Beyond the Arithmetic Mean: Extensions of Spectral Clustering and Semi-Supervised Learning for Signed and Multilayer Graphs via Matrix Power Means

A dissertation submitted towards the degree of
Doctor of Natural Sciences (Dr. rer. nat.)
of the Faculty of Mathematics and Computer Science
of Saarland University

by
Pedro Eduardo Mercado López, M.Sc.

Saarbrücken, 2021

Date of defense	30th of November, 2021
Dean of the faculty	Univ.-Prof. Dr. Thomas Schuster

Examination Committee

Chair	Prof. Dr. Markus Bläser
Reviewer, Advisor	Prof. Dr. Matthias Hein
Reviewer	Prof. Dr. Francesco Tudisco
Reviewer	Prof. Dr. Ulrike von Luxburg
Academic assistant	Dr. Amir Zandieh

ABSTRACT

In this thesis we present extensions of spectral clustering and semi-supervised learning to signed and multilayer graphs. These extensions are based on a one-parameter family of matrix functions called Matrix Power Means. In the scalar case, this family has the arithmetic, geometric and harmonic means as particular cases.

We study the effectivity of this family of matrix functions through suitable versions of the stochastic block model to signed and multilayer graphs. We provide provable properties in expectation and further identify regimes where the state of the art fails whereas our approach provably performs well. Some of the settings that we analyze are as follows: first, the case where each layer presents a reliable approximation to the overall clustering; second, the case when one single layer has information about the clusters whereas the remaining layers are potentially just noise; third, the case when each layer has only partial information but all together show global information about the underlying clustering structure.

We present extensive numerical verifications of all our results and provide matrix-free numerical schemes. With these numerical schemes we are able to show that our proposed approach based on matrix power means is scalable to large sparse signed and multilayer graphs.

Finally, we evaluate our methods in real world datasets. For instance, we show that our approach consistently identifies clustering structure in a real signed network where previous approaches failed. This further verifies that our methods are competitive to the state of the art.

ZUSAMMENFASSUNG

In dieser Arbeit stellen wir Erweiterungen von spektralem Clustering und teilüberwachtem Lernen auf signierte und mehrschichtige Graphen vor. Diese Erweiterungen basieren auf einer einparametrischen Familie von Matrixfunktionen, die Potenzmittel genannt werden. Im skalaren Fall hat diese Familie die arithmetischen, geometrischen und harmonischen Mittel als Spezialfälle.

Wir untersuchen die Effektivität dieser Familie von Matrixfunktionen durch Versionen des stochastischen Blockmodells, die für signierte und mehrschichtige Graphen geeignet sind. Wir stellen beweisbare Eigenschaften vor und identifizieren darüber hinaus Situationen in denen neueste, gegenwärtig verwendete Methoden versagen, während unser Ansatz nachweislich gut abschneidet. Wir untersuchen unter anderem folgende Situationen: erstens den Fall, dass jede Schicht eine zuverlässige Approximation an die Gesamtclusterung darstellt; zweitens den Fall, dass eine einzelne Schicht Informationen über die Cluster hat, während die übrigen Schichten möglicherweise nur Rauschen sind; drittens den Fall, dass jede Schicht nur partielle Informationen hat, aber alle zusammen globale Informationen über die zugrunde liegende Clusterstruktur liefern.

Wir präsentieren umfangreiche numerische Verifizierungen aller unserer Ergebnisse und stellen matrixfreie numerische Verfahren zur Verfügung. Mit diesen numerischen Methoden sind wir in der Lage zu zeigen, dass unser vorgeschlagener Ansatz, der auf Potenzmitteln basiert, auf große, dünnbesetzte signierte und mehrschichtige Graphen skalierbar ist.

Schließlich evaluieren wir unsere Methoden an realen Datensätzen. Zum Beispiel zeigen wir, dass unser Ansatz konsistent Clustering-Strukturen in einem realen signierten Netzwerk identifiziert, wo frühere Ansätze versagten. Dies ist ein weiterer Nachweis, dass unsere Methoden konkurrenzfähig zu den aktuell verwendeten Methoden sind.

ACKNOWLEDGEMENTS

I want to thank Prof. Dr. Matthias Hein. It is because of Matthias that this journey has been worthwhile. Thanks to him I learned the meaning of high scientific standards, of working hard to understand the problem that we had in our hands, and to always keep ourselves on track. Given the current hype and fast developments in our field, it has been of utmost relevance to have a working atmosphere like the one Matthias provides, where we constantly keep ourselves on the track of providing high quality scientific results.

I want to thank Dagmar Glaser for all her advice, assistance and orientation in our group. There were plenty of times where Dagmar was the keymaster for organisation tasks, cultural answers, and plenty of existencial questions about everything that can pop out in a floor full of international PhD students.

A key ingredient of all this experience are my colleagues in the group. I am thankful to all of you, for being joyful, cheerful, motivating and risk-oriented. In particular I would like to thank Quynh Nguyen Ngoc for being the best office mate in our two-person office in Saarland, Francesco Tudisco for always having an enormous roman sense of empathy, guidance and understanding, to Antoine Gautier for his endless creativity and for his example of hard-work, to Francesco Croce for his time cultivating me about the italian traditions of coffee and milk (before noon), Julian Bitterwolf for his unprecedented divergent creativity and for being my favourite office mate in our two-person office in Tübingen, Alexander Meike for initiating interesting debates in our coffee time (though he does not drink any coffee), Max Augustin for sharing with us his passion about sport hobbies in Tübingen, Laurenz Hemmen for his fresh view on the work we do in the group, and Maksym Andriushchenko for his incredible level of stamina and motivation pervading all of our group.

There are more members of our group that left an imprint in this time, but that departed while I was just starting: Syama Sundar Rangapuram, Thomas Buehler, Martin Slawski, Sahely Bhadra, Pratik Jawanpuria, Vikram Tankasali, Mahesh Mukkamala, Nikita Vedeneev, Shweta Mahajan, Cristian Caloian, Pramod Mudrakarta, Anastasia Podosinnikova, among others. Thanks to all of you for always being present.

I would like to thank Evgeny Sobaka, Mohamed Sa7by, Anna Khoreva, Jesus Calvillo, Georgios Arvanitidis, Pavel Kolev, Kailash, and Andreitaw Chihuahua for their friendship, help and orientation before, during and at the final phase of my studies in Saarland and Tübingen.

I want to thank to all my friends whom I met accidentally through the joyous project of Fabrizio Nunnari, and that lead to so deep and endless friendships. Learning how to walk in an embrace (which ideally would resemble what usually people call tango) ended up being the place to build everlasting friendships. Thanks Fabrizio, Nuria, Charalampos, Alina, Cheng Hua, Oxana, Ineschen, Eva K., Ursula, Eva

Csiz., Judith G., Tabea G., Vania, Ildar, Estrellita, Gitana, Jesus, Barbie, Susanne, Mirjana, Heesh, Charlotte, Eva Almenar, Dilafruz, Jussif, Detlef, Melina, Elie L'Hoyest, and many more.

My flatmates Bennet, Clara and Jonas have been instrumental on keeping the motivation high in the unexplored and mysterious streets and basements of Tübingen.

I am so grateful for my friends that still keep in touch even when the time difference makes things a bit difficult: Emmanuel, Jorge C. Febbles, Rodo Homöplato, Guillo, Juan Pablo, Daniel 8-5, Job, Rubelio, Mildred.

In its own way I would like to thank Alexis Arroyo, Javier Velasco, and Carmenchu for their very essential and relevant presence, independent of time, location and subject.

I want to thank Prof. Angel Kuri Morales. It was in a very casual conversation with him that I ever first heared of topics related to machine learning and data mining. Such a conversation in one of the corridors of ITAM has been key in the path I have taken in my life. I want to thank him as well for my very first course on neural networks.

This research adventure started way long time ago during my bachelor studies, on the first day when I met Araceli Reyes Guerrero. There is no way I can thank for all the passion that Araceli has shared along the time in the classrooms of ITAM, during her office hours and voluntary seminars held during summers having me as her only student, where my assignment was to pseudo-lecture her basic texts on linear algebra that she knew by heart. Araceli's example and altruism have been paramount in my development from the very first days.

Finally, I want to thank my family, for their inconditional support and love.

CONTENTS

1 INTRODUCTION

Interactions are probably one of the major footprints of our times. Nowadays interactions are permeating every aspect of our lives, taking unprecedented magnitudes and posing new scientific challenges. Some examples of the current relevance of interactions are:

- People who will never physically meet interact with each other by getting involved in discussions that potentially engage thousands of people from the most diverse countries and cultures.

- Researchers collaborate with colleagues who are based in other continents to potentially produce scientific outputs that would be impossible wihtout interactions, involving hundreds of scientific institutes around the world.

- International political treaties are discussed through virtual conferences, allowing interactions and dynamics to evolve. We reach agreements with worldwide consequences.

- Due to the present global pandemic, e-commerce is nowadays reaching unprecedented popularity, inducing transactions between costumers and producers that otherwise would have never happened.

- Interactions are enhanced by the unprecedented level of human mobility around the world. This became evident with the astonishing virus spread speed leading to our current pandemic crisis.

To get a first glimpse of the nature of observed interactions, a first step consists in the identification of sets of observations that present a similar behavior. To reach this goal several clustering methods have been proposed so far, with spectral clustering one of the most popular.

Spectral clustering is a graph-based method that identifies clusters based on observed interactions, by first providing a suitable embedding of the nodes from a graph operator, and later making cluster assignments. Spectral clustering has received a relevant amount of attention for several reasons, for instance: superiority in performance against other clustering approaches, a simple algorithmic description, and many mathematical properties.

Despite the success of spectral clustering, one of its major drawbacks is its assumption that all observed interactions are of the same nature. However, interactions

between entities are potentially of many kinds, and each of them is likely to provide information regarding the underlying clustering structure between the entities. Particular instances of multiple kinds of interactions are:

- **Signed Graphs**: graphs that encode both positive and negative kinds of interactions, where positive interactions represent similarity, trust, or friendship, and negative interactions encode dissimilarity, distrust, or conflicts.
- **Multilayer Graphs**: graphs that encode multiple kinds of interactions between a fixed set of entities. For instance, it is possible that between researchers several diverse interactions take place at the same time, such as interactions by citations, co-authorships, or by jointly organized symposiums. All these interactions, while informative of certain underlying clustering structure, convey different semantic meanings.

The main limitation of spectral clustering to signed and multilayer graphs comes from the fact that it relies on generating an informative embedding of the set of nodes which is obtained from a suitable graph operator, for instance, the graph Laplacian. Hence, when multiple kinds of interactions are observed, the question is how one can merge the information encoded from multiple kinds of interactions to obtain a useful embedding from the nodes.

Since most graph-based methods rely on some sort of a graph operator, the limitations of spectral clustering on signed and multilayer graphs permeate as well into other graph-based techniques, for instance graph-based semi-supervised learning.

The goal of graph-based semi-supervised learning is to build a classifier that takes into account both labeled and unlabeled observations, by considering a suitable loss function and the underlying graph structure of the observations. Similar to what we have observed with spectral clustering, graph-based semi-supervised learning is not clearly applicable to cases with multiple kinds of interactions, since it requires a graph operator that induces the underlying clustering structure.

A natural extension of spectral clustering and semi-supervised learning to networks with multiple kinds of interactions is to take the graph Laplacian per interaction and afterwards use an average like the arithmetic mean. Whereas this notion at first sight is a sensible one, it is not clear how effective this approach is. Moreover, if we think about the arithmetic mean, a natural task would be to consider other cases like the geometric or harmonic means.

Throughout this thesis we study extensions of spectral clustering and semi-supervised learning to the case where multiple kinds of interactions are observed. Our proposed extensions are based on a one-parameter family of matrix functions called Matrix Power Means, which in the scalar case contains as particular cases the arithmetic, geometric and harmonic means. We prove several properties of our proposed extensions under a suitable variation of the stochastic block model for signed and

multilayer graphs, and show that in expectation they outperform the state of the art. We verify our findings through extensive numerical experiments. We further perform experiments in real datasets, showing that our approach does not perform worse than the state of the art. Finally, we present matrix-free numerical schemes to show that our proposed approaches are scalable to large sparse signed and multilayer graphs.

It is worthwhile mentioning that in this work we focus on multilayer graph approaches. A popular related task is the one related to multi-view learning, based on the assbumption that several views of the same entities are available. While this approach sounds similar to the one of multilayer graphs, it is important to emphasize that multi-view approaches often assume that multiple sets of features are available, and hence the task is not strictly a graph-based task. Yet, several of these approaches make use of certain graph tools, like the graph Laplacian. For an overview of multi-view learning please see (Sun, 2013; Xu *et al.*, 2013). We emphasize that in this work we will only consider graph-based approaches that do not rely on any feature data.

1.1 RELATED WORK

In this section a brief overview is given about related work on the analysis of signed and multilayer graphs. We first start with the case of signed graphs and focus on clustering approaches, to later consider multilayer graphs and emphasize the tasks of clustering and semi-supervised learning.

1.1.1 Signed Graphs

The analysis of signed graphs can be traced back to the concept of social balance (Cartwright and Harary, 1956; Harary, 1953; Davis, 1967), where the goal is to identify a partition of the set of nodes so that positive interactions are mainly inside the clusters, and negative interactions are mainly between clusters. This notion is motivated by the concept of a k-balanced signed graph.

Definition 1.1 (Davis (1967), k-balance)**.** *A signed graph is k-balanced if the set of vertices can be partitioned into k sets such that within the subsets there are only positive edges, and between them only negative.*

The concept of a k-balanced signed graph has been predominant in the analysis of signed graphs. Several challenges are posed by signed networks in tasks like edge prediction, where the task is not only to predict if there will be an interaction, as in unsigned graphs, but to predict if it will be positive or negative. For instance (Falher *et al.*, 2017) propose to predict a directed signed edge under the Trust-Troll model, i.e. in the fraction of outgoing negative edges (trollness) and incoming positive edges (trustworthiness), (Kumar *et al.*, 2016) introduce the notions of fairness and goodness

of users in a signed social network to predict signed edges, and in (Leskovec *et al.*, 2010a) features build on the signed degree of nodes and from social balance theory are used to predict the sign of an edge.

Signed networks pose as well several challenges to the task of node embeddings. Node embeddings have received a relevant amount of attention as they allow the application of standard feature-based methods for the analysis of graphs. In the case of signed graphs, positive edges suggest that the corresponding nodes should be embedded close to each other, whereas negative edges push them apart. For instance (Chiang *et al.*, 2011) propose to consider long cycles in signed graphs, as these provide a criterion related to the notion of k-balance, (Derr *et al.*, 2018) extend the concept of graph convolutional networks to signed graphs and apply it to generate informative embeddings of the nodes, (Kim *et al.*, 2018) propose a novel embedding technique for the case of directed signed graphs, and in (Wang *et al.*, 2017) it is proposed to consider 2-hop networks and generate embeddings that follow the level of balance observed there.

Another line of work is the task of node classification. In this task the goal is to build a classifier that takes into account both labeled and unlabeled observations, based on a suitable loss function and a regularizer that induces the information encoded by positive and negative edges. Since traditional graph operators are not suitable to signed graphs, several challenges are posed by this task. For instance (Mercado *et al.*, 2019a) take an approached based on diffuse interface methods,(Tang *et al.*, 2016a) are motivated by matrix factorization approaches inspired by social balance, and (Goldberg *et al.*, 2007) propose a suitable graph operator together with an extension inspired by wrapped kernels.

Closer to our goal are methods that extend spectral clustering to signed graphs. As mentioned previously, the main challenge is to introduce an operator that merges the information encoded by both positive and negatives interactions such that the eigenvectors corresponding to the smallest eigenvalues are informative. Several efforts have been focused on this task, for instance (Kunegis *et al.*, 2010) propose a new signed graph Laplacian inspired by extending the notion of graph cuts to signed graphs. The proposed signed graph Laplacian is basically the addition of the standard graph Laplacian on positive edges, and the signless Laplacian (Desai and Rao, 1994) on negative edges. They show that the proposed signed graph Laplacian inherits properties of the unsigned case, like being positive-semidefinite and that the zero eigenvalue is observed if and only if the signed graph is 2-balanced.

Inspired by (Kunegis *et al.*, 2010), (Chiang *et al.*, 2012) proposed several notions of graph cuts for signed graphs, together with their continuous relaxation inducing different Laplacians for signed graphs, which are additions of suitable graph operators on positive and negative edges.

Based on the observation that previous approaches are basically additions of matrices, works like (Mercado *et al.*, 2016) and (Cucuringu *et al.*, 2019) have proposed novel extensions of spectral clustering. For instance, in (Mercado *et al.*, 2016) it is proposed to take the matrix geometric mean of suitable Laplacians of positive and

negative edges, whereas (Cucuringu *et al.*, 2019) proposed a continuous relaxation inspired by the notion of non-uniform cuts.

A closely related approach to spectral clustering of signed graphs is correlation clustering (Bansal *et al.*, 2004). Whereas in spectral clustering it is required to specify the number of clusters to identify, in correlation clustering the number of clusters is automatically identified such that the final output is as close as possible to be k-balanced. The case where the number of clusters is fixed a-priori has been explored in (Giotis and Guruswami, 2006). In the context of correlation clustering, in (Saade *et al.*, 2015) the Bethe Hessian matrix is introduced. The Bethe Hessian matrix need not be positive definite, but the number of negative eigenvalues is a good estimation of the number of clusters in a signed network, and the corresponding eigenvectors provide information of the clustering structure (Saade *et al.*, 2014).

Further works related to clustering in signed graphs are (Sedoc *et al.*, 2017; Doreian and Mrvar, 2009; Knyazev, 2018; Kirkley *et al.*, 2019; Cucuringu *et al.*, 2018). For a comprehensive survey on works related to signed networks we refer the reader to (Tang *et al.*, 2016b; Gallier, 2016).

1.1.2 Multilayer Graphs

Multilayer graphs pose several challenges due to the multiple kinds of interactions that are encoded. For instance, each kind of interaction can be seen as a network that constitutes a layer of the corresponding multilayer network. In this case, a multilayer graph can be seen as a set of graphs over the same set of nodes. Moreover, if there are interactions between nodes in different layers, then the ordering is relevant, for instance when the layers are time snapshots of the same network over time (Taylor *et al.*, 2016, 2017), and thus tensor representations are potentially more suitable for this case(Kivelä *et al.*, 2014, see Section 2.2). Hence, the representation of multiple kinds of interactions poses already certain aspects into consideration. Connections between different representations of multilayer graphs have been studied, showing, for instance, that the spectrum of supra-adjacency matrices interlace with the spectrum of the average adjacency matrix (Sánchez-García *et al.*, 2014). In this work we are going to assume that interactions happen only inside layers, and that there is no particular ordering of layers.

Further challenges of multilayer graphs are related to notions like node centrality (Tudisco *et al.*, 2018; Battiston *et al.*, 2014; Solá *et al.*, 2013), multi-scale synthesis (Lockerman *et al.*, 2016), node embeddings (Zhang *et al.*, 2018; Yang *et al.*, 2020), link prediction(De Bacco *et al.*, 2017; Ermis *et al.*, 2015; Koptelov *et al.*, 2020) among others.

In the following chapters we will give a brief overview of several approaches concerning clustering and semi-supervised learning in multilayer graphs. For a general discussion we refer the reader to (Boccaletti *et al.*, 2014; Kivelä *et al.*, 2014; Aleta and Moreno, 2019).

1.1.2.1 *Clustering with Multilayer Graphs*

The task of clustering in single layer graphs is one of the most popular tasks (Fortunato, 2010), yet, its analysis on multilayer graphs remains in its infancy (Kivelä *et al.*, 2014, see Section 4.5.1). Most of the current methods on multilayer graphs rely on joint matrix factorizations, graph-aggregation approaches, extensions of modularity, and co-training, among others. We give a brief overview of these methods.

Several approaches to clustering with multilayer graphs have been proposed based on the notion of matrix factorizations (Dong *et al.*, 2012, 2014; Tang *et al.*, 2009; Zhao *et al.*, 2017a; Rocklin and Pinar, 2011; Xia *et al.*, 2014) where the goal is to represent a multilayer graph with a small number of matrices that convey the most relevant information regarding the underlying clustering structure of a multilayer graph. (Tang *et al.*, 2009), for instance, propose a joint factorization approach applied to the adjacency matrices of the layers of a multilayer graph, whereas (Dong *et al.*, 2012) instead proposes to do this on the Laplacian spectrum of the layers.

Another approach is based on Bayesian inference (De Bacco *et al.*, 2017; Paul and Chen, 2016a; Peixoto, 2015; Schein *et al.*, 2015, 2016; Jenatton *et al.*, 2012). For this approach a certain assumption on the distribution of the interactions is made to later optimize a suitable likelihood function which frequently relies on proper optimization techniques. For a self contained introduction into Bayesian approaches related to the stochastic block model see (Peixoto, 2019).

Extensions of Newman's modularity (Newman, 2006) and related null models have been proposed to multilayer graphs (Mucha *et al.*, 2010; Paul and Chen, 2016b; Wilson *et al.*, 2017). In the traditional single layer graph the goal of Newman's modularity (Newman, 2006) is to identify a partition of the set of nodes such that the edge density per partition is larger than that of a certain reference (null) model. One of the characteristics of modularity approaches is that they as well estimate the number of clusters and hence it is not necessary to fix in advance the number of clusters to search. For instance, (Wilson *et al.*, 2017) propose a method that recovers overlapping clusters and identifies observations that are not related to the overall clustering structure. They further prove consistency in a suitable multilayer stochastic block model. Moreover, several null models for multilayer graphs are introduced in (Paul and Chen, 2016b; Bassett *et al.*, 2013) related to the degree corrected stochastic block model and Girvan-Newman modularity (Newman and Girvan, 2004).

A co-training approach is proposed in (Kumar and III, 2011) where the main assumption is that each layer in a multilayer graph is sufficient to produce meaningful clusters and that layers are compatible in the sense that they basically generate the same clustering. Hence, the goal is to produce a clustering that is consistent with the information encoded in each layer. This approach is further explored in (Kumar *et al.*, 2011) under the notion of co-regularization.

A line of research that is close to our approach goes by compressing a multilayer graph in a single matrix to later perform clustering on it. A natural first approach goes by adding the adjacency matrices or the Laplacians of the layers (Chen and Hero, 2017; Huang *et al.*, 2012; Taylor *et al.*, 2017; Zhou and Burges, 2007). For

instance, (Chen and Hero, 2017) aim at identifying the optimal convex combination of the layers of a multilayer graph, whereas (Zhou and Burges, 2007) presents a weighted arithmetic mean motivated by graph cuts. A systematic analysis under the stochastic block model has been presented in (Paul and Chen, 2020) providing a comparison of matrix factorization and matrix averaging approaches.

For recent overviews related to clustering with multilayer graphs we refer the reader to (Kim and Lee, 2015; Sun, 2013; Xu *et al.*, 2013; Zhao *et al.*, 2017b).

1.1.2.2 *Semi-Supervised Learning with Multilayer Graphs*

The task of semi-supervised learning on graphs is to build a classifier that takes into consideration both labeled and unlabeled observations. A well-established approach to this task is to take a suitable loss function on the labeled nodes and a regularizer which provides information encoded by the graph. For semi-supervised learning on single layer graphs (Zhu *et al.*, 2003) consider a Gaussian Markov random field together with harmonic functions whereas in (Zhou *et al.*, 2003) a label propagation approach is taken, and (Belkin *et al.*, 2004) propose a manifold regularization approach. Further approaches have been developed through deep learning (Yang *et al.*, 2016) and graph convolutional networks (Kipf and Welling, 2017; Wu *et al.*, 2019; Chami *et al.*, 2019).

Extensions of semi-supervised learning to multilayer graphs pose the challenge of building a classifier by taking several kinds of interaction into consideration. Inspired by the single layer approach by (Zhu *et al.*, 2003; Zhou *et al.*, 2003; Belkin *et al.*, 2004) several works have proposed different multilayer graph extensions, and hence different kinds of multilayer graph operators are implicitly considered as some sort of regularizers. Hence, a great amount of attention has been posed on finding some sort of weighted arithmetic mean of the adjacency or Laplacian matrices of the layers of a multilayer graph, such that layers with a higher weight are more informative. For instance, (Zhou and Burges, 2007) propose to take a suitable weighted arithmetic mean of adjacency matrices inspired by the notion of multilayer graph cuts, whereas in (Mostafavi *et al.*, 2008) this is achieved by taking only labeled nodes. Moreover (Tsuda *et al.*, 2005) propose a suitable arithmetic mean of the Laplacians of the layers as a regularizer, whereas (Argyriou *et al.*, 2006) proposed a convex combination of Laplacians via the pseudo inverse Laplacian kernel. (Kato *et al.*, 2009) combine Laplacians by a maximum a posterior estimation taking a Gamma distribution as a prior for the weight of each layer, and (Nie *et al.*, 2016) propose a parameter-free approach. Further, a sparse linear combination of layer Laplacians is proposed in (Karasuyama and Mamitsuka, 2013). Recently, (Viswanathan *et al.*, 2019) proposed a multi-component extension of Gaussian Markov random fields where observations on the vertices are modelled as jointly Gaussian with an inverse covariance matrix that is a weighted linear combination of multiple matrices.

Furthermore, based on improvements of belief propagation (Koutra *et al.*, 2011), a

scalable approximation to multilayer graphs has been proposed in (Eswaran *et al.*, 2017), whereas in (Gujral and Papalexakis, 2018) a tensor factorization method is designed for semi-supervised learning in multilayer graphs.

1.2 CONTRIBUTIONS AND OUTLINE

The contributions in this thesis are extensions of spectral clustering for signed and multilayer graphs, and semi-supervised learning for multilayer graphs. The proposed extensions are based on a one-parameter family of matrix means called Matrix Power Means which in the scalar case has as particular cases the arithmetic, geometric and harmonic means.

We study the effectivity of Matrix Power Means under a suitable stochastic block model for signed and multilayer graphs, and provably show that different matrix means perform well under different settings in expectation. For instance, for the limit case $+\infty$ we show that matrix means effectively blend the information of a signed/multilayer graph when each layer provides global information of the clustering/class structure, whereas the limit case $-\infty$ is effective in at least two cases: a) when at least one layer conveys global information and the remaining layers are potentially just noise; b) when each layer provides only local information but taking all layers together one obtains global information of the clustering/class structure. All our provable results are extensively verified through numerical experiments.

Further, we perform experiments on real datasets and show that our approach is competitive to the state of the art. Furthermore, in signed networks our approach is the first one identifying explicit clustering structure in a real-world dataset where it was conjectured that there was no clustering structure.

Finally, we propose matrix-free numerical schemes showing that our proposed approaches based on matrix power means are scalable to large sparse signed and multilayer graphs.

We now present a summary of the contributions per chapter and the thesis' outline.

- **Background Material** (Ch. 2) In this chapter we give a brief overview of spectral clustering and provide a motivation in terms of graph cuts. Further, we briefly describe k-means. This constitutes the last step in the algorithmic description of spectral clustering.

  We further introduce in Section 2.3 a well-known one-parameter family of means called the scalar power means, which includes as particular cases the arithmetic, geometric and harmonic mean. Moreover, we introduce a matrix extension called the Matrix Power Means. The family of matrix power means are the fundamental tools that we use to propose extensions of spectral clustering and semi-supervised learning on signed and multilayer graphs. The following chapters present an analysis on the effectivity of matrix power means to merge the information encoded by multiple kinds of interactions under different contexts.

- **Spectral Clustering of Signed Graphs via Matrix Power Means** (Ch. 3).
  This chapter presents and extension of spectral clustering to signed graphs by introducing the Signed Power Mean Laplacian, which is the matrix power mean applied to the Laplacian of positive edges and the signless Laplacian of negative

edges. We present an analysis based on a suitable stochastic block model, and show that our proposed approach provably outperforms the state of the art in expectation. We further show that the Signed Power Mean Laplacian concentrates around its mean under the stochastic block model. Moreover, we show that our approach is competitive to the state of the art in real-world datasets. Finally, we present a matrix-free numerical scheme showing that our proposed approach is scalable to large sparse signed networks.

The content of this chapter corresponds to the ICML 2019 publication: "Spectral Clustering of Signed Graphs via Matrix Power Means" (Mercado *et al.*, 2019c).

- **Spectral Clustering of Multilayer Graphs via Matrix Power Means** (Ch. 4). Based on the insights from the previous chapter, we present an extension of spectral clustering to multilayer graphs via the Power Mean Laplacian obtained from the Laplacian of each layer. We present an analysis under the stochastic block model and show that our proposed approach outperforms the state of the art under three different regimes: robustness under the presence of noise-layers; the case when none of the layers contain full information of the clusters, but only if one considers them all together; and when clusters present fluctuations between layers. For the first two cases we present formal guarantees, whereas for the last one we present numerical experiments. We further provide experiments on real datasets and show that our approach performs no worse than the state of the art.

  The content of this chapter corresponds to the AISTATS 2018 publication: "The Power Mean Laplacian for Multilayer Graph Clustering" (Mercado *et al.*, 2018).

- **Semi-Supervised Learning on Multilayer Graphs via Matrix Power Means** (Ch. 5). In this chapter we consider the task of building a classifier taking both labeled and unlabeled observations, by considering a suitable loss function and the underlying multilayer graph structure of the observations. To induce the underlying clustering structure we propose the Power Mean Laplacian as a regularizer. We present an analysis under the stochastic block model and show that our proposed approach yields a good classification performance when at least one of the layers provides information from the class structure of the nodes. Moreover, we present a weighted loss function that provably recovers the classes in expectation when the number labeled nodes per class is different. We verify our findings with extensive numerical experiments and further show that our proposed approach is competitive to the state of the art in real datasets. Apart from that, we present a matrix-free numerical scheme showing that our method is scalable to large sparse graphs, outperforming the time execution of several state of the art approaches.

  The content of this chapter corresponds to the NeurIPS 2019 publication: "Generalized Matrix Means for Semi-Supervised Learning with Multilayer Graphs" (Mercado *et al.*, 2019b).

BACKGROUND MATERIAL 2

In this chapter we briefly introduce two of the main methods related to this work: spectral clustering and k-means. In this first part we present spectral clustering, which is a well-established, graph-based clustering method, and which we aim to extend to signed and multilayer graphs in the remainder chapters of this work. In the second section we briefly introduce k-means clustering, which corresponds to the last algorithmic step of spectral clustering.

2.1 SPECTRAL CLUSTERING

Spectral clustering is a well-established technique which has proven to be useful in the identification of sets of observations that present a similar behaviour. This is achieved by finding a partition of the sets of nodes such that nodes belonging to the same partition are highly similar. In this section we provide a brief introduction to spectral clustering, broadly following the influential work of (von Luxburg, 2007).

