 - 2\exp[-\frac{\epsilon^2}{2 + \epsilon/\sqrt{d_i}}] \geq 1 - 2\exp[-\frac{\epsilon^2}{2 + \epsilon/\sqrt{\log n}}] \tag{3.57}$$

Meanwhile we have

$$||\hat{V}||_F^2 = ||\hat{D}^{-1/2}\hat{Y}||_F^2 \leq \frac{1}{\hat{d}_{min}}||\hat{Y}||_F^2 = \frac{K}{\hat{d}_{min}} \tag{3.58}$$

Thus with probability at least $1 - 2\exp[-\frac{\epsilon^2}{2+\epsilon/\sqrt{\log n}}]$

$$(b) \leq \frac{K\epsilon^2}{\hat{d}_{min}} \tag{3.59}$$

$$\frac{1}{2}||\hat{V} - V||_F^2 \times d_{min} \leq (a) \leq \frac{K\epsilon^2}{\hat{d}_{min}} + \frac{64K||\hat{L} - L||^2}{\lambda_K^2} \tag{3.60}$$

The above inequality gives us a bound for the perturbation of first K eigenvectors of P.

Denote the l_2 norm of the perturbation for each row of V is e_i, $i = 1, \cdots, n$. Denote the number of rows that have perturbation greater than $\frac{1}{2}\min_{i \neq j} ||V(i,:) - V(j,:)||$ is m. Using Proposition 4 , we have

$$\frac{K\epsilon^2}{\hat{d}_{min}d_{min}} + \frac{64K||\hat{L} - L||^2}{d_{min}\lambda_K^2} \geq \sum_{i=1}^{n} e_i^2 \geq \frac{m}{4d_{tot}} g_{row}. \tag{3.61}$$

Solving the above, and use proposition 10. With probability at least $(1-2\exp[-\frac{\epsilon^2}{2+\epsilon/\sqrt{\log n}}])(1-e^{-r})$, we get,

$$m \leq \frac{C_0 K \gamma^4 d_{tot}}{\lambda_K^2 d_{min} g_{row} \log n} + \frac{4K d_{tot} \epsilon^2}{g_{row} \hat{d}_{min} d_{min}} \tag{3.62}$$

$$p_{err} = m/n \leq \frac{K d_{tot}}{n d_{min} g_{row}} \left[\frac{C_0 \gamma^4}{\lambda_K^2 \log n} + \frac{4\epsilon^2}{\hat{d}_{min}} \right] \tag{3.63}$$

In the PFM case when $\lambda_K \neq 0$, we can modify S_0 to be $S_0 = (\frac{\lambda_K + \lambda_{K+1}}{2}, 2)$, $\Delta = \sigma/2$, and then when n is sufficiently large $\lambda_{S_0}(\hat{L}) = \{\hat{\lambda}_1, \cdots, \hat{\lambda}_K\}$ also holds. The rest of proof can be down similarly, and we obtain

$$p_{err} \leq \frac{K d_{tot}}{n d_{min} g_{row}} \left[\frac{C_0 \gamma^4}{\sigma^2 \log n} + \frac{4\epsilon^2}{\hat{d}_{min}} \right] \tag{3.64}$$

Chapter 4

MODEL FREE GUARANTEES FOR MODEL BASED COMMUNITY RECOVERY

We have discussed how model-based recovery in networks has been made spectacular progress in the last few years. The understanding of the block-models, including SBM, DC-SBM and PFM has been advanced, especially in understanding the conditions when recovery of the true clustering is possible with small or no error. The algorithms for recovery with guarantees have also been improved. However, the impact of the above results is limited by the assumption that the observed data comes from the model.

This chapter proposes a framework to provide *theoretical guarantees for the results of model based clustering algorithms, without making any assumption about the data generating process.* To describe the idea, we need some notation. Assume that a graph $\mathcal{G}$ on n nodes is observed. A model-based algorithm clusters $\mathcal{G}$, and outputs clustering $\mathcal{C}$ and parameters $\mathcal{M}(\mathcal{G}, \mathcal{C})$.

The framework is as follows: if $\mathcal{M}(\mathcal{G}, \mathcal{C})$ fits the data $\mathcal{G}$ well, then we shall prove that any other clustering $\mathcal{C}'$ of $\mathcal{G}$ that also fits $\mathcal{G}$ well will be a small perturbation of $\mathcal{C}$. If this holds, then $\mathcal{C}$ with model parameters $\mathcal{M}(\mathcal{G}, \mathcal{C})$ can be said to capture the data structure in a meaningful way.

We exemplify our approach by obtaining model-free guarantees with $\mathcal{M}(\mathcal{G}, \mathcal{C})$ as SBM and PFM models. Moreover, we show that model-free and model-based results are intimately connected.

We follow the notation convention from Chapter 1.

4.1 Main theorem: blueprint and results for PFM, SBM

Let $\mathcal{M}$ be a model class, such as SBM, DC-SBM, PFM, and denote $\mathcal{M}(\mathcal{G}, \mathcal{C}) \in \mathcal{M}$ to be a model that is compatible with $\mathcal{C}$ and is fitted in some way to graph $\mathcal{G}$ (we do not assume in general that this fit is optimal).

Theorem 13 (Generic Theorem) *We say that clustering $\mathcal{C}$* fits *$\mathcal{G}$* well w.r.t *$\mathcal{M}$ iff $\mathcal{M}(\mathcal{G}, \mathcal{C})$ is "close to" $\mathcal{G}$. If $\mathcal{C}$ fits $\mathcal{G}$ well w.r.t $\mathcal{M}$, then (subject to other technical conditions) any other clustering $\mathcal{C}'$ which also fits $\mathcal{G}$ well is close to $\mathcal{C}$, i.e. $\mathrm{dist}(\mathcal{C}, \mathcal{C}')$ is small.*

In what follows, we will instantiate this Generic Theorem, and the concepts therein; in particular the following will be formally defined. (1) Model construction, i.e an algorithm to fit a

model in $\mathcal{M}$ to $(\mathcal{G}, \mathcal{C})$. This is necessary since we want our results to be computable in practice. (2) A goodness of fit measure between $\mathcal{M}(\mathcal{C}, \mathcal{G})$ and the data $\mathcal{G}$. (3) A distance between clusterings. We adopt the widely used Misclassification Error (or Hamming) distance dist and the Weighted Misclassification Error dist_d defined in Section 1.2.1. While dist is more popular, we believe dist_d is more natural, especially when node degrees are dissimilar, as d can be seen as a natural measure on the set of nodes, and dist_d is equivalent to the *earth-mover's* distance.

4.2 Main result for PFM

Constructing a model Given a graph $\mathcal{G}$ and a clustering $\mathcal{C}$ of its nodes, we wish to construct a PFM compatible with $\mathcal{C}$, so that its Laplacian L satisfies that $||\hat{L} - L||$ is small.

Let the spectral decomposition of $\hat{L}$ be

$$\hat{L} \;=\; [\hat{Y}\ \hat{Y}_{low}] \begin{bmatrix} \hat{\Lambda} & 0 \\ 0 & \hat{\Lambda}_{low} \end{bmatrix} \begin{bmatrix} \hat{Y}^T \\ \hat{Y}_{low}^T \end{bmatrix} \;=\; \hat{Y}\hat{\Lambda}\hat{Y}^T + \hat{Y}_{low}\hat{\Lambda}_{low}\hat{Y}_{low}^T \tag{4.1}$$

where $\hat{Y} \in \mathbb{R}^{n\times K}, \hat{Y}_{low} \in \mathbb{R}^{n\times(n-K)}$, $\hat{\Lambda} = diag(\hat{\lambda}_1, \cdots, \hat{\lambda}_K)$, $\hat{\Lambda}_{low} = diag(\hat{\lambda}_{K+1}, \cdots, \hat{\lambda}_n)$. To ensure that the matrices $\hat{Y}, \hat{Y}_{low}$ are uniquely defined we assume throughout the paper that $\hat{L}$'s K-th eigengap, i.e, $|\lambda_K| - |\lambda_{K+1}|$, is non-zero.

Assumption 8 *The eigenvalues of $\hat{L}$ satisfy $\hat{\lambda}_1 = 1 \geq |\hat{\lambda}_2| \geq \ldots \geq |\hat{\lambda}_K| > |\hat{\lambda}_{K+1}| \geq \ldots |\hat{\lambda}_n|$.*

Denote the subspace spanned by the columns of M, for any M matrix, by $\mathcal{R}(M)$, and $||\,||$ the Euclidean or spectral norm.

PFM Construction Algorithm

Input Graph $\mathcal{G}$ with $\hat{A}, \hat{D}, \hat{L}, \hat{Y}, \hat{\Lambda}$, clustering $\mathcal{C}$ with indicator matrix Z.

Output $(A, L) = PFM(\mathcal{G}, \mathcal{C})$

1. Construct an orthogonal matrix derived from Z.

$$Y_Z = \hat{D}^{1/2} Z C^{-1/2}, \text{ with } C = Z^T \hat{D} Z \text{ the column normalization of } Z. \tag{4.2}$$

Note $C_{kk} = \sum_{i \in k} \hat{d}_i$ is the volume of cluster k.

2. Project Y_Z on $\hat{Y}$ and perform Singular Value Decomposition.

$$F \;=\; Y_Z^T \hat{Y} \;=\; U \Sigma V^T \tag{4.3}$$

3. Change basis in $\mathcal{R}(Y_Z)$ to align with $\hat{Y}$.

$$Y \;=\; Y_Z U V^T. \quad \text{Complete } Y \text{ to an orthonormal basis } [Y\ B] \text{ of } \mathbb{R}^n. \tag{4.4}$$

4. Construct Laplacian L and edge probability matrix A.

$$L \;=\; Y \hat{\Lambda} Y^T + (BB^T) \hat{L} (BB^T), \qquad A \;=\; \hat{D}^{1/2} L \hat{D}^{1/2}. \tag{4.5}$$

The PFMconstructed from the algorithm above is only a construction instead of an estimation. We cannot ensure the PFM!to be positive from this construction or ensure that it is the maximum likelihood estimation to the graph. The reason for this is that it is fundamentally a hard problem.

Proposition 1 *Let $\mathcal{G}, \hat{A}, \hat{D}, \hat{L}, \hat{Y}, \hat{\Lambda}$ and Z be defined as above, and $(A, L) = PFM(\mathcal{G}, \mathcal{C})$. Then,*

1. *$\hat{D}$ and L, or A define a PFM with degrees $\hat{d}_{1:n}$, whenever $A_{ij} \geq 0$.*

2. *The columns of Y are eigenvectors of L with eigenvalues $\hat{\lambda}_{1:K}$.*

3. $\hat{D}^{1/2}\mathbf{1}$ *is an eigenvector of both* L *and* $\hat{L}$ *with eigenvalue* $\hat{\lambda}_1 = 1$.

The proof is relegated to Secton 4.8, as are all the omitted proofs.

$PFM(\mathcal{G}, \mathcal{C})$ is an estimator for the PFM parameters given the clustering. It is evidently not the Maximum Likelihood estimator, but we can show that it is consistent in the following sense.

Proposition 2 (Informal) *Assume that* $\mathcal{G}$ *is sampled from a PFM with parameters* D^*, L^* *and compatible with* $\mathcal{C}^*$*, and let* $L = PFM(\mathcal{G}, \mathcal{C}^*)$*. Then, under standard recovery conditions for PFM (e.g [62])* $||L^* - L|| = o(1)$ *w.r.t.* n.

Assumption 9 (Goodness of fit for PFM) $||\hat{L} - L|| \leq \varepsilon$.

$PFM(\mathcal{G}, \mathcal{C})$ instantiates $\mathcal{M}(\mathcal{G}, \mathcal{C})$, and Assumption 9 instantiates the goodness of fit measure. It remains to prove an instance of Generic Theorem 13 for these choices.

Theorem 14 (Model free recovery guarantee for PFM) *Let* $\mathcal{G}$ *be a graph with* $\hat{d}_{1:n}, \hat{D}, \hat{L}, \hat{\lambda}_{1:n}$ *as defined, and* $\hat{L}$ *satisfy Assumption 8. Let* $\mathcal{C}, \mathcal{C}'$ *be two clusterings with* K *clusters, and* L, L' *be their corresponding Laplacians, defined as in* (4.5)*, and satisfy Assumption 9 respectively. Set* $\delta = \frac{(K-1)\varepsilon^2}{(|\hat{\lambda}_K| - |\hat{\lambda}_{K+1}|)^2}$ *and* $\delta_0 = \min_k C_{kk} / \max_k C_{kk}$ *with* C *defined as in* (4.2)*, where* k *indexes the clusters of* $\mathcal{C}$*. Then, whenever* $\delta \leq \delta_0$*,*

$$\mathrm{dist}_{\hat{d}}(\mathcal{C}, \mathcal{C}') \leq \frac{\max_k C_{kk}}{\sum_k C_{kk}} \delta, \tag{4.6}$$

with $\mathrm{dist}_{\hat{d}}$ *being the weighted ME distance* (1.3)*.*

In the remainder of this section we outline the proof steps, while the partial results of Proposition 3, 4, 5 are proved in the Supplement. First, we apply the perturbation bound called the Sinus Theorem of Davis and Kahan, in the form presented in Chapter V of [58].

Proposition 3 *Let* $\hat{Y}, \hat{\lambda}_{1:n}, Y$ *be defined as usual. If Assumptions 8 and 9 hold, then*

$$||\operatorname{diag}(\sin\theta_{1:K}(\hat{Y}, Y))|| \leq \frac{\varepsilon}{|\hat{\lambda}_K| - |\hat{\lambda}_{K+1}|} = \varepsilon' \tag{4.7}$$

where $\theta_{1:K}$ *are the canonical (or principal) angles between* $\mathcal{R}(\hat{Y})$ *and* $\mathcal{R}(Y)$ *(see e.g [13]).*

The next step concerns the closeness of $Y, \hat{Y}$ in Frobenius norm. Since Proposition 3 bounds the sinuses of the canonical angles, we exploit the fact that the cosines of the same angles are the singular values of $F = Y^T\hat{Y}$ of (4.3).

Proposition 4 *Let $M = YY^T$, $\hat{M} = \hat{Y}\hat{Y}^T$ and F, ε' as above. Assumptions 8 and 9 imply that*

1. $||F||_F^2 = \operatorname{trace} M\hat{M}^T \geq K - (K-1)\varepsilon'^2$.

2. $||M - \hat{M}||_F^2 \leq 2(K-1)\varepsilon'^2$.

Now we show that all clusterings which satisfy Proposition 4 must be close to each other in the weighted ME distance. For this, we first need an intermediate result. Assume we have two clusterings $\mathcal{C}, \mathcal{C}'$, with K clusters, for which we construct Y_Z, Y, L, M, respectively Y'_Z, Y', L', M' as above. Then, the subspaces spanned by Y and Y' will be close.

Proposition 5 *Let $\hat{L}$ satisfy Assumption 8 and let $\mathcal{C}, \mathcal{C}'$ represent two clusterings for which L, L' satisfy Assumption 9. Then, $||Y_Z^T Y'_Z||_F^2 \geq K - 4(K-1)\varepsilon'^2 = K - \delta$*

The main result now follows from Proposition 5 and Theorem 9 of [43], as shown in the Supplement. This proof approach is different from the existing perturbation bounds for clustering, which all use counting arguments. The result of [43] is a *local* equivalence, which bounds the error we need in terms of δ defined above ("local" meaning the result only holds for small δ).

4.3 Main Theorem for SBM

In this section, we offer an instantiation of Generic Theorem 13 for the case of the SBM. As before, we start with a model estimator, which in this case is the Maximum Likelihood estimator.

SBM Estimation Algorithm

Input Graph with $\hat{A}$, clustering $\mathcal{C}$ with indicator matrix Z.

Output $A = SBM(\mathcal{G}, \mathcal{C})$

1. Construct an orthogonal matrix derived from Z: $Y_Z = ZC^{-1/2}$ with $C = Z^T Z$.
2. Estimate the edge probabilities: $B = C^{-1} Z^T \hat{A} Z C^{-1}$.
3. Construct A from B by $A = ZBZ^T$.

Proposition 6 *Let $\tilde{B} = C^{1/2} B C^{1/2}$ and denote the eigenvalues of $\tilde{B}$, ordered by decreasing magnitude, by $\lambda_{1:K}$. Let the spectral decomposition of $\tilde{B}$ be $\tilde{B} = U \Lambda U^T$, with U an orthogonal matrix and $\Lambda = \mathrm{diag}(\lambda_{1:K})$. Then*

1. *A is a SBM.*
2. *$\lambda_{1:K}$ are the K principal eigenvalues of A. The remaining eigenvalues of A are zero.*
3. *$A = Y \Lambda Y^T$ where $Y = Y_Z U$.*

Assumption 10 (Eigengap) *B is non-singular (or, equivalently, $|\lambda_K| > 0$.*

Assumption 11 (Goodness of fit for SBM) $||\hat{A} - A|| \leq \varepsilon$.

With the model (SBM), estimator, and goodness of fit defined, we are ready for the main result.

Theorem 15 (Model free recovery theorem for SBM) *Let $\mathcal{G}$ be a graph with incidence matrix $\hat{A}$, and $\hat{\lambda}_K^A$ the K-th singular value of $\hat{A}$. Let $\mathcal{C}, \mathcal{C}'$ be two clusterings with K clusters, satisfying Assumptions 10 and 11. Set $\delta = \frac{4K\varepsilon^2}{|\hat{\lambda}_K^A|^2}$ and $\delta_0 = \min_k n_k / \max_k n_k$, where k indexes the clusters of $\mathcal{C}$. Then, whenever $\delta \leq \delta_0$, $\mathrm{dist}(\mathcal{C}, \mathcal{C}') \leq \delta \max_k n_k / n$, where* dist *represents the ME distance* (1.2).

Note that the eigengap of $\hat{A}$, $\hat{\Lambda}_K^A$ is not bounded above, and neither is ε. Since the SBM is less flexible than the PFM, we expect that for the same data $\mathcal{G}$, Theorem 15 will be more restrictive than Theorem 14.

4.4 The results in perspective

4.4.1 Cluster validation

Theorems like 14, 15 can provide model free guarantees for clustering. We exemplify this procedure in the experimental Section 4.6, using standard spectral clustering as described in e.g [57, 56, 52]. What is essential is that all the quantities such as ε and δ *are computable from the data*.

Moreover, if Y is available, then the bound in Theorem 14 can be improved.

Proposition 7 *Theorem 14 holds when δ is replaced by $\delta_Y = K - \langle \hat{M}, M \rangle_F + (K-1)(\varepsilon')^2 + 2\sqrt{2(K-1)}\varepsilon' ||\hat{M} - M||_F$, with $\varepsilon' = \varepsilon/(|\hat{\lambda}_K| - |\hat{\lambda}_{K+1}|)$ and $M, \hat{M}$ defined in Proposition 4.*

4.4.2 Using existing model-based recovery theorems to prove model-free guarantees

We exemplify this by using (the proof of) Theorem 3 of [62] to prove the following.

Theorem 16 (Alternative result based on [62] for PFM) *Under the same conditions as in Theorem 14,* $\mathrm{dist}_{\hat{d}}(\mathcal{C}, \mathcal{C}') \leq \delta_{WM}$, *with* $\delta_{WM} = 128 \frac{K\varepsilon^2}{(|\hat{\lambda}_K| - |\hat{\lambda}_{K+1}|)^2}$.

It follows, too, that with the techniques in this paper, the error bound in [62] can be improved by a factor of 128.

Similarly, if we use the results of [57] we obtain alternative model-free guarantee for the SBM.

Assumption 1 (Alternative goodness of fit for SBM) $||\hat{L}^2 - L^2||_F \leq \varepsilon$, *where $\hat{L}, L$ are the Laplacians of $\hat{A}$ and $A = SBM(\mathcal{G}, \mathcal{C})$ respectively.*

Theorem 17 (Alternative result based on [57] for SBM) *Under the same conditions as in Theorem 15, except for replacing Assumption 11 with 1,* $\mathrm{dist}(\mathcal{C}, \mathcal{C}') \leq \delta_{RCY}$ *with* $\delta_{RCY} = \frac{\varepsilon^2}{|\hat{\lambda}_K|^4} \frac{16 \max_k n_k}{n}$.

In comparison with Theorem 11, problem with this result is that Assumption 1 is much stronger than Assumption 11 (being in Frobenius norm). The more recent results of [56] (with unspecified constants) in conjunction with our original Assumptions 10, 11, and the assumption that all clusters have equal sizes, give a bound of $\mathcal{O}(K\varepsilon^2/\hat{\lambda}_K^2)$ for the SBM; hence our model-free Theorem 15 matches this more restrictive model-based theorem.

4.4.3 Sanity checks and Extensions

It can be easily verified that if indeed $\mathcal{G}$ is sampled from a SBM, or PFM, then for large enough n, and large enough model eigengap, Assumptions 8 and 9 (or 10 and 11) will hold.

Some immediate extensions and variations of Theorems 14, 15 are possible. For example, one could replace the spectral norm by the Frobenius norm in Assumptions 9 and 11, which would simplify some of the proofs. However, using the Frobenius norm would be a much stronger assumption [57] Theorem 14 holds not just for simple graphs, but in the more general case when $\hat{A}$ is a weighted graph (i.e. a *similarity matrix*). The theorems can be extended to cover the case when $\mathcal{C}'$ is a clustering that is α-worse than $\mathcal{C}$, i.e when $||L' - \hat{L}|| \geq ||L - \hat{L}||(1-\alpha)$.

4.4.4 Clusterability and resilience

Our Theorems also imply the stability of a clustering to perturbations of the graph $\mathcal{G}$. Indeed, let $\hat{L}'$ be the Laplacian of $\mathcal{G}'$, a perturbation of $\mathcal{G}$. If $||\hat{L}' - \hat{L}|| \leq \varepsilon$, then $||\hat{L}' - L|| \leq 2\varepsilon$, and (1) $\mathcal{G}'$ is well fitted by a PFM whenever $\mathcal{G}$ is, and (2) $\mathcal{C}$ is δ stable w.r.t $\mathcal{G}'$, hence $\mathcal{C}$ is what some authors [15] call *resilient*.

A graph $\mathcal{G}$ is *clusterable* when $\mathcal{G}$ can be fitted well by some clustering $\mathcal{C}^*$. Much work [6, 12] has been devoted to showing that clusterability implies that finding a $\mathcal{C}$ close to $\mathcal{C}^*$ is computationally efficient. Such results can be obtained in our framework, by exploiting existing recovery theorems such as [57, 56, 62], which give recovery guarantees for Spectral Clustering, under the assumption of sampling from the model. For this, we can simply replace the model assumption with the assumption that there is a $\mathcal{C}^*$ for which L (or A) satisfies Assumptions 8 and 9 (or 10 and

11).

4.5 Related work

To our knowledge, there is no work of the type of Theorem 1 in the literature on SBM, DC-SBM, PFM. The closest work is by [11] which guarantees approximate recovery *assuming* $\mathcal{G}$ is close to a DC-SBM. In Chapter 2 we have already shwon that the assumptiosn in [11] is much stronger than here.

Spectral clustering is also used for loss-based clustering in (weighted) graphs and some stability results exist in this context. Even though they measure clustering quality by different criteria, so that the ε values are not comparable, we review them here. The recent paper of [54], Theorem 1.2 states that if the K-way *Cheeger constant* of $\mathcal{G}$ is $\rho(k) \leq (1-\hat{\lambda}_{K+1})/(cK^3)$ then the clustering error[1] $\mathrm{dist}_{\hat{d}}(\mathcal{C}, \mathcal{C}^{opt}) \leq C/c = \delta_{PSZ}$. In the current proof, the constant $C = 2 \times 10^5$; moreover, $\rho(K)$ cannot be computed tractably. In [47], the bound δ_{MSX} depends on ε_{MSX}, the *Normalized Cut* scaled by the eigengap. Since both bounds refer to the result of spectral clustering, we can compare the relationship between δ_{MSX} and ε_{MSX}; for [47], this is $\delta_{MSX} = 2\varepsilon_{MSX}[1 - \varepsilon_{MSX}/(K-1)]$, which is about $K-1$ times larger than δ when $\epsilon = \epsilon_{MSX}$. In [7], $\mathrm{dist}(\mathcal{C}, \mathcal{C}')$ is defined in terms of $||Y_Z^T - Y_Z'||_F^2$, and the loss is (closely related) to $||\hat{A} - SBM(\mathcal{G}, \mathcal{C})||_F^2$. The bound does not take into account the eigengap, that is, the stability of the subspace $\hat{Y}$ itself.

Bootstrap for validating a clustering $\mathcal{C}$ was studied in [30] (see also references therein for earlier work). We will discuss this class of methods in the next chapter. In [3] the idea is to introduce a statistics, and large deviation bounds for it, *conditioned on sampling from a SBM* (with covariates) and on a given $\mathcal{C}$.

4.6 Experimental evaluation

Experiment Setup Given $\mathcal{G}$, we obtain a clustering $\mathcal{C}_0$ by spectral clustering [52]. Then we calculate clustering $\mathcal{C}$ by perturbing $\mathcal{C}_0$ with gradually increasing noise. For each $\mathcal{C}$, we construct

[1] The results is stronger, bounding the perturbation of each cluster individually by δ_{PSZ}, but it also includes a factor larger than 1, bounding the error of K-means algorithm.

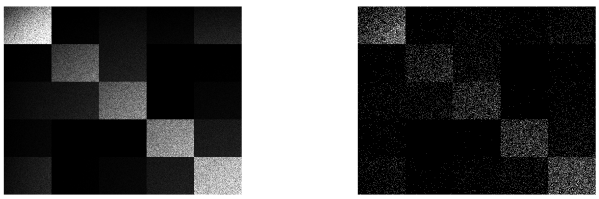
PFM $(\mathcal{C}, \mathcal{G})$and SBM$(\mathcal{C}, \mathcal{G})$ model, and further compute ϵ, δ and δ_0. If $\delta \leq \delta_0$, $\mathcal{C}$ is guaranteed to be stable by the theorems. In the remainder of this section, we describe the data generating process for the simulated datasets and the results we obtained.

PFM Datasets We generate from PFM model with $K = 5$, $n = 10000$, $\lambda_{1:K} = (1, 0.875, 0.75, 0.625, 0.5)$. $eigengap = 0.48$, $n_{1:K} = (2000, 2000, 2000, 2000, 2000)$. The stochastic matrix R and its stationary distribution ρ are shown below. We sample an adjacency matrix $\hat{A}$ from A (shown below).

$$\rho = \begin{bmatrix} .25 & .12 & .17 & .18 & .28 \end{bmatrix}$$

$$R = \begin{bmatrix} .79 & .02 & .06 & .03 & .10 \\ .03 & .71 & .23 & .00 & .02 \\ .09 & .16 & .69 & .00 & .06 \\ .04 & .00 & .00 & .80 & .16 \\ .10 & .01 & .03 & .11 & .76 \end{bmatrix}$$

A $\hat{A}$



Perturbed PFM Datasets A is obtained from the previous model by perturbing its principal subspace (details in Supplement). Then we sample $\hat{A}$ from A.

Lancichinetti-Fortunato-Radicchi (LFR) simulated matrix [35] The LFR benchmark graphs are widely used for community detection algorithms, due to heterogeneity in the distribution of node degree and community size. A LFR matrix is simulated with $n = 10000$, $K = 4$, $n_k = (2467, 2416, 2427, 2690)$ and $\mu = 0.2$, where μ is the mixing parameter indicating the fraction of edges shared between a node and the other nodes from outside its community.

Political Blogs Dataset A directed network $\vec{A}$ of hyperlinks between weblogs on US politics, compiled from online directories by Adamic and Glance [2], where each blog is assigned a political leaning, liberal or conservative, based on its blog content. The network A contains 1490 blogs. After erasing the disconnected nodes, $n = 983$. We study $\hat{A} = (\vec{A}^T \vec{A})^3$, which is a smoothed

undirected graph. For $\vec{A}^T \vec{A}$ we find no guarantees.

The first two data sets are expected to fit the PFM well, but not the SBM, while the LFR data is expected to be a good fit for a SBM. Since all bounds can be computed on weighted graphs as well, we have run the experiments also on the edge probability matrices A used to generate the PFM and perturbed PFM graphs.

The results of these experiments are summarized in Figure 5.5. For all of the experiments, the clustering $\mathcal{C}$ is ensured to be stable by Theorem 14 as the unweighted error grows to a breaking point, then the assumptions of the theorem fail. In particular, the $\mathcal{C}_0$ is always stable in the PFM framework.

Comparing δ from Theorem 15 to that from Theorem 14, we find that Theorem 15 (guarantees for SBM) is much harder to satisfy. All δ values from Theorem 9 are above 1, and not shown.[2] In particular, for the SBM model class, the $\mathcal{C}$ cannot be proved stable even for the LFR data.

Note that part of the reason why with the PFM model very little difference from the clustering $\mathcal{C}_0$ can be tolerated for a clustering to be stable is that the large eigengap makes $PFM(\mathcal{G}, \mathcal{C})$ differ from $PFM(\mathcal{G}, \mathcal{C}_0)$ even for very small perturbations.

By comparing the bounds for $\hat{A}$ with the bounds for the "weighted graphs" A, we can evaluate that the sampling noise on δ is approximately equal to that of the clustering perturbation. Of course, the sampling noise varies with n, decreasing for larger graphs. Moreover, from Political Blogs data, we see that "smoothing" a graph, by e.g. taking powers of its adjacency matrix, has a stability inducing effect.

4.7 Conclusion

This paper makes several contributions. At a high level, it poses the problem of model free validation in the area of community detection in networks. The stability paradigm is not entirely new, but using it explicitly with model-based clustering (instead of cost-based) is. So is "turning around"

[2] We also computed δ_{RCY} but the bounds were not informative.