Spectral clustering, based on the first eigenvectors of the (normalized) graph Laplacian, first provides a k-dimensional embedding of the nodes of the corresponding graph and then applies k-means to return a partition of the set of nodes. The corresponding pseudo code is shown in Algorithm 1.

Algorithm 1: Spectral clustering

Input: Symmetric adjacency matrix W, number k of clusters to construct.
Output: Clusters $C_1, \ldots, C_k$.

1 Compute Laplacian matrix L_{sym}.
2 Compute eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ corresponding to the k smallest eigenvalues of L_{sym}.
3 Let $U = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$ be the matrix containing eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ as columns.
4 Cluster the rows of U with k-means into clusters $C_1, \ldots, C_k$.

Spectral clustering relies on the (normalized) graph Laplacian, which we now introduce. For a given graph $G = (V, W)$ with a set of nodes V and adjacency matrix W, the graph Laplacian and its normalized symmetric version are defined as

$$L = D - W\,, \qquad L_{\text{sym}} = D^{1/2} L D^{1/2} \tag{2.1}$$

where D is a diagonal matrix with $D_{ii} = d_i$, and $d_i = \sum_{j=1}^{n} w_{ij}$. Among the properties of the graph Laplacians L and L_{sym} are the following (von Luxburg, 2007):

- L and L_{sym} are symmetric and positive semi-definite,
- the smallest eigenvalue of L and L_{sym} is 0,

- the quadratic form of the Laplacian L is: $x'Lx = \frac{1}{2}\sum_{i,j=1}^{n} w_{ij}(x_i - x_j)^2$
- the multiplicity of the eigenvalue 0 of L and L_{sym} is equal to the number of connected components in the graph.

The intuition behind spectral clustering can be seen from different perspectives. In what follows we provide a brief notion on how spectral clustering can be seen as a relaxation of the normalized graph cut problem. We would like to mention that we not only consider binary adjacency adjacency matrices but as well general weighted graphs.

2.1.1 Spectral Clustering from the normalized graph cut perspective

Spectral clustering can be seen as a continuous relaxation of a discrete optimization problem in terms of graph cuts. In particular, the discrete optimization problem can be casted as the normalized graph cut problem as follows:

$$\min_{(C_1,\ldots,C_k)\in P_k} \text{Ncut}(C_1,\ldots,C_k) := \sum_{i=1}^{k} \frac{\text{cut}(C_i,\bar{C}_i)}{\text{vol}(C_i)}. \tag{2.2}$$

where P_k is the set of all k-partitions of the vertex set V, and

$$\text{cut}(C,\bar{C}) = \sum_{v_i\in C, v_j\in\bar{C}} w_{ij} \tag{2.3}$$

$$\text{vol}(C) = \sum_{v_i\in C} d_i \tag{2.4}$$

where $d_i = \sum_{j=1}^{n} w_{ij}$ Since the discrete optimization problem (2.2) is NP-hard, a continuous relaxation is suitable. In particular, the following formulation leads to spectral clustering with the normalized graph Laplacian L_{sym}:

$$\min_{H\in\mathbb{R}^{n\times k}} \text{Tr}(H'L_{\text{sym}}H) \text{ subject to } H'H = I. \tag{2.5}$$

where the columns of the minimizer matrix H correspond to the k-eigenvectors with the smallest eigenvalues of L_{sym} (Lütkepohl, 1996, Section 5.2.2.(6)).

We now verify that (2.5) indeed is a relaxation of (2.2). We first observe that by defining $T = D^{-1/2}H$ we obtain

$$\text{Tr}(H'L_{\text{sym}}H) = \text{Tr}(T'D^{1/2}D^{-1/2}LD^{-1/2}D^{1/2}T) = \text{Tr}(T'LT) \tag{2.6}$$

together with the fact $H'H = T'D^{1/2}D^{1/2}T = T'DT$. Hence the optimization problem (2.5) can now be expressed as

$$\min_{T\in\mathbb{R}^{n\times k}} \text{Tr}(T'LT) \text{ subject to } T'DT = I. \tag{2.7}$$

Observe that from the solution matrix T of (2.7) one can recover the solution matrix H of (2.5) by the relation $H = D^{1/2}T$. We now introduce a specific family of matrices. Let $T \in \mathbb{R}^{n\times k}$ be defined entrywise as

$$t_{ij} = \begin{cases} \frac{1}{\sqrt{\text{vol}(C_j)}} & \text{if } v_i \in C_j \\ 0 & \text{else} \end{cases} \tag{2.8}$$

where one can see that the columns of T are weighted indicator vectors. Let t_i denote the i^{th} column of matrix T. Then, one can easily verify that $t_i' D t_i = \delta_{ij}$ and

$$t_i' L t_i = \frac{\text{cut}(C_i, \bar{C}_i)}{\text{vol}(C_i)} \tag{2.9}$$

and hence

$$\text{Tr}(T'LT) = \sum_{i=1}^{k} t_i' L t_i = \sum_{i=1}^{k} \frac{\text{cut}(C_i, \bar{C}_i)}{\text{vol}(C_i)} = \text{Ncut}(C_1, \ldots, C_k) \tag{2.10}$$

Hence, the graph normalized cut problem (2.2) can be restated in terms of matrices of the form (2.8), namely

$$\begin{aligned} &\min_{(C_1,\ldots,C_k)\in P_k} \text{Tr}(T'LT) \\ &\text{subject to: } T'DT = I, \\ &\qquad\qquad\quad T \text{ as defined in (2.8).} \end{aligned}$$

To further wrap-up the current exposition, observe that by relaxing the search space from sets to matrices with real entries, and recalling that $T = D^{-1/2}H$, we get the continuous relaxation initially described in (2.5).

2.2 k-MEANS

From the previous Section 2.1 we have seen that the last step of spectral clustering (see Algorithm 1) relies on applying k-means to the matrix composed by certain eigenvectors of the corresponding Laplacian. In this section we provide a brief introduction to k-means.

The k-means method (traditionally believed that the name was first coined in (MacQueen, 1967)) is a clustering method that identifies sets of observations such that the distances between elements in the same cluster are smaller than those between elements belonging to different clusters. This is achieved by identifying certain prototypes per cluster (for instance, the mean among observations belonging to a certain cluster) such that all observations belonging to a cluster share its nearest prototype.

The optimization problem of k-means can be stated as follows

$$\underset{\substack{(C_1,\dots,C_k)\in P_k \\ \mu_1,\dots,\mu_k\in\mathbb{R}^D}}{\arg\min} \sum_{i=1}^{k} \sum_{x_j\in C_i} \left\|x_j-\mu_i\right\|^2 := J(k) \tag{2.11}$$

where $C_1,\dots,C_k$ are the final clusters, P_k is the set of all k-partitions of the given observations $x_1,\dots,x_n\in\mathbb{R}^D$, and $\mu_1,\dots,\mu_k\in\mathbb{R}^D$ are the corresponding prototypes. This definition of the k-means problem has been shown to be NP-Hard (Mahajan *et al.*, 2012; Aloise *et al.*, 2009). Hence approaches providing an approximate solution have been proposed, among them Lloyd's algorithm (Lloyd, 1982), which is described in Algorithm 2.

Algorithm 2: Lloyd's algorithm

Input: Observations $x_1,\dots,x_n\in\mathbb{R}^D$, number of clusters k to construct.
Output: Clusters $C_1,\dots,C_k$, and prototypes $\mu_1,\dots,\mu_k\in\mathbb{R}^D$

Initialize $\mu_1,\dots,\mu_k$.
while *J(k) has not converged* **do**
  // Update clusters
  for $i\leftarrow 1$ **to** n **do**
    Assign x_i to C_{j^*} if $j^*=\arg\min_j \left\|x_i-\mu_j\right\|$
  end
  // Update prototypes
  for $i\leftarrow 1$ **to** k **do**
    $\mu_i=\frac{1}{|C_i|}\sum_{x_j\in C_i} x_j$
  end
end

In practice it is recommended to execute Lloyd's algorithm several times with different starting prototypes, and choose the output that reached the best solution in terms of Eq. (2.11).

We briefly mention that Lloyd's algorithm consists of two alternating optimization problems, each of them leading to the corresponding updating rule for clusters and corresponding prototypes. For each updating rule we have the following observations:

- Update clusters: for fixed prototypes one can minimize $J(k)$ with respect to $C_1,\dots,C_k$, leading to the corresponding update rule for clusters, which is obtained from the minimization of the entry $\left\|x_i-\mu_j\right\|^2$ of $J(k)$, by identifying the nearest prototype of the corresponding observation,
- Update prototypes: for fixed clusters, one can minimize $J(k)$ with respect to $\mu_1,\dots,\mu_k$, leading to the corresponding update rule for prototypes. This is particularly clear since

$$\frac{1}{|C_i|}\sum_{x_j\in C_i} x_j = \underset{\mu_i\in\mathbb{R}^D}{\arg\min} \sum_{x_j\in C_i} \left\|x_j-\mu_i\right\|^2 \tag{2.12}$$

Aiming to improve the performance of k-means by considering suitable prototype initialization approaches, several techniques have been proposed, for instance, k-means++(Arthur and Vassilvitskii, 2007), one-shot coresets (Bachem *et al.*, 2018), and k-medoids (Newling and Fleuret, 2017).

2.3 POWER MEANS

This section introduces a general family of scalar means called scalar power means. We will show that scalar power means depends on one parameter, and that for particular values it yields well known means as the arithmetic, geometric and harmonic means. Based on this scalar family of means we will introduce a matrix extension called the matrix power means. This will be our main tool to extend spectral clustering and semi-supervised learning to signed and multilayer graphs.

2.3.1 Scalar Power Means

The scalar power mean of a set of non-negative scalars $x_1, \ldots, x_T$ is a general one-parameter family of means defined for $p \in \mathbb{R}$ as

$$m_p(x_1, \ldots, x_T) = \Big(\frac{1}{T}\sum_{i=1}^{T} x_i^p\Big)^{1/p}.$$

It is easy to see that the scalar power mean includes some well-known means as special cases, among them the harmonic, geometric and arithmetic means:

name	minimum	harmonic mean	geometric mean	arithmetic mean	maximum
p	$p \to -\infty$	$p = -1$	$p \to 0$	$p = 1$	$p \to \infty$
$m_p(a,b)$	$\min\{a,b\}$	$2\big(\frac{1}{a}+\frac{1}{b}\big)^{-1}$	$\sqrt{ab}$	$(a+b)/2$	$\max\{a,b\}$

Table 2.1: Particular cases of scalar power means

While the arithmetic mean is a well-known mean (Lovric, 2011, see p.788-791), the task of choosing a suitable mean in a given context is not trivial, as stated in (United Nations Development Programme, 1997, see p.117-121):

> There is an inescapable arbitrariness in the choice of p. The right way to deal with this issue is to explain clearly what is being assumed [...]

Several examples of applying different power means are available. For instance, the geometric mean ($p \to 0$) is used in the financial context to estimate the average return of an investment based on compound interest over time, whereas the harmonic mean ($p = -1$) is used to compare different indexes composed by multiple price-earning ratios. The harmonic mean is preferred in this setting partly because it avoids overestimations that are commonly seen with the arithmetic mean (Agrrawal *et al.*, 2010). For further discussions on when the harmonic mean is preferred over

the arithmetic mean we refer to (Ferger, 1931; Hand, 1994; Haans, 2008). Further, the power mean with parameter $p = 3$ is used by the United Nations Development Program to calculate the Gender-Related Development Index (GDI), as it "places greater weight on those dimensions in which deprivation is larger" (United Nations Development Programme, 1997, see p.117-121).

The scalar power mean is monotone in the parameter p as stated by the following Theorem.

Theorem 2.1 ((Hardy *et al.*, 1934, Theorem 16), (Bullen, 2013, Ch. 3, Theorem 1))**.** *Let $p < q$ then $m_p(a,b) \leq m_q(a,b)$ with equality if and only if $a = b$.*

From this Theorem we can see that for powers $p \in \{-1, 0, 1\}$ yields the well-known harmonic-geometric-arithmetic mean inequality:

$$m_{-1}(a,b) \leq m_0(a,b) \leq m_1(a,b)\,.$$

This inequality has been widely studied. For instance, there are well over 70 proofs of the arithmetic-geometric mean inequality in (Bullen, 2013).

2.3.2 Matrix Power Means

Since matrices do not commute, the scalar power mean can be extended to positive definite matrices in a number of different ways, all of them coinciding when applied to commuting matrices (Bhatia, 2009, Chapter 4). In this work we use the following matrix power mean.

Definition 2.1 ((Bhagwat and Subramanian, 1978))**.** *Let $A_1, \ldots, A_T$ be symmetric positive definite matrices, and $p \in \mathbb{R}$. The matrix power mean of $A_1, \ldots, A_T$ with exponent p is*

$$M_p(A_1, \ldots, A_T) = \left(\frac{1}{T} \sum_{i=1}^{T} A_i^p \right)^{1/p} \tag{2.13}$$

where $A^{1/p}$ is the unique positive definite solution of the matrix Equation $X^p = A$.

The previous definition can be extended to positive semi-definite matrices. For $p > 0$, $M_p(A_1, \ldots, A_T)$ exists for positive semi-definite matrices, whereas for $p \leq 0$ it is necessary to add a suitable diagonal shift to $A_1, \ldots, A_T$ to enforce them to be positive definite (see (Bhagwat and Subramanian, 1978) for details).

We call the matrix above *matrix power mean* and we recover for $p = 1$ the standard arithmetic mean of the matrices. Note that for $p \to 0$, the power mean (2.13) converges to the Log-Euclidean matrix mean (Bhagwat and Subramanian, 1978; Arsigny *et al.*, 2007)

$$M_0(A_1, \ldots, A_T) = \exp\left(\frac{1}{T} \sum_{i=1}^{T} \log A_i \right)\,, \tag{2.14}$$

which is a popular form of matrix geometric mean used, for instance, in diffusion tensor imaging or quantum information theory (see (Arsigny *et al.*, 2006; Petz, 2007)).

Based on the Karcher mean, a different one-parameter family of matrix power means has been discussed for instance in (Lim and Pálfia, 2012). When the parameter goes to zero, the Karcher-based power mean of two matrices A and B converges to the geometric mean

$$A\#B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{-1/2}$$

The mean $A\#B$ has been used for instance for clustering in signed networks (Fasi and Iannazzo, 2018; Mercado *et al.*, 2016), for metric learning (Zadeh *et al.*, 2016) and for geometric optimization (Sra and Hosseini, 2016). However, when more than two matrices are considered, the Karcher-based power mean is defined as the solution of a set of nonlinear matrix equations without any known closed-form solution (Bini and Iannazzo, 2011). For an overview of matrix means we refer the reader to (Bhatia, 2009, Chapter 4).

In the following sections we will present several extensions of spectral clustering and semi-supervised learning to signed and multilayer graph. Recall that spectral clustering is a graph-based method relying on the spectrum of certain graph operators. Hence, we now introduce a result that gives a first glimpse into the spectrum of the matrix power means, showing the effect of the matrix power mean when the given matrices have a common eigenvector.

Lemma 2.1. *Let* $\mathbf{u}$ *be an eigenvector of* $A_1, \dots, A_T$ *with corresponding eigenvalues* $\lambda_1, \dots, \lambda_T$. *Then* $\mathbf{u}$ *is an eigenvector of* $M_p(A_1, \dots, A_T)$ *with eigenvalue* $m_p(\lambda_1, \dots, \lambda_T)$.

Proof. Observe that for any positive definite matrix M, if $Mx = \lambda(M)x$, then $M^p = \lambda(M)^p x$. Thus, we can see that as $A_i u = \lambda_i u$ for $i = 1, \dots, T$. then, $A_i^p u = \lambda_i^p u$. Hence,

$$M_p^p(A_1, \dots, A_T)u = \left(\frac{1}{T}\sum_{i=1}^{T} A_i^p\right) u = \left(\frac{1}{T}\sum_{i=1}^{T} \lambda_i^p\right) u = m_p^p(\lambda_1, \dots, \lambda_T)u$$

Thus u is an eigenvector of $M_p(A_1, \dots, A_T)$ with eigenvalue $m_p(\lambda_1, \dots, \lambda_T)$. □

3 SPECTRAL CLUSTERING OF SIGNED GRAPHS VIA MATRIX POWER MEANS

In this chapter we extend spectral clustering to signed graphs via the one-parameter family of *Signed Power Mean Laplacians*, defined as the matrix power mean of normalized standard and signless Laplacians of positive and negative edges. We provide a thorough analysis of the proposed approach in the setting of a general Stochastic Block Model that includes models such as the Labeled Stochastic Block Model and the Censored Block Model. We show that in expectation the signed power mean Laplacian captures the ground truth clusters under reasonable settings where state-of-the-art approaches fail. Moreover, we prove that the eigenvalues and eigenvectors of the signed power mean Laplacian concentrate around their expectation under reasonable conditions in the general Stochastic Block Model. Extensive experiments on random graphs and real-world datasets confirm the theoretically predicted behavior of the signed power mean Laplacian and show that it compares favorably with state-of-the-art methods.

3.1 INTRODUCTION

The analysis of graphs has received a significant amount of attention due to their capability to encode interactions that naturally arise in social networks. Yet, the vast majority of graph methods focuses on the case where interactions are of the same type, leaving aside the case where different kinds of interactions are available (Leskovec *et al.*, 2010b). Graphs and networks with both positive and negative edge weights arise naturally in a number of social, biological and economic contexts. Social dynamics and relationships are intrinsically positive and negative: users of online social networks such as Slashdot and Epinions, for example, can express positive interactions, like friendship and trust, and negative ones, like enmity and distrust. Other important application settings are the analysis of gene expressions in biology (Fujita *et al.*, 2012) or the analysis of financial and economic time sequences (Ziegler *et al.*, 2010; Pavlidis *et al.*, 2006), where similarity and variable dependence measures commonly used may attain both positive and negative values (e.g. the Pearson correlation coefficient).

As briefly stated in Subsection 1.1.1 the analysis of signed graphs can be traced back to social balance theory (Cartwright and Harary, 1956; Harary, 1953; Davis, 1967) where the concept of a k-balance signed graph is introduced. The analysis of signed networks has been then pushed forward through the study of a variety of tasks in signed graphs, as for example edge prediction (Kumar *et al.*, 2016; Leskovec *et al.*, 2010a; Falher *et al.*, 2017), node classification (Mercado *et al.*, 2019a; Tang *et al.*,

2016a), node embeddings (Chiang *et al.*, 2011; Derr *et al.*, 2018; Kim *et al.*, 2018; Wang *et al.*, 2017; Yuan *et al.*, 2017), node ranking (Chung *et al.*, 2013; Shahriari and Jalili, 2014), and clustering (Chiang *et al.*, 2012; Kunegis *et al.*, 2010; Mercado *et al.*, 2016; Sedoc *et al.*, 2017; Doreian and Mrvar, 2009; Knyazev, 2018; Kirkley *et al.*, 2019; Cucuringu *et al.*, 2019; Cucuringu *et al.*, 2018). For recent surveys on the topic see (Tang *et al.*, 2016b; Gallier, 2016) .

In this chapter we present a novel extension of spectral clustering for signed graphs. For a brief introduction to spectral clustering please see Section 2.1. We introduce the family of *Signed Power Mean (SPM) Laplacians*: a one-parameter family of graph matrices for signed graphs that blends the information from positive and negative interactions through the matrix power mean, a general class of matrix means that contains the arithmetic, geometric, and harmonic mean as special cases. The family of Signed Power Mean Laplacians is inspired by recent extensions of spectral clustering which merge the information encoded by positive and negative interactions through different types of arithmetic (Chiang *et al.*, 2012; Kunegis *et al.*, 2010) and geometric (Mercado *et al.*, 2016) means of the standard and signless graph Laplacians.

We analyze the performance of the signed power mean Laplacian in a general Signed Stochastic Block Model. We first provide an analysis in expectation showing that the smaller, the parameter of the signed power mean Laplacian, the less restrictive are the conditions that ensure to recover the ground truth clusters. In particular, we show that the limit cases $+\infty$ and $-\infty$ are related to the boolean operators AND and OR, respectively, in the sense that for the limit case $+\infty$ clusters are recovered only if both positive *and* negative interactions are informative, whereas for $-\infty$ clusters are recovered if positive *or* negative interactions are informative. Second, we show that the eigenvalues and eigenvectors of the signed power mean Laplacian concentrate around their mean, so that our results hold also for the case where one samples from the stochastic block model. To our knowledge these are the first concentration results for matrix power means under any stochastic block model for signed graphs.

Finally, we show that the signed power mean Laplacian compares favorably with state-of-the-art approaches through extensive numerical experiments on diverse real world datasets.

Notation. A signed graph is a pair $G^{\pm} = (G^+, G^-)$, where $G^+ = (V, W^+)$ and $G^- = (V, W^-)$ encode positive and negative edges, respectively, with positive symmetric adjacency matrices W^+ and W^-, and a common vertex set $V = \{v_1, \dots, v_n\}$. Note that this definition allows the simultaneous presence of both positive and negative interactions between the same two nodes. This is a major difference with respect to the alternative point of view where $G^{\pm}$ is associated to a single symmetric matrix W with positive and negative entries. In this case $W = W^+ - W^-$, with $W^+_{ij} = \max\{0, W_{ij}\}$ and $W^-_{ij} = -\min\{0, W_{ij}\}$, implying that every interaction is either positive or negative, but not both at the same time. We denote by $D^+_{ii} = \sum_{j=1}^n w^+_{ij}$ and $D^-_{ii} = \sum_{j=1}^n w^-_{ij}$ the diagonal matrix of the degrees of G^+ and G^-, respectively, and $\bar{D} = D^+ + D^-$.

3.2 RELATED WORK

The study of clustering of signed graphs can be traced back to the theory of social balance (Cartwright and Harary, 1956; Harary, 1953; Davis, 1967), where a signed graph is called k-balanced if the set of vertices can be partitioned into k sets such that within the subsets there are only positive edges, and between them only negative.

Inspired by the notion of k-balance, different approaches for signed graph clustering have been introduced. In particular, many of them aim to extend spectral clustering to signed graphs by proposing novel signed graph Laplacians. A related approach is correlation clustering (Bansal *et al.*, 2004). Unlike spectral clustering, where the number of clusters is fixed a-priori, correlation clustering approximates the optimal number of clusters by identifying a partition that is as close as possible to be k-balanced. In this setting, the case where the number of clusters is constrained has been considered in (Giotis and Guruswami, 2006).

We briefly introduce the standard and signless Laplacian and review different definitions of Laplacians on signed graphs. The final clustering algorithm to find k clusters is the same for all of them: compute the smallest k eigenvectors of the corresponding Laplacian, use the eigenvectors to embed the nodes into $\mathbb{R}^k$, obtain the final clustering by doing k-means in the embedding space. However, we will see below that in some cases we have to slightly deviate from this generic principle by using the $k-1$ smallest eigenvectors instead.

Laplacians of Unsigned Graphs: In the following all weight matrices are non-negative and symmetric. Given an assortative graph $G = (V, W)$, standard spectral clustering is based on the Laplacian and its normalized version defined as:

$$L = D - W \qquad L_{\text{sym}} = D^{-1/2} L D^{-1/2} \tag{3.1}$$

where $D_{ii} = \sum_{j=1}^{n} w_{ij}$ is the diagonal matrix of the degrees of G. Both Laplacians are symmetric positive semidefinite and the multiplicity of the eigenvalue 0 is equal to the number of connected components in G. For a more detailed introduction please see Section 2.1.

For disassortative graphs, i.e. when edges carry only dissimilarity information, the goal is to identify clusters such that the amount of edges between clusters is larger than the one inside clusters. Spectral clustering is extended to this setting by considering the signless Laplacian matrix and its normalized version (see e.g. Liu (2015); Mercado *et al.* (2016)), defined as:

$$Q = D + W \qquad Q_{\text{sym}} = D^{-1/2} Q D^{-1/2} \tag{3.2}$$

Both Laplacians are positive semi-definite, and the smallest eigenvalue is zero if and only if the graph has a bipartite component (Desai and Rao, 1994).

Laplacians of Signed Graphs: Signed graphs encode both positive and negative interactions. In the ideal k-balanced case positive interactions present an assortative behavior, whereas negative interactions present a disassortative behavior. With this in mind, several novel definitions of *signed Laplacians* have been proposed. We briefly review them for later reference.

In (Chiang *et al.*, 2012) the balance ratio Laplacian and its normalized version are defined as:

$$L_{BR} = D^+ - W^+ + W^-, \quad L_{BN} = \bar{D}^{-1/2} L_{BR} \bar{D}^{-1/2} \tag{3.3}$$

whereas in (Kunegis *et al.*, 2010) the signed ratio Laplacian and its normalized version have been defined as:

$$L_{SR} = \bar{D} - W^+ + W^-, \quad L_{SN} = \bar{D}^{-1/2} L_{SR} \bar{D}^{-1/2} \tag{3.4}$$

The signed Laplacians L_{BR} and L_{BN} need not be positive semidefinite, while the signed Laplacians L_{SR} and L_{SN} are positive semidefinite with eigenvalue zero if and only if the graph is 2-balanced.

In the context of correlation clustering, in (Saade *et al.*, 2015) the Bethe Hessian matrix is defined as:

$$H = (\alpha - 1)I - \sqrt{\alpha}(W^+ - W^-) + \overline{D} \tag{3.5}$$

where α is the average node degree $\alpha = \frac{1}{n}\sum_{i=1}^{n} \overline{D}_{ii}$. The Bethe Hessian H need not be positive definite. In fact, eigenvectors with negative eigenvalues bring information of clustering structure (Saade *et al.*, 2014).

Let $L^+ = D^+ - W^+$ and $Q^- = D^- + W^-$ be the Laplacian and signless Laplacian of G^+ and G^-, respectively. As noted in (Mercado *et al.*, 2016), $L_{SR} = L^+ + Q^-$ i.e. it coincides with twice the arithmetic mean of L^+ and Q^-. Note that the same holds for H when the average degree α is equal to one, i.e. $H = L_{SR}$ when $\alpha = 1$. In (Mercado *et al.*, 2016), the arithmetic mean and geometric mean of the normalized Laplacian and its signless version are used to define new Laplacians for signed graphs:

$$L_{AM} = L^+_{\text{sym}} + Q^-_{\text{sym}}, \quad L_{GM} = L^+_{\text{sym}} \# Q^-_{\text{sym}} \tag{3.6}$$

where $A\#B = A^{-1/2}(A^{1/2}BA^{1/2})^{1/2}A^{-1/2}$ is the geometric mean of A and B, with $L^+_{\text{sym}} = (D^+)^{-1/2}L^+(D^+)^{-1/2}$ and $Q^-_{\text{sym}} = (D^-)^{-1/2}Q^-(D^-)^{-1/2}$. While the computation of L_{GM} is more challenging, (Mercado *et al.*, 2016) have shown that the clustering assignment obtained with the geometric mean Laplacian L_{GM} outperforms all other signed Laplacians.

Both the arithmetic and the geometric means are special cases of a much richer one-parameter family of means known as power means. Based on this observation, we introduce the *Signed Power Mean Laplacian* in Section 3.3, defined via the matrix version of the family of power means which we briefly reviewed in Section 2.3.

3.3 THE SIGNED POWER MEAN LAPLACIAN

Given a signed graph $G^{\pm} = (G^+, G^-)$ we define the Signed Power Mean (SPM) Laplacian L_p of $G^{\pm}$ as

$$L_p = M_p(L^+_{\text{sym}}, Q^-_{\text{sym}}) = \left(\frac{(L^+_{\text{sym}})^p + (Q^-_{\text{sym}})^p}{2} \right)^{1/p}. \tag{3.7}$$

Algorithm 3: Spectral clustering of signed graphs with L_p

Input: Symmetric matrices W^+, W^-, number k of clusters to construct.
Output: Clusters $C_1, \dots, C_k$.
1 Let $k' = k - 1$ if $p \geq 1$ and $k' = k$ if $p < 1$.
2 Compute eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_{k'}$ corresponding to the k' smallest eigenvalues of L_p.
3 Set $U = (\mathbf{u}_1, \dots, \mathbf{u}_{k'})$ and cluster the rows of U with k-means into clusters $C_1, \dots, C_k$.

For the case $p < 0$ the matrix power mean requires positive definite matrices, hence we use in this case the matrix power mean of diagonally shifted Laplacians, i.e. $L^+_{\text{sym}} + \varepsilon I$ and $Q^-_{\text{sym}} + \varepsilon I$. Our theoretical analysis that is following holds for all possible shifts $\varepsilon > 0$, whereas in Section 3.6.3 we discuss the numerical robustness with respect to ε. The clustering algorithm for identifying k clusters in signed graphs is given in Algorithm 3. Please note that for $p \geq 1$ we deviate from the usual scheme and use the first $k - 1$ eigenvectors rather than the first k. The reason is a result of the analysis in the stochastic block model in Section 3.4. In general, the main influence of the parameter p of the power mean is on the ordering of the eigenvalues. In Section 3.4 we will see that this effect on the ordering of eigenvalues significantly influences the performance of different instances of SPM Laplacians, in particular, the arithmetic and geometric mean discussed in (Mercado *et al.*, 2016) are suboptimal for the recovery of the ground truth clusters. For the computation of the matrix power mean we adapt the scalable Krylov subspace-based algorithm proposed in (Mercado *et al.*, 2018).

3.4 STOCHASTIC BLOCK MODEL ANALYSIS

In this section we analyze the signed power mean Laplacian L_p under a general Signed Stochastic Block Model. Our results here are twofold. First, we derive new conditions in expectation that guarantee that the eigenvectors corresponding to the smallest eigenvalues of L_p recover the ground truth clusters. These conditions reveal that, in this setting, the state-of-the-art signed graph matrices are suboptimal as compared to L_p for negative values of p. Second, we show that our result in expectation transfers to sampled graphs as we prove conditions that ensure that both eigenvalues and eigenvectors of L_p concentrate around their expected value with high probability. We verify our results by several experiments where the clustering performance of state-of-the-art matrices and L_p are compared on random graphs following the Signed Stochastic Block Model. All proofs hold for an arbitrary diagonal shift $\varepsilon > 0$, whereas the shift is set to $\varepsilon = \log_{10}(1 + |p|) + 10^{-6}$ in the numerical experiments.