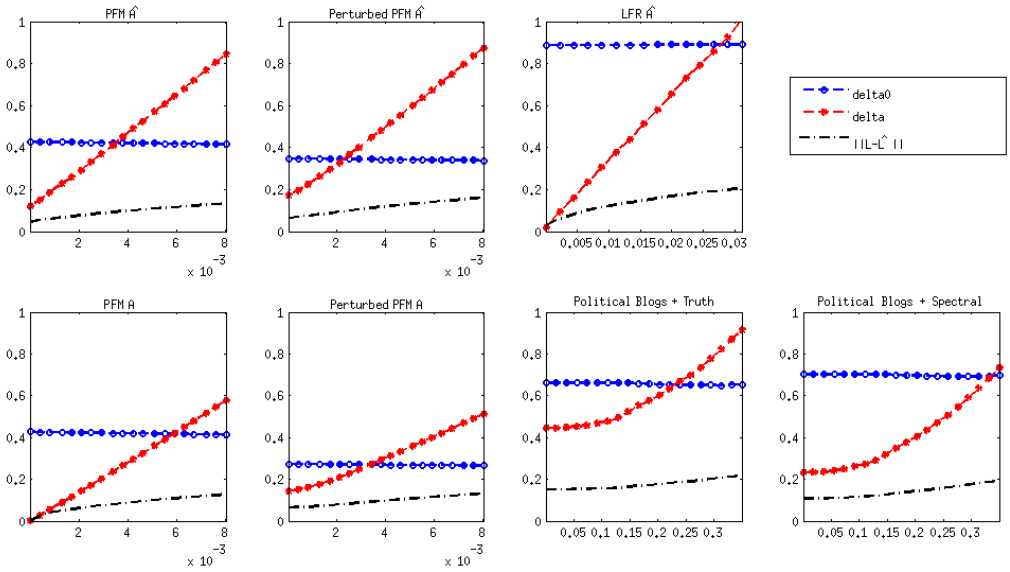


Figure 4.1: Quantities $\epsilon, \delta, \delta_0$ from Theorem 14 plotted vs $\mathrm{dist}(\mathcal{C}, \mathcal{C}_0)$ for various datasets: $\hat{A}$ denotes a simple graph, while A denotes a weighted graph (i.e. a non-negative matrix). For the Political Blogs: Truth means $\mathcal{C}_0$ is true clustering of [2], spectral means $\mathcal{C}_0$ is obtained from spectral clustering. For SBM, δ is always greater than δ_0.

the model-based recovery theorems to be used in a model-free framework.

All quantities in our theorems are computable from the data and the clustering $\mathcal{C}$, i.e do not contain undetermined constants, and do not depend on parameters that are not available. As with distribution-free results in general, making fewer assumptions allows for less confidence in the conclusions, and the results are not always informative. Sometimes this should be so, e.g when the data does not fit the model well. But it is also possible that the fit is good, but not good enough to satisfy the conditions of the theorems as they are currently formulated. This happens with the SBM bounds, and we believe tighter bounds are possible for this model. It would be particularly interesting to study the non-spectral, sharp thresholds of [1] from the point of view of model-free recovery. A complementary problem is to obtain *negative guarantees* (i.e that $\mathcal{C}$ is *not* unique up to perturbations).

At the technical level, we obtain several different and model-specific stability results, that bound the perturbation of a clustering by the perturbation of a model. They can be used both

in model-free and in existing or future model-based recovery guarantees, as we have shown in Section 4.1 and in the experiments. The proof techniques that lead to these results are actually simpler, more direct, and more elementary than the ones found in previous papers.

4.8 Proofs

Proof of Proposition 2

1. Proof by verification.

2. $LY = Y\hat{\Lambda}Y^TY + (BB^T)\hat{L}(BB^T)Y = Y\hat{\Lambda}$. Since B is the orthogonal complement of Y, it follows that it is a stable subspace as well.

3. This is a well known result; see for example [58].

The celebrated Sinus Theorem is reproduced here for completeness.

Theorem 18 (Sinus Theorem of Davis-Kahan, from [58], Theorem V.3.6) *Let $\hat{L}$ be a Hermitian matrix with spectral resolution given by* (4.1)*, Y be any $n \times K$ matrix with orthonormal columns, and M any symmetric $K \times K$ matrix with eigenvalues $\mu_{1:K}$. Let $R = \hat{L}Y - YM$ and $\Delta = \min_{\lambda \in \hat{\lambda}_{K+1:n}, \mu \in \mu_{1:K}} |\lambda - \mu| > 0$. Then, for any unitarily invariant norm $||\,||$, $||\operatorname{diag}(\sin\theta_{1:K}(\hat{Y}, Y))|| \leq \frac{||R||}{\Delta}$, where $\theta_{1:K}$ are the canonical angles between $\mathcal{R}(\hat{Y})$ and $\mathcal{R}(Y)$.*

Proof of Proposition 3 This is a corollary of Theorem 3.6 in [58]. If eigenvalues are sorted by their absolute values, then $\hat{\lambda}_{K+1:n} \in [-|\hat{\lambda}_{K+1}|, |\hat{\lambda}_{K+1}|]$ and $\mu_{1:K} \in \mathbb{R} \setminus (-|\hat{\lambda}_{K+1}| - \Delta, |\hat{\lambda}_{K+1}| + \Delta)$. If we set $M = \hat{\Lambda}$, so that $\hat{\lambda}_{1:K} \in \mathbb{R} \setminus (-|\hat{\lambda}_{K+1}| - \Delta, |\hat{\lambda}_{K+1}| + \Delta)$. Now we view Y as a perturbation of $\hat{Y}$, hence

$$R = \hat{L}Y - Y\hat{\Lambda} = \hat{L}Y - LY + (LY - Y\hat{\Lambda}) = (\hat{L} - L)Y \tag{4.8}$$

$$||R|| = ||(\hat{L} - L)Y|| \leq ||\hat{L} - L||||Y|| \leq \varepsilon. \tag{4.9}$$

From Theorem 18 the result follows. □

Proof of Proposition 4 For 1:

$$\begin{aligned} ||F||_F^2 &= \operatorname{trace} FF^T = \operatorname{trace} U\Sigma V^T V\Sigma U^T = \operatorname{trace} U^T U\Sigma V^T V\Sigma = \operatorname{trace}\Sigma^2 \\ &= 1+\sum_{k=2}^{K}\cos^2\theta_k = 1+\sum_{k=2}^{K}(1-\sin^2\theta_k) = K-\sum_{k=2}^{K}\sin^2\theta_k \text{ since } \theta_1=0 \quad (4.10) \\ &\geq K-(K-1)\varepsilon'^2 \quad (4.11) \end{aligned}$$

For 2: Denote $\operatorname{trace}\hat{M}^T M = \langle\hat{M}, M\rangle_F$. Then $||M-\hat{M}||_F^2 = ||M||_F^2 + ||\hat{M}||_F^2 - 2\angle\hat{M}, M\rangle_F \leq K+K-2(K-(K-1)\varepsilon'^2) = 2(K-1)\varepsilon'^2$. □

Proof of Proposition 5 We have that $|\langle M-\hat{M}, M'-\hat{M}\rangle_F| \leq ||M-\hat{M}||_F||M'-\hat{M}||_F$. From Proposition 4 the r.h.s is no larger than $2(K-1)\varepsilon'^2$.

$$\begin{aligned} -\langle M-\hat{M}, M'-\hat{M}\rangle_F &\leq ||M-\hat{M}||_F||M'-\hat{M}||_F \leq 2(K-1)\varepsilon'^2 \quad (4.12) \\ -\langle M, M'\rangle_F + \langle\hat{M}, M\rangle_F + \langle\hat{M}, M'\rangle_F - ||\hat{M}||_F^2 &\leq 2(K-1)\varepsilon'^2 \quad (4.13) \\ \langle M, M'\rangle_F &\geq \langle\hat{M}, M\rangle_F + \langle\hat{M}, M'\rangle_F - K - 2(K-1)\varepsilon'^2 \quad (4.14) \\ &\geq 2K-2(K-1)\varepsilon'^2 - K - 2(K-1)\varepsilon'^2 = K-4(K-1)\varepsilon'^2 \quad (4.15) \end{aligned}$$

Now, note that $\operatorname{trace} MM' = \operatorname{trace} YY^T Y'(Y')^T = \operatorname{trace}((Y')^T Y))(Y^T Y') = ||Y^T Y'||_F^2$. Moreover, by (4.4), Y_Z and Y differ by a unitary transformation. Since $||\,||_F$ is unitarily invariant, the result follows.

Proof of Theorem 14 We apply Theorem 9 of [43] with $A_X = Z, A_{X'} = Z'$, and $\tilde{A}_X = Y$, $\tilde{A}_{X'} = Y'$. It follows that $p_{XY_{kk'}} = \sum_{i\in k\cap k'}\hat{d}_i / \sum_{i=1}^n \hat{d}_i$. Hence, the point weights are proportional to $\hat{d}_{1:n}$. Also, evidently, $p_{min}/p_{max} = \delta_0$, and the result follows.

Note that we use the fact that both PFM's have degrees equal to $\hat{d}_{1:n}$ to obtain this proof. □

Proposition 8 *Assumptions 10 and 11, imply* $||\operatorname{diag}(\sin\theta_{1:K}(\hat{Y}, Y))|| \leq \varepsilon/|\hat{\lambda}_K^A| = \varepsilon'$*, where* $\hat{\lambda}_K^A$ *is the* K*-th eigenvalue of* $\hat{A}$*.*

Proof of Proposition 8 We consider $\hat{A}$ a perturbation of A, its eigenvectors $\hat{Y}$ as the perturbed

eigenvectors of A and $M = \hat{\Lambda}$. Then, $R = A\hat{Y} - \hat{Y}\hat{\Lambda}$

$$\begin{aligned} ||R|| &= ||A\hat{Y} - \hat{Y}\hat{\Lambda}|| && (4.16)\\ &= ||(A\hat{Y} - \hat{A}\hat{Y}) + (\hat{A}\hat{Y} - \hat{Y}\hat{\Lambda})|| && (4.17)\\ &\leq ||(A - \hat{A})\hat{Y}|| && (4.18)\\ &\leq ||A - \hat{A}||||\hat{Y}|| \ \leq \ \varepsilon. && (4.19) \end{aligned}$$

The separation between $\hat{\Lambda}$ and the residual spectrum of A is $|\hat{\lambda}_K|$. From the main Davis-Kahan theorem 18 the result follows. □

Proof of Proposition 6 The proofs of 1 and 2 are straightforward. To show 3, note that $A = ZC^{-1}Z^T\hat{A}ZC^{-1}Z^T = Y_Z C^{1/2}BC^{1/2}Y_Z^T = Y_Z U\Lambda U^T Y_Z^T = Y\Lambda Y^T$. The definition of B above shows that this is the Maximum Likelihood estimator of B given the clustering $\mathcal{C}$.

$$\Leftrightarrow \quad B_{kl} \ = \ \frac{\#\text{edges from cluster } k \text{ to cluster } l}{n_k n_l} \tag{4.20}$$

Proof of Theorem 15 We now follow the steps outlined in section 4.1 with ε' from Proposition 8 to obtain our main stability result.

Proof of Proposition 7 In the Proof of Proposition 5, we replace the bounds corresponding to $\langle \hat{M}, M\rangle_F, ||\hat{M} - M||_F$ by the actual values computed from $M, \hat{M}$. We obtain

$$\langle M, M'\rangle_F \geq \langle \hat{M}, M\rangle_F - (K-1)(\varepsilon')^2 - 2\sqrt{2(K-1)}\varepsilon'||\hat{M} - M||_F. \tag{4.21}$$

Proof of Proposition 2

From the Proof of this theorem, we have that $||L^* - \hat{L}|| = o(1)$, $||(D^*)^{1/2} - \hat{D}^{1/2}|| = o(1)$, $||\lambda^* - \hat{\Lambda}|| = o(1)$, and $||\hat{Y} - Y^*|| = o(1)$. Let Z be the indicator matrix of $\mathcal{C}^*$. The principal eigenvectors of L^* are $Y^* = (D^*)^{1/2}Z(C^*)^{-1/2}$. It follows then that $||Z^T\hat{D}Z - Z^T D^* Z|| = o(1)$, and since $C = Z^T\hat{D}Z$, $Y_Z = \hat{D}^{1/2}ZC^{-1/2}$ we have that $||Y_Z - Y^*|| = o(1)$, $||F^* - F|| = o(1)$ where $F^* = Y^T Y^*$. Moreover, since $||\hat{Y} - Y^*|| = o(1)$, $||F - I|| = o(1)$ Hence $||UV^T - I|| = o(1)$. Since the choice of B depends only on $\mathcal{R}(Y_Z)$, it follows immediately that $||BB^T\hat{L}B^T B - B^*(B^*)^T L^*(B^*)^T B^*|| = o(1)$. Now, $L = Y_Z UV^T\hat{\Lambda}VU^T Y_Z^T + BB^T\hat{L}B^T B$, and $L^* = Y^*\Lambda^*(Y^*)^T + B^*(B^*)^T L^*(B^*)^T B^*$, which completes the proof. □

perturbation of the PFM model To obtain a noisy PFM model A, we calculate the first K piecewise constant [47] eigenvectors V of the transition matrix $P = D^{-1}A$, from which we obtain V^* by perturbing each entry in V with a noise $\epsilon \sim unif(0, 10^{-4})$. The perturbed similarity matrix A is then obtained as $A = D^{1/2}(D^{1/2}V^*\hat{\Lambda}V^{*T}D^{1/2} + \hat{Y}_{low}\hat{\Lambda}_{low}\hat{Y}_{low}^T)D^{1/2}$. An adjacency matrix $\hat{A}$ is generated from A. In figure 4.2, we show the perturbed graphs A and $\hat{A}$.

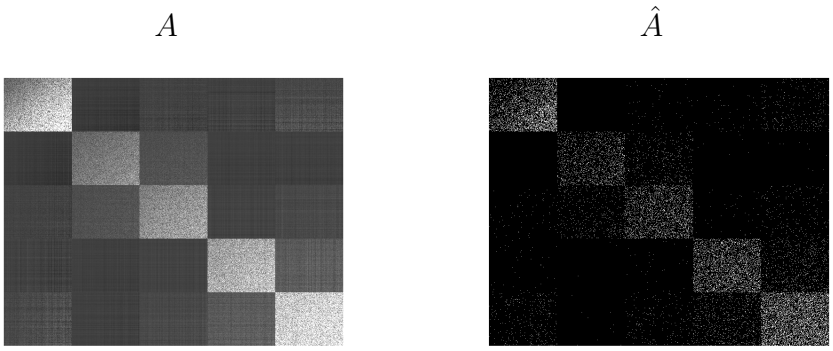


Figure 4.2: Left: the visualization of the perturbed A. Right: the visualization of the perturbed $\hat{A}$

Chapter 5

MEASURING THE ROBUSTNESS OF GRAPH PROPERTIES

In data science it is natural to consider an observed graph, or network, $\mathcal{G}$ as a realization of a random process. Consequently, the properties of $\mathcal{G}$ (such as diameter, conductance, clustering) and the inferences drawn from them are incomplete without some measure of variability or confidence.

We propose a new methodology for evaluating the robustness of graph properties by measuring the effect of small perturbations of the graph on the respective property. This methodology is based on concepts from robust statistics [26] and exploits the technique introduced in [44] to augment a graph Laplacian with user-defined weights. Let $\mathcal{G} = (\mathcal{V}, A)$ be a graph with $n = |\mathcal{V}|$ nodes and adjacency matrix A; A can be symmetric or asymmetric (corresponding to a directed network), and we assume $A_{ij} \in \{0, 1\}$, $A_{ii} = 0$ (simple graph) or $A_{ij} \geq 0$ (weighted graph). Our methodology applies to both scenarios, and makes no assumptions on how the graph was generated.

Our methodology consists of four key components:

1. *Perturb* the nodes of the graph by assigning them multiplicative weights, with w_i the weight of node i, $i = 1, 2, \ldots n$. If $w_i = 1$ for all i, we have the original graph $\mathcal{G}$.

2. Express the desired graph property $f(\mathcal{G})$ as a *smooth function of the weights*.

3. Construct *measures of robustness* inspired by the robust statistics literature, such as *influence function (IF)*, and *breakdown point (BP)*.

4. *Evaluate* these measures on the current graph $\mathcal{G}$.