The Stochastic Block Model (**SBM**) is a well-established generative model for graphs and a canonical tool for studying clustering methods (Holland *et al.*, 1983; Rohe *et al.*, 2011; Abbe, 2018). Graphs drawn from the SBM show a prescribed clustering structure, as the probability of an edge between two nodes depends only on the clustering membership of each node. We introduce our SBM for signed Graphs

(**SSBM**): we consider k ground truth clusters $\mathcal{C}_1, \dots, \mathcal{C}_k$, all of them of size $|\mathcal{C}| = \frac{n}{k}$, and parameters $p^+_{\text{in}}, p^+_{\text{out}}, p^-_{\text{in}}, p^-_{\text{out}} \in [0,1]$ where p^+_{in} (resp. p^-_{in}) is the probability of observing an edge inside clusters in G^+ (resp. G^-) and p^+_{out} (resp. p^-_{out}) is the probability of observing an edge between clusters in G^+ (resp. G^-). Calligraphic letters are used for the expected adjacency matrices: $\mathcal{W}^+$ and $\mathcal{W}^-$ are the expected adjacency matrix of G^+ and G^-, respectively, where $\mathcal{W}^+_{i,j} = p^+_{\text{in}}$ and $\mathcal{W}^-_{i,j} = p^-_{\text{in}}$ if v_i, v_j belong to the same cluster, whereas $\mathcal{W}^+_{i,j} = p^+_{\text{out}}$ and $\mathcal{W}^-_{i,j} = p^-_{\text{out}}$ if v_i, v_j belong to different clusters.

Other extensions of the SBM to the signed setting have been considered. Particularly relevant examples are the Labelled Stochastic Block Model (**LSBM**) (Heimlicher *et al.*, 2012) and the Censored Block Model (**CBM**) (Abbe *et al.*, 2014). In the context of signed graphs, both LSBM and CBM assume that an observed edge can be either positive or negative, but not both. Our SSBM, instead, allows the simultaneous presence of both positive and negative edges between the same pair of nodes, as the parameters $p^+_{\text{in}}, p^+_{\text{out}}, p^-_{\text{in}}, p^-_{\text{out}}$ in SSBM are independent. Moreover, the edge probabilities defining both the LSBM and the CBM can be recovered as special cases of the SSBM. In particular, the LSBM corresponds to the SSBM for the choices

$$\begin{aligned} p^+_{\text{in}} &= \overline{p}_{\text{in}}\mu^+, \quad p^-_{\text{in}} = \overline{p}_{\text{in}}\mu^- && \text{(within clusters)} \\ p^+_{\text{out}} &= \overline{p}_{\text{out}}\nu^+, \; p^-_{\text{out}} = \overline{p}_{\text{out}}\nu^- && \text{(between clusters)} \end{aligned}$$

where $\overline{p}_{\text{in}}$ and $\overline{p}_{\text{out}}$ are edge probabilities within and between clusters, respectively, whereas μ^+ and $\mu^- = 1 - \mu^+$ (resp. ν^+ and $\nu^- = 1 - \nu^+$) are the probabilities of assigning a positive and negative label to an edge within (resp. between) clusters. Similarly, the CBM corresponds to the SSBM for the particular choices $\overline{p}_{\text{in}} = \overline{p}_{\text{out}}$, $\mu^+ = \nu^- = (1-\eta)$ and $\mu^- = \nu^+ = \eta$ where η is a noise parameter.

3.4.1 SBM Analysis in Expectation

In this section our goal is to identify conditions in expectation in terms of k, $p^+_{\text{in}}, p^+_{\text{out}}, p^-_{\text{in}}$, and p^-_{out}, such that $\mathcal{C}_1, \dots, \mathcal{C}_k$ are recovered by the smallest eigenvectors of the signed power mean Laplacian. Consider the following k vectors:

$$\boldsymbol{\chi}_1 = \mathbf{1}, \qquad \boldsymbol{\chi}_i = (k-1)\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}_{\overline{\mathcal{C}_i}} \; .$$

$i = 2, \dots, k$. The node embedding given by $\{\boldsymbol{\chi}_i\}_{i=1}^k$ is informative in the sense that applying k-means on $\{\boldsymbol{\chi}_i\}_{i=1}^k$ trivially recovers the ground truth clusters $\mathcal{C}_1, \dots, \mathcal{C}_k$ as all nodes of a cluster are mapped to the same point. Note that the constant vector $\boldsymbol{\chi}_1$ could be omitted as it does not add clustering information. We derive conditions for the SSBM such that $\{\boldsymbol{\chi}_i\}_{i=1}^k$ are the smallest eigenvectors of the signed power mean Laplacian in expectation. er

Theorem 3.1. *Let $\mathcal{L}_p = M_p(\mathcal{L}^+_{\text{sym}}, \mathcal{Q}^-_{\text{sym}})$ and let $\varepsilon > 0$ be the diagonal shift.*

- *If $p \geq 1$, then $\{\boldsymbol{\chi}_i\}_{i=2}^k$ correspond to the $(k\text{-}1)$-smallest eigenvalues of $\mathcal{L}_p$ if and only if $m_p(\rho^+_\varepsilon, \rho^-_\varepsilon) < 1 + \varepsilon$;*

- *If $p < 1$, then $\{\chi_i\}_{i=1}^k$ correspond to the k-smallest eigenvalues of $\mathcal{L}_p$ if and only if $m_p(\rho_\varepsilon^+, \rho_\varepsilon^-) < 1 + \varepsilon$;*

with $\rho_\varepsilon^+ = 1 - (p_{\text{in}}^+ - p_{\text{out}}^+)/(p_{\text{in}}^+ + (k-1)p_{\text{out}}^+) + \varepsilon$ and $\rho_\varepsilon^- = 1 + (p_{\text{in}}^- - p_{\text{out}}^-)/(p_{\text{in}}^- + (k-1)p_{\text{out}}^-) + \varepsilon$.

Proof. We first show that $\chi_1, \ldots, \chi_k$ are eigenvectors of $\mathcal{W}^+$ and $\mathcal{W}^-$. For χ_1 we have,

$$\mathcal{W}^+\chi_1 = \mathcal{W}^+\mathbf{1} = |\mathcal{C}|(p_{\text{in}}^+ + (k-1)p_{\text{out}}^+)\mathbf{1} = d^+\mathbf{1} = \lambda_1^+\mathbf{1}$$

For the remaining vectors $\chi_2, \ldots, \chi_k$ we have

$$\begin{aligned}
\mathcal{W}^+\chi_i &= \mathcal{W}^+\big((k-1)\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}_{\overline{\mathcal{C}_i}}\big) \\
&= \mathcal{W}^+\big(k\mathbf{1}_{\mathcal{C}_i} - (\mathbf{1}_{\mathcal{C}_i} + \mathbf{1}_{\overline{\mathcal{C}_i}})\big) \\
&= \mathcal{W}^+\big(k\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}\big) \\
&= k|\mathcal{C}|(p_{\text{in}}^+\mathbf{1}_{\mathcal{C}_i} + p_{\text{out}}^+\mathbf{1}_{\overline{\mathcal{C}_i}}) - d^+\mathbf{1} \\
&= k|\mathcal{C}|(p_{\text{in}}^+\mathbf{1}_{\mathcal{C}_i} + p_{\text{out}}^+\mathbf{1}_{\overline{\mathcal{C}_i}}) - d^+(\mathbf{1}_{\mathcal{C}_i} + \mathbf{1}_{\overline{\mathcal{C}_i}}) \\
&= |\mathcal{C}|(kp_{\text{in}}^+ - d^+)\mathbf{1}_{\mathcal{C}_i} + |\mathcal{C}|(kp_{\text{out}}^+ - d^+)\mathbf{1}_{\overline{\mathcal{C}_i}} \\
&= |\mathcal{C}|(k-1)(p_{\text{in}}^+ - p_{\text{out}}^+)\mathbf{1}_{\mathcal{C}_i} - |\mathcal{C}|(p_{\text{in}}^+ - p_{\text{out}}^+)\mathbf{1}_{\overline{\mathcal{C}_i}} \\
&= |\mathcal{C}|(p_{\text{in}}^+ - p_{\text{out}}^+)\big((k-1)\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}_{\overline{\mathcal{C}_i}}\big) \\
&= |\mathcal{C}|(p_{\text{in}}^+ - p_{\text{out}}^+)\chi_i \\
&= \lambda_i\chi_i
\end{aligned}$$

The same procedure holds for $\mathcal{W}^-$. Thus, we have shown that $\chi_1, \ldots, \chi_k$ are eigenvectors of both $\mathcal{W}^+$ and $\mathcal{W}^-$. In particular, we have seen that

$$\begin{aligned}
\lambda_1^+ &= |\mathcal{C}|(p_{\text{in}}^+ + (k-1)p_{\text{out}}^+), \quad \lambda_i^+ = |\mathcal{C}|(p_{\text{in}}^+ - p_{\text{out}}^+) \\
\lambda_1^- &= |\mathcal{C}|(p_{\text{in}}^- + (k-1)p_{\text{out}}^-), \quad \lambda_i^- = |\mathcal{C}|(p_{\text{in}}^- - p_{\text{out}}^-)
\end{aligned}$$

for $i = 2, \ldots, k$. Further, as both matrices $\mathcal{W}^+$ and $\mathcal{W}^-$ share all their eigenvectors, they are simultaneously diagonalizable, that is there exists a non-singular matrix Σ such that $\Sigma^{-1}\mathcal{W}^\pm\Sigma = \Lambda^\pm$, where Λ^+ and Λ^- are diagonal matrices $\Lambda^\pm = \text{diag}(\lambda_1^\pm, \ldots, \lambda_k^\pm, 0, \ldots, 0)$.

As we assume that all clusters are of the same size $|C|$, the expected signed graph is a regular graph with degrees d^+ and d^-. Hence, the normalized Laplacian and normalized signless Laplacian of the expected signed graph can be expressed as

$$\begin{aligned}
\mathcal{L}_{\text{sym}}^+ &= \Sigma(I - \frac{1}{d^+}\Lambda^+)\Sigma^{-1} \\
\mathcal{Q}_{\text{sym}}^- &= \Sigma(I + \frac{1}{d^-}\Lambda^-)\Sigma^{-1}
\end{aligned}$$

Thus, we can observe that

$$\begin{aligned}
\lambda_1^+(\mathcal{L}_{\text{sym}}^+) &= 0, & \lambda_1^-(\mathcal{Q}_{\text{sym}}^-) &= 2 \\
\lambda_i^+(\mathcal{L}_{\text{sym}}^+) &= 1-\rho^+, & \lambda_i^-(\mathcal{Q}_{\text{sym}}^-) &= 1+\rho^- \\
\lambda_j^+(\mathcal{L}_{\text{sym}}^+) &= 1, & \lambda_j^-(\mathcal{Q}_{\text{sym}}^-) &= 1
\end{aligned}$$

for $i = 2, \ldots, k$, and $j = k+1, \ldots, |V|$, where

$$\begin{aligned}
\rho^+ &= (p_{\text{in}}^+ - p_{\text{out}}^+)/(p_{\text{in}}^+ + (k-1)p_{\text{out}}^+) \\
\rho^- &= (p_{\text{in}}^- - p_{\text{out}}^-)/(p_{\text{in}}^- + (k-1)p_{\text{out}}^-)
\end{aligned}$$

By obtaining the signed power mean Laplacian on diagonally shifted matrices,

$$\mathcal{L}_p = M_p(\mathcal{L}_{\text{sym}}^+ + \varepsilon I, \mathcal{Q}_{\text{sym}}^- + \varepsilon I)$$

we have by Lemma 2.1

$$\begin{aligned}
\lambda_1(\mathcal{L}_p) &= m_p(\lambda_1^+ + \varepsilon, \lambda_1^- + \varepsilon) = m_p(\varepsilon, 2+\varepsilon) \\
\lambda_i(\mathcal{L}_p) &= m_p(1-\rho^+ + \varepsilon, 1+\rho^- + \varepsilon) \\
\lambda_j(\mathcal{L}_p) &= m_p(\lambda_j^+ + \varepsilon, \lambda_j^- + \varepsilon) = 1+\varepsilon
\end{aligned} \tag{3.8}$$

Observe that $\lambda_j(\mathcal{L}_p)$, with $j = k+1, \ldots, |V|$, corresponds to eigenvectors that do not yield an informative embedding. Hence, we do not want this eigenvalue to belong to the bottom of the spectrum of $\mathcal{L}_p$. Thus, for the case of $\chi_2, \ldots, \chi_k$, we can see that they will be located at the bottom of the spectrum if the following condition holds:

$$\lambda_i(\mathcal{L}_p) = m_p(1-\rho^+ + \varepsilon, 1+\rho^- + \varepsilon) = m_p(\rho_\varepsilon^+, \rho_\varepsilon^-) < 1+\varepsilon = \lambda_j(\mathcal{L}_p)$$

It remains to analyze the case of the constant eigenvector χ_1. Note that its associated eigenvalue $\lambda_1(\mathcal{L}_1)$ has the following relationship to the non-informative eigenvectors:

$$\lambda_1(\mathcal{L}_1) = m_1(\varepsilon, 2+\varepsilon) = 1+\varepsilon = \lambda_j(\mathcal{L}_p)$$

By Theorem 2.1 we know that the scalar power mean is monotone in its parameter p, and thus, for the case $p < 1$ we observe

$$\lambda_1(\mathcal{L}_p) = m_p(\varepsilon, 2+\varepsilon) < m_1(\varepsilon, 2+\varepsilon) = \lambda_1(\mathcal{L}_1) = \lambda_j(\mathcal{L}_p)$$

and for the case $p \geq 1$ we observe

$$\lambda_1(\mathcal{L}_p) = m_p(\varepsilon, 2+\varepsilon) \geq m_1(\varepsilon, 2+\varepsilon) = \lambda_1(\mathcal{L}_1) = \lambda_j(\mathcal{L}_p)$$

This means that for positive powers $p \geq 1$, the constant eigenvector χ_1 does not belong to the bottom of the spectrum, whereas for $p < 1$ it always does. With this in mind, we reach the desired result. □

Note that Theorem 3.1 is the reason why Alg. 3 uses only the first $k-1$ eigenvectors for $p \geq 1$. The problem is that the constant eigenvector need not be among the first k eigenvectors in the SSBM for $p \geq 1$. However, as it is constant and thus uninformative in the embedding, this does not lead to any loss of information. The following Corollary shows that the limit cases of L_p are related to the boolean operators AND and OR.

Corollary 3.1. *Let* $\mathcal{L}_p = M_p(\mathcal{L}^+_{\text{sym}}, \mathcal{Q}^-_{\text{sym}})$.

- $\{\chi_i\}_{i=2}^k$ *correspond to the* $(k\text{-}1)$*-smallest eigenvalues of* $\mathcal{L}_\infty$ *iff* $p^+_{\text{in}} > p^+_{\text{out}}$ ***and*** $p^-_{\text{in}} < p^-_{\text{out}}$,
- $\{\chi_i\}_{i=1}^k$ *correspond to the k-smallest eigenvalues of* $\mathcal{L}_{-\infty}$ *iff* $p^+_{\text{in}} > p^+_{\text{out}}$ ***or*** $p^-_{\text{in}} < p^-_{\text{out}}$.

Proof. Following the proof from Theorem 3.1, we can observe that $\lim_{p\to\infty} m_p(x) = \max\{x_1, \ldots, x_T\}$ and $\lim_{p\to-\infty} m_p(x) = \min\{x_1, \ldots, x_T\}$.

Thus, $m_\infty(\rho^+_\varepsilon, \rho^-_\varepsilon) = \max(1-\rho^+ + \varepsilon, 1+\rho^- + \varepsilon)$, and hence $m_\infty(\rho^+_\varepsilon, \rho^-_\varepsilon) < 1+\varepsilon$ if and only if $\rho^+ > 0$ and $\rho^- < 0$, yielding the desired conditions.

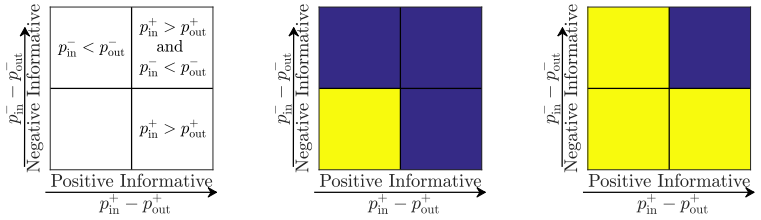
The case for $p \to -\infty$ is analogous: $m_\infty(\rho^+_\varepsilon, \rho^-_\varepsilon) = \min(1-\rho^+ + \varepsilon, 1+\rho^- + \varepsilon)$ and thus $m_{-\infty}(\rho^+_\varepsilon, \rho^-_\varepsilon) < 1+\varepsilon$ if and only if $\rho^+ > 0$ or $\rho^- < 0$, yielding the desired conditions. □

The conditions for $\mathcal{L}_\infty$ are the most conservative ones, as they require that G^+ and G^- are informative, i.e. G^+ has to be assortative and G^- disassortative. Under these conditions every clustering method for signed graphs should be able to identify the ground truth clusters in expectation. On the other hand, the less restrictive conditions for the recovery of the ground truth clusters correspond to the limit case $\mathcal{L}_{-\infty}$. If G^+ or G^- are informative, then the ground truth clusters are recovered, that is, $\mathcal{L}_{-\infty}$ only requires that G^+ is assortative or G^- is disassortative. In particular, the following corollary shows that smaller values of p require less restrictive conditions to ensure the identification of the informative eigenvectors.

Corollary 3.2. *Let* $q \leq p$. *If* $\{\chi_i\}_{i=\theta(p)}^k$ *correspond to the k-smallest eigenvalues of* $\mathcal{L}_p$, *then* $\{\chi_i\}_{i=\theta(q)}^k$ *correspond to the k-smallest eigenvalues of* $\mathcal{L}_q$, *where* $\theta(x) = 1$ *if* $x \leq 0$ *and* $\theta(x) = 2$ *if* $x > 0$.

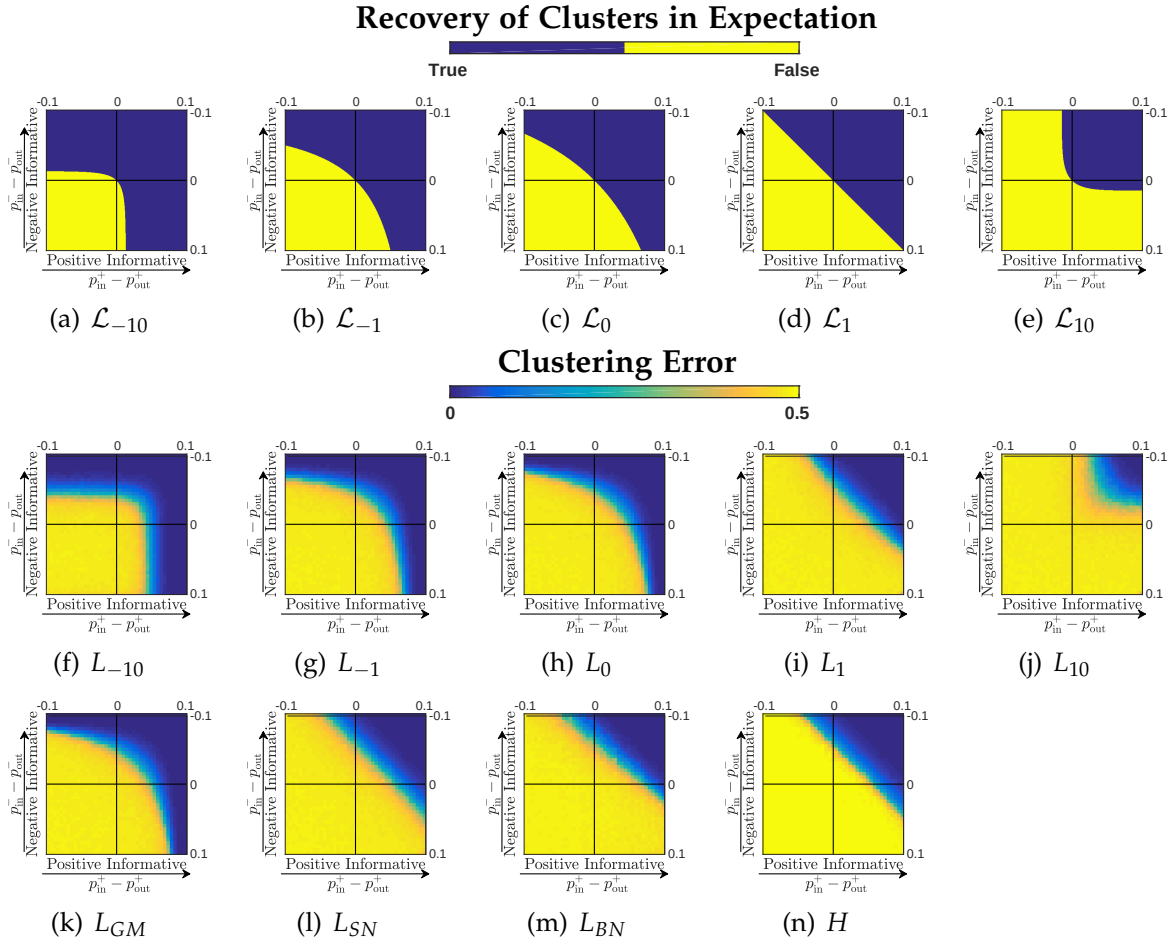
Proof. If $\lambda_1, \ldots, \lambda_k$ resp. $(\lambda_2, \ldots, \lambda_k)$ are among the k (resp. $k-1$)-smallest eigenvalues of $\mathcal{L}_p$, then by Theorem 3.1, we have $m_p(\rho^+_\varepsilon, \rho^-_\varepsilon) < 1+\epsilon$. By Theorem 2.1 we have $m_q(\rho^+_\varepsilon, \rho^-_\varepsilon) \leq m_p(\rho^+_\varepsilon, \rho^-_\varepsilon)$, Theorem 3.1 concludes the proof. □

To better understand the different conditions we have derived, we visualize them in Fig. 3.1, where the x-axis corresponds to how assortative G^+ is, while the y-axis corresponds to how disassortative G^- is. The conditions of the limit case $\mathcal{L}_\infty$, i.e. the case where G^+ and G^- have to be informative, correspond to the upper-right, dark blue region in Fig. 3.1(c), and correspond to the 25% of all possible configurations of the SBM. The conditions for the limit case $\mathcal{L}_{-\infty}$, i.e. the case where G^+ or G^- has to be informative, instead correspond to all possible configurations



(a) SBM Diagram (b) $L_{-\infty}$ (OR) (c) L_{∞} (AND)

Figure 3.1: Stochastic Block Model (SBM) for signed graphs. From left to right: Fig. 3.1(a) SBM Diagram. Fig. 3.1(b) SBM for $L_{-\infty}$(OR), Fig. 3.1(c) SBM for L_{∞}(AND). according to Corollary 3.1.



(a) $\mathcal{L}_{-10}$ (b) $\mathcal{L}_{-1}$ (c) $\mathcal{L}_{0}$ (d) $\mathcal{L}_{1}$ (e) $\mathcal{L}_{10}$

(f) L_{-10} (g) L_{-1} (h) L_{0} (i) L_{1} (j) L_{10}

(k) L_{GM} (l) L_{SN} (m) L_{BN} (n) H

Figure 3.2: Performance visualization for two clusters for different parameters of the SBM. **Top row**: The settings where the signed power mean Laplacians $\mathcal{L}_p$ identify the ground truth clusters in expectation for the SBM, see Theorem 3.1, are highlighted in dark blue, whereas yellow indicates failure. **Middle/Bottom row**: average clustering error (dark blue: small error, yellow: large error) of the signed power mean Laplacian L_p and $L_{GM}, L_{SN}, L_{BN}, H$ for 50 samples from the SBM.

of the SBM except for the bottom-left region. This is depicted in Fig. 3.1(b) and corresponds to the 75% of all possible configurations under the SBM. In Fig. 3.2 we present the corresponding conditions for recovery in expectation for the cases

$p \in \{-10, -1, 0, 1, 10\}$. We can visually verify that the larger the value of p the smaller is the region where the conditions of Theorem 3.1 hold. In particular, one can compare the change of conditions as one moves from the signed harmonic ($\mathcal{L}_{-1}$), geometric ($\mathcal{L}_0$), to the arithmetic ($\mathcal{L}_1$) mean Laplacians verifying the ordering described in Corollary 5.2. Moreover, we clearly observe that $\mathcal{L}_{-10}$ and $\mathcal{L}_{10}$ are already quite close to the conditions necessary for the limit cases $\mathcal{L}_{-\infty}$ and $\mathcal{L}_{\infty}$, respectively.

In the middle row of Fig. 3.2 we show the average clustering error for each power mean Laplacian when sampling 50 times from the SSBM following the diagram presented in Fig. 3.1(a) and fixing the sparsity of G^+ and G^- by setting $p^+_{\text{in}} + p^+_{\text{out}} = 0.1$ and $p^-_{\text{in}} + p^-_{\text{out}} = 0.1$ with two clusters each of size 100. We observe that the areas with low clustering error qualitatively match the regions where in expectation we have recovery of the clusters. However, due to the sampling which can make one of the graphs G^+ and G^- quite sparse and as we just consider graphs with 200 nodes, together with the sampling variance in the stochastic block model, the area of low clustering error is smaller in comparison to the region of guaranteed recovery in expectation.

In the bottom row of Fig. 3.2 we show the clustering error for the state of the art methods L_{GM}, L_{SN}, L_{BM} and H. We can see that L_{GM} presents a similar performance as the signed power mean Laplacian L_0. The next Theorem shows that the geometric mean Laplacian $\mathcal{L}_{GM}$ and the limit $p \to 0$ of the signed power mean Laplacian agree in expectation for the SSBM. This implies via Corollary 5.2 that this operator is inferior to the signed power mean Laplacian for $p < 0$. This is why we use in the experiments on real world graphs later on always $p < 0$.

Theorem 3.2. *Let $\mathcal{L}_{GM} = \mathcal{L}^+_{\text{sym}} \# \mathcal{Q}^-_{\text{sym}}$ and $\mathcal{L}_0$ be the signed power mean Laplacian with $p \to 0$ of the expected signed graph. Then, $\mathcal{L}_0 = \mathcal{L}_{GM}$.*

Proof. Following the proof from Theorem 3.1 we can see that $\mathcal{L}^+_{\text{sym}}$ and $\mathcal{Q}^-_{\text{sym}}$ share all of their eigenvectors. Let $\mathbf{u}$ be an eigenvector of $\mathcal{L}^+_{\text{sym}}$ and $\mathcal{Q}^-_{\text{sym}}$ with eigenvalues α and β, respectively.

By Lemma 2.1 we have

$$\mathcal{L}_0 \mathbf{u} = m_0(\alpha, \beta)\mathbf{u}$$

Moreover, from (Mercado *et al.*, 2016, Theorem 1) we know that

$$(\mathcal{L}^+_{\text{sym}} \# \mathcal{Q}^-_{\text{sym}})\mathbf{u} = \sqrt{\alpha\beta}\mathbf{u}$$

Further, $m_0(\alpha, \beta) = \sqrt{\alpha\beta}$. Hence, as $\mathcal{L}_0$ and $\mathcal{L}^+_{\text{sym}} \# \mathcal{Q}^-_{\text{sym}}$ have in common all eigenvectors and eigenvalues, we conclude that $\mathcal{L}_0 = \mathcal{L}^+_{\text{sym}} \# \mathcal{Q}^-_{\text{sym}}$. □

In the bottom row of Fig. 3.2 we can observe that L_{SN}, L_{BN} and H present a similar behaviour to the arithmetic mean Laplacian L_1. A quick computation shows that for the case where both G^+, G^- have the same node degree in expectation, the conditions of Theorem 3.1 for $\mathcal{L}_1$ reduce to $p^-_{\text{in}} + p^+_{\text{out}} < p^+_{\text{in}} + p^-_{\text{out}}$. It turns out that this condition is also required by $\mathcal{L}_{SN}, \mathcal{L}_{BN}$ and $\mathcal{H}$, as the following shows.

Theorem 3.3 ((Mercado *et al.*, 2016, Theorem 2)). *Let $\mathcal{L}_{BN}$ and $\mathcal{L}_{SN}$ be the balanced normalized Laplacian and signed normalized Laplacian of the expected signed graph. The following statements are equivalent:*

- $\{\chi_i\}_{i=1}^k$ *are the eigenvectors corresponding to the k-smallest eigenvalues of* $\mathcal{L}_{BN}$.
- $\{\chi_i\}_{i=1}^k$ *are the eigenvectors corresponding to the k-smallest eigenvalues of* $\mathcal{L}_{SN}$.
- *inequalities* $p_{\text{in}}^- + (k-1)p_{\text{out}}^- < p_{\text{in}}^+ + (k-1)p_{\text{out}}^+$ *and* $p_{\text{in}}^- + p_{\text{out}}^+ < p_{\text{in}}^+ + p_{\text{out}}^-$ *hold.*

Finally, we present conditions in expectation for the Bethe Hessian to identify the ground truth clustering.

Theorem 3.4. *Let $\mathcal{H}$ be the Bethe Hessian of the expected signed graph. Then $\{\chi_i\}_{i=2}^k$ are the eigenvectors corresponding to the $(k-1)$-smallest negative eigenvalues of $\mathcal{H}$ if and only if the following conditions hold:*

1. $\max\{0, \frac{2(d^+ + d^-) - 1}{\sqrt{d^+ + d^-}|\mathcal{C}|}\} < (p_{\text{in}}^+ - p_{\text{out}}^+) - (p_{\text{in}}^- - p_{\text{out}}^-)$
2. $p_{\text{out}}^+ < p_{\text{out}}^-$

Moreover, for the limit case $|V| \to \infty$ the first condition reduces to $p_{\text{in}}^- + p_{\text{out}}^+ < p_{\text{in}}^+ + p_{\text{out}}^-$.