We exemplify our approach by examining the robustness of weighted cut (WCut), number of weakly Connected Components (wCC's), eigengap and clustering. In the rest of this chapter, we make use of the following notations. We define $w_i, i = 1, 2, \cdots, n$, the weight associated with node i. Since we are interested in both directed and undirected graphs, we redefine $d_i = \sum_{j=1}^{n} A_{ij}$ as the out-degree of node i. We use the definition in [44] and redefine $L = I - \frac{1}{2} D^{-1/2}(A + A^T)D^{-1/2}$ as the Laplacian matrix associated with A. This definition is consistent with the usual definition of

$L = I - D^{-1/2}AD^{-1/2}$ when the graph is symmetric.

Note that in the previous chapters, we have been using A for the model and $\hat{A}$ for the observed graph. In this chapter, we do not focus on model classes. For simplicity we redefine A, d, D, L as the properties of the observed graph. The above quantities will be marked with a symbol $\tilde{}$ when they are perturbed. This is consistent with the notation in matrix perturbation literature. To demonstrate, $\tilde{A}$ represents the adjacency matrix of the perturbed graph.

In the rest of this chapter, we proceed as follows. In Section 5.1, we describe our method of bootstrapping and perturbing the networks. In Section 5.2, we talk about the graph properties of interest. In Section 5.3, we discuss measures for evaluating robustness of graph properties. In Section 5.6, we use both synthetic and real datasets to analyze our methods. We conclude our findings in Section 5.7.

5.1 Perturbing the network

Existing methods of perturbing networks can be found in recent works [30, 24, 14, 4]. They mostly involve removing or duplicating edges or nodes randomly and independently in the graph. [24] randomly removes edges from the graph. [30] maintains the the number of vertices and edges of the original graph, and perturbs the graph by moving the edges to the other locations. On the other hand, [14] and [4] bootstrap the network by subsampling the nodes. Although their approach is straightforward, the perturbation cannot be well controlled. Firstly, the topology of the graphs can change dramatically for sparse graphs. Randomly removing or adding the nodes and edges makes the strength of perturbation hard to control. For example, assume we have two densely clusters $\mathcal{C}_1$ and $\mathcal{C}_2$ connected by a single edge e. The effect of removing e versus removing an edge within $\mathcal{C}_1$ or $\mathcal{C}_2$ is very different, since the former will change the graph structure drastically by making it disconnected. Moreover, removing or adding edges (nodes) are discrete moves, therefore the perturbation cannot be arbitrarily small.

In order to have a fine control on the amount of perturbation and make it as smooth as possible, in our approach, we preserve the graph topology by keeping all the nodes and edges, and we perturb the graph by assigning random weights to the nodes and then distribute these weights to edges. Specifically, we assign weight w_i to node i, where w_i is generated i.i.d from a distribution with $E(w_i) = 1$, standard deviation σ_w, and support on $(0, \infty)$. The perturbation can be controlled smoothly with σ_w. Since $w_i > 0$, no nodes or edges is removed or added, and the topology of the graph is preserved.

We propose two ways of constructing $\tilde{A}$ from weighted nodes.

- Asymmetric perturbation: perturb outgoing edges of node i, so that $\tilde{A}_{ij} = w_i A_{ij}$ (whereas $\tilde{A}_{ji} = w_j A_{ji}$). The out-degree becomes $\tilde{d}_i = d_i w_i$. The perturbed Laplacian is,

$$\tilde{L}_{ij} = 1 - \frac{\tilde{A}_{ij} + \tilde{A}_{ji}}{2\sqrt{\tilde{d}_i \tilde{d}_j}} = 1 - \frac{w_i A_{ij} + w_j A_{ji}}{2\sqrt{w_i w_j d_i d_j}} \tag{5.1}$$

- Symmetric perturbation: place w_i on both outgoing edges and incoming edges of node i. $\tilde{A}_{ij} = (w_i + w_j - 1)A_{ij}$. Since this method leads to complicated $\tilde{L}$ in form, we discuss perturbing one node at a time. If node t is perturbed, the Laplacian becomes,

$$\tilde{L}_{ij} = \begin{cases} 1 - \frac{A_{tj} w_t + A_{jt} w_t}{2\sqrt{d_t w_t (\sum_{m \neq t} A_{jm} + A_{jt} w_t)}} & i = t, j \neq t \\ 1 - \frac{A_{ij} + A_{ji}}{2\sqrt{(\sum_{m \neq t} A_{jm} + A_{jt} w_t)(\sum_{m \neq t} A_{im} + A_{it} w_t)}} & i, j \neq t \\ 0 & i, j = t \end{cases} \tag{5.2}$$

Both of the above methods ensure $E(\tilde{A}_{ij}) = A_{ij}$. Notice that in the current methods, it can be easily shown that bias is introduced in A_{ij}, in which case one cannot separate the perturbation from structural change and the change of values in adjacency matrix.

Although both perturbation methods seem to be reasonable, they have graph-specific advantages. The asymmetric method provides a very simple way for calculating a graph property like $\tilde{d}$ after perturbation. On the other hand, it is not interesting for the perturbation of weakly Connected Components (wCC's), as we shall see in Section 5.3. On the other hand, the symmetric method maintains a symmetric perturbation of adjacency matrix. If $A_{ij} = A_{ji}$, then $\tilde{A}_{ij} = \tilde{A}_{ji}$ after perturbation. However, $\tilde{L}$ is complicated in form. The symmetric perturbation is not very interpretable for some graph properties (WCut), but it is useful for evaluating the robustness in eigengap and wCC's.

In the above methods, we put weights on nodes and then distribute the weights to edges. The reason for doing this instead of directly perturbing the edges is that it is more natural to maintain the association between nodes and edges. The edges connected to the same node should be dependent rather than independent. Perturbing the weights on edges [30] independently neglects these relationships, which has been found both unrealistic and disrupts the topology of sparse network. Moreover, in many applications, nodes are more meaningful than edges. Firstly, a node often carries its own attributes. Secondly, a node is described by the multiple nodes it is connected to. On the other hand, it is more difficult to glean information from an edge. For example, in a Facebook network, a node is a person with complex information including age, birth place, school, the friends he connected to, etc. While an edge is formed when two people befriend each other, and we cannot even learn how well these two people know each other and it tells us little on other information.

5.1.1 The bias in $\tilde{L}$

The graph properties we shall study are closely related to L. It would be nice if the changes in graph properties are only from the noise introduced in graph structure and keeping the entires in $\tilde{L}$ to be the same with that in L in expectation. Since having bias in L introduces another factor for the change in graph property, of which the strength of perturbation is hard to control. Therefore in principle, we would want $E(\tilde{L}_{ij}) = L_{ij}$. In the following paragraphs we prove that this is impos-

sible under all the current perturbation methods. For the current methods, where people move the edges and subsample the nodes, the bias in $\tilde{L}$ is apparent. We show in Proposition 9 that both of our perturbation methods introduce bias too.

Proposition 9 *Assume $w_i, i = 1 : n$ are generated i.i.d from a distribution with $E(w_i) = 1$, $\sigma_w \neq 0$, $w_i > 0$. For asymmetric perturbation, $E(\tilde{L}_{ij}) < L_{ij}$.*
If we further assume there exists a triangle $< t, p, q >$ in the graph. For symmetric perturbation, with node t perturbed, $E(\tilde{L}_{ij}) < L_{ij}$.

Proof:

In the asymmetric perturbation, $\tilde{L}$ is shown in Equation 5.1. Assume $L_{ij} \neq 0$. In order to have $E(\tilde{L}_{ij}) = L_{ij}$, we need

$$E(1 - \frac{w_i A_{ij} + w_j A_{ji}}{2\sqrt{w_i w_j d_i d_j}}) = 1 - \frac{A_{ij} + A_{ji}}{2\sqrt{d_i d_j}} \tag{5.3}$$

$$A_{ij}(1 - E(\sqrt{\frac{w_i}{w_j}})) + A_{ji}(1 - E(\sqrt{\frac{w_j}{w_i}})) = 0 \tag{5.4}$$

Since w_i and w_j are i.i.d, we then have,

$$E(\sqrt{\frac{w_i}{w_j}}) = E(\sqrt{\frac{w_j}{w_i}}) = E(\sqrt{w_i})E(\frac{1}{\sqrt{w_j}}) = E(\sqrt{w_i})E(\frac{1}{\sqrt{w_i}}) = 1 \tag{5.5}$$

Assume $x = \sqrt{w_i}$. Equivalently we have,

$$E(\frac{1}{x}) = \frac{1}{E(x)} \tag{5.6}$$

Assume $f(x) = \frac{1}{x}$. Since $f(x)$ is strictly convex and Jensen's inequality yields, $E(f(x)) \geq f(E(x))$, that is, $E(\frac{1}{x}) \geq \frac{1}{E(x)}$. where equality holds only when $\sigma(x) = 0$. Therefore $E(\sqrt{\frac{w_i}{w_j}}) > 1$. We have reached a contradiction. $E(\tilde{L}) < L$.

In the symmetric perturbation, $\tilde{L}$ after perturbing node t is shown in Equation 5.2. In order have $E(\tilde{L}_{ij}) = L_{ij}$, we need,

$$E(1 - \frac{A_{tj} w_t + A_{jt} w_t}{2\sqrt{d_t w_t (\sum_{m \neq t} A_{jm} + A_{jt} w_t)}}) = 1 - \frac{A_{tj} + A_{jt}}{2\sqrt{d_t d_j}} \text{ when } i = t, j \neq t \tag{5.7}$$

$$E(1-\frac{A_{ij}+A_{ji}}{2\sqrt{(\sum_{m\neq t}A_{jm}+A_{jt}w_t)(\sum_{m\neq t}A_{im}+A_{it}w_t)}})=1-\frac{A_{ij}+A_{ji}}{2\sqrt{d_id_j}} \text{ when } i,j\neq t \quad (5.8)$$

That is,

$$E(\sqrt{w_t})E(\frac{1}{\sqrt{a_j+(1-a_j)w_j}})=1, \text{ when } i=t, j\neq t, \quad (5.9)$$

$$E(\frac{1}{\sqrt{a_i+(1-a_i)w_i}})E(\frac{1}{\sqrt{a_j+(1-a_j)w_j}})=1, \text{ when } i,j\neq t, \quad (5.10)$$

where $a_i=\frac{\sum_{m\neq t}A_{jm}}{d_j}$ and $0\leq a_i\leq 1$.

Since there is a triangle $<t,p,q>$ in the graph. We can easily derive from Equation 5.9 and 5.10 that $E(\sqrt{w_t})=1$, and $E(\frac{1}{\sqrt{a_i+(1-a_i)w_i}})=1, \forall i$.

$E(\frac{1}{\sqrt{a_i+(1-a_i)w_i}})$ is a continous and differentiable function of a_i. When $a_i=1$, $E(\frac{1}{\sqrt{a_i+(1-a_i)w_i}})=1$. When $a_i=0$, $E(\frac{1}{\sqrt{a_i+(1-a_i)w_i}})=E(\frac{1}{\sqrt{w_i}})>\frac{1}{E(\sqrt{w_i})}=1$. Since $\frac{\partial}{\partial a_i}E(\frac{1}{\sqrt{a_i+(1-a_i)w_i}})<0$, $E(\frac{1}{\sqrt{a_i+(1-a_i)w_i}})>1$ for $0<a_i<1$. We have a contradiction. Therefore in conclusion, $E(\tilde{L})<L$.■

We have proved that in theory there will be bias introduced in $\tilde{L}$, it would be a good idea to examine how much bias will actually be introduced in practice. Under asymmetric perturbation, one can easily derive from the proof in Proposition 9 that $E(\sqrt{w})E(\frac{1}{\sqrt{w}})=\frac{\tilde{L}_{ij}}{L_{ij}}$. Therefore, the more $E(\sqrt{w})E(\frac{1}{\sqrt{w}})$ deviates from 1, the more bias introduced in $\tilde{L}_{ij}$. Since there are very limited number of conventional distributions that allow support on $(0,\infty)$ and mean equals 1, we propose a class of mixture distributions $Mixture(a,b,\sigma_-^2,\sigma_+^2,T_-,T_+,p)$, where $ap+b(1-p)=1$, T_- is a distribution with $E(x)=a$, $\sigma(x)=\sigma_-^2$, with support on $(0,1)$, T_+ is a distribution with $E(x)=b$, $\sigma(x)=\sigma_+^2$, with support on $(1,\infty)$. Assume $w\sim Mixture(a,b,\sigma_-^2,\sigma_+^2,T_-,T_+,p)$, then $p(w\sim T_-)=p$, $p(w\sim T_+)=1-p$. We can easily show that $E(w)=1$, $\sigma^2(w)=p_-\sigma_-^2+p_+\sigma_+^2$. The mixture distribution subsumes all the distributions concentrated on 1, with support on $(0,\infty)$.

In the following experiment, we generate w_i from the distributions below accordingly.

1. Node resampling: w_i is obtained from resampling the nodes with replacement to form a sample of size N. Assume node i appears m times in the sample, $w_i=\frac{mn}{N}$. $\sigma_w=\sqrt{\frac{n-1}{N}}$.

We can easily see as N goes up, σ_w decreases. We can then control σ_w by varying the size of N.

2. Binary distribution: w_i follows binary distribution on $\{a, b\}$, with $P(w_i = a) = p$, $P(w_i = b) = 1 - p$. $b = \frac{1-ap}{1-p}$. $E(w_i) = 1$, $\sigma_w^2 = p(a-1)^2 + (1-p)(b-1)^2$.

3. Gamma distribution: $w_i \sim Gamma(a, \frac{1}{a})$. $E(w_i) = 1$, $\sigma_w^2 = \frac{1}{a}$.

4. Mixture-Gamma-Uniform distrbution: $a = 0.5, T_- = uniform(0, 1), T_+ = Gamma(\frac{b-1}{\sigma_+}, \sigma_+) + 1$.

5. Mixture-Lognormal-Uniform distribution: $a = 0.5, T_- = uniform(0, 1), T_+ = lognormal(b - 1, \sigma_+)$.

The results are shown in Figure 5.1. We observe that the bias is sensitive to the choice of weight distribution. In specific, the bias introduced by Gamma distribution is growing exponentially with σ_w, thus not a good option. Node resampling and Binary distribution generate smallest bias. However node resampling can only allow the perturbation strength σ_w to be as much as 1, otherwise the topology of the graph is changed. Binary distribution only allows two choices of weight, which is very limited. The mixture distributions behave similarly in terms of introducing bias. Although the bias is nontrivial, the fact that it is upper bounded and that w_i can be generated in a continuous manner make them good candidates for performing perturbations on graphs.

We also evaluate the bias introduced in $\tilde{L}$ in the current methods including [30] and [14]. Since they move the edges or delete nodes, the bias depends on the graph. We generate a graph from DC-SBM model [55] with $n = 800$, $w_i \sim 0.5 + uniform(0.5, 1)$, and evaluate $\frac{E(\tilde{L}_{ij})}{E(L_{ij})}$. As a reminder, $\frac{E(\tilde{L}_{ij})}{E(L_{ij})} = E(\sqrt{w})E(\frac{1}{\sqrt{w}})$. In Figure 5.2, we show that both methods introduce large bias compare to our perturbation methods.

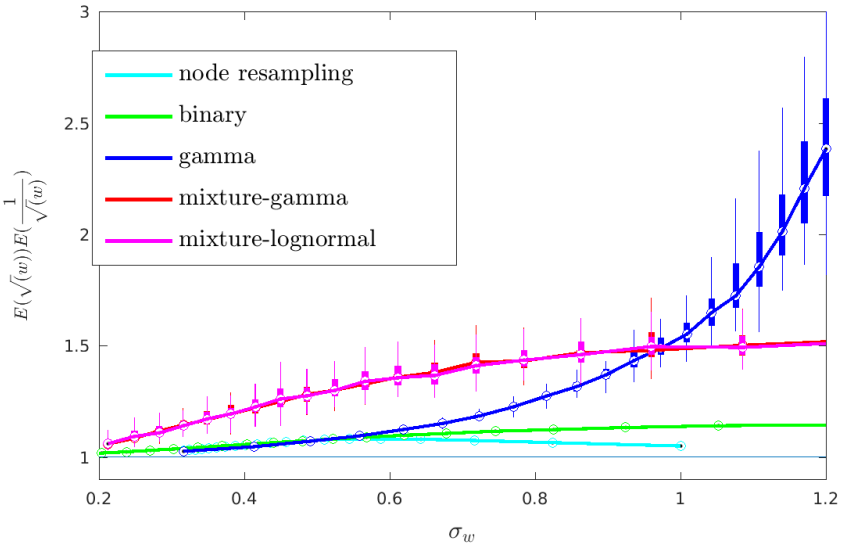
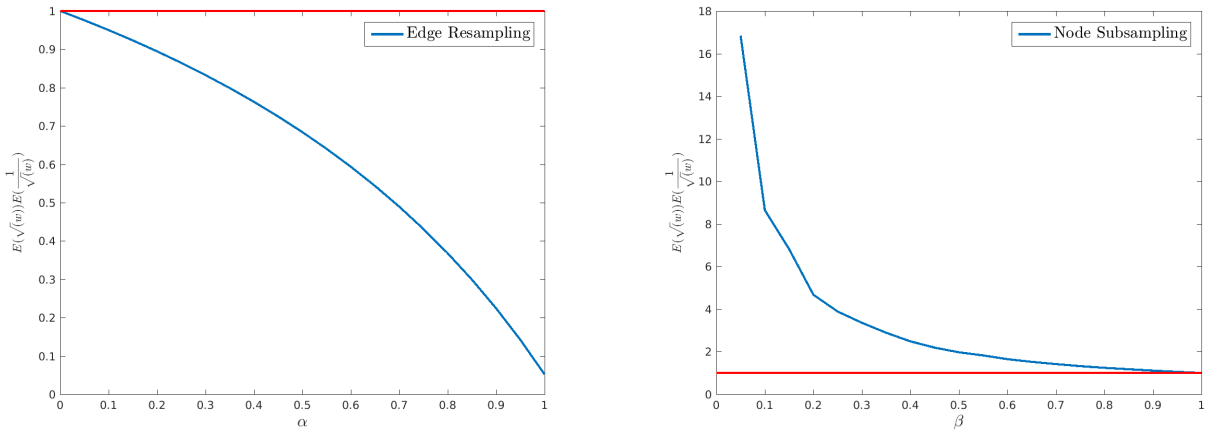


Figure 5.1: $E(\sqrt{w})E(\frac{1}{\sqrt{w}})$ under asymmetric perturbation. Each boxplot represents 100 repetitions.

5.1.2 Partial perturbation and full perturbation

In this section we discussed perturbing the graph by adding i.i.d weights to the nodes. Although we make i.i.d assumption on w_i, it can be relaxed. In the rest of this chapter, we consider two approaches to utilize the weight perturbation. First, in order to evaluate the sensitivity of graph properties, we perturb all the nodes i.i.d with different perturbation strength σ_w, and then investigate how the graph property changes with respect to it. In specific, we fix $E(w) = 1$ and vary σ_w for all the nodes. Alternatively, in order to probe the source of sensitivity, we can perturb a subset of nodes. For these nodes, we perturb their weights i.i.d by fixing σ_w and vary $E(w)$, from which we can discover the source of the robustness in graph properties. The use of these two perturbation approaches will be shown in Section 5.6.



perturbation strength proportion of nodes sampled

Figure 5.2: Left: the bias from the method in [30]. α indicates the strength of perturbation. Right: the bias from the method in [14]. β indicates the proprotion of nodes that are subsampled. Larger β indicates larger perturbation strength.

5.2 Expressing graph properties with weights

The success of our approachs depend on our ability to express properties of interest as functions $f(\mathcal{G}, w)$, which are continuous and differentiable w.r.t. the node weights $w_{1:n}$. In this project we focus on graph properties that depend on graph Laplacians. Many important graph properties depend on Laplacians. For instance, the *mixing time* of the graph depends on the second smallest eigenvalue $\lambda_2(L)$. Diffusion distances [50] between nodes can also be approximated by the principal eigenvectors and values of L. The number of connected components C of $\mathcal{G}$ is equal to the multiplicity of 0 in the spectrum of L, etc. There are four graph properties that we are particularly interested in, which will be used as examples in the following sections: the weighted cut of the graph (WCut), the number of weakly connected components, the eigengap and clustering.

5.2.1 Weighted Cut(WCut)

Weighted cut is defined as a graph property associated with clustering in the general graph setting [44], which can be used for both directed and undirected graphs. It is similar in motivation to the normalized cut for undirected graph. Both aim in finding a cut of low weight in the graph while balancing the sizes of the clusters. The multiway version of the normalized cut MNCut of [44] is a special case of WCut.

Formally, WCut with respect to clustering $\mathcal{C}$ with K clusters is defined as

$$WCut(\mathcal{G}, w, \mathcal{C}) = \sum_{k=1}^{K} \frac{1}{\tilde{D}_k} \sum_{i \in \mathcal{C}_k} (\tilde{d}_i - \sum_{j \in \mathcal{C}_k} \tilde{A}_{ij}) \tag{5.11}$$

where $\tilde{D}_k = \sum_{i \in \mathcal{C}_k} \tilde{d}_i$. Further define $\tilde{D}_{kk} = \sum_{i \in \mathcal{C}_k} \sum_{j \in \mathcal{C}_k} \tilde{A}_{ij}$, we can then equivalently write Wcut as,

$$WCut(\mathcal{G}, w, \mathcal{C}) = \sum_{k=1}^{K} (1 - \frac{\tilde{D}_{kk}}{\tilde{D}_k}) \tag{5.12}$$

Small WCut suggests sparse connections between clusters, thus better quality of clustering.

5.2.2 Number of weakly Connected Components (wCC's) and eigengap

It is already know that the number of connected components (CC's) is not robust, since randomly adding a node or removing some edges can easily change the number of CC's. Instead, we study the number of weakly Connected Components (wCC's), where "weakly" means sparse connections between CC's.

We propose a pair of functions $f_u(\mathcal{G}, w, K) = \lambda_{K+1}(L(\tilde{\mathcal{G}})) - \sum_{k=1}^{K} \lambda_i(L(\tilde{\mathcal{G}}))$, $f_l(\mathcal{G}, w, K) = \sum_{k=1}^{K} \lambda_k(L(\tilde{\mathcal{G}}))$ to describe the number of wCC's. The reason we use f_l is that, if there are K number of CC's, then $f_l = 0$. We would then expect f_l to be close to 0 for K number of wCC's. We choose f_u because we expect a significant gap between f_l and $\lambda_{K+1}(L(\tilde{\mathcal{G}}))$ for a stable number of wCC's. When f_u is away from 0 (respectively f_l near 0) there are exactly K "weakly connected"

components in $\tilde{\mathcal{G}}$. If this holds for large perturbations, then K can be considered robust.

The Kth eigengap is defined as $f_e(\tilde{\mathcal{G}}, w, K) = \lambda_{K+1}(L(\tilde{\mathcal{G}})) - \lambda_K(L(\tilde{\mathcal{G}}))$. It indicates K principal subspace when f_e is large compared to the other eigengaps.

Notice that function f defined for number of wCC's and eigengap are only meaningful for undirected graphs. Since the eigenvalues is not very interpretable for the directed graphs.

5.3 Influence functions

In order to evaluate the graph properties, we construct the target properties as differentiable functions $f(\mathcal{G}, w)$ and see how much $f(\mathcal{G}, w)$ changes with respect to the size of perturbation of w. In this section, we present tools for quantifying robustnes including *Influence Function* (IF) and *Breakdown Points*. We firstly talk about Influence function (IF), which was invented by Hampel in [26]. The importance of IF lies in its interpretation: it gives a picture of the infinitesimal behavior of the asymptotic value. While numerous studies have been done on methods for analysis of data sampled from a known distribution i.i.d, there has not been much work on using these tools to evaluate the robustness of graph properties. Here we define for a perturbed graph $\tilde{\mathcal{G}}$,

$$IF_t = \left.\frac{\partial f(\mathcal{G}, w)}{\partial w_t}\right|_{w_{1:n}=1}, \tag{5.13}$$

which measures the local influence of w_t on f.

5.3.1 Influence function of WCut

Proposition 10 *Assume a graph $\mathcal{G}$ with $\mathcal{C}$, d_i defined as usual. Assume node $t \in \mathcal{C}_{k_0}$. $d_{ik} = \sum_{j\in\mathcal{C}_k} A_{ij}$, $d_{ki} = \sum_{j\in\mathcal{C}_k} A_{ji}$. $D_{k_0k_0^-} = \sum_{i\in\mathcal{C}_{k_0}, j\notin\mathcal{C}_{k_0}} A_{ij}$. Then using asymmetric perturbation, the influence function for node t is*

$$IF_t^{WCut}(\mathcal{G}, \mathcal{C}) = \frac{d_t D_{kk} - d_{tk} D_k}{D_k^2}. \tag{5.14}$$

Using symmetric perturbation.

$$IF_t^{WCut}(\mathcal{G},\mathcal{C}) = \sum_{k=1}^{n} \frac{D_{kk}d_{kt}}{D_k^2} - \frac{D_{k_0}d_{k_0t} + D_{k_0\bar{k_0}}D_{tk_0} - D_{k_0k_0}\sum_{j\notin\mathcal{C}_{k_0}} A_{tj}}{D_{k_0}^2} \tag{5.15}$$

Proof:

We firstly perturb the graph using the asymmetric method. Assume $t \in \mathcal{C}_k$, we have

$$\frac{\partial WCut(\mathcal{G},w)}{\partial w_t} = \frac{\partial}{\partial w_t}\frac{1}{\sum_{j\in\mathcal{C}_k} d_j w_j}\sum_{j\in\mathcal{C}_k} w_j(d_j - d_{jk}) \tag{5.16}$$

$$= \frac{d_t\sum_{j\in\mathcal{C}_k} w_j d_{jk} - d_{tk}\sum_{j\in\mathcal{C}_k} d_j w_j}{(\sum_{j\in\mathcal{C}_k} d_j w_j)^2}. \tag{5.17}$$

The influence function is then derived as

$$\frac{\partial WCut(\mathcal{G},w)}{\partial w_t}|_{w_{1:n}=1} = \frac{d_t\sum_{j\in\mathcal{C}_k} d_{jk} - d_{tk}\sum_{j\in\mathcal{C}_k} d_j}{(\sum_{j\in\mathcal{C}_k} d_j)^2} = \frac{d_t D_{kk} - d_{tk}D_k}{D_k^2} \tag{5.18}$$

$$\tag{5.19}$$

In the symmetric perturbation, since the perturbation can be explained as perturbing one node at a time, one can easily show that the influence function is same for perturbing one node or multiple nodes. For simplicity, we assume perturbing node t with weight w_t. WCut can be written as

$$\begin{aligned} WCut(\mathcal{G},w) = &\sum_{k,t\notin\mathcal{C}_k}(1 - \frac{D_{kk}}{\sum_{i\in C_k}[(\sum_{j\neq t} A_{ij}) + A_{it}w_t]})+ \\ &(1 - \frac{\sum_{i\in\mathcal{C}_{k_0}} A_{it}w_t + \sum_{j\in\mathcal{C}_{k_0}} A_{tj}w_t + \sum_{i,j\neq t,i,j\in\mathcal{C}_{k_0}} A_{ij}}{d_t w_t + \sum_{i\in\mathcal{C}_{k_0},i\neq t}[(\sum_{j\neq t} A_{ij}) + A_{it}w_t]}) \end{aligned} \tag{5.20}$$

$$\frac{\partial WCut(\mathcal{G},w)}{\partial w_t}|_{w_{1:n}=1} = \sum_{k,t\notin\mathcal{C}_k} \frac{D_{kk}\sum_{i\in\mathcal{C}_k} A_{it}}{D_k^2} - \frac{(\sum_{i\in\mathcal{C}_{k_0}} A_{it} + \sum_{j\in\mathcal{C}_{k_0}} A_{tj})D_{k_0} - D_{k_0k_0}(d_t + \sum_{i\in\mathcal{C}_{k_0}} A_{it})}{D_{k_0}^2} \tag{5.21}$$

$$= \sum_{k=1}^{n} \frac{D_{kk}\sum_{i\in\mathcal{C}_k} A_{it}}{D_k^2} - \frac{D_{k_0}\sum_{i\in\mathcal{C}_{k_0}} A_{it} + D_{k_0\bar{k_0}}\sum_{j\in\mathcal{C}_{k_0}} A_{tj} - D_{k_0k_0}\sum_{j\notin\mathcal{C}_{k_0}} A_{tj}}{D_{k_0}^2} \tag{5.22}$$

$$= \sum_{k=1}^{n} \frac{D_{kk}d_{kt}}{D_k^2} - \frac{D_{k_0}d_{k_0t} + D_{k_0\bar{k_0}}d_{tk_0} - D_{k_0k_0}\sum_{j\notin\mathcal{C}_{k_0}} A_{tj}}{D_{k_0}^2} \tag{5.23}$$

■

In the asymmetric perturbation, IF has an intuitive interpretation when a point has no influence, i.e, $IF_t = 0$,

$$\frac{d_{tk}}{d_i} = \frac{D_{kk}}{D_k} = \frac{mean_{i\in\mathcal{C}_k} d_{ik}}{mean_{i\in\mathcal{C}_k} d_i} \tag{5.24}$$

It means that, node t has 0 influence in WCut when the proportion of edges that goes to $\mathcal{C}_K$ equals the cluster level ratio of averages. If $IF_t > 0$, node t tends to make WCut larger when more weight is put upon t, the quality of clustering decreases since the clustering becomes less well separated. Node t is therefore considered unstable to the clustering, or not well clustered. If $IF_t < 0$, node t is well clustered since WCut will decrease if t is weighted more. Hence IF measures the robustness of clustering in node level. It is worth noticing that the influence of a node depends only on the cluster that the node belongs to, and is independent of the rest of the clusters. Moreover, the influences of the nodes within a cluster always cancel each other. In a well separated clustering, we would expect the influences of the nodes the be concentrated around 1. When the clusters are completed separated, that is, there is no edge between clusters, it can be easily shown that all the node influence equals 0.

$$\sum_{i\in\mathcal{C}_k} \frac{\partial WCut(\mathcal{G}, w)}{\partial w_i}|_{w_{1:n}=1} = \frac{\sum_{i\in\mathcal{C}_k} d_i \sum_{j\in\mathcal{C}_k} d_{jk} - \sum_{i\in\mathcal{C}_k} d_{ik} \sum_{j\in\mathcal{C}_k} d_j}{(\sum_{j\in\mathcal{C}_k} d_j)^2} = 0, \tag{5.25}$$

For the symmetric perturbation, the meaning for the above results is not very interpretable, since its form is very complicated and $IF_t = 0$ does not provide us with any clear interpretation in its balance state.