Proof. In our framework the we can see that $J = W^+ - W^-$. In the proof of Theorem 3.1 we can see that expected adjacency matrices $\mathcal{W}^+$ and $\mathcal{W}^-$ have three distinct eigenvalues:

$$\lambda_1^+ = |\mathcal{C}|(p_{\text{in}}^+ + (k-1)p_{\text{out}}^+), \quad \lambda_i^+ = |\mathcal{C}|(p_{\text{in}}^+ - p_{\text{out}}^+)$$
$$\lambda_1^- = |\mathcal{C}|(p_{\text{in}}^- + (k-1)p_{\text{out}}^-), \quad \lambda_i^- = |\mathcal{C}|(p_{\text{in}}^- - p_{\text{out}}^-)$$

for $i = 2, \ldots, k$, with corresponding eigenvectors $\chi_1, \ldots, \chi_k$. Remaining eigenvalues are equal to zero. Further, as both matrices $\mathcal{W}^+$ and $\mathcal{W}^-$ share all their eigenvectors, then the expected matrix $\mathcal{J} = \mathcal{W}^+ - \mathcal{W}^-$ has the same eigenvectors with eigenvalues being the difference between the positive and negative counterparts, i.e. $\mathcal{J}\chi_i = \mu_i \chi_i$ where

$$\mu_i = \lambda_i^+ - \lambda_i^- .$$

As we assume that all clusters are of the same size $|C|$, the expected signed graph is a regular graph with degrees d^+ and d^-. Thus, in expectation $\widehat{\alpha} = d^+ + d^-$, where $d^+ = |\mathcal{C}|(p_{\text{in}}^+ + (k-1)p_{\text{out}}^+)$ and $d^- = |\mathcal{C}|(p_{\text{in}}^- + (k-1)p_{\text{out}}^-)$. Hence, the Bethe hessian of the expected signed graph can be expressed as

$$\begin{aligned}\mathcal{H} &= (\widehat{\alpha} - 1)I - \sqrt{\widehat{\alpha}}\mathcal{J} + \overline{\mathcal{D}} \\ &= (\widehat{\alpha} - 1)I - \sqrt{\widehat{\alpha}}\mathcal{J} + \widehat{\alpha}I \\ &= (2\widehat{\alpha} - 1)I - \sqrt{\widehat{\alpha}}\mathcal{J}\end{aligned}$$

It is easy to see that the matrix $\mathcal{H}$ is some sort of a diagonal shift of $\mathcal{J}$, and thus they have the same eigenvectors. In particular we can observe that:

$$\begin{aligned}\mathcal{H}\chi_i &= \big((2\widehat{\alpha}-1)I - \sqrt{\widehat{\alpha}}\mathcal{J}\big)\chi_i \\ &= (2\widehat{\alpha}-1)\chi_i - \sqrt{\widehat{\alpha}}\mathcal{J}\chi_i \\ &= (2\widehat{\alpha}-1)\chi_i - \sqrt{\widehat{\alpha}}\mu_i\chi_i \\ &= \big((2\widehat{\alpha}-1) - \sqrt{\widehat{\alpha}}\mu_i\big)\chi_i\end{aligned}$$

Hence, the corresponding eigenvalues of $\mathcal{H}$ are:

$$\lambda_i = (2\widehat{\alpha}-1) - \sqrt{\widehat{\alpha}}\mu_i\,. \tag{3.9}$$

All in all, the corresponding eigenvalues of the expected Bethe hessian matrix $\mathcal{H}$ are:

$$\begin{aligned}\lambda_1 &= (2\widehat{\alpha}-1) - \sqrt{\widehat{\alpha}}(d^+ - d^-)\,, \\ \lambda_i &= (2\widehat{\alpha}-1) - \sqrt{\widehat{\alpha}}|\mathcal{C}|\big((p^+_{\text{in}} - p^+_{\text{out}}) - (p^-_{\text{in}} - p^-_{\text{out}})\big)\,, \\ \lambda_j &= (2\widehat{\alpha}-1)\,.\end{aligned}$$

for $i = 2,\dots,k$ and $j = k+1,\dots,n$.

We now focus on the conditions that are necessary so that eigenvectors $\chi_2,\dots,\chi_k$ have the smallest negative eigenvalues. This is based on the fact that informative eigenvectors of the Bethe Hessian H have the smallest negative eigenvalue. From Eq.3.9 we can see that the general condition for eigenvalues of the Bethe Hessian in expectation $\mathcal{H}$ to be negative is

$$\lambda_i < 0 \iff \frac{2\widehat{\alpha}-1}{\sqrt{\widehat{\alpha}}} < \mu_i\,. \tag{3.10}$$

Hence the conditions to be analyzed are:

$$\begin{aligned}&\lambda_i < \lambda_1, \quad \text{for } i = 2,\dots,k \\ &\lambda_i < 0, \quad \text{for } i = 2,\dots,k \\ &\lambda_i < \lambda_j, \quad \text{for } i = 2,\dots,k \ \text{ and } j = k+1,\dots,n\end{aligned}$$

Therefore we can easily see that the corresponding condition $\lambda_i < \lambda_1$ boils down to

$$p^+_{\text{out}} < p^-_{\text{out}}$$

whereas condition $\lambda_i < 0$ is equivalent to

$$\frac{2(d^+ + d^-) - 1}{\sqrt{d^+ + d^-}|\mathcal{C}|} < \big((p^+_{\text{in}} - p^+_{\text{out}}) - (p^-_{\text{in}} - p^-_{\text{out}})\big)$$

and for the remaining condition $\lambda_i < \lambda_j$ the equivalent condition is

$$0 < (p_{\text{in}}^+ - p_{\text{out}}^+) - (p_{\text{in}}^- - p_{\text{out}}^-)$$

By taking together conditions for $\lambda_i < 0$ and $\lambda_i < \lambda_j$ we get the desired result.

For the limit case $|V| \to \infty$ we have the following. Let

$$\begin{aligned} c_2 &= p_{\text{in}}^+ + (k-1)p_{\text{out}}^+ + p_{\text{in}}^- + (k-1)p_{\text{out}}^- \\ c_3 &= (p_{\text{in}}^+ - p_{\text{out}}^+) - (p_{\text{in}}^- - p_{\text{out}}^-) \end{aligned}$$

The first condition of Theorem 3.4 can be expressed as follows:

$$\begin{aligned} &\frac{2(d^+ + d^-) - 1}{\sqrt{d^+ + d^-}|\mathcal{C}|} < \left((p_{\text{in}}^+ - p_{\text{out}}^+) - (p_{\text{in}}^- - p_{\text{out}}^-)\right) \iff \\ &\frac{2c_2^{1/2}}{|\mathcal{C}|^{1/2}} - \frac{1}{|\mathcal{C}|^{3/2}c_2^{1/2}} < c_3\,. \end{aligned}$$

Hence, in the limit where $|C| \to \infty$ the above condition turns into

$$0 < c_3 \iff p_{\text{in}}^- + p_{\text{out}}^+ < p_{\text{in}}^+ + p_{\text{out}}^-\,. \tag{3.11}$$

yielding the desired conditions.

□

The following Lemma states the interesting fact that the Bethe Hessian works better for large graphs

Lemma 3.1. *Let $\mathcal{H}_n$ be the Bethe Hessian of the expected signed graph under the SBM with n nodes. Let $\chi^n = \{\chi_i\}_{i=2}^k$ where $\chi_2, \ldots, \chi_k \in \mathbb{R}^n$. Let $\frac{3}{2} < d^+ + d^-$. Let $n < m$. If χ^n are eigenvectors corresponding to the $(k-1)$-smallest negative eigenvalues of $\mathcal{H}_n$, then χ^m are eigenvectors corresponding to the $(k-1)$-smallest negative eigenvalues of $\mathcal{H}_m$.*

Proof. In this proof we show that if for a given signed graph with n nodes the conditions of Theorem 3.4 hold, then conditions of Theorem 3.4 hold for expected signed graphs with a larger number of nodes.

By Theorem 3.4, we know that for a given graph in expectation with n nodes, eigenvectors $\chi^n = \{\chi_i\}_{i=2}^k$ correspond to the $(k-1)$-smallest negative eigenvalues of $\mathcal{H}_n$ if and only the following conditions hold:

1. $\max\{0, \frac{2(d^+ + d^-) - 1}{\sqrt{d^+ + d^-}|\mathcal{C}|}\} < (p_{\text{in}}^+ - p_{\text{out}}^+) - (p_{\text{in}}^- - p_{\text{out}}^-)$
2. $p_{\text{out}}^+ < p_{\text{out}}^-$

Observe that the right hand side of the above conditions does not depend on the number of nodes in the graph. We proceed by analyzing the left hand side of the first condition:

$$\frac{2(d^+ + d^-) - 1}{|\mathcal{C}|\sqrt{d^+ + d^-}}. \tag{3.12}$$

Note that under the Stochastic Block Model in consideration, all k clusters are of size $|\mathcal{C}| = \frac{n}{k}$. We now identify conditions such that the Equation 3.12 decreases with larger values of $|\mathcal{C}|$.

Let $x, \alpha \in \mathbb{R}$. Define the scalar function $g : \mathbb{R}_{>0} \to \mathbb{R}$ as

$$g(x) = \frac{2\alpha x - 1}{\sqrt{\alpha x^3}}$$

Observe that we recover Equation 3.12 by letting $x = |\mathcal{C}|$ and $\alpha = p_{\text{in}}^+ + (k-1)p_{\text{out}}^+ + p_{\text{in}}^- + (k-1)p_{\text{out}}^-$ where $\alpha x = d^+ + d^-$.

The corresponding derivative is

$$g'(x) = \frac{3 - 2\alpha x}{2x\sqrt{\alpha x^3}}$$

Then

$$g'(x) < 0 \iff \frac{3}{2} < \alpha x. \tag{3.13}$$

Hence, if $\frac{3}{2} < \alpha x$ then $g(y) < g(x)$ if and only if $x < y$. We now apply this result to our setting.

Let $|\mathcal{C}_n| := |\mathcal{C}| = \frac{n}{k}$ and $|\mathcal{C}_m| = \frac{m}{k}$ denote the cluster size of the expected signed graphs with n and m nodes, respectively. Let $\alpha = p_{\text{in}}^+ + (k-1)p_{\text{out}}^+ + p_{\text{in}}^- + (k-1)p_{\text{out}}^-$. Let $\frac{3}{2} < d^+ + d^-$. Then

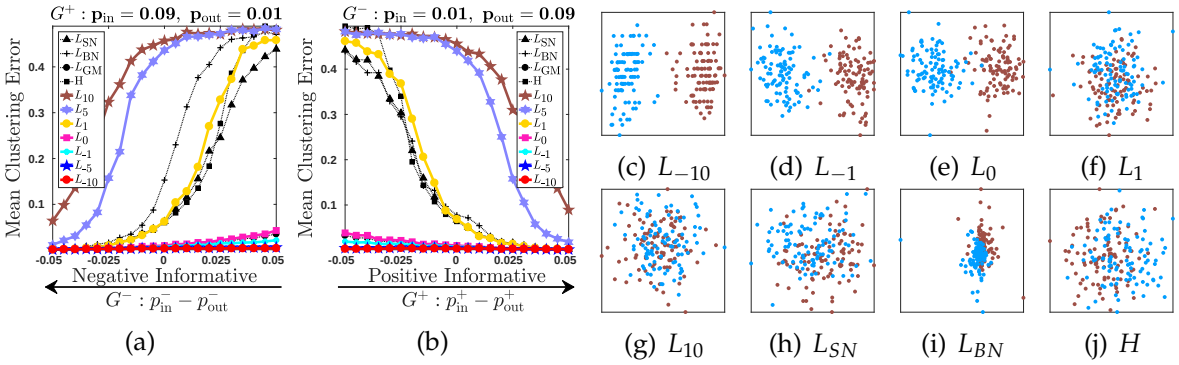
$$g(|\mathcal{C}_m|) < g(|\mathcal{C}_n|) = \frac{2(d^+ + d^-) - 1}{|\mathcal{C}|\sqrt{d^+ + d^-}} \tag{3.14}$$

if and only if $n < m$. Hence, if conditions 1 and 2 hold for the expected graph G with n nodes and its expected absolute degree is larger than $\frac{3}{2}$, i.e. $\frac{3}{2} < d^+ + d^-$, then conditions 1 and 2 hold for expected graphs with a larger number of nodes, leading to the desired result. □

We can observe that the first condition in Theorem 3.4 is related to conditions of $\mathcal{L}_1$ and $\mathcal{L}_{SN}, \mathcal{L}_{BN}$ through the inequality $p_{\text{in}}^- + p_{\text{out}}^+ < p_{\text{in}}^+ + p_{\text{out}}^-$. This explains why the performance of the Bethe Hessian H resembles the one of arithmetic Laplacians L_{SN}, L_{BN}, L_1.

3.4.2 SBM random graphs

We now zoom in on a particular setting of Fig. 3.2. Namely, the case where G^+ (resp.G^-) is fixed to be informative, whereas the remaining graph transitions from informative to uninformative. The corresponding results are in Fig. 3.3.



(a) (b) (c) L_{-10} (d) L_{-1} (e) L_0 (f) L_1 (g) L_{10} (h) L_{SN} (i) L_{BN} (j) H

Figure 3.3: **Left**: Mean clustering error under the SSBM, with two clusters of size 100 and 50 runs. In Fig. 3.3(a): G^+ is informative, i.e. assortative with $p^+_{\text{in}} = 0.09$ and $p^+_{\text{out}} = 0.01$. In Fig. 3.3(b): G^- is informative, i.e. disassortative with $p^-_{\text{in}} = 0.01$ and $p^-_{\text{out}} = 0.09$. **Right**: Node embeddings induced by eigenvectors of different signed Laplacians for a random graph drawn from SSBM for 2 clusters of size 100, $p^+_{\text{in}} = 0.025, p^+_{\text{out}} = 0.075, p^-_{\text{in}} = 0.01, p^-_{\text{out}} = 0.09$.

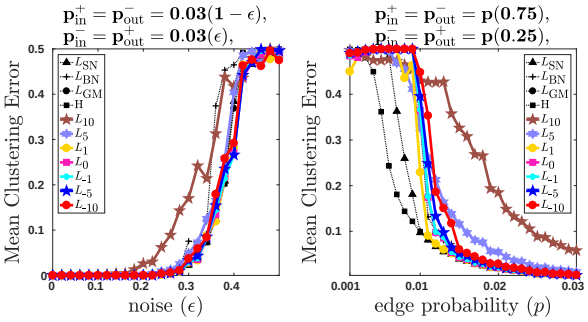
In Fig. 3.3(a) we consider the case where G^+ is informative with parameters $p^+_{\text{in}} = 0.09$ and $p^+_{\text{out}} = 0.01$ (this corresponds to $p^+_{\text{in}} - p^+_{\text{out}} = 0.08$ in Fig. 3.2), and G^- goes from being informative ($p^-_{\text{in}} < p^-_{\text{out}}$) to non-informative ($p^-_{\text{in}} \geq p^-_{\text{out}}$). We confirm that the power mean Laplacian L_p presents smaller clustering errors for smaller values of p. Moreover, it is clear that in the case $p < 0$, L_p is able to recover clusters even in the case where G^- is not informative, whereas for $p > 0$, L_p requires both G^+ and G^- to be informative. We observe that the smallest (resp. largest) clustering errors correspond to L_{-10} (resp. L_{10}), corroborating Corollary 5.2. Further, we can observe that L_{GM} and L_0 have a similar performance, as well as L_{SN}, L_{BN}, L_1, H, as observed before, confirming Theorem 3.2 and Theorem 3.4, respectively.

In Fig. 3.3(b) similar observations hold for the case where G^- is informative with parameters $p^-_{\text{in}} = 0.01$ and $p^-_{\text{out}} = 0.09$ (this corresponds to $p^-_{\text{in}} - p^-_{\text{out}} = -0.08$ in Fig. 3.2), and G^+ goes from being non-informative ($p^+_{\text{in}} \leq p^+_{\text{out}}$) to informative ($p^+_{\text{in}} > p^+_{\text{out}}$). Within this setting we present the eigenvector-based node embeddings of each method for the case $p^+_{\text{in}} = 0.025, p^+_{\text{out}} = 0.075, p^-_{\text{in}} = 0.01, p^-_{\text{out}} = 0.09$, in right hand side of Fig. 3.3. For L_{-10}, L_{-1}, L_0 the embeddings split the clusters properly, whereas remaining embeddings are not informative, verifying the effectivity of L_p with $p < 0$.

3.4.3 Experiments with the Censored Block Model

In this section we present a numerical evaluation of different methods under the Stochastic Block Model following the parameters corresponding to the Censored Block Model (**CBM**), following (Saade *et al.*, 2015).

Observe that the CBM is a particular case of the Stochastic Block Model for signed graphs as introduced in Section 3.4. Following (Saade *et al.*, 2015), the CBM has two



(a) $p = 0.03, \epsilon \in [0, 0.5]$

(b) $p \in [0.001, 0.03], \epsilon = 0.25$

Figure 3.4: Mean clustering error under the Censored Block Model (Saade *et al.*, 2015), with two clusters of size 500 and 20 runs. Fig. 3.4(a): probability of observing and edge is fixed to $p = 0.03$, and $\epsilon \in [0, 0.5]$. Fig. 3.4(b): probability of flipping sign of an edge is fixed to $\epsilon = 0.25$, and $p \in [0.001, 0.03]$

parameters: probability of observing an edge ($\overline{p}$), and the probability of flipping the sign of an edge (ϵ). The CBM can be recovered from the SSBM introduced in Section 3.4 by setting

$$p_{\text{in}}^{+} = p_{\text{out}}^{-} = \overline{p}(1-\epsilon), \qquad p_{\text{in}}^{-} = p_{\text{out}}^{+} = \overline{p}\epsilon$$

Observe that the parameter ϵ works as a noise parameter: the noiseless setting corresponds to $\epsilon = 0$, where positive and negative edges are only inside and between clusters, respectively. The case where $\epsilon = 0.5$ corresponds to the case where no clustering structure is conveyed by the sign of the edges.

We present a numerical evaluation under the SSBM with parameters from CBM in Fig. 3.4. We consider two clusters and fix a priori its size to be of 500 nodes each. We present the clustering error out of 20 realizations from the SSM with parameters following the CBM. We consider two settings:

First setting: we fix the probability of observing an edge to $\overline{p} = 0.03$, and evaluate over different values of $\epsilon \in [0, 0.5]$. In Fig. 3.4(a) we can observe that there is no relevant difference in clustering error between methods. Further, as expected we can see that for small values of ϵ all methods perform well, and for larger values of ϵ the clustering error increases;

Second setting: we fix the probability of flipping the sign of an edge to $\epsilon = 0.25$, and evaluate over different values of $\overline{p} \in [0.001, 0.03]$. In Fig. 3.4(b) we can observe that the performance of the Bethe Hessian is best for small values of $\overline{p}$, i.e. for sparser graphs. Following the Bethe Hessian are the arithmetic mean Laplacian L_1 together with the signed normalized Laplacian L_{SN}.

Hence we have observed that for sufficiently dense graphs following the Censored Block Model, the performance of different methods is rather similar, whereas for sparser graphs the Bethe Hessian performs best, confirming the analysis presented in (Saade *et al.*, 2015).

So far we have presented an analysis in expectation with verifications on sampled random graphs following the stochastic and censored block models. In the following section we continue with an analysis under the stochastic block model and focus on an analysis of consistency.

3.4.4 Consistency of the Signed Power Mean Laplacian for the Stochastic Block Model

In this section we prove two novel concentration bounds for signed power mean Laplacians of signed graphs drawn from the SSBM. The bounds show that, for large graphs, our previous results in expectation transfer to sampled graphs with high probability. We first show in Theorem 3.5 that L_p is close to $\mathcal{L}_p$. Then, in Theorem 3.6, we show that eigenvalues and eigenvectors of L_p are close to those of $\mathcal{L}_p$. We derive this result by tracing back the consistency of the matrix power mean to the consistency of the standard and signless Laplacian established in (Chung and Radcliffe, 2011).

The consistency of spectral clustering on unsigned graphs for the SBM has been studied in (Lei and Rinaldo, 2015; Sarkar and Bickel, 2015; Rohe *et al.*, 2011). More recently also the consistency of several variants of spectral clustering has been shown (Qin and Rohe, 2013; Joseph and Yu, 2016; Chaudhuri *et al.*, 2012; Le *et al.*, 2017; Fasino and Tudisco, 2018; Davis and Sethuraman, 2018). Moreover, while the case of multilayer graphs under the SBM has been previously analyzed (Han *et al.*, 2015; Heimlicher *et al.*, 2012; Jog and Loh, 2015; Paul and Chen, 2020; Xu *et al.*, 2014, 2020; Yun and Proutiere, 2016), there are no consistency results for matrix power means for signed graphs. While our main emphasis is on the analysis of the SPM Laplacian, our proofs are general enough to cover also the consistency of the matrix power means for unsigned multilayer graphs (Mercado *et al.*, 2018). In Thm. 3.5 we show that the SPM Laplacian L_p for the SSBM is concentrated around $\mathcal{L}_p$, with high probability for large n. The following results hold for general shifts ε.

Theorem 3.5. *Let p be a non-zero integer, let*

$$C_p = \begin{cases} (2p)^{1/p}(2+\varepsilon)^{1-1/p} & p \geq 1 \\ |2p|^{1/|p|}\varepsilon^{-(3+1/|p|)} & p \leq -1 \end{cases}$$

Choose $\epsilon > 0$. If $\frac{n}{k}(p^+_{\text{in}}+(k-1)p^+_{\text{out}}) > 3\ln(8n/\epsilon)$, and $\frac{n}{k}(p^-_{\text{in}}+(k-1)p^-_{\text{out}}) > 3\ln(8n/\epsilon)$, then with probability at least $1-\epsilon$, we have

$$\left\|L_p - \mathcal{L}_p\right\| \leq C_p\, m_{|p|}^{1/|p|}\left(\sqrt{\frac{3\ln(8n/\epsilon)}{\frac{n}{k}(p^+_{\text{in}}+(k-1)p^+_{\text{out}})}}, \sqrt{\frac{3\ln(8n/\epsilon)}{\frac{n}{k}(p^-_{\text{in}}+(k-1)p^-_{\text{out}})}}\right)$$

Proof. Please see Appendix A.1. □

In Thm 3.5 we take the spectral norm. A more general version of Theorem 3.5 for the inhomogeneous Erdős-Rényi model, where edges are formed independently with probabilities p^+_{ij}, p^-_{ij} is given in Theorem A.1. Theorem 3.5 builds on top of concentration results of (Chung and Radcliffe, 2011) proven for the unsigned case $||L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}||$. We can see that the deviation of L_p from $\mathcal{L}_p$ depends on the power mean of the individual deviations of L^+_{sym} and Q^-_{sym} from $\mathcal{L}^+_{\text{sym}}$ and $\mathcal{Q}^-_{\text{sym}}$, respectively. Note that the larger the size n of the graph is, the stronger is the concentration of L_p around $\mathcal{L}_p$.

The next Theorem shows that the eigenvectors corresponding to the smallest eigenvalues of L_p are close to the corresponding eigenvectors of $\mathcal{L}_p$. This is a key result showing consistency of our spectral clustering technique with L_p for signed graphs drawn from the SSBM.

Theorem 3.6. *Let $p \neq 0$ be an integer. Let $V_k, \mathcal{V}_k \in \mathbb{R}^{n\times k}$ be orthonormal matrices whose columns are the eigenvectors of the k smallest eigenvalues of L_p and $\mathcal{L}_p$, respectively. Let ρ_ε^+, ρ_ε^- and C_p be defined as in Theorems 3.1 and 3.5, respectively. Define $\tilde{k} = k-1$, if $p \geq 1$, and $\tilde{k} = k$, if $p \leq -1$ and choose $\epsilon > 0$.*
If $m_p(\rho_\varepsilon^+, \rho_\varepsilon^-) < 1+\varepsilon$, $\delta^+ := \frac{n}{k}(p_{\text{in}}^+ + (k-1)p_{\text{out}}^+) > 3\ln(8n/\epsilon)$, and $\delta^- := \frac{n}{k}(p_{\text{in}}^- + (k-1)p_{\text{out}}^-) > 3\ln(8n/\epsilon)$, then there exists an orthogonal matrix $O_{\tilde{k}} \in \mathbb{R}^{\tilde{k}\times\tilde{k}}$ such that, with probability at least $1-\epsilon$, we have

$$\|V_{\tilde{k}} - \mathcal{V}_{\tilde{k}} O_{\tilde{k}}\| \leq \frac{\sqrt{8\tilde{k}}\, C_p\, m_{|p|}^{1/|p|}\Big(\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^+}}, \sqrt{\frac{3\ln(8n/\epsilon)}{\delta^-}}\Big)}{(1+\varepsilon) - m_p(\rho_\varepsilon^+, \rho_\varepsilon^-)}$$

Proof. Please see Appendix A.3. □

Note that the main difference compared to Theorem 3.5 is the corresponding spectral gap $\gamma_p = (1+\varepsilon) - m_p(\rho_\varepsilon^+, \rho_\varepsilon^-)$ of $\mathcal{L}_p$, which is the difference of the eigenvalues corresponding to the informative versus non-informative eigenvectors of $\mathcal{L}_p$. Thus the stronger the clustering structure the tighter is the concentration of the eigenvectors. Moreover, from the monotonicity of m_p we have $\gamma_p \geq \gamma_q$ for $p < q$, and thus for $p \leq -1$ the spectral gap increases with $|p|$, ensuring a stronger concentration of eigenvectors for smaller values of p.

3.5 COMPUTATION OF THE SMALLEST EIGENVALUES AND EIGENVECTORS OF $M_p(L_{\text{sym}}^+, Q_{\text{sym}}^-)$

For the computation of the eigenvectors corresponding to the smallest eigenvalues of the signed power mean Laplacian L_p with $p < 0$, we take the Polynomial Krylov Subspace Method. The corresponding numerical scheme is presented in Algorithms 4 and 5.

We briefly explain Algorithm 4. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of $L_p = M_p(L_{\text{sym}}^+, Q_{\text{sym}}^-)$. Let $p < 0$. Then the eigenvalues of L_p^p are $\lambda_1^p \geq \cdots \geq \lambda_n^p$, that is, the eigenvectors corresponding to the smallest eigenvalues of L_p correspond to the largest eigenvalues of L_p^p. Thus, in order to obtain the eigenvectors corresponding to the smallest eigenvalues of L_p we have to apply the power method to L_p^p. This is depicted in Algorithm 4 . However, the main computational task now is the matrix-vector multiplications $(L_{\text{sym}}^+)^p\mathbf{x}$ and $(Q_{\text{sym}}^-)^p\mathbf{x}$. This is approximated through the Polynomial Krylov Subspace Method (PKSM). This approximation method allows to obtain $(L_{\text{sym}}^+)^p\mathbf{x}$ and $(Q_{\text{sym}}^-)^p\mathbf{x}$ without ever computing the matrices $(L_{\text{sym}}^+)^p$ and $(Q_{\text{sym}}^-)^p$, respectively. This is depicted in Algorithm 5.

Algorithm 4: PM applied to $M_p(L^+_{\text{sym}}, Q^-_{\text{sym}})$.

Input: $\mathbf{x}_0$, $p < 0$
Output: Eigenpair $(\lambda, \mathbf{x})$ of $M_p(L^+_{\text{sym}}, Q^-_{\text{sym}})$

repeat
  $\mathbf{u}_k^{(1)} \leftarrow (L^+_{\text{sym}})^p \mathbf{x}_k$ (Compute with Alg. 5)
  $\mathbf{u}_k^{(2)} \leftarrow (Q^-_{\text{sym}})^p \mathbf{x}_k$ (Compute with Alg. 5)
  $\mathbf{y}_{k+1} \leftarrow \frac{1}{2}(\mathbf{u}_k^{(1)} + \mathbf{u}_k^{(2)})$
  $\mathbf{x}_{k+1} \leftarrow \mathbf{y}_{k+1} / \|\mathbf{y}_{k+1}\|_2$
until *tolerance reached*
$\lambda \leftarrow (\mathbf{x}_{k+1}^T \mathbf{x}_k)^{1/p}$, $\mathbf{x} \leftarrow \mathbf{x}_{k+1}$

Algorithm 5: PKSM for the computation of $A^p\mathbf{y}$.

Input: $\mathbf{u}_0 = \mathbf{y}$, $V_0 = [\,\cdot\,]$, $p < 0$
Output: $\mathbf{x} = A^p\mathbf{y}$

$\mathbf{v}_0 \leftarrow \mathbf{y}/\|\mathbf{y}\|_2$
for $s = 0, 1, 2, \ldots, n$ **do**
  $\tilde{V}_{s+1} \leftarrow [V_s, \mathbf{v}_s]$
  $V_{s+1} \leftarrow$ Orthogonalize columns of $\tilde{V}_{s+1}$
  $H_{s+1} \leftarrow V_{s+1}^T A V_{s+1}$
  $\mathbf{x}_{s+1} \leftarrow V_{s+1}(H_{s+1})^p \mathbf{e}_1 \|\mathbf{y}\|_2$
  if *tolerance reached* **then** *break*
  $\mathbf{v}_{s+1} \leftarrow A\mathbf{v}_s$
end
$\mathbf{x} \leftarrow \mathbf{x}_{s+1}$

The main idea of PKSM s-step is to project a given matrix A onto the space $\mathbb{K}^s(A, \mathbf{y}) = \{\mathbf{y}, A\mathbf{y}, \ldots, A^{s-1}\mathbf{y}\}$ and solve the corresponding problem there. The projection on to $\mathbb{K}^s(A, \mathbf{y})$ is done by means of the Lanczos process, producing a sequence of matrices V_s with orthogonal columns where the first column of V_s is $\mathbf{y}/\|\mathbf{y}\|$ and range$(V_s) = \mathbb{K}^s(A, \mathbf{y})$. Moreover, at each step we have $AV_s = V_s H_s + \mathbf{v}_{s+1}\mathbf{e}_s^T$ where H_s is $s \times s$ symmetric tridiagonal, and $\mathbf{e}_i$ is the i-th canonical vector. The matrix product vector $\mathbf{x} = A^p\mathbf{y}$ is approximated by $\mathbf{x}_s = V_s(H_s)^p\mathbf{e}_1 \|\mathbf{y}\| \approx A^p\mathbf{y}$.

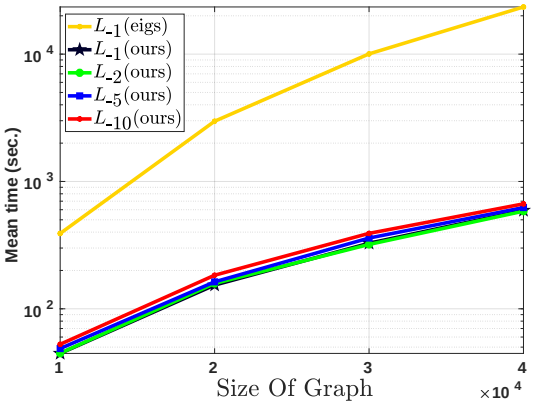


Figure 3.5: Median execution time of 10 runs for different Laplacians. Graphs have two perfect clusters and 2.5% of edges among nodes. $L_{GM}(ours)$ uses Algs 4 and 5, whereas we used Matlab's eigs for the other matrices. The use of eigs on L_{GM} is prohibitive as it needs the matrix L_{GM} to be built (we use the toolbox provided in Bini and Ianazzo (2015)), destroying the sparsity of the original graphs. Experiments are performed using one thread.