5.3.2 Influence function of number of wCC's and eigengap

Here we consider the property of wCC's and eigengap described by f_u, f_l and f_e, which are proposed in Section 5.2.2. Since they both depend on λ, we firstly study $\frac{\partial \lambda_k}{\partial w_t}$ for a single λ_k in Proposition 13.. The influence functions of interest can be easily derived from there. This is because the

influence functions can be written as,

$$IF_t^{f_u} = \frac{\partial \lambda_{K+1}}{\partial w_t} - \sum_{i=1}^{K} \frac{\partial \lambda_i}{\partial w_t} \tag{5.26}$$

$$IF_t^{f_l} = \sum_{i=1}^{K} \frac{\partial \lambda_i}{\partial w_t} \tag{5.27}$$

$$IF_t^{f_e} = \frac{\partial \lambda_{K+1}}{\partial w_t} - \frac{\partial \lambda_K}{\partial w_t} \tag{5.28}$$

Unfortunately, asymmetric perturbation is not very interesting for undirected graphs, since the properties remain untouched despite the perturbation, as will be shown in Proposition 11. This is because the eigengap of L is equal to the eigengap of the transition matrix $P = D^{-1}A$, and P stays unchanged after the asymmetric perturbation. This motivates the use of symmetric perturbation, the results of which are shown in Proposition 13.

Proposition 11 *Using asymmetric perturbation, assume A symmetric and λ_k of multiplicity 1. $\frac{\partial \lambda_k}{\partial w_t}|_{w_{1:n}=1} = 0$. v_i is the i-th element of the k-th eigenvector of L.*

Proof:

$$\frac{\partial \tilde{\lambda}_k}{\partial w_t} = \sum_{ij} \frac{\partial \tilde{\lambda}_k}{\partial \tilde{L}_{ij}} \frac{\partial \tilde{L}_{ij}}{\partial w_t} = \sum_{ij} v_i v_j \frac{\partial \tilde{L}_{ij}}{\partial w_t} \tag{5.29}$$

Since $\tilde{L}_{ij} = 1 - \frac{A_{ij}w_i + A_{ji}w_j}{2\sqrt{w_i w_j d_i d_j}}$ We then obtain,

$$\frac{\partial \tilde{L}_{ij}}{\partial w_i} = \frac{-2A_{ij}\sqrt{w_i w_j d_i d_j} + A_{ij}\sqrt{w_i w_j d_i d_j} + A_{ji}\sqrt{w_j^3 d_i d_j / w_i}}{4 w_i w_j d_i d_j} \tag{5.30}$$

For an undirected graph, $\frac{\partial \tilde{L}_{ij}}{\partial w_i}|_{w_{1:n}=1} = 0$ always. ■

Proposition 12 *[39] $\frac{\partial \lambda_k}{\partial L ij} = \sum_{i,j \neq t} v_i v_j$, where v_i is the i-th element of the k-th eigenvector of L.*

Proposition 13 *Define A, d, λ_k, L, $\tilde{L}$, w as usual. Assume λ_k is of multiplicity 1. Define transition matrix $P = D^{-1}A$, v_i is the i-th element of the k-th eigenvector of L.*

In symmetric perturbation,

$$\frac{\partial \tilde{\lambda}_k}{\partial w_t}|_{w_{1:n}=1} = (1 - \lambda_k)(\sum_i v_i^2 P_{it} - v_t^2) \tag{5.31}$$

Proof: After performing the symmetric perturbation, we obtain $\tilde{L}$ from Equation 5.2. We then calculate the influence function $\frac{\partial \tilde{L}_{ij}}{\partial w_t}|_{w_{1:n}=1}$,

$$\frac{\partial \tilde{L}_{ij}}{\partial w_t}|_{w_{1:n}=1} = \begin{cases} -\frac{A_{tj}+A_{jt}}{4\sqrt{d_t d_j}}(1-\frac{A_{jt}}{d_j}) & i=t, j\neq t \\ \frac{A_{ij}+A_{ji}}{4} \times \frac{A_{jt}d_i + A_{it}d_j}{(d_i d_j)^{3/2}} & i,j \neq t \\ 0 & i,j=t \end{cases} \tag{5.32}$$

Then we can derive $\frac{\partial \lambda_k}{\partial w_t}|_{w_{1:n}=1}$ using Proposition 12,

$$\frac{\partial \tilde{\lambda}_k}{\partial w_t}|_{w_{1:n}=1} = \sum_{ij} \frac{\partial \tilde{\lambda}_k}{\partial \tilde{L}_{ij}} \frac{\partial \tilde{L}_{ij}}{\partial w_t} \tag{5.33}$$

$$= \sum_{i,j\neq t} v_i v_j \frac{A_{ij}+A_{ji}}{2} \times \frac{A_{it}d_j + A_{jt}d_i}{2(d_i d_j)^{3/2}} - v_t \sum_{j=1}^{n} v_j \times \frac{A_{tj}+A_{jt}}{2}(\frac{1}{\sqrt{d_j d_t}} - \frac{A_{jt}}{d_j\sqrt{d_j d_t}}) \tag{5.34}$$

$$= (1-\lambda_k)(\sum_i v_i^2 P_{it} - v_t^2), \tag{5.35}$$

■

5.3.3 Clustering

Clustering, defined as a partition of nodes, cannot be written as a smooth function of weights, since the clustering can only be changed discretely. Therefore we cannot use IF to measure the robustness of clustering. However, a clustering can be evaluated by breakdown points (BP), which will be discussed in the next section.

5.4 Breakdown points (BP)

While IF measures local infinitesimal influences, the *breakdown point* (BP) measures the global reliability of the graph properties. It is developed by [26] and widely used in robust statistics literature. It informs the range of perturbations that can be tolerated before the "structural information

in the data" is lost. Similar with Influence function, the use of BP is limited by the assumption that the data is generated i.i.d from known distributions, thus not applicable to the graph properties.

We are the first to extend the definition of BP to the graph properties. For a graph property $f(\mathcal{G})$, a BP is defined as,

$$\sigma_w^* := max\{\sigma_w; |f(\tilde{\mathcal{G}}) - f(\mathcal{G})| \leq \epsilon \text{ with probability } 1-\alpha\} \tag{5.36}$$

where ϵ and α are defined by users. Through finding BP, we define meaningful and computable thresholds where information about a specific graph property is lost. For instance, for $f(\mathcal{G}, w) = \lambda_2(L(\mathcal{G}_w))$, a BP can be defined when the eigengap between λ_2 and the next largest eigenvalue vanishes.

Another advantage of BP is that it allows robustness measure for non-differentiable graph property functions, such as clustering. It does not require the graph properties to be written as smooth functions of weights. It is a descriptive measure that allows global measure of robustness.

5.5 Related Work

In the past literature, there has been some work in perturbing the social networks. In [30], they restrict their perturbed networks to maintain the same number of vertices and edges as the original unperturbed network, and the perturbation is meant for the position of the edges only. The amount of perturbation is controlled by the number of edges being moved. In [14] and [4], they focus on subsampling the nodes of the networks. [14] propose uniform subsampling bootstrap scheme, in which they iteratively select a subset of vertices without replacement and consider the graph induced by the subset of vertices. They also consider a subgraph subsampling bootstrap, where they use a enumeration scheme to find all possible subgraphs with a fixed vertice size, and they selecte the subgraph with a fixed probability p.

We have also seen work in detecting dense communities and largest connected component in

[59]. They formalize tests for the existence of a dense random subgraph based on a variant of scan statistics. Although they offer sharpe detection bounds, their theorems only make a judgement on the existence of the subgraph instead of finding out where the subgraph is. Moreover, one has to go through all the subgraphs in order to calculate the test statistics, which in application, could be computationally intensive.

5.6 Experiments

In this section, we perform experiments to test the robustness of clustering, WCut, number of wCC's and eigengap. We employ both synthetic datasets and a Facebook dataset. Because symmetric perturbation is less interpretable, we only apply asymmetric perturbation when studying connected components and eigengap, and apply asymmetric perturbation for the remainder of the study.

5.6.1 Datasets

Synthetic datasets The synthetic datasets are generated from the DC-SBM model with the number of clusters $K = 5$, and the distribution of the cluster sizes $\frac{n_k}{n} = (0.1, 0.2, 0.3, 0.2, 0.2)$. DC-SBM is defined in Chapter 1.2.2. The clustering is guaranteed to be recovered with small errors by spectral clustering algorithm if the graph is generated from DC-SBM [55, 56, 61]. The weights of DC-SBM, $w^i_{DC-SBM} \sim 0.5 + 0.5 \times Uniform(0, 1)$ if not otherwise specified. The graphs are generated with different n and spectrum in the following experiments. The visualization of the graphs and the model parametrization are shown in Figure 5.3.

Facebook dataset The Facebook dataset [38] is an undirected connected graph which consists of 10 anonymized ego networks. It has 4039 nodes and 88234 edges. The data was collected from survey participants using a Facebook application called Social Circles. Each cluster is consists of the members within an ego network. The visualization of the Facebook dataset is shown in Figure 5.4.

In our experiments, we only examines undirected graphs. Since the definition of L is universal for all graphs, the robustness of WCut and clustering can be measured in the same way. The graph properties we defined for the number of wCC's and eigengap are only meaningful for undirected graphs, and we do not know yet how to measure their robustness for directed graphs.

We say a dataset or a graph is hard if there are many edges between clusters, and graph properties calculated from them are expected to be sensitive. A harder graph usually has smaller n, larger eigengap, denser connections between clusters and sparser connections within clusters. From the theory of the recovery of clustering described in Chapter 2, it can be indicated that the graph properties in the harder graphs wil be less robustness.

Notice that in the real datasets, e.g. Facebook, one may argue that the perturbation may not be meaningful since the edges are given continuous weights after the perturbation, while the edges can only take the values 0 and 1, e.g, two people are either friends or not. It is true that in reality, the change of this kind of social networks is restricted to adding or deleting nodes or edges. Our perturbation methods do allow for deletion of nodes and edges when the weight are generated node resampling mechanism. They can be viewed as deletions from the graph. We can also utilize node resampling to generate discrete weights for the nodes and edges to make it more meaningful. One potentail concern is that we do have the constraint that no new edges or nodes can be added from our perturbation framework.

5.6.2 Robustness of clustering

We design this experiment to answer two questions. Firstly, For fixed σ_w, does BP depend on weight distribution? Secondly, Is BP informative? We perform the experiment using the following steps. Firstly, we generate w from the four weight distributions: node resampling, binary, gamma and mixture distributions. with $E(w) = 1$ and σ_w, as described in Section 5.1. We assign various magnitude of σ_w to capture the amount of perturbation the clusterings can tolerate.

Secondly, for each perturbed graph $\tilde{\mathcal{G}}$, we perform spectral clustering algorithm and obtain $\tilde{\mathcal{C}}$. We then compute misclassification errors $\mathrm{dist}(\mathcal{C}_{true}, \tilde{\mathcal{C}})$, and $\mathrm{dist}(\mathcal{C}_{spc}, \tilde{\mathcal{C}})$, where $\mathcal{C}_{true}$ is the true clustering of the underlying model, and $\mathcal{C}_{spc}$ is the clustering obtained from the unperturbed graph through spectral clustering. The misclassification error is define in Section 1.2.1. If there is a value of σ_w where $\tilde{\mathcal{C}}$ becomes very unstable, that the misclassification error starts to have high variance, we will call it the breakdown point.

We employ two undirected synthetic datasets from DC-SBM. They both have $n = 800$, and $\lambda_{1:K} = (0, 0.2, 0.4, 0.6, 0.8)$. The difference is that the first dataset is generated with $w^i_{DC-SBM} \sim 0.5 + 0.5 \times Uniform(0, 1)$, and the second one is generated with $w^i_{DC-SBM} \sim 0.4 + 0.6 \times Uniform(0, 1)$. The results are shown in Figure 5.5.

For all the cases, we observe a significant break in the variance of misclassification errors, which we define as BP. For example, the BP for binary distribution in the easy dataset is around 0.8, which becomes 0.6 in the hard dataset. Notice that with all weight distributions, BP's in the harder dataset are always smaller comparing to that in the eaiser dataset. This confirms that BP is informative, since it predicts the sensitivity of graph properties in harder (less robust) dataset. For the same dataset, across different weight distributions, the misclassification errors break at different σ_w, which suggests that BP varies with different weight distributions.

5.6.3 Robustness of WCut

In the experiments here we want to verify that IF of WCut is informative, and then probe the source of robustness of WCut at the node level.

For the synthetic dataset with $n = 800$, $w \sim 0.5 + 0.5 \times Uniform(0, 1)$, We firstly obtain its clustering $\mathcal{C}$ from spectral clustering algorithm. We then add noise to $\mathcal{C}$ by randomly picking 200 nodes and reassigning them to other clusters randomly, in which way we obtain $\tilde{\mathcal{C}}$. The WCut of

$\tilde{\mathcal{C}}$ is expected to be less robust comparing to that of $\mathcal{C}$, since there are already more noises in $\tilde{\mathcal{C}}$. After computing IF^{WCut} for both $\mathcal{C}$ and $\tilde{\mathcal{C}}$, we plot the histograms in Figure 5.6. We observe that in the histogram for $\mathcal{C}$, IF^{WCut} is more concentrated on 0 with fewer large influence nodes, thus indicating a robust clustering. On the other hand, the histogram for $\tilde{\mathcal{C}}$ indicates the clustering being sensitive. These findings correspond with our expectations and show that IF^{WCut} is informative.

The other experiment we do is to probe where the sensitivity comes from through a partial perturbation. We generate synthetic dataset with $n = 800$, and $\lambda_{1:K} = (0, 0.2, 0.4, 0.6, 0.8)$. We also perform the experiment on the Facebook dataset. We select the nodes with $IF_i^{WCut}(\mathcal{C}_{true}) > 0$. These nodes are considered not well clustered because increasing the weights on them is expected to increase WCut, thus worse quality of clustering. For these nodes, we generate w from $Gamma(0.1/\mu, 10\mu^2)$, where $E(w) = \mu$, $Var(w) = 0.1$. The rest of the nodes have $w = 1$. We then assign the weights to the edges through asymmetric perturbation. We also examine the perturbed clustering from spectral clustering algorithm.

The results are shown in Figure 5.7, we observe that with both datasets, WCut increases as the weights on the nodes with bad influence increase, which indicates that the nodes with IF^{WCut} causes the quality of clustering to drop, and could potentially be not well clustered. In the synthetic dataset, $\mathrm{dist}(\tilde{\mathcal{C}}, \mathcal{C}_{true})$ increases as $E(w)$ increases, and decreases slightly but steadily as $E(w)$ decreases. This is because spectral clustering is equivalent to optimizing WCut [44], when less weight is imposed on nodes with bad influence for $IF_i^{WCut}(\mathcal{C}_{true})$, WCut gets smaller and $\tilde{\mathcal{C}}$ becomes closer to $\mathcal{C}_{true}$. In the Facebook dataset, $\mathrm{dist}(\tilde{\mathcal{C}}, \mathcal{C}_{true})$ decreases as $E(w)$ increases. This is because the underlying clustering $\mathcal{C}_{true}$ in Facebook does not correspond to the clustering that minimizing WCut. Therefore minimizing WCut does not lead to $\tilde{\mathcal{C}}$ being close to $\mathcal{C}_{true}$.

5.6.4 Robustness of wCC's and eigengap

We firstly probe the sensitivity of wCC's and eigengap. We calculate $IF_i^{f_u}$, $IF_i^{f_l}$ and IF^{f_e} for the nodes and select the nodes with bad influence. We call the nodes with $IF_i^{f_u} < 0$ or $IF_i^{f_l} > 0$

or $IF^{f_e} < 0$ nodes with bad influence, since these nodes make the distinction between wCC's or eigengap less significant. We do a partial perturbation on the graphs by perturbing the weights of these nodes with $w_i \sim Uniform(a, a + 0.25)$, $a \in [0.5, 0.6, \cdots, 1.5]$.