Time Execution Analysis. We present a time execution analysis in Fig. 3.5. We depict the mean time execution out of 10 runs of the power mean Laplacian L_p with $p \in \{-1, -2, -5 - 10\}$. In particular L_{-1}(ours), L_{-2}(ours), L_{-5}(ours) and L_{-10}(ours) depict the time execution using our proposed method based on Algorithm 4 together with the polynomial Krylov subspace method described in Algorithm 5. For comparison we consider L_{-1}(eigs) which is computed with the function eigs from MATLAB instead of using Algorithm 5. All experiments are performed using one thread. For evaluation random signed graphs following the SSBM are generated, with parameters $p^+_{\text{in}} = p^-_{\text{out}} = 0.05$ and $p^-_{\text{in}} = p^+_{\text{out}} = 0.025$ with two equal sized clusters, and graph size $|V| \in \{10000, 20000, 30000, 40000\}$. We

observe that our matrix-free approach based on the polynomial Krylov subspace method systematically outperforms the natural approach based on the explicit computation of power matrices per layer.

3.6 EXPERIMENTS

In this section we present experiments on different datasets. In Subsection 3.6.1 we present experiments on UCI datasets, and in Subsection 3.6.2 we present experiments on the Wikipedia-Elections dataset. Observe that in the datasets considered here it is unlikely that the corresponding signed graph follow the stochastic block model. Yet, we will observe that the proposed approach based on the Signed Power Mean Laplacian does present a competitive performance with the state of the art.

3.6.1 Experiments on UCI datasets

We evaluate the signed power mean Laplacian with L_{-10}, L_{-1} against L_{SN}, L_{BN}, L_{AM}, L_{GM} and H using datasets from the UCI repository. We build W^+ from the k^+ nearest neighbor graph, whereas W^- is obtained from the k^- farthest neighbor graph. For each dataset we evaluate all clustering methods over all possible choices of $k^+, k^- \in \{3, 5, 7, 10, 15, 20, 40, 60\}$, yielding in total 64 cases. We present the following statistics: Best(%): proportion of cases where a method yields the smallest clustering error. Strictly Best(%): proportion of cases where a method is the *only one* yielding the smallest clustering error. Results are shown in Table 3.1.

Observe that in 4 datasets H and L_{GM} present a competitive performance. For the remaining cases we can see that the best performance is obtained by the signed power mean Laplacians L_{-1}, L_{-10}. This verifies the superiority of negative powers ($p < 0$) to positive ($p > 0$) powers of L_p and related approaches like L_{SN}, L_{BN}. Moreover, although the Bethe Hessian is known to be optimal under the sparse transition theoretic limit under the Censored Block Model (Saade *et al.*, 2015), in the context where graphs unlikely follow a SBM distribution we can see that it is outperformed by the signed power mean Laplacian L_p.

We emphasize that the eigenvectors of L_p are calculated without ever computing the matrix itself, by using the method described in Sec.3.5.

3.6.2 Experiments on Wikipedia-Elections

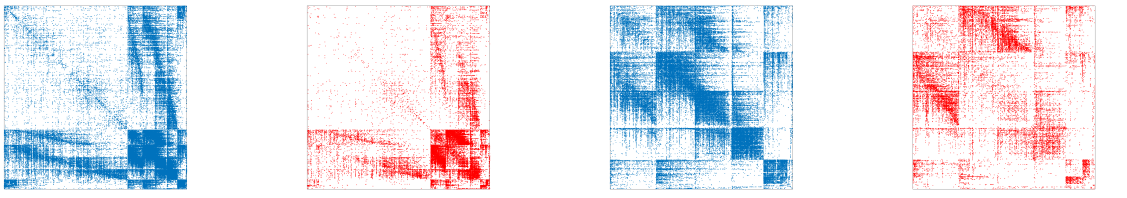
We now evaluate the Signed Power Mean Laplacian L_p with $p \in \{-10, -5, -2, -1, 0, 1\}$ on Wikipedia-Elections dataset (Leskovec and Krevl, 2014). In this dataset each node represents an editor requesting to become administrator and positive (resp. negative) edges represent supporting (resp. against) votes to the corresponding admin candidate.

		iris	wine	ecoli	australian	cancer	vehicle	german	image	optdig	isolet	USPS	pendig	20new	MNIST
	# vertices	150	178	336	690	699	846	1000	2310	5620	7797	9298	10992	18846	70000
	# classes	3	3	8	2	2	4	2	7	10	26	10	10	20	10
H	Best (%)	14.1	14.1	10.9	7.8	0.0	20.3	34.4	0.0	3.1	0.0	0.0	0.0	**45.3**	0.0
	Str. best (%)	4.7	3.1	7.8	3.1	0.0	14.1	17.2	0.0	3.1	0.0	0.0	0.0	**45.3**	0.0
	Avg. error	16.8	32.2	23.9	41.7	13.2	58.4	**29.5**	64.0	32.5	54.0	41.3	39.2	88.5	48.2
L_{SN}	Best (%)	10.9	10.9	14.1	4.7	15.6	12.5	12.5	17.2	26.6	7.8	14.1	7.8	26.6	14.1
	Str. best (%)	1.6	3.1	9.4	1.6	15.6	7.8	0.0	17.2	26.6	7.8	14.1	7.8	26.6	14.1
	Avg. error	17.5	32.2	24.5	42.8	8.8	57.2	29.9	53.9	24.9	51.2	38.6	37.8	89.0	45.8
L_{BN}	Best (%)	4.7	12.5	0.0	6.3	1.6	6.3	40.6	0.0	0.0	0.0	0.0	0.0	1.6	0.0
	Str. best (%)	1.6	1.6	0.0	0.0	1.6	4.7	17.2	0.0	0.0	0.0	0.0	0.0	1.6	0.0
	Avg. error	26.6	33.6	30.5	42.5	10.2	61.6	29.6	57.2	41.1	67.4	50.1	50.5	92.5	58.6
L_{AM}	Best (%)	6.3	20.3	7.8	6.3	0.0	20.3	15.6	9.4	0.0	0.0	0.0	0.0	4.7	0.0
	Str. best (%)	1.6	9.4	6.3	1.6	0.0	7.8	1.6	6.3	0.0	0.0	0.0	0.0	4.7	0.0
	Avg. error	19.0	32.7	24.4	42.7	11.6	58.1	29.7	47.7	33.5	49.6	44.7	48.3	89.7	56.1
L_{GM}	Best (%)	32.8	35.9	34.4	32.8	7.8	17.2	**46.9**	6.3	29.7	28.1	12.5	0.0	1.6	**82.8**
	Str. best (%)	1.6	7.8	**21.9**	23.4	6.3	14.1	**25.0**	6.3	28.1	28.1	9.4	0.0	1.6	**82.8**
	Avg. error	14.1	31.9	20.4	39.3	11.3	57.6	**29.5**	46.8	13.0	42.6	27.6	45.0	89.9	**26.7**
L_{-1}	Best (%)	25.0	**45.3**	**39.1**	**42.2**	0.0	12.5	15.6	**39.1**	4.7	**37.5**	4.7	9.4	12.5	1.6
	Str. best (%)	0.0	**14.1**	18.8	**31.3**	0.0	9.4	1.6	29.7	4.7	**37.5**	4.7	9.4	12.5	1.6
	Avg. error	13.8	**29.8**	**20.3**	**38.2**	8.3	56.2	29.8	39.7	16.3	**42.1**	25.2	32.9	**88.3**	32.3
L_{-10}	Best (%)	**73.4**	43.8	25.0	34.4	**76.6**	**31.3**	20.3	**39.1**	**37.5**	26.6	**71.9**	**82.8**	7.8	1.6
	Str. best (%)	**42.2**	7.8	10.9	18.8	**75.0**	**25.0**	4.7	**31.3**	**35.9**	26.6	**68.8**	**82.8**	7.8	1.6
	Avg. error	**12.7**	30.2	20.8	38.6	**5.7**	**55.9**	29.7	**39.4**	**12.1**	42.3	**21.9**	**26.9**	89.8	28.6

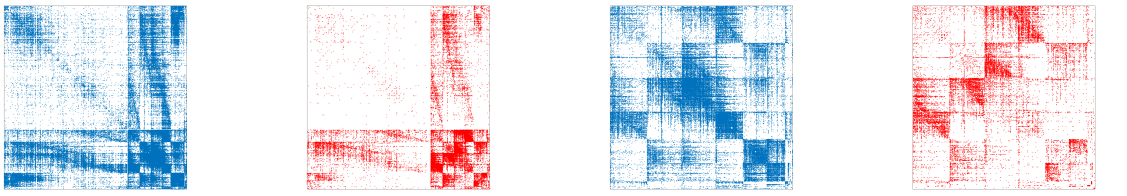
Table 3.1: Experiments on UCI datasets. Positive edges generated by k-nearest neighbours, and negative edges generated by k-farthest neighbours. We report the percentage of cases where each method achieves the smallest and strictly smallest clustering error, and the average clustering error.

While (Chiang *et al.*, 2012) conjectured that this dataset has no clustering structure, recent works (Mercado *et al.*, 2016; Cucuringu *et al.*, 2019) have shown that indeed there is clustering structure. As noted in (Mercado *et al.*, 2016), using the geometric mean Laplacian L_{GM} and looking for k clusters unveils the presence of a large non-informative cluster and $k-1$ remaining smaller clusters which show relevant clustering structure.

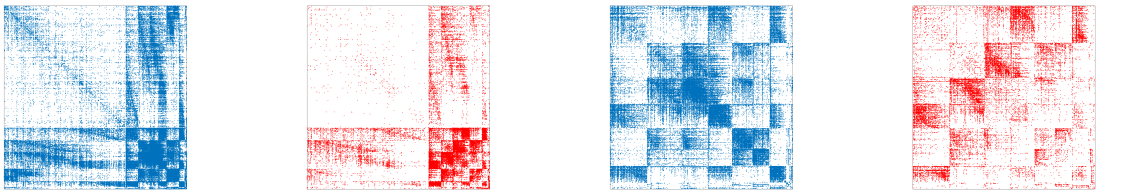
Our results verify these recent findings. We set the number of clusters to identify to $k = 30$ and in Fig. 3.6 we present the sorted adjacency matrices according to the identified clusters. In the first two columns (left to right) we can see that there is a large cluster (upper-left corner of each adjacency matrix) that does not resemble any structure, whereas the remaining part of the graph does present certain clustering structure. The following third and fourth columns zoom in into this region. We can see that when $p \leq 0$ the Signed Power Mean Laplacian L_p identifies clustering structure, whereas this structure is overlooked by the arithmetic mean case $p = 1$. Moreover, we can see that different powers identify slightly different clusters: this happens as this dataset does not necessarily follow the Signed Stochastic Block Model, and hence we do not fully retrieve the same behavior that was studied in Section 3.4.



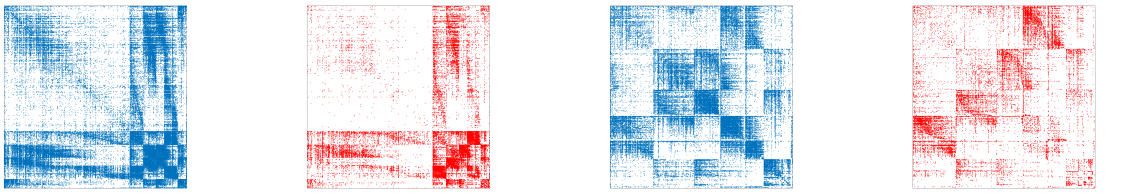
(a) $L_{-10} : W^+$ (b) $L_{-10} : W^-$ (c) $L_{-10} : W^+$ (Zoom) (d) $L_{-10} : W^-$ (Zoom)



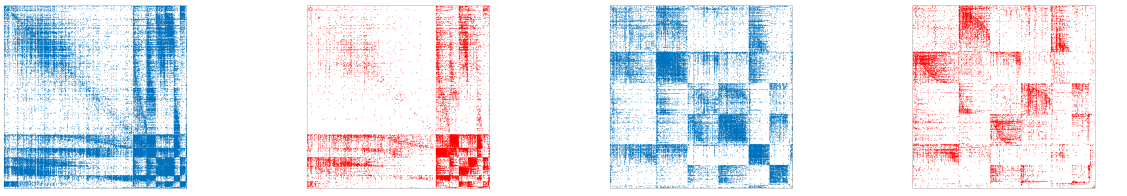
(e) $L_{-5} : W^+$ (f) $L_{-5} : W^-$ (g) $L_{-5} : W^+$ (Zoom) (h) $L_{-5} : W^-$ (Zoom)



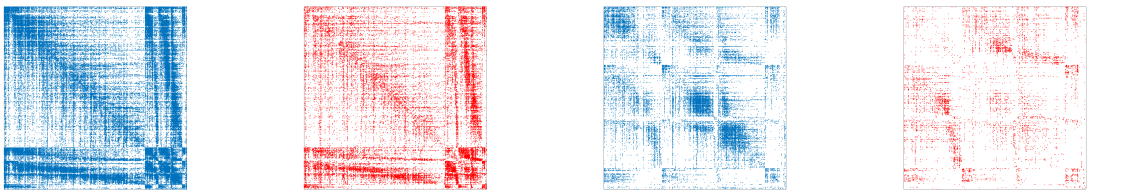
(i) $L_{-2} : W^+$ (j) $L_{-2} : W^-$ (k) $L_{-2} : W^+$ (Zoom) (l) $L_{-2} : W^-$ (Zoom)



(m) $L_{-1} : W^+$ (n) $L_{-1} : W^-$ (o) $L_{-1} : W^+$ (Zoom) (p) $L_{-1} : W^-$ (Zoom)



(q) $L_0 : W^+$ (r) $L_0 : W^-$ (s) $L_0 : W^+$ (Zoom) (t) $L_0 : W^-$ (Zoom)



(u) $L_1 : W^+$ (v) $L_1 : W^-$ (w) $L_1 : W^+$ (Zoom) (x) $L_1 : W^-$ (Zoom)

Figure 3.6: Sorted adjacency matrices according to clusters identified by the Power Mean Laplacian L_p with $p \in \{-10, -5, -2, -1, 0, 1\}$. Columns from left to right: First two columns depict adjacency matrices W^+ and W^- sorted through the corresponding clustering. Third and fourth columns depict the portion of adjacency matrices W^+ and W^- corresponding to the $k-1$ identified clusters. Rows from top to bottom: Clustering corresponding to $L_{-10}, L_{-5}, L_{-2}, L_{-1}, L_0, L_{-1}$.

3.6.3 On Diagonal Shift

In this section we briefly discuss the effect of the diagonal shift on the power mean Laplacian L_p for $p \leq 0$. In the definition of power mean Laplacian in Eq. 3.7 it is mentioned that for negative powers $p \leq 0$ a diagonal shift is necessary.

To evaluate the influence of the magnitude of the diagonal shift we perform numerical evaluations on two different kinds of signed graphs: on the one side we consider signed graphs generated through the Signed Stochastic Block Model introduced in Section 3.4, and on the other side we consider signed graphs built from standard machine learning benchmark datasets following Section 3.6.1.

Experiments with benchmark datasets. We now perform a numerical evaluation on different real-world networks, following the procedure of Section 3.6.1. Moreover, we perform this analysis for $p \in \{-1, -10\}$ and diagonal shifts $\{10^{-10}, 10^{-9}, \ldots, 10^{3}\}$. The corresponding results are presented in Fig. 3.7, where we show the average clustering error taken across all values of k^+ and k^- (for more details on the construction of the corresponding signed graphs please see Section 3.6.1).

We can observe a general behavior for L_{-10} across datasets where for a small diagonal shift, the clustering error is high, and decreases for larger shifts, generally reaching its minimum clustering error around diagonal shifts equal to one, to later present a slight increase in clustering error. This confirms the proposed approach to set the diagonal shift to $log_{10}(1 + |p|) + 10^{-6}$ which for the case of $p = -10$ is ≈ 1.04.

For the case of the harmonic mean Laplacian L_{-1} we can observe that it presents a more stable behavior that slightly resembles the one of L_{-10}. In particular, we can observe that there is a region from 10^{-6} to 10^{-1} where the smallest average clustering error is achieved. Hence, L_{-1} is relatively more robust to different diagonal shifts. This confirms our observations made based on signed graphs following the SBM.

Experiments with SSBM. We begin with experiments based on signed graphs following the SSBM. The corresponding results are presented in Fig. 3.8. We study the performance of the power mean Laplacians $L_{-1}, L_{-2}, L_{-5}, L_{-10}$ with diagonal shifts $\{10^{-10}, 10^{-9}, \ldots, 10^{3}\}$. Moreover, the case where either G^+ or G^- are informative i.e. assortative and disassortative, respectively. In particular, in top (resp. bottom) row of Fig. 3.8 the results correspond to the case where G^+ (resp.G^-) is fixed to be assortative (resp. disassortative).

We can observe that the larger the value of p, the more robust the performance of the corresponding power mean Laplacian L_p to the values of the diagonal shift. For instance, we can see for L_{-1} (see Figs. 5.14(d) and 3.8(e)) that the smaller the diagonal shift, the better the smaller the clustering error, whereas for diagonal shifts $10^0, 10^1, 10^2, 10^3$ its performance clearly deteriorates.

On the other side we can see that the power mean Laplacian L_{-10} presents a high sensibility towards the value of the diagonal shift (see Figs. 5.6(d) and 3.8(h)) where the diagonal shift should be neither too large nor too small, being the values $\{10^{-2}, 10^{-1}, 10^{0}\}$ the more suitable for this particular case. Thiese observations are a verification for the setting with sparse graphs, as it is observed in Fig. 3.9.

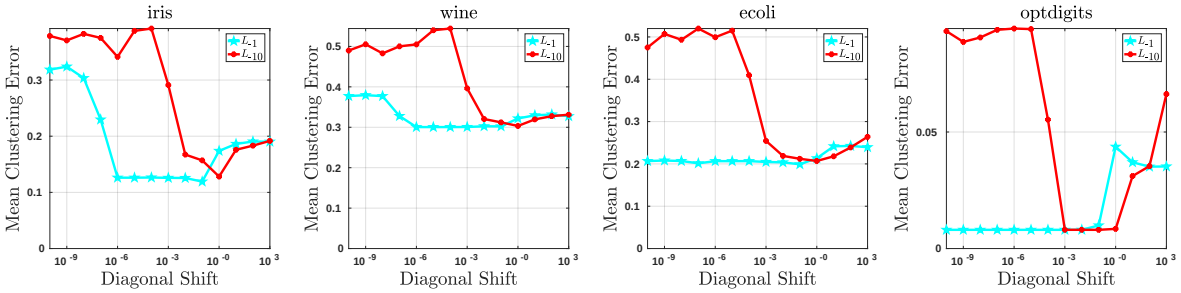
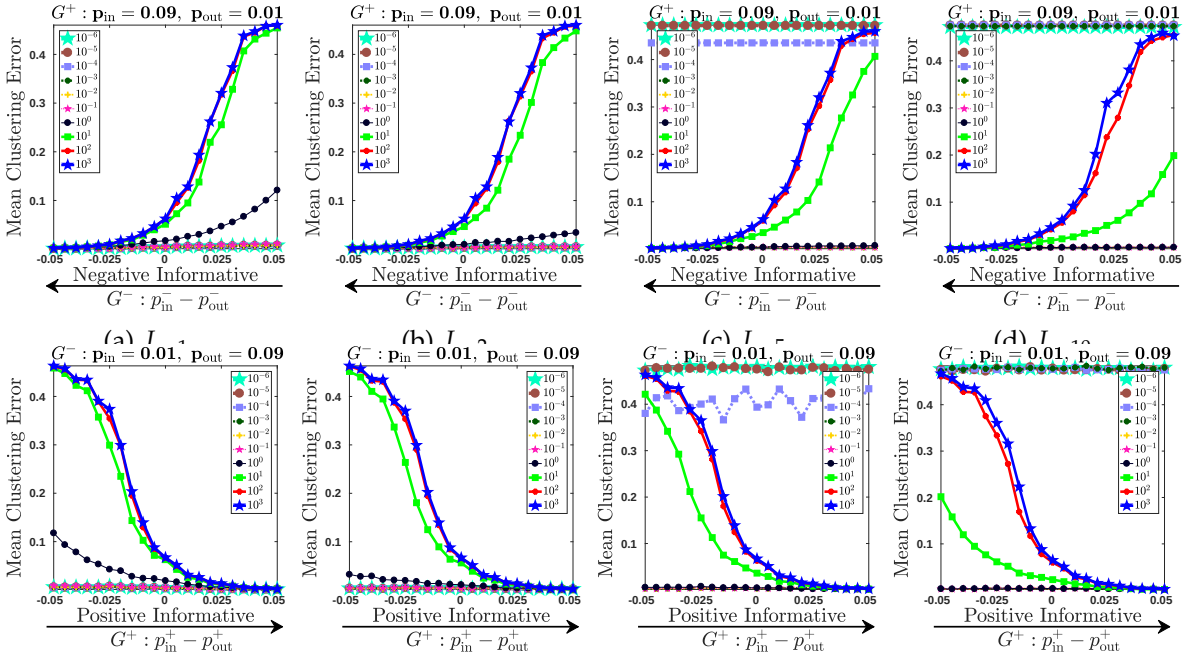


Figure 3.7: Mean clustering error of the power mean Laplacians L_{-1} and L_{-10} with diagonal shifts $\{10^{-10}, 10^{-9}, \ldots, 10^{3}\}$.

On condition number. We now consider a condition number approach to study the effect of the diagonal shift. Recall that the eigenvalue computation scheme considered in this chapter and the corresponding Algorithm 4 are described in Section 3.5. We can observe that the main computation steps are related to the matrix vector operations $(L^{+}_{\text{sym}})^{p}\mathbf{x}_k$ and $(Q^{-}_{\text{sym}})^{p}\mathbf{x}_k$ with $p < 0$. We highlight that this framework considers only the case where $p < 0$.

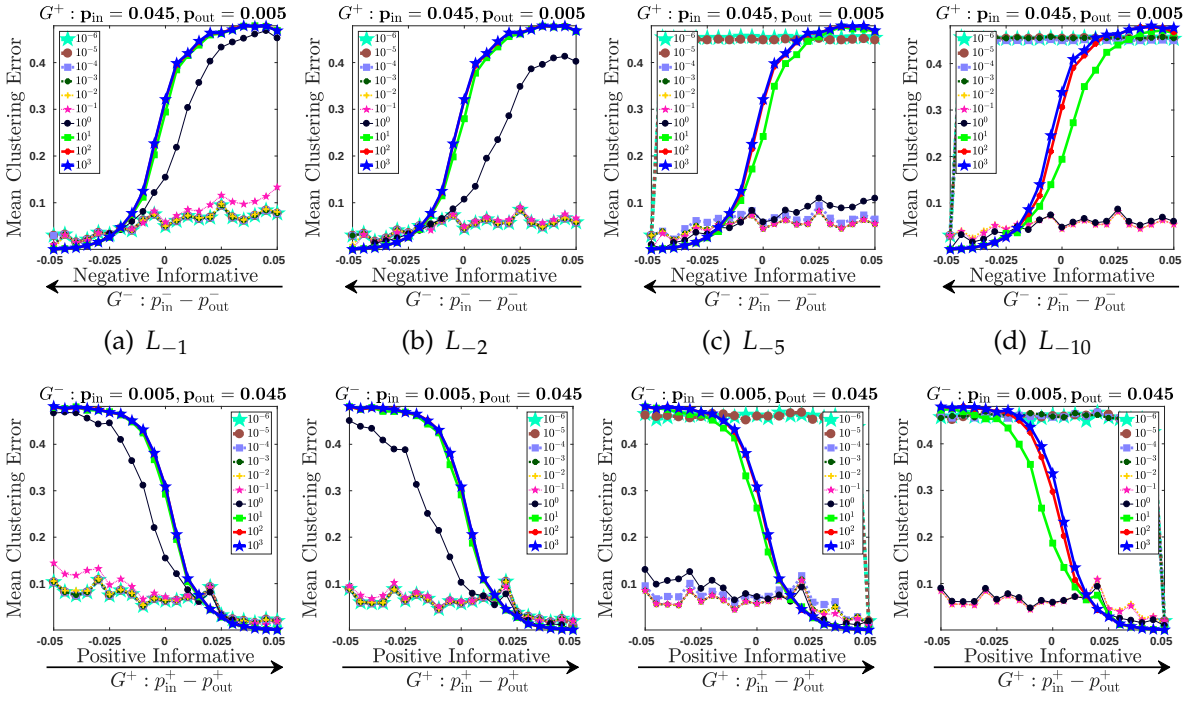
Observe that in the operation $(L^{+}_{\text{sym}})^{p}\mathbf{x}_k$, with $p < 0$, the condition number plays an influential role due to the inverse operation implied by the negativity of p. Note that the eigenvalues of the normalized Laplacians L^{+}_{sym} are contained in the interval $[0, 2]$, hence, it is a singular matrix. As mentioned in the definition of the power mean Laplacian in Eq. 3.7, a suitable diagonal shift is necessary for the case where $p < 0$. Hence, the eigenvalues of the shifted Laplacian $L^{+}_{\text{sym}} + \mu I$ are contained in the interval $[\mu, 2 + \mu]$, therefore, condition number is equal to $\frac{\lambda_{\max}(L^{+}_{\text{sym}})}{\lambda_{\min}(L^{+}_{\text{sym}})}$ which in this case reduces to $\frac{2+\mu}{\mu}$. Thus, it follows that the condition number of $(L^{+}_{\text{sym}} + \mu I)^{p}$ is $g(\mu, p) := \left(\frac{2+\mu}{\mu}\right)^{|p|}$. It is easy to see that $\frac{2+\mu}{\mu} > 1$ and hence $g(\mu, p)$ grows with larger values of $|p|$, hence the condition number is larger for smaller values of the power mean Laplacian. Moreover, the growth rate of $g(\mu, p)$ is larger for smaller values of μ, suggesting that the shift μ should be set as large as possible. Yet, very large values of μ overcome the information contained in the Laplacian matrix. Hence, the diagonal shift should not be too small (due to numerical stability) and should not be too large (due to information obfuscation). This confirms the behavior presented in Figs. 3.8, 3.9 and 3.7.



(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

(e) L_{-1} (f) L_{-2} (g) L_{-5} (h) L_{-10}

Figure 3.8: Mean clustering error under SBM for different diagonal shifts with sparsity 0.1. Details in Sec. 3.6.3.



(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

(e) L_{-1} (f) L_{-2} (g) L_{-5} (h) L_{-10}

Figure 3.9: Mean clustering error under SBM for different diagonal shifts with sparsity 0.05. Details in Sec. 3.6.3.

3.7 CONCLUSION

In this chapter we have addressed the task of clustering signed graphs that encode both positive and negative edges. We have introduced the Signed Power Mean Laplacian, which is a one-parameter family of matrix means that has as particular cases the arithmetic, log-euclidean and harmonic matrix means of the Laplacian of positive edges and the signless Laplacian of negative edges. We have discussed that state of the art approaches can be seen as some sort of arithmetic mean of suitable Laplacians, and we have shown that the arithmetic mean is suboptimal under a version of the stochastic block model for signed graphs. We have presented an analysis in expectation with verifications on sampled random graphs following a suitable stochastic block model. Moreover, we have shown that the eigenvalues and eigenvectors of the Signed Power Mean Laplacian concentrate around their expectation. Finally, through extensive numerical experiments we have verified that our approach is competitive with the state of the art.

In the next chapter we will consider the case of clustering on multilayer graphs, where multiple kinds of interactions are encoded, yet, contrary to signed graphs, all interactions encode some sort of similarity.

4 SPECTRAL CLUSTERING OF MULTILAYER GRAPHS VIA MATRIX POWER MEANS

In the previous chapter we have studied the case of clustering in signed graphs that encode two particular kinds of interactions. In this chapter we study clustering on multilayer graphs that encode different kinds of interactions between the same set of entities. In a similar way as in the signed graph case, one of the challenges with multilayer graphs is how to merge the information from different layers in a meaningful way. In this chapter we introduce a one-parameter family of matrix power means for merging the Laplacians from different layers and analyze it in expectation with several multilayer graph extensions of the stochastic block model. We show that this family allows to recover ground truth clusters under different settings and verify this in real-world data. While computing the matrix power mean can be very expensive for large graphs, we introduce a numerical scheme to efficiently compute its eigenvectors for the case of large sparse graphs.

4.1 INTRODUCTION

Multilayer graphs have received an increasing amount of attention due to their capability to encode different kinds of interactions between the same set of entities (Boccaletti *et al.*, 2014; Kivelä *et al.*, 2014). This kind of graphs arise naturally in diverse applications such as transportation networks (Gallotti and Barthelemy, 2015), financial-asset markets (Bazzi *et al.*, 2016), temporal dynamics (Taylor *et al.*, 2017, 2016), semantic world clustering (Sedoc *et al.*, 2017), multi-video face analysis (Cao *et al.*, 2015), mobile phone networks (Kiukkonen *et al.*, 2010), social balance (Cartwright and Harary, 1956), citation analysis (Tang *et al.*, 2009), and many others.

The extension of clustering techniques to multilayer graphs is a challenging task and several approaches have been proposed so far. For an overview see (Kim and Lee, 2015; Sun, 2013; Xu *et al.*, 2013; Zhao *et al.*, 2017b). For instance, (Dong *et al.*, 2012, 2014; Tang *et al.*, 2009; Zhao *et al.*, 2017a) rely on matrix factorizations, whereas (De Bacco *et al.*, 2017; Paul and Chen, 2016a; Peixoto, 2015; Schein *et al.*, 2015, 2016) take a Bayesian inference approach, and (Kumar and III, 2011; Kumar *et al.*, 2011) enforce consistency among layers in the resulting clustering assignment. In (Mucha *et al.*, 2010; Paul and Chen, 2016b; Wilson *et al.*, 2017) Newman's modularity (Newman, 2006) is extended to multilayer graphs. Recently (De Domenico *et al.*, 2015; Stanley *et al.*, 2016) proposed to compress a multilayer graph by combining sets of similar layers (called 'strata') to later identify the corresponding communities. Of

particular interest to our work is the popular approach (Argyriou *et al.*, 2006; Chen and Hero, 2017; Huang *et al.*, 2012; Taylor *et al.*, 2017; Zhou and Burges, 2007) that first blends the information of a multilayer graph by finding a suitable weighted arithmetic mean of the layers and then applies standard clustering methods to the resulting mono-layer graph.

In this chapter we focus on extensions of spectral clustering to multilayer graphs. For a brief introduction to spectral clustering please refer to Section 2.1. We propose to blend the information of a multilayer graph by taking certain matrix power means of Laplacians of the layers.

The power mean of scalars is a general family of means that includes as special cases, the arithmetic, geometric and harmonic means (see Section 2.3.1). The arithmetic mean of Laplacians has been used before in the case of signed networks (Kunegis *et al.*, 2010) and thus our family of matrix power means is a natural extension of that approach. One of our main contributions is to show that the arithmetic mean is actually suboptimal to merge information from different layers.

We analyze the family of matrix power means in the Stochastic Block Model (SBM) for multilayer graphs in two settings, see Section 4.4. In the first one all the layers are informative, whereas in the second setting none of the individual layers contains the full information except for the case when considered all together. We show that as the parameter of the family of Laplacian means tends to $-\infty$, in expectation one can recover perfectly the clusters in both situations. We provide extensive experiments which show that this behavior is stable when one samples sparse graphs from the SBM. Moreover, in Section 4.6, we provide additional experiments on real-world graphs which confirm our finding in the SBM.

Similarly to the previous chapter, a main challenge for our approach is that the matrix power mean of sparse matrices is in general dense and thus does not scale to large sparse networks in a straightforward fashion. Thus a further contribution of this chapter in Section 4.5 is to show that the first few eigenvectors of the matrix power mean can be computed efficiently. Our algorithm combines the power method with a Krylov subspace approximation technique and allows to compute the extremal eigenvalues and eigenvectors of the power mean of matrices without ever computing the matrix itself.

4.2 RELATED WORK

A common assumption among clustering methods for multilayer graphs is that a sensible clustering can be obtained by taking any single layer, and that clustering information across layers is consistent, in the sense that each layer taken individually basically generates the same clustering. This notion has led to different extensions of clustering to multilayer graphs, like those based on co-training (Kumar and III, 2011) and co-regularization (Kumar *et al.*, 2011), where the goal is to generate an embedding of the set of nodes that is enforced to be aligned to the embedding that each layer provides.

Joint matrix factorizations provide another approach related to the assumption previously described, where the goal was to find a factorization of the layers of a multilayer graph, such that the underlying information shared in common by the layers is obtained, to later apply a suitable clustering method. For instance, a joint matrix factorization approach can be performed on the adjacency matrices of the layers Tang *et al.* (2009) or on the eigenvectors corresponding to the smallest eigenvalues of the Laplacians of the layers Dong *et al.* (2012).

Modularity-based approaches are closely related to spectral clustering. In these techniques the goal is to identify clusters such that their edge density is larger than certain reference null model. Several extensions of the notion of modularity for multilayer graphs have been recently introduced (Mucha *et al.*, 2010; Paul and Chen, 2016b; Wilson *et al.*, 2017). In contrast to spectral clustering, modularity-based approaches do not require to pre-specify the number of clusters.

Another line of work that recently has gained a great amount of attention is based on Bayesian inference (De Bacco *et al.*, 2017; Peixoto, 2015; Schein *et al.*, 2015, 2016; Jenatton *et al.*, 2012). This methods rely on suitable assumptions of the distribution of the interactions encoded in a multilayer graph, leading to a likelihood function whose optimization not only outputs a clustering, but tells as well to what degree the observed clustering structure steps away from randomness, together with model insights regarding the dynamics of the multilayer graph (Peixoto, 2019).

Finally, extensions of spectral clustering in general aim at identifying a multilayer graph operator that blends the information encoded in a multilayer graph such that the eigenvectors corresponding to the smallest eigenvalues are informative about the clustering structure. Several of these extensions rely on some sort of arithmetic mean, for instance the Laplacian of the average adjacency matrix, or the average Laplacian matrix (Paul and Chen, 2020). Further examples are (Zhou and Burges, 2007) which is motivated through the notion of multilayer graphs cuts, and (Chen and Hero, 2017) which identifies optimal convex combinations of layers based on graph noise models.

The arithmetic mean can be seen as a particular case of a more general family of means. In the next section we introduce the Power Mean Laplacian, which is based on a one-parameter family of matrix means called matrix power means, and which in the scalar case has the arithmetic, geometric and harmonic means as particular cases (see Section 2.3 for a brief overview on power means).

4.3 THE POWER MEAN LAPLACIAN

Let $V = \{v_1, \dots, v_n\}$ be a set of nodes and let T the number layers, represented by adjacency matrices $\mathbb{W} = \{W^{(1)}, \dots, W^{(T)}\}$. For each non-negative weight matrix $W^{(t)} \in \mathbb{R}_+^{n \times n}$ we have a graph $G^{(t)} = (V, W^{(t)})$ and a multilayer graph is the set $\mathbb{G} = \{G^{(1)}, \dots, G^{(T)}\}$.

In this chapter our main focus are assortative graphs. This kind of graphs are used to model the situation where edges carry *similarity* information of pairs of vertices

Algorithm 6: Spectral clustering of multilayer graphs with L_p

Input: Symmetric matrices $W^{(1)}, \ldots, W^{(T)}$, number k of clusters to construct.
Output: Clusters $C_1, \ldots, C_k$.
1 Compute eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_{k'}$ corresponding to the k smallest eigenvalues of L_p.
2 Set $U = (\mathbf{u}_1, \ldots, \mathbf{u}_k)$ and cluster the rows of U with k-means into clusters $C_1, \ldots, C_k$.

and thus are indicative for vertices being in the same cluster. For an assortative graph $G = (V, W)$, spectral clustering is typically based on the Laplacian matrix and its normalized version, defined respectively as

$$L = D - W \qquad L_{\text{sym}} = D^{-1/2} L D^{-1/2}$$

where $D_{ii} = \sum_{j=1}^{n} w_{ij}$ is the diagonal matrix of the degrees of G. Both Laplacians are symmetric positive semidefinite and the multiplicity of eigenvalue 0 is equal to the number of connected components in G. For more details please see Section 2.1.

Given a multilayer graph with all assortative layers $G^{(1)}, \ldots, G^{(T)}$, our goal is to come up with a clustering of the vertex set V. We point out that in this chapter a clustering is a partition of V, that is each vertex is uniquely assigned to one cluster.

We consider the multilayer graph $\mathbb{G} = (G^{(1)}, \ldots, G^{(T)})$ and define the **power mean Laplacian** L_p of $\mathbb{G}$ as

$$L_p = M_p(L_{\text{sym}}^{(1)}, \ldots, L_{\text{sym}}^{(T)}) = \left(\frac{1}{T} \sum_{i=1}^{T} (L_{\text{sym}}^{(i)})^p \right)^{1/p} \tag{4.1}$$

where $L_{\text{sym}}^{(t)}$ is the normalized Laplacian of the graph $G^{(t)}$. Note that Definition 2.1 of the matrix power mean $M_p(A_1, \ldots, A_T)$ requires $A_1, \ldots, A_T$ to be positive definite. As the normalized Laplacian is positive semi-definite, in the following, for $p \leq 0$ we add to $L_{\text{sym}}^{(t)}$ in Equation (4.1) a small diagonal shift which ensures positive definiteness, that is we consider $L_{\text{sym}}^{(t)} + \varepsilon I$ throughout this chapter. For all numerical experiments we set $\varepsilon = \log(1 + |p|)$ for $p < 0$ and $\varepsilon = 10^{-6}$ for $p = 0$. Abusing notation slightly, we always mean the shifted versions in the following, unless the shift is explicitly stated.

Similar to spectral clustering for a single graph, we propose Alg. 6 for the spectral clustering of multilayer graphs based on the matrix power mean of Laplacians. As in standard spectral clustering, see (von Luxburg, 2007), our Algorithm 6 uses the eigenvectors corresponding to the k smallest eigenvalues of the power mean Laplacian L_p. Thus the relative ordering of the eigenvalues of L_p is of utmost importance. By Lemma 2.1 we know that if $A_i \mathbf{u} = \lambda(A_i)\mathbf{u}$, for $i = 1, \ldots, n$, then the corresponding eigenvalue of the matrix power mean is $m_p(\lambda(A_1), \ldots, \lambda(A_T))$. Hence, the ordering of eigenvalues strongly depends on the choice of the parameter p. In the next section

we study the effect of the parameter p on the ordering of the eigenvectors of L_p for multilayer graphs following the stochastic block model.

4.4 STOCHASTIC BLOCK MODEL ANALYSIS

In this section we present an analysis of the eigenvectors and eigenvalues of the power mean Laplacian under the Stochastic Block Model (SBM) for multilayer graphs. The SBM is a widespread random graph model for single-layer networks with a prescribed clustering structure (Rohe *et al.*, 2011). Studies on community detection for multilayer networks following the SBM can be found in (Han *et al.*, 2015; Heimlicher *et al.*, 2012; Jog and Loh, 2015; Xu *et al.*, 2014, 2020; Yun and Proutiere, 2016).

In order to grasp how different methods identify communities in multilayer graphs following the SBM we will analyze three different settings. In the first setting all layers follow the same node partition (see f.i. (Han *et al.*, 2015)). In this case we study the robustness of the spectrum of the power mean Laplacian when the first layer is informative and the other layers are noise or even contain contradicting information. In the second setting we consider the particularly interesting situation where multilayer-clustering is superior over each individual clustering. More specifically, we consider the case where three clusters are to be found but each layer contains only information about one of them and only considering all of the layers together reveals the information about the underlying cluster structure. In a third setting we go beyond the standard SBM and consider the case where we have a graph partition for each layer, but this partition changes from layer to layer according to a generative model (see f.i.(Bazzi *et al.*, 2016)). However, for the last setting we only provide an empirical study, whereas for the first two settings we analyze the spectrum also analytically.

In the following we denote by $\mathcal{C}_1, \dots, \mathcal{C}_k$ the ground truth clusters that we aim to recover. All the $\mathcal{C}_i$ are assumed to have the same size $|\mathcal{C}|$. Calligraphic letters are used for the expected matrices in the SBM. In particular, for a layer $G^{(t)}$ we denote by $\mathcal{W}^{(t)}$ its expected adjacency matrix, by $\mathcal{D}^{(t)} = \operatorname{diag}(\mathcal{W}^{(t)}\mathbf{1})$ the expected degree matrix and by $\mathcal{L}^{(t)}_{\text{sym}} = I - (\mathcal{D}^{(t)})^{-1/2}\mathcal{W}^{(t)}(\mathcal{D}^{(t)})^{-1/2}$ the expected normalized Laplacian.

4.4.1 Case 1: Robustness to noise where all layers have the same cluster structure

The case where all layers follow a given node partition is a natural extension of the mono-layer SBM to the multilayer setting. This is done by having different edge probabilities for each layer (Han *et al.*, 2015), while fixing the same node partition in all layers. We denote by $p^{(t)}_{\text{in}}$ (resp. $p^{(t)}_{\text{out}}$) the probability that there exists an edge in layer $G^{(t)}$ between nodes that belong to the same (resp. different) clusters. Then $\mathcal{W}^{(t)}_{ij} = p^{(t)}_{\text{in}}$ if v_i, v_j belong to the same cluster and $\mathcal{W}^{(t)}_{ij} = p^{(t)}_{\text{out}}$ if v_i, v_j belong to

different clusters. Consider the following k vectors:

$$\chi_1 = \mathbf{1}, \qquad \chi_i = (k-1)\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}_{\overline{\mathcal{C}_i}}$$

The use of k-means on the embedding induced by the vectors $\{\chi_i\}_{i=1}^k$ identifies the ground truth communities $\{\mathcal{C}_i\}_{i=1}^k$. It turns out that in expectation $\{\chi_i\}_{i=1}^k$ are eigenvectors of the power mean Laplacian L_p. We look for conditions so that they correspond to the k smallest eigenvalues as this implies that our spectral clustering Algorithm 6 recovers the ground truth.

Before addressing the general case, we discuss the case of two layers. For this case we want to illustrate the effect of the power mean by simply studying the extreme limit cases

$$\mathcal{L}_\infty := \lim_{p\to\infty} \mathcal{L}_p \quad \text{and} \quad \mathcal{L}_{-\infty} := \lim_{p\to-\infty} \mathcal{L}_p\,.$$

where $\mathcal{L}_p = M_p(\mathcal{L}_{\text{sym}}^{(1)}, \mathcal{L}_{\text{sym}}^{(2)})$. The next Lemma shows that $\mathcal{L}_\infty$ and $\mathcal{L}_{-\infty}$ are related to the logical operators AND and OR, respectively, in the sense that in expectation $\mathcal{L}_\infty$ recovers the clusters if and only if $G^{(1)}$ **and** $G^{(2)}$ have both clustering structure, whereas in expectation $\mathcal{L}_{-\infty}$ recovers the clusters if and only if $G^{(1)}$ **or** $G^{(2)}$ has clustering structure.

Lemma 4.1. *Let* $\mathcal{L}_p = M_p(\mathcal{L}_{\text{sym}}^{(1)}, \mathcal{L}_{\text{sym}}^{(2)})$.

- $\{\chi_i\}_{i=1}^k$ *correspond to the k smallest eigenvalues of* $\mathcal{L}_\infty$ *iff* $p_{\text{in}}^{(1)} > p_{\text{out}}^{(1)}$ ***and*** $p_{\text{in}}^{(2)} > p_{\text{out}}^{(2)}$.
- $\{\chi_i\}_{i=1}^k$ *correspond to the k smallest eigenvalues of* $\mathcal{L}_{-\infty}$ *iff* $p_{\text{in}}^{(1)} > p_{\text{out}}^{(1)}$ ***or*** $p_{\text{in}}^{(2)} > p_{\text{out}}^{(2)}$.

Proof. The result follows directly from Theorem 4.1 with $T = 2$. □

The following theorem gives general conditions on the recovery of the ground truth clusters in dependency on p and the size of the shift in $\mathcal{L}_p$. Note that, in analogy with Lemma 4.1, as $p \to -\infty$ the recovery of the ground truth clusters is achieved if at least one of the layers is informative, whereas if $p \to \infty$ all of them have to be informative in order to recover the ground truth.

Theorem 4.1. *Let* $\mathcal{L}_p = M_p(\mathcal{L}_{\text{sym}}^{(1)}, ..., \mathcal{L}_{\text{sym}}^{(T)})$ *then* $\chi_1, \ldots, \chi_k$ *correspond to the k-smallest eigenvalues of* $\mathcal{L}_p$ *if and only if*

$$m_p(\boldsymbol{\rho}_\varepsilon) < 1 + \varepsilon,$$

where $(\boldsymbol{\rho}_\varepsilon)_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \varepsilon$, *and* $t = 1, \ldots, T$.
In particular, for $p \to \pm\infty$, *we have*

1. $\chi_1, \ldots, \chi_k$ *correspond to the k-smallest eigenvalues of* $\mathcal{L}_\infty$ *if and only if all layers are informative, i.e.* $p_{\text{in}}^{(t)} > p_{\text{out}}^{(t)}$ *holds for all* $t \in \{1, \ldots, T\}$.
2. $\chi_1, \ldots, \chi_k$ *correspond to the k-smallest eigenvalues of* $\mathcal{L}_{-\infty}$ *if and only if there is at least one informative layer, i.e. there exists a* $t \in \{1, \ldots, T\}$ *such that* $p_{\text{in}}^{(t)} > p_{\text{out}}^{(t)}$.

Proof. We first show that $\chi_1, \ldots, \chi_k$ are eigenvectors of $\mathcal{W}^1, \ldots, W^{(T)}$. For χ_1 we have,

$$\mathcal{W}^{(t)}\chi_1 = \mathcal{W}^{(t)}\mathbf{1} = |\mathcal{C}|(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)})\mathbf{1} = d^{(t)}\mathbf{1} = \lambda_1^{(t)}\mathbf{1}$$

For the remaining vectors $\chi_2, \ldots, \chi_k$ we have

$$\begin{aligned}
\mathcal{W}^{(t)}\chi_i &= \mathcal{W}^{(t)}\big((k-1)\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}_{\overline{\mathcal{C}_i}}\big) \\
&= \mathcal{W}^{(t)}\big(k\mathbf{1}_{\mathcal{C}_i} - (\mathbf{1}_{\mathcal{C}_i} + \mathbf{1}_{\overline{\mathcal{C}_i}})\big) \\
&= \mathcal{W}^{(t)}\big(k\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}\big) \\
&= k|\mathcal{C}|(p_{\text{in}}^{(t)}\mathbf{1}_{\mathcal{C}_i} + p_{\text{out}}^{(t)}\mathbf{1}_{\overline{\mathcal{C}_i}}) - d^{(t)}\mathbf{1} \\
&= k|\mathcal{C}|(p_{\text{in}}^{(t)}\mathbf{1}_{\mathcal{C}_i} + p_{\text{out}}^{(t)}\mathbf{1}_{\overline{\mathcal{C}_i}}) - d^{(t)}(\mathbf{1}_{\mathcal{C}_i} + \mathbf{1}_{\overline{\mathcal{C}_i}}) \\
&= |\mathcal{C}|(kp_{\text{in}}^{(t)} - d^{(t)})\mathbf{1}_{\mathcal{C}_i} + |\mathcal{C}|(kp_{\text{out}}^{(t)} - d^{(t)})\mathbf{1}_{\overline{\mathcal{C}_i}} \\
&= |\mathcal{C}|(k-1)(p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})\mathbf{1}_{\mathcal{C}_i} - |\mathcal{C}|(p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})\mathbf{1}_{\overline{\mathcal{C}_i}} \\
&= |\mathcal{C}|(p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})\big((k-1)\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}_{\overline{\mathcal{C}_i}}\big) \\
&= |\mathcal{C}|(p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})\chi_i \\
&= \lambda_i\chi_i
\end{aligned}$$

Thus, we have shown that $\chi_1, \ldots, \chi_k$ are eigenvectors of $\mathcal{W}^{(1)}, \ldots, W^{(T)}$. In particular, we have seen that

$$\lambda_1^{(t)} = |\mathcal{C}|(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}), \quad \lambda_i^{(t)} = |\mathcal{C}|(p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})$$

for $i = 2, \ldots, k$. Further, as matrices $\mathcal{W}^{(1)}, \ldots, W^{(T)}$ share all their eigenvectors, they are simultaneously diagonalizable, that is there exists a non-singular matrix Σ such that for $t = 1, \ldots, T$ we have $\Sigma^{-1}\mathcal{W}^{(t)}\Sigma = \Lambda^{(t)}$, where $\Lambda^{(t)}$ are diagonal matrices $\Lambda^{(t)} = \text{diag}(\lambda_1^{(t)}, \ldots, \lambda_k^{(t)}, 0, \ldots, 0)$.

As we assume that all clusters are of the same size $|C|$, the expected multilayer graph is a regular graph with degrees $d^{(1)}, \ldots, d^{(T)}$. Hence, the normalized Laplacians of the expected multilayer graph can be expressed as

$$\mathcal{L}_{\text{sym}}^{(t)} = \Sigma(I - \frac{1}{d^{(t)}}\Lambda^{(t)})\Sigma^{-1}$$

Thus, we can observe that

$$\begin{aligned}
\lambda_1^{(t)} &:= \lambda_1(\mathcal{L}_{\text{sym}}^{(t)}) = 0, \\
\lambda_i^{(t)} &:= \lambda_i(\mathcal{L}_{\text{sym}}^{(t)}) = 1 - \rho_t, \\
\lambda_j^{(t)} &:= \lambda_j(\mathcal{L}_{\text{sym}}^{(t)}) = 1,
\end{aligned}$$

for $i = 2, \ldots, k$, and $j = k+1, \ldots, |V|$, where

$$\rho_t = (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)}) / (p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)})$$

By obtaining the power mean Laplacian on diagonally shifted matrices,

$$\mathcal{L}_p = M_p(\mathcal{L}_{\text{sym}}^{(1)} + \varepsilon I, ..., \mathcal{L}_{\text{sym}}^{(T)} + \varepsilon I)$$

we have by Lemma 2.1

$$\begin{aligned} \lambda_1(\mathcal{L}_p) &= m_p(\lambda_1^{(1)} + \varepsilon, \ldots, \lambda_1^{(T)} + \varepsilon) = m_p(\varepsilon) \\ \lambda_i(\mathcal{L}_p) &= m_p(1 - \rho_1 + \varepsilon, \ldots, 1 - \rho_T + \varepsilon) \\ \lambda_j(\mathcal{L}_p) &= m_p(\lambda_j^{(1)} + \varepsilon, \ldots, \lambda_j^{(T)} + \varepsilon) = 1 + \varepsilon \end{aligned} \tag{4.2}$$

Observe that $\lambda_j(\mathcal{L}_p)$, with $j = k+1, \ldots, |V|$, corresponds to eigenvectors that do not yield an informative embedding. Hence, we do not want this eigenvalue to belong to the bottom of the spectrum of $\mathcal{L}_p$. Thus, for the case of $\boldsymbol{\chi}_2, \ldots, \boldsymbol{\chi}_k$, we can see that they will be located at the bottom of the spectrum if the following condition holds:

$$\lambda_i(\mathcal{L}_p) = m_p(1 - \rho_1 + \varepsilon, \ldots, 1 - \rho_T + \varepsilon) = m_p(\boldsymbol{\rho}_\varepsilon) < 1 + \varepsilon = \lambda_j(\mathcal{L}_p)$$

It remains to analyze the case of the constant eigenvector $\boldsymbol{\chi}_1$. Note that its associated eigenvalue $\lambda_1(\mathcal{L}_1)$ has the following relationship to the non-informative eigenvectors:

$$\lambda_1(\mathcal{L}_1) = \varepsilon < 1 + \varepsilon = \lambda_j(\mathcal{L}_p)$$

which trivially holds, leading to the desired result.

For the limit cases we can observe that $\lim_{p\to\infty} m_p(x) = \max\{x_1, \ldots, x_T\}$ and $\lim_{p\to-\infty} m_p(x) = \min\{x_1, \ldots, x_T\}$. Thus, $m_\infty(\boldsymbol{\rho}_\varepsilon) < 1 + \varepsilon$ if and only if $\rho_t > 0$ for all $t = 1, \ldots, T$. The case for $p \to -\infty$ is analogous: $m_{-\infty}(\boldsymbol{\rho}_\varepsilon) < 1 + \varepsilon$ if and only if $\rho_t > 0$ for at least one t in $t = 1, \ldots, T$. □

Theorem 4.1 shows that the informative eigenvectors of $\mathcal{L}_p$ are at the bottom of the spectrum if and only if the scalar power mean of the corresponding eigenvalues is small enough. Since the scalar power mean is monotonically decreasing with respect to p, this explains why the limit case $p \to \infty$ is more restrictive than $p \to -\infty$. The corollary below shows that the coverage of parameter settings in the SBM for which one recovers the ground truth becomes smaller as p grows.

Corollary 4.1. *Let $q \leq p$. If $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_k$ correspond to the k-smallest eigenvalues of $\mathcal{L}_p$, then $\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_k$ correspond to the k-smallest eigenvalues of $\mathcal{L}_q$.*

Proof. If $\lambda_1, \ldots, \lambda_k$ are among the k-smallest eigenvalues of $\mathcal{L}_p$, then by Theorem 4.1, we have $m_p(\boldsymbol{\rho}_\varepsilon) < 1 + \varepsilon$. As m_p is monotone in the parameter p (see Theorem 2.1) we have $m_q(\boldsymbol{\rho}_\varepsilon) \leq m_p(\boldsymbol{\rho}_\varepsilon)$, Theorem 4.1 concludes the proof.

□

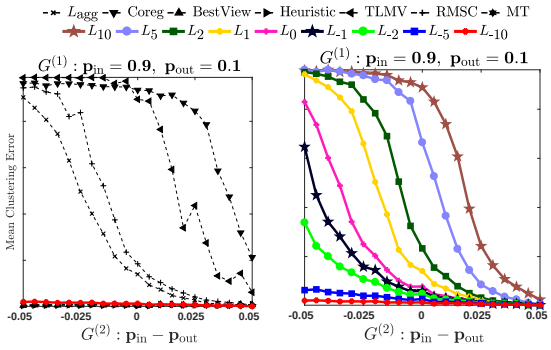


Figure 4.1: Mean Clustering Error under the SBM with two clusters. First layer $G^{(1)}$ is *assortative*. Second layer $G^{(2)}$ transitions from disassortative to assortative. Fig. 4.1(a): Comparison of L_{-10} with state of art. Fig. 4.1(b): Performance of L_p with $p \in \{0, \pm 1, \pm 2, \pm 5, \pm 10\}$.

The previous results hold in expectation. The following experiments show that these findings generalize to the case where one samples from the SBM. In Fig. 4.1 we present experiments on sparse sampled multilayer graphs from the SBM. We consider two clusters of size $|\mathcal{C}| = 100$ and show the mean of clustering error of 50 runs. We evaluate the power mean Laplacian L_p with $p \in \{0, \pm 1, \pm 2, \pm 5, \pm 10\}$ and compare with other methods described in Section 4.6.

In Fig. 4.1 we fix the first layer $G^{(1)}$ to be strongly assortative and let the second layer $G^{(2)}$ run from a disassortative to an assortative configuration. In Fig.4.1(a) we can see that the power mean Laplacian L_{-10} returns the smallest clustering error, together with the multitensor method, the best single view and the heuristic approach across all parameter settings. The latter two work well by construction in this setting. However, we will see that they fail for the second setting we consider next. All the other competing methods fail as the second graph $G^{(2)}$ becomes non-informative respesctively even violates the assumption to be assortative. In Fig. 4.1(b) we can see that the smaller the value of p, the smaller the clustering error of the power mean Laplacian L_p, as stated in Corollary 4.1.

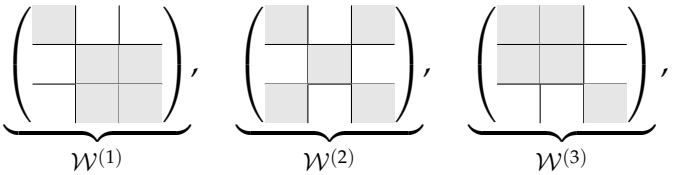
4.4.2 Case 2: No layer contains full information on the clustering structure

We consider a multilayer SBM setting where each individual layer contains only information about one of the clusters and only considering all the layers together reveals the complete cluster structure. For this particular instance, all power mean Laplacians $\mathcal{L}_p$ allow to recover the ground truth for any non-zero integer p.

For the sake of simplicity, we limit ourselves to the case of three layers and three clusters, showing an assortative behavior in expectation. Let the expected adjacency matrix $\mathcal{W}^{(t)}$ of layer $G^{(t)}$ be defined by

$$\mathcal{W}^{(t)}_{i,j} = \begin{cases} p_{\text{in}}, & v_i, v_j \in \mathcal{C}_t \text{ or } v_i, v_j \in \overline{\mathcal{C}_t} \\ p_{\text{out}}, & \text{else} \end{cases} \tag{4.3}$$

for $t = 1, 2, 3$. Note that the three expected adjacency matrices have the form



where each (block) row and column corresponds to a cluster $\mathcal{C}_i$ and gray blocks correspond to nodes whose probability of connections is p_{in}, whereas white blocks correspond to nodes whose probability of connections is p_{out}. Let us assume an assortative behavior on all the layers, that is $p_{\text{in}} > p_{\text{out}}$. In this case spectral clustering applied on a single layer $\mathcal{W}^{(t)}$ would return cluster $\mathcal{C}_t$ and a random partition of the complement, failing to recover the ground truth clustering $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$. This is shown in the following Theorem.

Theorem 4.2. *If $p_{\text{in}} > p_{\text{out}}$, then for any $t = 1, 2, 3$, there exist scalars $\alpha > 0$ and $\beta > 0$ such that the eigenvectors of $\mathcal{L}_{\text{sym}}^{(t)}$ corresponding to the two smallest eigenvalues are*

$$\chi_1 = \alpha \mathbf{1}_{\mathcal{C}_t} + \mathbf{1}_{\overline{\mathcal{C}_t}} \quad \textit{and} \quad \chi_2 = -\beta \mathbf{1}_{\mathcal{C}_t} + \mathbf{1}_{\overline{\mathcal{C}_t}}$$

whereas any vector orthogonal to both χ_1 and χ_2 is an eigenvector for the third smallest eigenvalue.

Proof. Please see Appendix B. □

On the other hand, it turns out that the power mean Laplacian L_p is able to merge the information of each layer, obtaining the ground truth clustering, for all integer powers different from zero. This is formally stated in the following.

Theorem 4.3. *Let $p_{\text{in}} > p_{\text{out}}$ and for $\varepsilon > 0$ define*

$$\tilde{\mathcal{L}}_{\text{sym}}^{(t)} = \mathcal{L}_{\text{sym}}^{(t)} + \varepsilon I, \quad t = 1, 2, 3.$$

Then the eigenvectors of $\mathcal{L}_p = M_p(\tilde{\mathcal{L}}_{\text{sym}}^{(1)}, \tilde{\mathcal{L}}_{\text{sym}}^{(2)}, \tilde{\mathcal{L}}_{\text{sym}}^{(3)})$ corresponding to its three smallest eigenvalues are

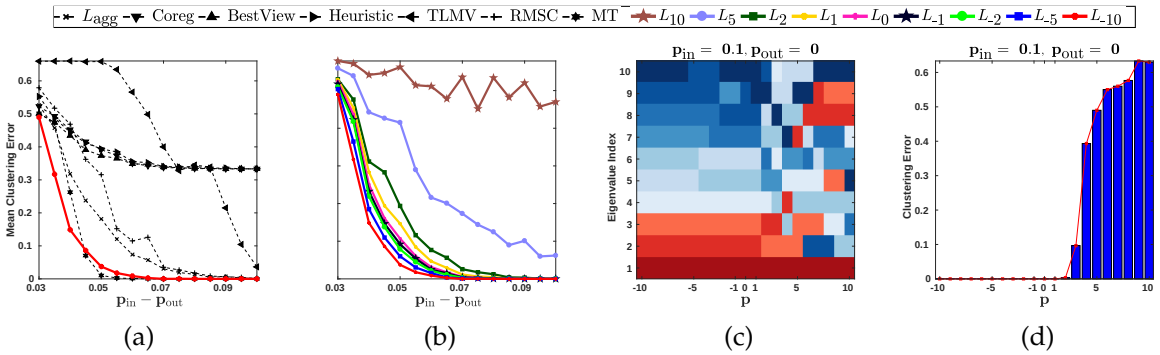
$$\chi_1 = \mathbf{1}, \quad \chi_2 = \mathbf{1}_{\mathcal{C}_2} - \mathbf{1}_{\mathcal{C}_1}, \quad \textit{and} \quad \chi_3 = \mathbf{1}_{\mathcal{C}_3} - \mathbf{1}_{\mathcal{C}_1}$$

for any nonzero integer p.

Proof. Please see Appendix B. □

The proof of Theorem 4.3 is more delicate than the one of Theorem 4.1, as it involves the addition of powers of matrices that do not have the same eigenvectors.