We also examine their robustness by making full perturbation on the entire graph. We generate $w_i \sim Mixture(0.5, b, 0, 0.1, T_-, Gamma, p)$ for all the nodes, where T_- is a distribution centered around 0.51 with probability 1, p ranges from 0.1 to 0.9. We choose T_- to maintain the weights on the edges to be positive. Since the strength of perturbation is mostly coming from T_+ and the choice of the base binary distribution, we do not lose generosity. In this way $E(w) = 1$ and $\sigma(w) = \sigma_w$ varies.

The synthetic dataset for testing the robustness of wCC's is generated with $n = 2000$, and $\lambda_{1:K} = (0, 0.1, 0.11, 0.12, 0.13)$. In the synthetic dataset, because of the nature of the model, the largest f_l in the graph appears when $K = 5$. In the Facebook dataset, through calculation ,we find that the largest f_l appears when $K = 7$. Therefore, we assume initially there are 5 wCC's in the synthetic dataset and 7 connected components in Facebook dataset. The results are shown in Figure 5.8.

The synthetic dataset for testing the robustness of eigengap is generated with $n = 800$ and $\lambda_{1:K} = (0, 0.1, 0.2, 0.3, 0.4)$. We call $f_e(i) = \lambda_{i+1} - \lambda_i$ the ith eigengap. Through calculation, we find that the largest eigengap is the 5th in synthetic dataset, and 7th in the Facebook dataset.

We observe that, f_l increases while f_u and f_e decrease as $E(w)$ increases in both datasets, which indicates that, as the bad influence increases, there is greater connectivity between the connected components and the eigengap becomes less significant. BP is defined as the σ_w where f_u, f_l and f_e is dominated with another K. We observe the BP's for both the number of wCC's and eigengap. In the Facebook dataset, they seem to be robust to perturbation. Note that the Facebook dataset is consist of 10 ego-networks while our experiments indicate that there are only 7 robust-

ness wCC's inside, meaning there are 3 more connected users who are grouped to one wCC.

5.7 Discussion

This chapter makes several contributions. Firstly, it provides an innovative way to perturb the network by assigning the weights on the nodes and edges. The strength of perturbation can be well controlled and can be arbitrarily small. The topology of the graph is also preserved after the perturbation. Secondly, it extends the definitions of influence function and breakdown point to the graph properties. Although these measures are widely used in the robust statistics literature, using them on graph properties is the first time. Last but not the least, we are also able to probe the source of robustness by quantifying the influence of nodes on the robustness of graph properties, which provides a deeper insight into the problem.

Our perturbation framework also have its limitations. For example, no new edges or nodes can be added to the graphs through our perturbation methods, and this may not be natural evaluating the graph properties in some social networks. Moreover, the perturbation methods is not suitable for evaluating some graph properties, e.g. properties related to graph distances. These could be the area for future explorations.

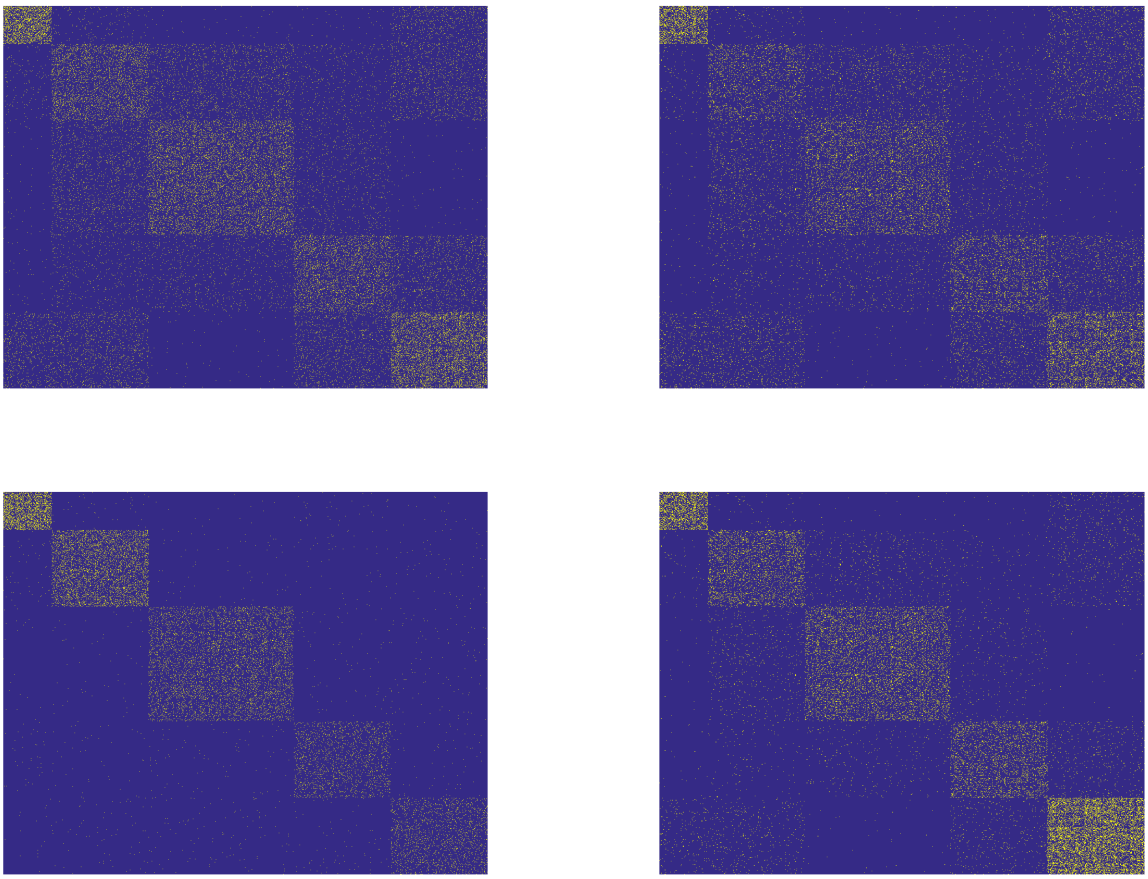


Figure 5.3: The synthetic datasets. Top left: $n = 800$, $\lambda_{1:K} = (0, 0.2, 0.4, 0.6, 0.8)$, $w^i_{DC-SBM} \sim 0.5 + 0.5 \times Uniform(0, 1)$. Top right: n and $\lambda_{1:K}$ the same with top left, $w^i_{DC-SBM} \sim 0.4 + 0.6 \times Uniform(0, 1)$. Down left: $n = 2000$, $\lambda_{1:K} = (0, 0.1, 0.11, 0.12, 0.13)$. Down right: $n = 800$, $\lambda_{1:K} = (0, 0.1, 0.2, 0.3, 0.4)$

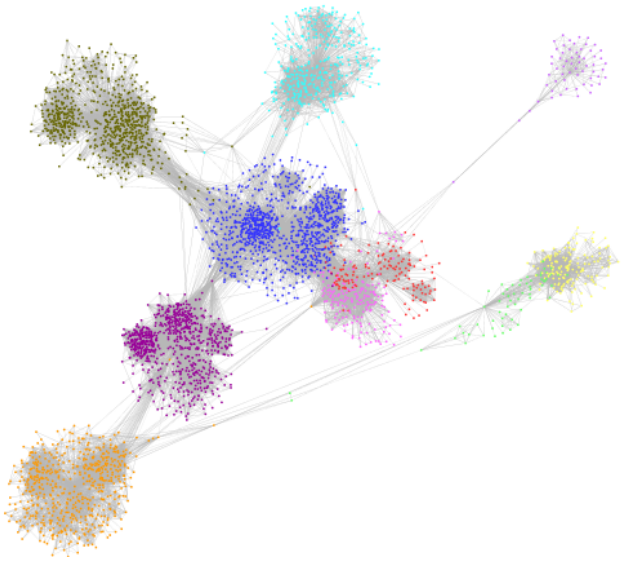


Figure 5.4: Facebook dataset

Figure 5.5: Left: easy dataset with $w^i_{DC-SBM} \sim 0.5 + 0.5 \times Uniform(0, 1)$. Right: hard dataset with $w^i_{DC-SBM} \sim 0.4 + 0.6 \times Uniform(0, 1)$. 1st row: Node resampling. 2And row: Binary . 3rd row: gamma distribution. 4th row: mixture-uniform gamma distirbution. $\tilde{\mathcal{C}}$ is the weighted version of clustering obtained from spectral clustering algorithm. Each boxplot is consist of 100 repetitions.

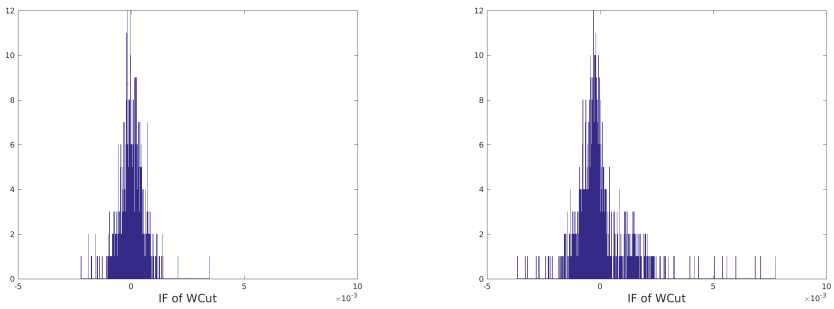


Figure 5.6: The histograms of IF^{WCut} for synthetic datasets. Left: $\mathcal{C}$. Right: $\tilde{\mathcal{C}}$.

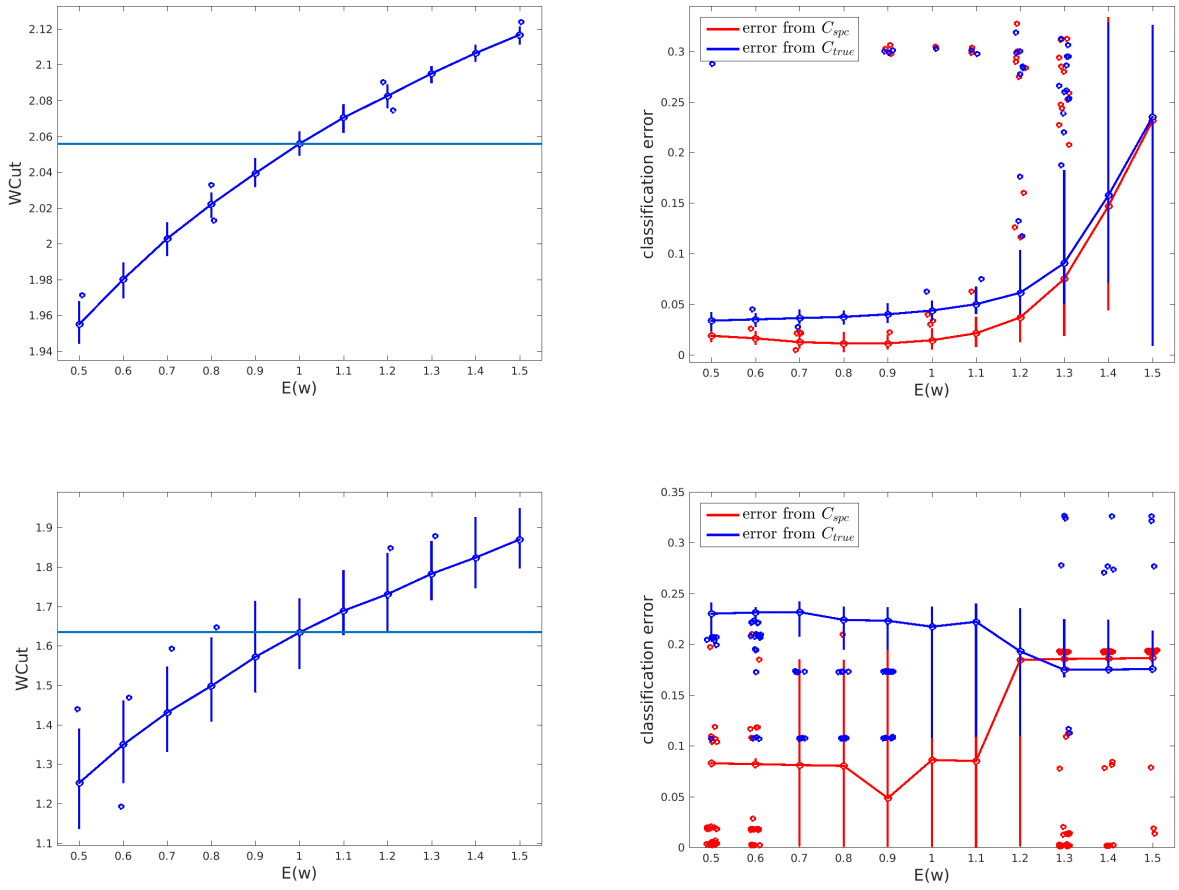


Figure 5.7: Left column: the change of WCut with respect to $E(w)$. Right column: the change of classification error with respect of $E(w)$. First row: synthetic datset. Second row: Facebook datset. Each boxplot is consist of 100 repetitions.

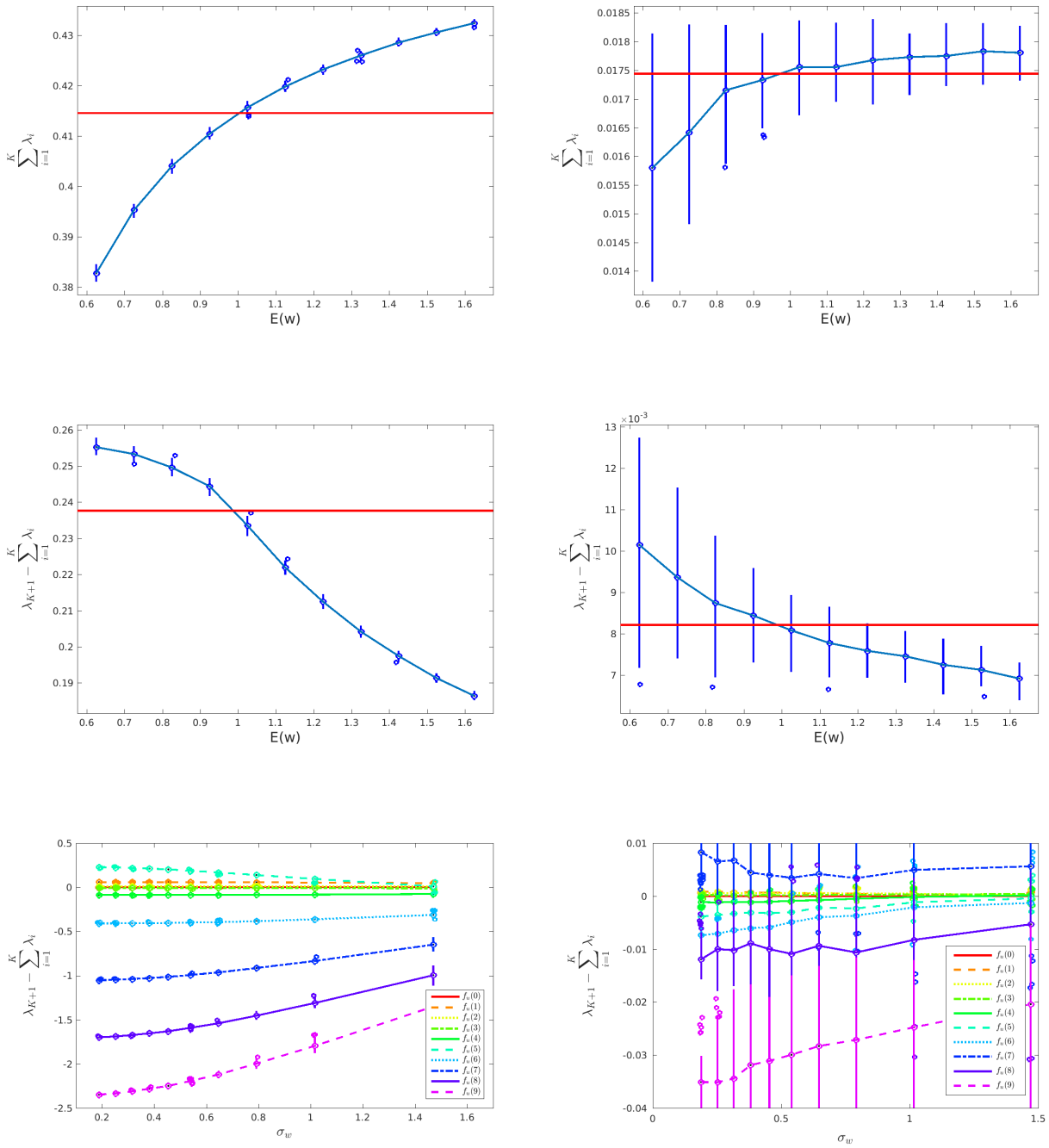


Figure 5.8: Topleft: $IF(\lambda_2, w)$. Left column: synthetic datset. Right column: Facebook dataset. First row: $IF(f_l)$. Second row: $IF(f_u)$. Third row: breakdown point of f_u. Each boxplot is consist of 100 repetitions.