Note that Theorem 4.3 does not distinguish the behavior for distinct values of p. In expectation all nonzero integer values of p work the same. This is different to Theorem 4.1, where the choice of p had a relevant influence on the eigenvector embedding even in expectation. However, we see in the experiments on graphs sampled from the SBM (Figure 4.2) that the choice of p has indeed a significant influence on the performance even though they are the same in expectation. This



(a) (b) (c) (d)

Figure 4.2: SBM experiments with three layers. Each layer is informative with respect to one cluster. 4.2(a): Comparison of L_{-10} with state of art. 4.2(b): Performance of L_p with $p \in \{0, \pm 1, \pm 2, \pm 5, \pm 10\}$. 4.2(c): Eigenvalue ordering of power mean Laplacian L_p across different powers. The ordering clearly changes for powers $p \geq 2$, inducing non-informative eigenvectors to the bottom of the spectrum. 4.2(d): Clustering error of the power mean Laplacian L_p. Clustering error increases with $p \geq 2$, as suggested by ordering changes depicted in 4.2(c).

suggests that the smaller p, the smaller the variance in the difference to the expected behavior in the SBM. We leave as an open problem if such a dependency can be shown analytically.

In Figs. 4.2(a) and 4.2(b) we present the mean clustering error out of ten runs. In Fig. 4.2(a) one can see that BestView and Heuristic, which rely on clusterings determined by single views, return high clustering errors which correspond to the identification of only a single cluster. The result of Theorem 4.3 explains this failure. The reason for the increasing clustering error with p can be seen in Fig. 4.2(c) where we analyze how the ordering of eigenvectors changes for different values of p. We can see that for negative powers, the informative eigenvectors belong to the bottom three eigenvalues (denoted in red). For the cases where $p \geq 2$ the ordering changes, pushing non-informative eigenvectors to the bottom of the spectrum and thus resulting into a high clustering error, as presented in Fig. 4.2(d). However, we conclude that also for this second case a strongly negative power mean Laplacian as L_{-10} works best.

4.4.3 Case 3: Non-consistent partitions between layers

We now consider the case where all the layers follow the same node partition (as in Section 4.4.1), but the partitions may fluctuate from layer to layer with a certain probability. We use the multilayer network model introduced in (Bazzi *et al.*, 2016). This generative model considers a graph partition for each layer, allowing the partitions to change from layer to layer according to an interlayer dependency tensor. For the sake of clarity we consider a one-parameter interlayer dependency tensor with parameter $\tilde{p} \in [0, 1]$ (i.e. a uniform multiplex network according to the

	$\tilde{p}$					
	0.5	0.6	0.7	0.8	0.9	1.0
L_{agg}	0.3	1.3	3.0	8.0	22.3	100.0
Coreg	0.3	0.0	0.3	0.0	0.0	64.7
BestView	9.7	1.0	0.3	0.0	0.7	77.3
Heuristic	0.0	0.0	0.0	0.0	0.3	59.3
TLMV	0.7	0.7	4.0	6.0	24.7	100.0
RMSC	1.0	1.7	4.0	7.0	19.7	100.0
MT	1.3	0.3	0.7	3.0	17.0	100.0
L_{10}	0.0	0.0	0.0	0.0	1.0	100.0
L_5	0.0	0.0	0.0	0.0	5.0	100.0
L_2	0.0	0.0	0.3	2.3	18.3	100.0
L_1	1.0	1.0	3.0	7.0	30.3	100.0
L_0	4.3	4.3	9.7	15.3	38.3	100.0
L_{-1}	6.7	7.7	15.7	16.3	42.3	100.0
L_{-2}	8.0	13.0	20.3	20.7	42.7	100.0
L_{-5}	22.3	23.0	36.3	37.7	50.0	100.0
L_{-10}	**69.0**	**76.3**	**68.0**	**67.3**	**59.7**	100.0

	μ					
	0.0	0.1	0.2	0.3	0.4	0.5
L_{agg}	24.7	21.7	21.3	21.7	24.3	21.3
Coreg	16.7	16.7	13.3	11.7	6.0	1.0
BestView	16.7	17.0	17.0	17.7	11.7	9.0
Heuristic	16.7	16.3	15.0	9.0	2.0	0.7
TLMV	25.7	24.3	21.7	23.3	21.0	20.0
RMSC	26.3	22.0	23.0	21.7	20.3	20.0
MT	19.7	19.7	21.0	20.7	20.7	20.7
L_{10}	16.7	17.3	17.0	16.7	16.7	16.7
L_5	17.0	18.0	17.3	17.7	18.0	17.0
L_2	23.0	21.3	19.3	19.0	20.3	18.0
L_1	26.3	25.3	24.0	23.0	22.3	21.3
L_0	33.3	30.3	28.7	28.0	28.0	23.7
L_{-1}	36.3	33.0	33.3	32.0	29.0	25.0
L_{-2}	37.3	36.3	36.7	34.0	31.3	29.0
L_{-5}	48.0	45.0	49.0	44.3	43.0	40.0
L_{-10}	**71.7**	**72.3**	**72.7**	**74.7**	**76.3**	**72.7**

Table 4.1: Percentage of cases where the minimum clustering error is achieved by different methods. Left: Columns correspond to a fixed value of $\tilde{p}$ and we aggregate over $\mu \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Right: Columns correspond to a fixed value of μ and we aggregate over $\tilde{p} \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$.

notation used in Section 3.B in (Bazzi *et al.*, 2016)), where for $\tilde{p} = 0$ the partitions between layers are independent, and for $\tilde{p} = 1$ the partitions between layers are identical. Once the partitions are obtained, edges are generated following a multilayer degree-corrected SBM (DCSBM in Section 4 of (Bazzi *et al.*, 2016)), according to a one-parameter affinity matrix with parameter $\mu \in [0, 1]$, where for $\mu = 0$ all edges are within communities whereas for $\mu = 1$ edges are assigned ignoring the community structure.

We choose $\tilde{p} \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and $\mu \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ and consider all possible combinations of $(\tilde{p}, \mu)$. For each pair we count how many times, out of 50 runs, each method achieves the smallest clustering error. The remaining parameters of the DCSBM are set as follows: exponent $\gamma = -3$, minimum degree and maximum degree $k_{min} = k_{max} = 10$, $|V| = 100$ nodes, $T = 10$ layers and $K = 2$ communities. As partitions between layers are not necessarily the same, we take the most frequent node assignment among all 10 layers as ground truth clustering.

In Table 4.1, left side, we show the results for fixed values of $\tilde{p}$ and average over all values of μ. In the right table we show the corresponding results for fixed values of μ and average over all values of $\tilde{p}$. In the left table we can see that for $\tilde{p} = 1$, where the partition is the same in all layers, all methods recover the clustering, while, as one would expect, the performance decreases with smaller values of $\tilde{p}$. Further, we note that the performance of the power mean Laplacian improves as $\tilde{p}$ decreases and L_{-10} again achieves the best result. In the right table we see that performance is degrading with larger values of μ. This is expected as for larger values of μ the edges inside the clusters are less concentrated. Again the performance of the power mean Laplacian improves as p decreases and L_{-10} performs best.

4.5 COMPUTING THE SMALLEST EIGENVALUES AND EIGENVECTORS OF MATRIX POWER MEANS

We present an efficient method for the computation of the smallest eigenvalues of $M_p(A_1, \dots, A_T)$ which does not require the computation of the matrix $M_p(A_1, \dots, A_T)$. This is particularly important when dealing with potential large-scale problems as $M_p(A_1, \dots, A_T)$ is typically dense even though each A_i is a sparse matrix. We restrict our attention to the case $p < 0$ which is the most interesting one in practice. The positive case $p > 0$ as well as the limit case $p \to 0$ deserve a different analysis and are not considered here.

Let $A_1, \dots, A_T$ be positive definite matrices. If $\lambda_1 \leq \cdots \leq \lambda_n$ are the eigenvalues of $M_p(A_1, \dots, A_T)$ corresponding to the eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_n$, then $\mu_i = (\lambda_i)^p$, $i = 1, \dots, n$, are the eigenvalues of $M_p(A_1, \dots, A_T)^p$ corresponding to the eigenvectors $\mathbf{u}_i$. However, the function $f(x) = x^p$ is order reversing for $p < 0$. Thus, the relative ordering of the μ_i's changes into $\mu_1 \geq \cdots \geq \mu_n$. Thus, the smallest eigenvalues and eigenvectors of $M_p(A_1, \dots, A_T)$ can be computed by addressing the largest ones of $M_p(A_1, \dots, A_T)^p$. To this end we propose a power method type outer-scheme, combined with a Krylov subspace approximation inner-method. The pseudo code is presented in Algs. 7 and 8. Each step of the outer iteration in Alg. 7 requires to compute the pth power of T matrices times a vector. Computing $A^p \times \mathit{vector}$, reduces to the problem of computing the product of a matrix function times a vector. Krylov methods are among the most efficient and most studied strategies to address such a computational issue. As A^p is a polynomial in A, we apply a Polynomial Krylov Subspace Method (PKSM), whose pseudo code is presented in Alg. 8 and which we briefly describe in the following. For further details we refer to (Higham, 2008) and the references therein. For the sake of generality, below we describe the method for a general positive definite matrix A.

The general idea of PKSM s-th iteration is to project A onto the subspace $\mathbb{K}^s(A, \mathbf{y}) = \text{span}\{\mathbf{y}, A\mathbf{y}, \dots, A^{s-1}\mathbf{y}\}$ and solve the problem there. The projection onto $\mathbb{K}^s(A, \mathbf{y})$ is realized by means of the Lanczos process, producing a sequence of matrices V_s with orthogonal columns, where the first column of V_s is $\mathbf{y}/\,||\mathbf{y}||_2$ and $\text{range}(V_s) = \mathbb{K}^s(A, \mathbf{y})$. Moreover at each step we have $AV_s = V_s H_s + \mathbf{v}_{s+1}\mathbf{e}_s^T$ where H_s is $s \times s$ symmetric tridiagonal, and $\mathbf{e}_i$ is the i-th canonical vector. The matrix vector product $\mathbf{x} = A^p\mathbf{y}$ is then approximated by $\mathbf{x}_s = V_s(H_s)^p\mathbf{e}_1\|\mathbf{y}\| \approx A^p\mathbf{y}$.

Clearly, if operations are done with infinite precision, the exact $\mathbf{x}$ is obtained after n steps. However, in practice, the error $\|\mathbf{x}_s - \mathbf{x}\|$ decreases very fast with s and often very few steps are enough to reach a desirable tolerance. Two relevant observations are in order: first, the matrix $H_s = V_s^T A V_s$ can be computed iteratively alongside the Lanczos method, thus it does not require any additional matrix multiplication; second, the p power of the matrix H_s can be computed directly without any notable increment in the algorithm cost, since H_s is tridiagonal of size $s \times s$.

Algorithm 7: PM applied to M_p.

Input: $\mathbf{x}_0$, $p < 0$
Output: Eigenpair $(\lambda, \mathbf{x})$ of M_p

repeat
  $\mathbf{u}_k^{(1)} \leftarrow (A_1)^p \mathbf{x}_k$
  $\vdots$
  $\mathbf{u}_k^{(T)} \leftarrow (A_T)^p \mathbf{x}_k$
  $\mathbf{y}_{k+1} \leftarrow \frac{1}{T} \sum_{i=1}^{T} \mathbf{u}_k^{(i)}$
  $\mathbf{x}_{k+1} \leftarrow \mathbf{y}_{k+1} / \|\mathbf{y}_{k+1}\|_2$
until *tolerance reached*
$\lambda \leftarrow (\mathbf{x}_{k+1}^T \mathbf{x}_k)^{1/p}$, $\mathbf{x} \leftarrow \mathbf{x}_{k+1}$

Algorithm 8: PKSM for $A^p\mathbf{y}$

Input: $\mathbf{u}_0 = \mathbf{y}$, $V_0 = [\,\cdot\,]$, $p < 0$
Output: $\mathbf{x} = A^p\mathbf{y}$

$\mathbf{v}_0 \leftarrow \mathbf{y} / \|\mathbf{y}\|_2$
for $s = 0, 1, 2, \ldots, n$ **do**
  $\tilde{V}_{s+1} \leftarrow [V_s, \mathbf{v}_s]$
  $V_{s+1} \leftarrow$ Orthogonalize columns of $\tilde{V}_{s+1}$
  $H_{s+1} \leftarrow V_{s+1}^T A V_{s+1}$
  $\mathbf{x}_{s+1} \leftarrow V_{s+1}(H_{s+1})^p \mathbf{e}_1 \|\mathbf{y}\|_2$
  if *tolerance reached* **then** *break*
  $\mathbf{v}_{s+1} \leftarrow A\mathbf{v}_s$
end
$\mathbf{x} \leftarrow \mathbf{x}_{s+1}$

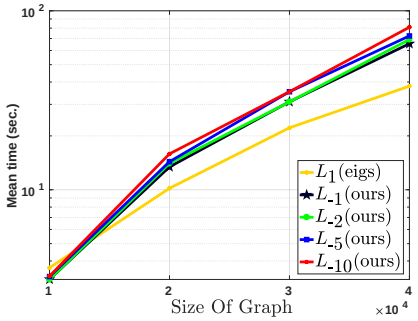


Figure 4.3: Mean execution time of 10 runs for the power mean Laplacian L_p. L_p(ours) stands for the power mean Laplacian together with our proposed Power Method (Alg. 7) based on the Polynomial Krylov Approximation Method (Alg. 8). L_1(eigs) stands for the arithmetic mean Laplacian with eigenvectors computed with Matlab's eigs function. Experiments are performed using one thread.

Several eigenvectors can be simultaneously computed with Algs. 7 and 8 by orthonormalizing the current eigenvector approximation at every step of the power method (Alg. 7) (see f.i. algorithm 5.1 Subspace iteration in (Saad, 2011)). Moreover, the outer iteration in Alg. 7 can be easily run in parallel as the vectors $\mathbf{u}_k^{(i)}$, $i = 1, \ldots, T$ can be built independently of each other.

A numerical evaluation of Algs. 7 and 8 is presented in Fig. 4.3. We consider graphs of sizes $|V| \in \{1\times10^4, 2\times10^4, 3\times10^4, 4\times10^4\}$. Further, for each multilayer graph we generate two assortative graphs with parameters $p_{\text{in}} = 0.05$ and $p_{\text{in}} = 0.025$, following the SBM. Moreover, we consider the power mean Laplacian $L_p = M_p(L_{\text{sym}}^{(1)}, L_{\text{sym}}^{(2)})$ with parameter $p \in \{-1, -2, -5, -10\}$. As a baseline we take the arithmetic mean Laplacian $L_1 = M_1(L_{\text{sym}}^{(1)}, L_{\text{sym}}^{(2)})$ and use Matlab's eigs function. For all cases, we compute the two eigenvectors corresponding to the smallest eigenvalues. We present the mean execution time of 10 runs. Experiments are performed using one thread.

4.6 EXPERIMENTS

We take the following baseline approaches of spectral clustering applied to: the average adjacency matrix ($\mathbf{L_{agg}}$), the arithmetic mean Laplacian ($\mathbf{L_1}$), the layer with

	3Sources	**BBC**	**BBCS**	**Wiki**	**UCI**	**Citeseer**	**Cora**	**WebKB**
# vertices	169	685	544	693	2000	3312	2708	187
# layers	3	4	2	2	6	2	2	2
# classes	6	5	5	10	10	6	7	5
L_{agg}	0.194	0.156	0.152	0.371	0.162	0.373	0.452	**0.277**
Coreg	0.215	0.196	0.164	0.784	0.248	0.395	0.659	0.444
Heuristic	**0.192**	0.218	0.198	0.697	0.280	0.474	0.515	0.400
TLMV	0.284	0.259	0.317	0.412	0.154	0.363	0.533	0.430
RMSC	0.254	0.255	0.194	0.407	0.173	0.422	0.507	0.279
MT	0.249	**0.133**	0.158	0.544	0.103	0.371	0.436	0.298
L_1	0.194	0.154	0.148	0.373	0.163	0.285	**0.367**	0.440
$\boldsymbol{L_{-10}}$ **(ours)**	0.200	0.159	**0.144**	**0.368**	**0.095**	**0.283**	0.374	0.439

Table 4.2: Average Clustering Error

the largest spectral gap (**Heuristic**), and to the layer with the smallest clustering error (**BestView**). Further, we consider: Pairwise Co-Regularized Spectral Clustering (Kumar *et al.*, 2011) with parameter $\lambda = 0.01$ (**Coreg**) which proposes a spectral embedding generating a clustering consistent among all graph layers, Robust Multi-View Spectral Clustering (Xia *et al.*, 2014) with parameter $\lambda = 0.005$ (**RMSC**) which obtains a robust consensus representation by fusing noiseless information present among layers, spectral clustering applied to a suitable convex combination of normalized adjacency matrices (Zhou and Burges, 2007) (**TLMV**), and a tensor factorization method (De Bacco *et al.*, 2017) (**MT**), which considers a multi-layer mixed membership SBM.

We take several datasets: *3sources*(Liu *et al.*, 2013), *BBC*(Greene and Cunningham, 2005) and *BBC Sports*(Greene and Cunningham, 2009) news articles, a dataset of Wikipedia articles(Rasiwasia *et al.*, 2010), the hand written *UCI* digits dataset with six different features and citations datasets *CiteSeer*(Lu and Getoor, 2003), *Cora*(McCallum *et al.*, 2000) and *WebKB*(Craven *et al.*, 2011), (from WebKB we only take the subset Texas). For each layer we build the corresponding adjacency matrix from the k-nearest neighbor graph based on the Pearson linear correlation between nodes, i.e. the higher the correlation the nearer the nodes are. We test all clustering methods over all choices of $k \in \{20, 40, 60, 80, 100\}$ and present the average clustering error in Table 4.2. The datasets CiteSeer, Cora and WebKB have two layers: one is a fixed citation network, whereas the second one is the k-nearest neighbor graph built on documents features. We can see that in four out of eight datasets the power mean Laplacian L_{-10} gets the smallest clustering error. The largest difference in clustering error is present in the UCI dataset, where MT turns out to be the the second best. Further, L_1 presents the smallest clustering error in Cora, being L_{-10} close to it. The smallest clustering error in WebKB is achieved by L_{agg}. This dataset is particularly challenging, due to conflictive layers(He *et al.*, 2017).

4.7 CONCLUSION

In this chapter we have studied the task of clustering on graphs that encode multiple kinds of interactions. We have introduced the Power Mean Laplacian and analyzed it under the stochastic block model for multilayer graphs under different settings. We have shown that in expectation it recovers the ground truth clusters under suitable conditions, and verified them through extensive numerical experiments on random multilayer graphs. Moreover, we have observed that our proposed approach performs no worse than state of the art approaches on real world datasets.

In the next chapter we continue our study of multilayer graphs, yet we will leave the unsupervised task of clustering to enter into semi-supervised learning.

5 SEMI-SUPERVISED LEARNING ON MULTILAYER GRAPHS VIA MATRIX POWER MEANS

In previous chapters we have studied the unsupervised learning task of clustering on signed and multilayer graphs. In this chapter we study the task of semi-supervised learning on multilayer graphs by taking into account both labeled and unlabeled observations together with the information encoded by each individual graph layer. We propose a regularizer based on the matrix power mean, which is a one-parameter family of matrix means that includes the arithmetic, geometric and harmonic means as particular cases. We analyze it in expectation under a Multilayer Stochastic Block Model and verify numerically that it outperforms state of the art methods. Moreover, we introduce a matrix-free numerical scheme based on contour integral quadratures and Krylov subspace solvers that scales to large sparse multilayer graphs.

5.1 INTRODUCTION

The task of graph-based Semi-Supervised Learning (**SSL**) is to build a classifier that takes into account both labeled and unlabeled observations, together with the information encoded by a given graph (Subramanya and Talukdar, 2014; Chapelle *et al.*, 2010). A common and successful approach to this task is to take a suitable loss function on the labeled nodes and a regularizer which provides information encoded by the graph (Zhou *et al.*, 2003; Zhu *et al.*, 2003; Belkin *et al.*, 2004; Yang *et al.*, 2016; Kipf and Welling, 2017). Whereas this task is well studied, traditionally these methods assume that the graph is composed by interactions of one single kind, i.e. only one graph is available.

For the case where multiple graphs, or equivalently, multiple layers are available, the challenge is to boost the classification performance by merging the information encoded in each graph. The arguably most popular approach for this task consists of finding some form of convex combination of graph matrices, where more informative graphs receive a larger weight (Tsuda *et al.*, 2005; Zhou and Burges, 2007; Argyriou *et al.*, 2006; Nie *et al.*, 2016; Karasuyama and Mamitsuka, 2013; Kato *et al.*, 2009; Viswanathan *et al.*, 2019; Ye and Akoglu, 2018).

Note that a convex combination of graph matrices can be seen as a weighted arithmetic mean of graph matrices. In the context of multilayer graph clustering, we have shown in Chapters 3 and 4 that weighted arithmetic means are suboptimal under certain benchmark generative graph models, whereas other matrix means are able to discover clustering structures that the arithmetic means overlook.

In this chapter we study the task of semi-supervised learning with multilayer graphs with a novel regularizer based on the power mean Laplacian. For a brief introduction to matrix power means we refer to Section 2.3. We show that in expectation under a Multilayer Stochastic Block Model, our approach provably correctly classifies unlabeled nodes in settings where state of the art approaches fail. In particular, a limit case of our method is provably robust against noise, yielding good classification performance as long as one layer is informative and the remaining layers are potentially just noise. We verify the analysis in expectation with extensive experiments with random graphs, showing that our approach compares favorably with state of the art methods, yielding a good classification performance on several relevant settings where state of the art approaches fail.

Moreover, our approach scales to large datasets: even though the computation of the power mean Laplacian is in general prohibitive for large graphs, we present a matrix-free numerical scheme based on integral quadrature methods and Krylov subspace solvers which allows us to apply the power mean Laplacian regularizer to large sparse graphs. Finally, we perform numerical experiments on real world datasets and verify that our approach is competitive to state of the art approaches.

Notation. Recall that we define a multilayer graph with T layers as the set $\mathbb{G} = \{G^{(1)}, \dots, G^{(T)}\}$, with each graph layer defined as $G^{(t)} = (V, W^{(t)})$, where $V = \{v_1, \dots, v_n\}$ is the node set and $W^{(t)} \in \mathbb{R}_+^{n \times n}$ is the corresponding adjacency matrix, which we assume symmetric and nonnegative. We further denote the layers' normalized Laplacians as $L_{\text{sym}}^{(t)} = I - (D^{(t)})^{-1/2} W (D^{(t)})^{-1/2}$, where $D^{(t)}$ is the degree diagonal matrix with $(D^{(t)})_{ii} = \sum_{j=1}^{n} W_{ij}^{(t)}$. For a brief introduction to the graph Laplacian please see Section 2.1.

5.2 RELATED WORK

In this section, we give a brief overview of graph-based semi-supervised methods for multilayer graphs. A well-established approach to this task is, for the single-layer case, based on building a classifier that takes into account both labeled and unlabeled observations, together with the information encoded in the graph. The information encoded in the graph is frequently induced through a regularized based on a graph operator, like the graph Laplacian. When taking a quadratic loss, this leads to a linear system of equations (Zhou *et al.*, 2003; Zhu *et al.*, 2003).

This approach has been the motivation of several extensions to multilayer graphs. For instance, (Tsuda *et al.*, 2005) proposes a regularization approach on each of the Laplacians of the layers:

$$\min_{f \in \mathbb{R}^n, \gamma \in \mathbb{R}_+} \|f - y\|_2^2 + c\gamma \qquad \text{s.t. } \; f^T L^{(i)} f \leq \gamma, \text{ for } \; i = 1, \dots, T$$

where the vector $y \in \mathbb{R}^n$ contains the classes of labeled observations and zero otherwise. This approach further leads to a dual problem which is expressed in terms of a weighted arithmetic mean of Laplacians (Tsuda *et al.*, 2005). Further, in (Kato *et al.*, 2009) an approach is proposed which is able to identify informative layers of a multilayer graph using as a regularizer a weighted arithmetic mean of Laplacians:

$$\min_{f \in \mathbb{R}^n} \ \beta_1 \sum_{i=1}^{l} (f_i - y_i)^2 + \beta_2 \left\| f \right\|_2^2 + \beta_3 f^T L(u) f$$

where l is the number of labeled nodes and $L(u) = \sum_{i=1}^{T} u_i L(i)$. The optimal weights of $L(u)$ are identified through an updating rule derived from an Expectation Maximization approach.

In (Karasuyama and Mamitsuka, 2013) the goal is to build a classifier while identifying a sparse linear combination of the layers of the graph by taking a suitable regularization scheme based on a weighted arithmetic mean of Laplacians:

$$\min_{f \in \mathbb{R}^n, \mu \in \mathbb{R}^n_+} \ \sum_{i=1}^{T} \frac{\mu_i}{||L^{(i)}_{\text{sym}}||_F} f^T L^{(i)} f + \lambda_1 \left\| f - y \right\|_2^2 + \lambda_2 \left\| \mu \right\|_2^2 \qquad \text{s.t.} \ \sum_{i=1}^{T} \mu_i = 1$$

(Argyriou *et al.*, 2006) propose an optimal linear combination by taking a convex combination of the pseudo inverse Laplacian into consideration, whereas (Nie *et al.*, 2016), propose a parameter-free method inspired by the graph cut problem, and (Zhou and Burges, 2007) introduce a weighted arithmetic mean of normalized adjacency matrices to later classify node labels by solving a linear system of equations.

A slightly different approach is proposed in (Viswanathan *et al.*, 2019) where the goal is to extend the notion of Gaussian Markov random fields to multilayer graphs through an inverse covariance matrix that is a weighted linear combination of suitable matrices.

Moreover, based on improvements of belief propagation (Koutra *et al.*, 2011), a scalable approximation to multilayer graphs have been proposed in (Eswaran *et al.*, 2017) where the belief corresponds to node labels, whereas in (Gujral and Papalexakis, 2018) a tensor factorization method is designed for semi-supervised learning in multilayer graphs.

We have observed that several extensions to multilayer graphs rely on weighted arithmetic means of suitable matrices. In the next section we introduce a multilayer graph operator that we will use as a regularizer. This multilayer graph operator is a one-parameter family of matrix functions, which has, in the scalar case, the arithmetic, geometric and harmonic means as particular cases.

5.3 SSL WITH THE POWER MEAN LAPLACIAN

In this section we now introduce our multilayer graph regularizer that is based on the power mean Laplacian.

The **Power Mean Laplacian**, as introduced in Chapter 4, is a matrix extension of the scalar power mean (see Section 2.3) applied to the Laplacians of a multilayer graph and proposed as a more robust way to blend the information encoded across the layers. It is defined as

$$L_p = \left(\frac{1}{T} \sum_{i=1}^{T} (L_{\text{sym}}^{(i)})^p \right)^{1/p}$$

where $A^{1/p}$ is the unique positive definite solution of the matrix equation $X^p = A$. For the case $p \leq 0$ a small diagonal shift $\varepsilon > 0$ is added to each Laplacian, i.e. we replace $L_{\text{sym}}^{(i)}$ with $L_{\text{sym}}^{(i)} + \varepsilon$, to ensure that L_p is well defined as suggested in (Bhagwat and Subramanian, 1978). In what follows all the proofs hold for an arbitrary shift. Following Chapter 4, we set $\varepsilon = \log_{10}(1 + |p|) + 10^{-6}$ for $p \leq 0$ in the numerical experiments.

We consider the following optimization problem for the task of semi-supervised learning in multilayer graphs: Given k classes $r = 1, \ldots, k$ and membership vectors $Y^{(r)} \in \mathbb{R}^n$ defined by $Y_i^{(r)} = 1$ if node v_i belongs to class r and $Y_i^{(r)} = 0$ otherwise, we let

$$f^{(r)} = \underset{f \in \mathbb{R}^n}{\arg\min} \, \|f - Y^{(r)}\|^2 + \lambda f^T L_p f \,. \tag{5.1}$$

The final class assignment for an unlabeled node v_i is $y_i = \arg\max\{f_i^{(1)}, \ldots, f_i^{(k)}\}$. Note that the solution f of (5.1), for a particular class r, is such that $(I + \lambda L_p) f = Y^{(r)}$. Equation (5.1) has two terms: the first term is a loss function based on the labeled nodes whereas the second term is a regularization term based on the power mean Laplacian L_p, which accounts for the multilayer graph structure. It is worth noting that the Local-Global approach of (Zhou *et al.*, 2003) is a particular case of our approach when only one layer ($T = 1$) is considered. Moreover, note that when $p = 1$ we obtain a regularizer term based on the arithmetic mean of Laplacians $L_1 = \frac{1}{T} \sum_{i=1}^{T} L_{\text{sym}}^{(i)}$. In the following section we analyze our proposed approach (5.1) under the Multilayer Stochastic Block Model.

5.4 STOCHASTIC BLOCK MODEL ANALYSIS

In this section we provide an analysis of semi-supervised learning for multilayer graphs with the power mean Laplacian as a regularizer under the Multilayer Stochastic Block Model (**MSBM**). The MSBM is a generative model for graphs showing certain prescribed clusters/classes structures via a set of membership parameters $p_{\text{in}}^{(t)}$ and $p_{\text{out}}^{(t)}$, $t = 1, \ldots, T$. These parameters designate the edge probabilities: given the nodes v_i and v_j the probability of observing an edge between them on layer t is $p_{\text{in}}^{(t)}$ (resp. $p_{\text{out}}^{(t)}$), if v_i and v_j belong to the same (resp. different) cluster/class. Note that, unlike the Labeled Stochastic Block Model (Heimlicher *et al.*, 2012), the MSBM

allows multiple edges between the same pairs of nodes across the layers. For SSL with one layer under the SBM we refer the reader to (Saade *et al.*, 2018; Kanade *et al.*, 2016; Mossel and Xu, 2016).

We present an analysis in expectation. We consider k clusters/classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ of equal size $|\mathcal{C}| = n/k$. We denote the layers of a multilayer graph in expectation with calligraphic letters $E(\mathbb{G}) = \{E(G^{(1)}, \ldots, E(G^{(T)})\}$, i.e. $\mathcal{W}^{(t)}$ is the expected adjacency matrix of the t^{th}-layer. We assume that our multilayer graphs are non-weighted, i.e. edges are zero or one, and hence we have $\mathcal{W}_{ij}^{(t)} = p_{\text{in}}^{(t)}$, (resp. $\mathcal{W}_{ij}^{(t)} = p_{\text{out}}^{(t)}$) for nodes v_i, v_j belonging to the same (resp. different) cluster/class.