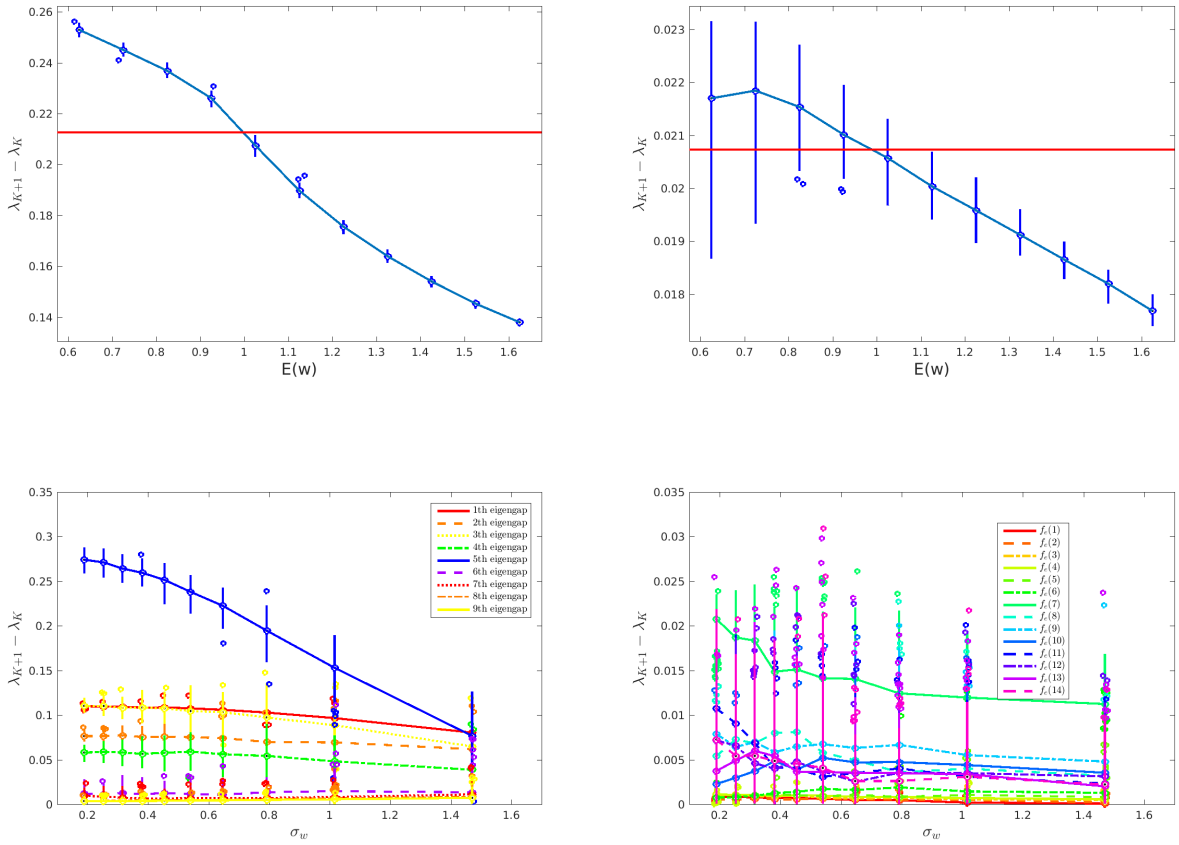


Figure 5.9: Left column: synthetic dataset. Right column: Facebook dataset. First row: $IF(f_e)$. Second row: breakdown point of f_e. Each boxplot is consist of 100 repetitions.

Chapter 6

CONCLUSION

In this thesis, we work on providing solutions for two problems in social networks: community recovery and robustness of graph properties. From Chapter 2 to Chapter 4, we focus on providing guarantees for community recovery, where we tackle the problem from both model-based and model-free perspectives. In Chapter 5, we propose a perturbation framework to measure the robustness of graph properties. In the rest of this chapter, we will summarize our findings and contributions in both problems.

6.1 Our contributions in community recovery problem

We firstly discuss our work in community recovery. In Chapter 2, we benchmark the current model-based recovery theorems, where the graphs are assumed to be generated from block models. We compare the current recovery theorems on DC-SBM and show that most of the assumptions in the recovery theorems are far from being necessary, suggesting that the number of cases that are predicted by the theorems is only a small proportion. The only recovery theorem that is shown to be useful in practice is [62], by which the clusterings from non-trivial cases are guaranteed. This recovery theorem is proven by us and wil be discussed in the next paragraph. We also create a software tool for researchers, which can be easily extended to test and compare new recovery theorems.

We then discuss our community recovery results proven in [62], and it is presented in Chapter 3. We firstly propose a broader class of network models called Preference Frame Model. PFM has its advantages in several ways: firstly, it subsumes the current block models including SBM and DC-SBM. Secondly, it separates the high level parameters for recovery of clusterings from the nuisance parameters and provide weaker constraints. Thirdly, its parametrization is more meaningful and more interpretable. Our recovery theorem is proven with respect to PFM. Although the model is more relaxed comparing to SBM and DC-SBM, the constraints for recovery is no more restricted.

The model-based guarantees for recovery discussed above are limited by the assumption that the graph needs to be generated from a model. In Chapter 4, we propose a model-free framework which provides theoretical guarantees for the results of model-based clustering algorithms without

making assumptions on data generation process. We also instantiate this framework and obtain the model-free guarantees for SBM and PFM models. This work formulates the problem of model free validation in the area of community detection in social networks. It is also closely related to the model-based recovery theorems, which can be easily modified to be model-free results through our framework. Meanwhile, the proof techniques we use in the model-free results are more elementary and more direct. They can be used for obtaining sharper thresholds in model-based theorems.

6.2 Our contributions in robustness of graph properties problem

In Chapter 5, we propose a perturbation framework for measuring the robustness of graph properties. We are not the first to explore this approach. Current methods [30, 24, 14, 4] have done perturbation on graphs by randomly deleting or adding nodes and edges, in which the magnitude of perturbation cannot be well controlled. The goal in this research is to provide model-free perturbation framework that allows for arbitrarily small perturbation. Our perturbation method is done through assigning weights to the nodes, where the strength of perturbation is controlled continuously by the variance of weight distribution. We suggest a mixture distribution for generating the weights, which can balance the bias introduced in Laplacian and provide a good variability of weights. Our perturbation methods have the advantages that they can be easily controlled to provide arbitrarily small perturbation and preserve the topology of the graphs.

We then measure the robustness of graph properties by extending the classical robustness measure, i.e, influence function and breakdown point, to the context of graphs. Although the are widely used in robust statistics literature, using them in the graphs is the first time. Moreover, our perturbation framework also enables one to probe the source of robustness at the node level.

BIBLIOGRAPHY

[1] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015.

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

[3] Edoardo M. Airoldi, David S. Choi, and Patrick J. Wolfe. Confidence sets for network structure. Technical Report arXiv:1105.6245, 2011.

[4] Waqar Ali, Anatol E Wegner, Robert E Gaunt, Charlotte M Deane, and Gesine Reinert. Comparison of large networks with sub-sampling strategies. *Scientific reports*, 6:28955, 2016.

[5] Sanjeev Arora, Rong Ge, Sushant Sachdeva, and Grant Schoenebeck. Finding overlapping communities in social networks: toward a rigorous approach. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 37–54. ACM, 2012.

[6] Pranjal Awasthi. Clustering under stability assumptions. In *Encyclopedia of Algorithms*, pages 331–335. 2016.

[7] Francis Bach and Michael I. Jordan. Learning spectral clustering with applications to speech separation. *Journal of Machine Learning Research*, 7:1963–2001, 2006.

[8] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 954–962, 2011.

[9] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise thresholds for spectral clustering. In John Shawe-Taylor, Richard S. Zemel, Peter L. Bartlett, Fernando C. N. Pereira, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 954–962, 2011.

[10] Maria-Florina Balcan, Christian Borgs, Mark Braverman, Jennifer Chayes, and Shang-Hua Teng. Finding endogenously formed communities. *arxiv preprint arXiv:1201.4899v2*, 2012.

[11] Maria-Florina Balcan, Christian Borgs, Mark Braverman, Jennifer Chayes, and Shang-Hua Teng. Finding endogenously formed communities. In *Proceedings of the Twenty-Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 767–783. SIAM, 2013.

[12] Shai Ben-David. Computational feasibility of clustering under clusterability assumptions. *CoRR*, abs/1501.00437, 2015.

[13] Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[14] Sharmodeep Bhattacharyya, Peter J Bickel, et al. Subsampling bootstrap of count features of networks. *The Annals of Statistics*, 43(6):2384–2411, 2015.

[15] Yonatan Bilu and Nathan Linial. Are stable instances easy? *CoRR*, abs/0906.3162, 2009.

[16] Bela Bollobas. *Random Graphs*. Cambridge University Press, second edition, 2001.

[17] K. Chaudhuri, F. Chung, and A. Tsiatas. Spectral clustering of graphs with general degrees in extended planted partition model. *Journal of Machine Learning Research*, pages 1–23, 2012.

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

[19] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

[20] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

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

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

[23] Paul Erdos and Alfréd Rényi. On the evolution of random graphs. *Bull. Inst. Internat. Statist*, 38(4):343–347, 1961.

[24] David Gfeller, Jean-Cédric Chappelier, and Paolo De Los Rios. Finding instabilities in the community structure of complex networks. *Physical Review E*, 72(5):056135, 2005.

[25] Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, and Edoardo M Airoldi. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

[26] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 114. John Wiley & Sons, 2011.

[27] P.D. Hoff, A.E. Raftery, and M.S. Handcock. Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.*, 97(460):1090–1098, 2002.

[28] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[29] M. O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

[30] B. Karrer, E. Levina, and M.E.J. Newman. Robustness of community structure in networks. *Physical Revieww E*, 77(4)::046119, 2008.

[31] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83:016107, Jan 2011.

[32] G. Karypis and V. Kumar. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM Journal on Scientific Computing*, 20:359392, 1998.

[33] Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. Efficient active algorithms for hierarchical clustering. *arXiv preprint arXiv:1206.4672*, 2012.

[34] Florent Krzakala, Cristopher Moore, Elchanan Mossel, Joe Neeman, Allan Sly, Lenka Zdeborová, and Pan Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

[35] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[36] Can M. Le and Roman Vershynin. Concentration and regularization of random graphs. *arXiv preprint arXiv:1506.00669*, 2015.

[37] Marc Lelarge, Laurent Massoulié, and Jiaming Xu. Reconstruction in the labeled stochastic block model. In *Information Theory Workshop (ITW), 2013 IEEE*, pages 1–5. IEEE, 2013.

[38] Jure Leskovec and Julian J Mcauley. Learning to discover social circles in ego networks. In *Advances in neural information processing systems*, pages 539–547, 2012.

[39] Jan R Magnus, Heinz Neudecker, et al. Matrix differential calculus with applications in statistics and econometrics. 1995.

[40] Brendan McKay. Asymptotics for symmetric 0-1 matrices with prescribed row sums. *Ars Combinatoria*, 19A:15–26, 1985.

[41] Brendan McKay and Nicholas Wormald. Uniform generation of random regular graphs of moderate degree. *Journal of Algorithms*, 11:52–67, 1990.

[42] Brendan McKay and Nicholas Wormald. Asymptotic enumeration by degree sequence of graphs with degrees $o(n^{1/2}$. *Combinatorica*, 11(4):369–382, 1991.

[43] Marina Meilă. Local equivalence of distances between clusterings – a geometric perspective. *Machine Learning*, 86(3):369–389, 2012.

[44] Marina Meilă and William Pentney. Clustering by weighted cuts in directed graphs. In Chid Apte, David Skillicorn, and Vipin Kumar, editors, *Proceedings of the SIAM Data Mining Conference, SDM*. SIAM, 2007.

[45] Marina Meilă and Jianbo Shi. Learning segmentation by random walks. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 873–879, Cambridge, MA, 2001. MIT Press.

[46] Marina Meilă and Jianbo Shi. A random walks view of spectral segmentation. In T. Jaakkola and T. Richardson, editors, *Artificial Intelligence and Statistics AISTATS*, 2001.

[47] Marina Meilă, Susan Shortreed, and Liang Xu. Regularized spectral learning. In Robert Cowell and Zoubin Ghahramani, editors, *Proceedings of the Artificial Intelligence and Statistics Workshop(AISTATS 05)*, 2005.

[48] E. Mossel, J. Neeman, and A. Sly. Belief propagation, robust reconstruction and optimal recovery of block models. In *Proceedings of The 27th Conference on Learning Theory*, volume 35, pages 356–370, 2014.

[49] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for binary symmetric block models. 2014.

[50] Boaz Nadler, Stephane Lafon, Ronald Coifman, and Ioannis Kevrekidis. Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 955–962, Cambridge, MA, 2006. MIT Press.

[51] M.E.J. Newman and Travis Martin. Equitable random graphs. 2014.

[52] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002. MIT Press.

[53] J.R. Norris. *Markov Chains*. Cambridge University Press, 1997.

[54] Richard Peng, He Sun, and Luca Zanetti. Partitioning well-clustered graphs: Spectral clustering works! In Peter Grünwald and Elad Hazan, editors, *Proceedings of The 28th Conference on Learning Theory (COLT)*, volume 40, pages 1–33, 2015.

[55] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, 2013.

[56] Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

[57] Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, pages 1878–1915, 2011.

[58] Gilbert W. Stewart and Ji-guang Sun. *Matrix perturbation theory*. Academic Press, San Diego, CA, 1990.

[59] Nicolas Verzelen, Ery Arias-Castro, et al. Community detection in sparse random networks. *The Annals of Applied Probability*, 25(6):3465–3510, 2015.

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

[61] Yali Wan and Marina Meila. Benchmarking recovery theorems for the DC-SBM. In *International Symposium on Artificial Intelligence and Mathematics (ISAIM)*, 2015.

[62] Yali Wan and Marina Meila. A class of network models recoverable by spectral clustering. In Daniel Lee and Masashi Sugiyama, editors, *Advances in Neural Information Processing Systems (NIPS)*, 2015.

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