In order to grasp how different methods classify the nodes in multilayer graphs following the MSBM, we analyze three different settings. In the first setting (Section 5.4.1) all layers have the same class structure and we study the conditions for different regularizers L_p to correctly predict class labels. We further show that our approach is robust against the presence of noise layers, in the sense that it achieves a small classification error when at least one layer is informative and the remaining layers are potentially just noise. In the second setting (Section 5.4.2) we consider the case where different classes of the same size have different number of labels. In the third setting (Section 5.4.3) we consider the case where each layer taken alone would lead to a large classification error whereas considering all the layers together can lead to a small classification error.

5.4.1 Case 1: Robustness to Noise

A common assumption in multilayer semi-supervised learning is that at least one layer encodes relevant information in the label prediction task. The next theorem discusses the classification error of the expected power mean Laplacian regularizer in this setting.

Theorem 5.1. *Let $E(\mathbb{G})$ be the expected multilayer graph with T layers following the multilayer SBM with k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ of equal size and parameters $\left(p_{\text{in}}^{(t)}, p_{\text{out}}^{(t)}\right)_{t=1}^{T}$. Assume the same number of labeled nodes are available per class. Then, the solution of (5.1) yields zero test error if and only if*

$$m_p(\boldsymbol{\rho}_\varepsilon) < 1 + \varepsilon\,, \tag{5.2}$$

where $(\boldsymbol{\rho}_\varepsilon)_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \varepsilon$, and $t = 1, \ldots, T$.

Proof. Please see Appendix C.1. □

This theorem shows that the power mean Laplacian regularizer allows to correctly classify the nodes if p is such that condition (5.2) holds. In order to better understand how this condition changes when p varies, we analyze in the next corollary the limit cases $p \to \pm\infty$.

Corollary 5.1. *Let $E(\mathbb{G})$ be an expected multilayer graph as in Theorem 5.1. Then,*

- *For $p \to \infty$, the test error is zero if and only if $p_{\text{out}}^{(t)} < p_{\text{in}}^{(t)}$ for all $t = 1, \dots, T$.*
- *For $p \to -\infty$, the test error is zero if and only there exists a $t \in \{1, \dots, T\}$ such that $p_{\text{out}}^{(t)} < p_{\text{in}}^{(t)}$.*

Proof. Observe that the limit cases of the scalar power means are

$$\lim_{p \to -\infty} m_p(x_1, \dots, x_T) = \min\{x_1, \dots, x_T\}$$
$$\lim_{p \to +\infty} m_p(x_1, \dots, x_T) = \max\{x_1, \dots, x_T\}$$

Applying this to condition

$$m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) < 1 + \varepsilon\,,$$

where $(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}})_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \varepsilon$, and $t = 1, \dots, T$ yields the desired result.

□

This corollary implies that the limit case $p \to \infty$ requires that *all layers* convey information regarding the clustering/class structure of the multilayer graph, whereas the case $p \to -\infty$ requires that *at least one layer* encodes clustering/class information, and hence it is clear that conditions for the limit $p \to -\infty$ are less restrictive than the conditions for the limit case $p \to \infty$. The next Corollary shows that the smaller the power parameter p is, the less restrictive are the conditions to yield a zero test error.

Corollary 5.2. *Let $E(\mathbb{G})$ be an expected multilayer graph as in Theorem 5.1. Let $p \leq q$. If $\mathcal{L}_q$ yields zero test error, then $\mathcal{L}_p$ yields a zero test error.*

Proof. By Theorem 2.1 we have that if $p \leq q$ then $m_p(x_1, \dots, x_T) \leq m_q(x_1, \dots, x_T)$. Therefore, applying this to our case we can see that

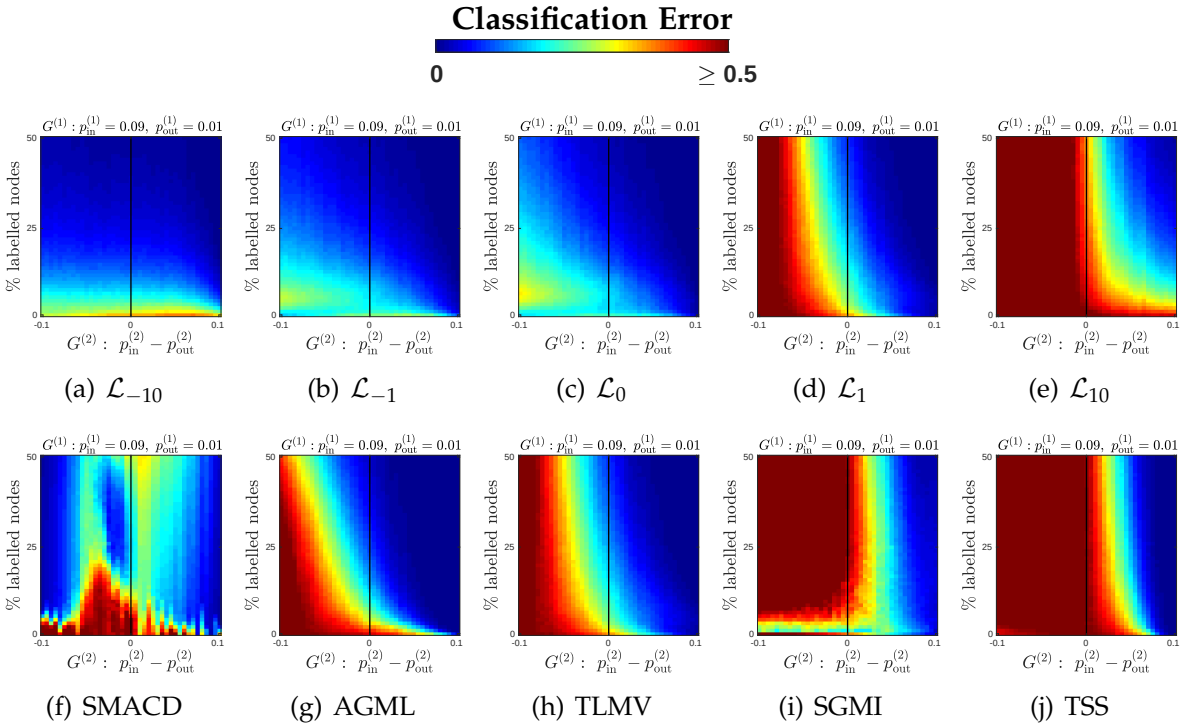
$$m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) \leq m_q(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) < 1 + \varepsilon$$

A zero test classification error with parameter q is achieved if and only if $m_q(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) < 1 + \varepsilon$, hence we can see that zero test classification error with parameter p is achieved if it is achieved with parameter q and $p \leq q$.

□

The previous results show the effectivity of the power mean Laplacian regularizer in expectation. We now present a numerical evaluation based on Theorem 5.1 and Corollaries 5.1 and 5.2 on random graphs sampled from the SBM. The corresponding results are presented in Fig. 5.1 for classification with regularizers $L_{-10}, L_{-1}, L_0, L_1, L_{10}$ and $\lambda = 1$.

We first describe the setting we consider: we generate random multilayer graphs with two layers ($T = 2$) and two classes ($k = 2$) each composed by 100 nodes ($|\mathcal{C}| =$



(a) $\mathcal{L}_{-10}$ (b) $\mathcal{L}_{-1}$ (c) $\mathcal{L}_0$ (d) $\mathcal{L}_1$ (e) $\mathcal{L}_{10}$

(f) SMACD (g) AGML (h) TLMV (i) SGMI (j) TSS

Figure 5.1: Average classification error under the Stochastic Block Model computed from 100 runs. **Top Row:** Particular cases with the power mean Laplacian. **Bottom Row:** State of the art models.

100). For each parameter configuration $(p_{\text{in}}^{(1)}, p_{\text{out}}^{(1)}, p_{\text{in}}^{(2)}, p_{\text{out}}^{(2)})$ we generate 10 random multilayer graphs and 10 random samples of labeled nodes, yielding a total of 100 runs per parameter configuration, and report the average test error. Our goal is to evaluate the classification performance under different SBM parameters and different amounts of labeled nodes. To this end, we fix the first layer $G^{(1)}$ to be informative of the class structure ($p_{\text{in}}^{(1)} - p_{\text{out}}^{(1)} = 0.08$), i.e. one can achieve a low classification error by taking this layer alone, provided sufficiently many labeled nodes are given. The second layer will go from non-informative (noisy) configurations ($p_{\text{in}}^{(2)} < p_{\text{out}}^{(2)}$, left half of x-axis) to informative configurations ($p_{\text{in}}^{(2)} > p_{\text{out}}^{(2)}$, right half of x-axis), with $p_{\text{in}}^{(t)} + p_{\text{out}}^{(t)} = 0.1$ for both layers. Moreover, we consider different amounts of labeled nodes: going from 1% to 50% (y-axis). The corresponding results are presented in Figs. 5.1(a),5.1(b),5.1(c),5.1(d), and 5.1(e).

In general one can expect a low classification error when both layers $G^{(1)}$ and $G^{(2)}$ are informative (right half of the x-axis). We can see that this is the case for all power mean Laplacian regularizers here considered here (see top row of Fig. 5.1). In particular, we can see in Fig. 5.1(e) that L_{10} performs well only when **both** layers are informative and completely fails when the second layer is not informative, regardless of the amount of labeled nodes. On the other side we can see in Fig. 5.1(a) that L_{-10} achieves in general a low classification error, regardless of the configuration of the

second layer $G^{(2)}$, i.e. when $G^{(1)}$ **or** $G^{(2)}$ are informative. Moreover, we can see that all areas with low classification error (dark blue) increase when the parameter p decreases, verifying the result from Corollary 5.2. In the bottom row of Fig. 5.1 we present the performance of state of the art methods. We can observe that most of them present a classification performance that resembles the one of the power mean Laplacian regularizer L_1. In general their classification performance drops when the level of noise increases, i.e. for non-informative configurations of the second layer $G^{(2)}$, and they are outperformed by the power mean Laplacian regularizer for small values of p.

5.4.2 Case 2: Unbalanced Class Labels

In the previous analysis we assumed that we had the same amount of labeled nodes per class. Now we consider the case where the number of labeled nodes per class is different. This setting was considered in (Zhu *et al.*, 2003), where the goal was to overcome unbalanced class proportions in labeled nodes. To this end, they propose a Class Mass Normalization (CMN) strategy, whose performance was also tested in (Zhu and Ghahramani, 2002). In the following result we show that, provided the ground truth classes have the same size, different amounts of labeled nodes per class affect the conditions in expectation for zero classification error of (5.1). For simplicity, we consider here only the case of two classes.

Theorem 5.2. *Let $E(\mathbb{G})$ be the expected multilayer graph with T layers following the multilayer SBM with two classes $\mathcal{C}_1, \mathcal{C}_2$ of equal size and parameters $\left(p_{\text{in}}^{(t)}, p_{\text{out}}^{(t)}\right)_{t=1}^{T}$. Assume n_1, n_2 nodes from $\mathcal{C}_1, \mathcal{C}_2$ are labeled, respectively. Let $\lambda = 1$. Then* (5.1) *yields zero test error if*

$$m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) < \min\left\{\frac{n_1}{n_2}, \frac{n_2}{n_1}\right\} \tag{5.3}$$

where $(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}})_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \varepsilon$, and $t = 1, \ldots, T$.

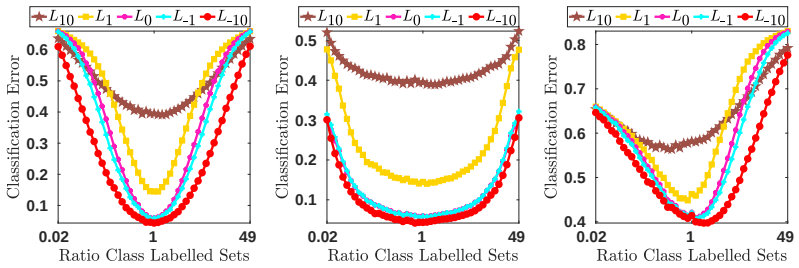
Proof. Please see Appendix C.2 □

Observe that Theorem 5.2 provides only a sufficient condition. A necessary and sufficient condition for a zero test error is given in Appendix C.2.

A different objective function can be employed for the case of classes with a different number of labels per class. Let C be the diagonal matrix defined by $C_{ii} = n/n_r$ if node v_i has been labeled to belong to class $\mathcal{C}_r$. Consider the following modification of (5.1)

$$\arg\min_{f \in \mathbb{R}^n} \|f - CY\|^2 + \lambda f^T L_p f \tag{5.4}$$

The next Theorem shows that using (5.4) in place of (5.1) allows us to retrieve the same condition of Theorem 5.1 for a zero test error in expectation in the setting where the number of labeled nodes per class is not equal.



(a) Uniform loss (b) Weighted loss (c) CMN

Figure 5.2: Different weighted loss strategies. From left to right: Fig. 5.2(a) uniform loss, Fig. 5.2(b) weighted loss, and Fig. 5.2(c) Class Mass Normalization.

Theorem 5.3. *Let $E(\mathbb{G})$ be the expected multilayer graph with T layers following the multilayer SBM k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ of equal size and parameters $\left(p_{\text{in}}^{(t)}, p_{\text{out}}^{(t)}\right)_{t=1}^{T}$. Let $n_1, \ldots, n_k$ be the number of labeled nodes per class. Let $C \in \mathbb{R}^{n \times n}$ be a diagonal matrix with $C_{ii} = n/n_r$ for $v_i \in \mathcal{C}_r$. The solution to (5.4) yields a zero test classification error if and only if*

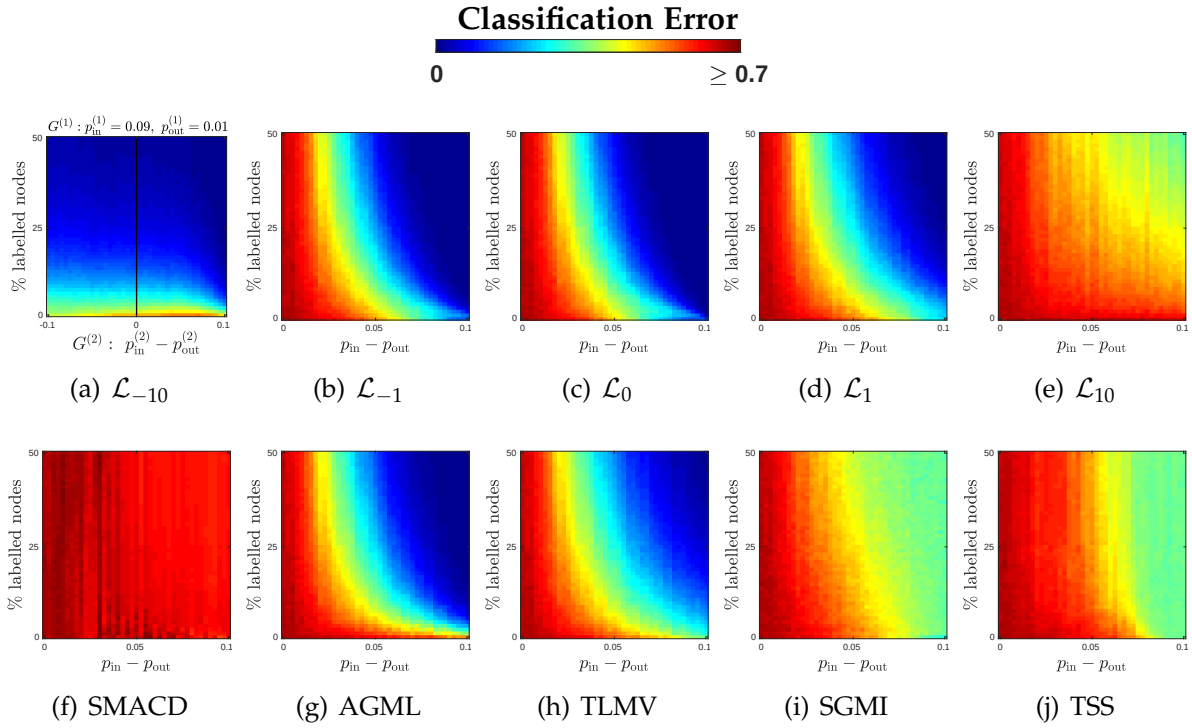
$$m_p(\boldsymbol{\rho}_\varepsilon) < 1 + \varepsilon\,, \tag{5.5}$$

where $(\boldsymbol{\rho}_\varepsilon)_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \varepsilon$, and $t = 1, \ldots, T$.

Proof. The proof is similar to the one of Theorem 5.1 (see Appendix C.1). The only change is in the terms $c_r \frac{n_r}{n}$. Since we have by definition that $c_r = \frac{n}{n_r}$ we have that $c_r \frac{n_r}{n} = 1$, leading to the conditions obtained by Theorem 5.1. □

In Figs. 5.2(a), 5.2(b), and 5.2(c). we present a numerical experiment with random graphs of our analysis in expectation. We consider the following setting: we generate multilayer graphs with two layers ($T = 2$) and two classes ($k = 2$) each composed by 100 nodes ($|\mathcal{C}| = 100$). We fix $p_{\text{in}}^{(1)} - p_{\text{out}}^{(1)} = 0.08$ and $p_{\text{in}}^{(2)} - p_{\text{out}}^{(2)} = 0$, with $p_{\text{in}}^{(t)} + p_{\text{out}}^{(t)} = 0.1$ for both layers. We fix the total amount of labeled nodes to be $n_1 + n_2 = 50$ and let $n_1, n_2 = 1, \ldots 49$. For each setting we generate 10 multilayer graphs and 10 sets of labeled nodes, yielding a total of 100 runs per setting, and report the average test classification error. In Fig. 5.2(a) we can see the performance of the power mean Laplacian regularizer without modifications. We can observe how different proportions of labeled nodes per class affect the performance. In Fig. 5.2(b), we present the performance of the modified approach (5.4) and observe that it yields a better performance against different class label proportions. Finally in Fig. 5.2(c) we present the performance based on Class Mass Normalization [1], where we can see that its effect is slightly skewed to one class and its overall performance is larger than the proposed approach.

[1] We follow the authors' implementation: http://pages.cs.wisc.edu/~jerryzhu/pub/harmonic_function.m



(a) $\mathcal{L}_{-10}$ (b) $\mathcal{L}_{-1}$ (c) $\mathcal{L}_0$ (d) $\mathcal{L}_1$ (e) $\mathcal{L}_{10}$

(f) SMACD (g) AGML (h) TLMV (i) SGMI (j) TSS

Figure 5.3: The Average test error under the SBM.Multilayer graph with 3 layers and 3 classes.**Top Row:** Particular cases with the power mean Laplacian. **Bottom Row:** State of the art models.

5.4.3 Case 3: No Layer Contains Full Information

In the previous section we considered the case where at least one layer had enough information to correctly estimate the node class labels. In this section we now consider the case where single layers taken alone obtain a large classification error, whereas if all the layers are taken together it is possible to obtain a good classification performance. For this setting we consider multilayer graphs with 3 layers ($T = 3$) and three classes ($k = 3$) $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$, each composed of 100 nodes ($|\mathcal{C}| = 100$) with the following expected adjacency matrix per layer:

$$\mathcal{W}_{i,j}^{(t)} = \begin{cases} p_{\text{in}}, & v_i, v_j \in \mathcal{C}_t \text{ or } v_i, v_j \in \overline{\mathcal{C}_t} \\ p_{\text{out}}, & \text{else} \end{cases} \tag{5.6}$$

for $t = 1, 2, 3$, i.e. layer $G^{(t)}$ is informative of class $\mathcal{C}_t$ but not of the remaining classes, and hence any classification method using one single layer will provide a poor classification performance. In Fig. 5.3 we present numerical experiments: for each parameter setting $(p_{\text{in}}, p_{\text{out}})$ we generate 5 multilayer graphs together with 5 samples of labeled nodes yielding a total of 25 runs per setting, and report the average test classification error. Also in this case we observe that the power mean Laplacian regularizer does identify the global class structure and that it leverages the information provided by labeled nodes, particularly for smaller values of p.

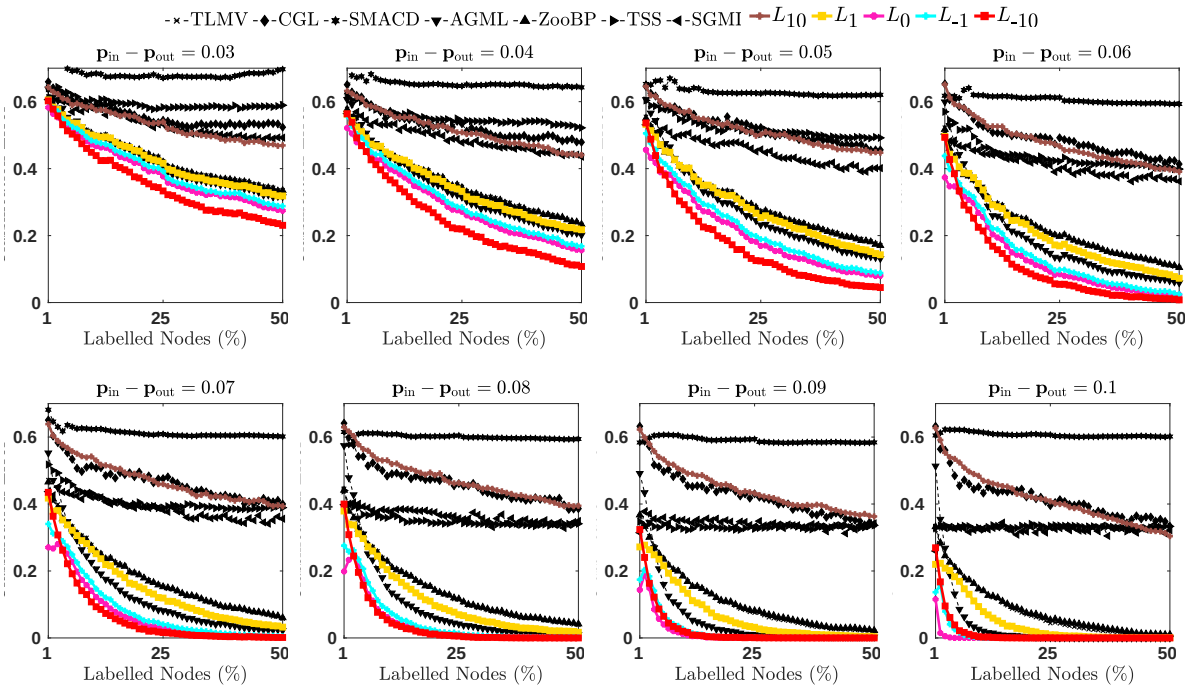


Figure 5.4: Average test error under the SBM.Multilayer graph with 3 layers and 3 classes where $p_{\text{in}} - p_{\text{out}} \in \{0.03, 0.04, \ldots, 0.1\}$.

On the other hand, this is not the case for all other state of the art methods. In fact, we can see that SGMI and TSS performs similarly to L_{10} which has the largest classification error. Moreover, we can see that AGML and TLMV perform similarly to the arithmetic mean of Laplacians L_1, which in turn is outperformed by the power mean Laplacian regularizer L_{-10}.

For a more detailed analysis we further consider the cases where $p_{\text{in}} - p_{\text{out}} \in \{0.03, 0.04, \ldots, 0.1\}$, which are depicted in Fig.5.4. On the x-axis we have the amount of labeled nodes and on the y- we have the classification error. We can see that in general there is a trend between the performance of our proposed method (colorful curves) and state of the art methods (black curves). We can see that the larger the gap $p_{\text{in}} - p_{\text{out}}$ the larger the difference is between our proposed method and state of the art methods. Moreover, one can see that the smaller the value of p, the better the performance of our proposed method. Moreover, there is a set of state of the art methods that do not improve their performance with larger amounts of labeled nodes. Yet, one can observe that there are three methods from the state of the art that perform close to our methods: TLMV, ZooBP and AGML, which perform similarly to our method L_1 (i.e. the arithmetic mean of Laplacians).

5.5 A MATRIX-FREE NUMERICAL METHOD FOR $(I+\lambda L_p)f = Y$

In this section we introduce a matrix-free method for the solution of the system $(I + \lambda L_p)f = Y$ based on contour integrals and Krylov subspace methods. The method exploits the sparsity of the Laplacians of each layer and is matrix-free, in the sense that it requires only to compute the matrix-vector product $L^{(i)}_{\text{sym}} \times \textit{vector}$, without requiring to store the matrices. Thus, when the layers are sparse, the method scales to large datasets. Observe that this is a critical requirement as L_p is in general a dense matrix, even for very sparse layers, and thus computing and storing L_p is very prohibitive for large multilayer graphs. We present a method for negative integer values $p < 0$, leaving aside the limit case $p \to 0$ as it requires a particular treatment.

Let $A_1, \dots, A_T$ be symmetric positive definite matrices, $S_p = A_1^p + \cdots + A_T^p$, $\varphi : \mathbb{C} \to \mathbb{C}$ be the complex function $\varphi(z) = z^{1/p}$ and let L_p be the matrix function $L_p = T^{-1/p}\varphi(S_p)$. The proposed method essentially transforms the original problem into a series of subproblems which thus allow us to solve the linear system $(I + \lambda L_p)^{-1}\mathbf{y}$ by solving several different linear systems with A_i as coefficient matrices. The method consists of three main nested inner–steps which we present below.

1. First, we solve the linear system $(I + \lambda L_p)^{-1}\mathbf{y}$ by a Krylov method (GMRES in our case (Saad and Schultz, 1986)). This method projects, at each iteration, the problem into the Krylov subspace spanned by $\{\mathbf{y}, \lambda L_p\mathbf{y}, (\lambda L_p)^2\mathbf{y}, \dots, (\lambda L_p)^h\mathbf{y}\}$. If $\kappa = \lambda_{\max}(L_p)/\lambda_{\min}(L_p)$, then the method converges as (Saad and Schultz, 1986)

$$O\left(\left(\frac{\kappa^2 - 1}{\kappa^2}\right)^{h/2}\right).$$

Thus, if L_p is well conditioned, a relatively small h is required. In order to build the appropriate Krylov subspace, we need to efficiently perform one matrix–vector product $L_p\mathbf{y}$ at each iteration.

2. Second, in order to compute $L_p\mathbf{y} = T^{-1/p}\varphi(S_p)\mathbf{y}$ we use the Cauchy integral form of the function φ, transformed via a conformal map, to approximate $\varphi(S_p)$ via the trapezoidal rule, as proposed in (Hale *et al.*, 2008). Let $m, M > 0$ be such that the interval $[m, M]$ contains the whole spectrum of S_p and let $t_1, \dots, t_N$ be N equally spaced contour points to be used in the trapezoidal rule. As φ has a singularity at $z = 0$ but just a brunch cut on $(-\infty, 0)$, we can approximate $\varphi(S_p)\mathbf{y}$ via (Hale *et al.*, 2008)

$$\varphi_N(S_p)\mathbf{y} = \frac{-8K(mM)^{1/4}}{\pi N k} S_p \ \text{Im}\left\{\sum_{i=1}^{N} \frac{\varphi(z_i^2)c_i d_i}{z_i(k^{-1} - s_i)^2}(z_i^2 I - S_p)^{-1}\mathbf{y}\right\}$$

where Im denotes the imaginary part, $k = \left((M/m)^{1/4} - 1\right)/\left((M/m)^{1/4} + 1\right)$, K is the value of the complete elliptic integral of the first kind, evaluated at ke^2, $s_i = \text{sn}(t_i)$ is the Jacobi elliptic sine function evaluated on the i-th contour point t_i, and

$$z_i = (mM)^{1/4}\left(\frac{k^{-1} + s_i}{k^{-1} - s_i}\right), \quad c_i = \sqrt{1 - s_i^2}, \quad d_i = \sqrt{1 - k^2 s_i^2},$$

for $i = 1, \ldots, N$. This approximation converges geometrically as the number of points increases. Precisely, it holds (Hale *et al.*, 2008)

$$\|\varphi(S_p)\mathbf{y} - \varphi_N(S_p)\mathbf{y}\| = O(e^{-2\pi^2 N/(\ln(M/m)+6)}) .$$

Thus, the computation of $\varphi(S_p)\mathbf{y}$ is reduced to N linear systems $(z_i^2 I - S_p)^{-1}\mathbf{y}$. Note that these systems are independent and thus they can be solved in parallel.

3. Finally, in order to solve the linear system $(zI - S_p)^{-1}\mathbf{y}$ we employ again a Krylov method. In order to build the Krylov space for $(zI - S_p)$ and $\mathbf{y}$ we need to efficiently perform one multiplication S_p times a vector per iteration. As $S_p = \sum_{i=1}^{T} A_i^p = \sum_{i=1}^{T} (A_i^{-1})^{|p|}$, this problem reduces to solving q linear systems with A_i as coefficient matrix, for $i = 1, \ldots, T$. As the matrices A_i are assumed sparse and positive definite, we can very efficiently solve each of these systems via the Preconditioned Conjugate Gradient method with an incomplete Cholesky preconditioner.

The pseudocode for the proposed algorithm is presented in Algorithms 9–11.

Input: $A_1, \ldots, A_T, p, \mathbf{y}, \lambda$
1 Compute preconditioners $P_1, \ldots, P_T$ for $A_1, \ldots, A_T$
2 Compute estimates for m and M such that eigenvalues$(S_p) \subseteq [m, M]$
3 Choose number of contour points N
4 Compute contour coefficients z_i, s_i, K, k
5 Solve $(I + \lambda L_p)^{-1}\mathbf{y}$ with GMRES, using Alg.10 as subroutine
Output: $\mathbf{u} = (I + \lambda L_p)^{-1}\mathbf{y}$

Algorithm 9: Solve $(I + \lambda L_p)^{-1}\mathbf{y}$

Input: $A_1, \ldots, A_T, p, \mathbf{y}, N, m, M$, contour coefficients z_i, s_i, c_i, d_i, k, K
1 $\mathbf{u} \leftarrow S_p\mathbf{y}$, using Alg.11
2 **for** $i = 1, \ldots, N$ **do**
3 $\quad \mathbf{u} \leftarrow$ solve$(z_i I - S_p, \mathbf{y})$ with GMRES, using Alg.11 as subroutine
4 $\quad \mathbf{u} \leftarrow \frac{(z_i^2)^{1/p} c_i d_i}{z_i (k^{-1} - s_i)^2}\mathbf{u}$
5 $\quad \mathbf{u}_{k+1} = \|\mathbf{v}_{k+1}\|_q^{1-q} |\mathbf{v}_{k+1}|^{q-2}\mathbf{v}_{k+1}$
6 **end**
7 $\mathbf{u} \leftarrow \frac{1}{T^{1/p}} \frac{-8K(mM)^{1/4}}{\pi N k} \mathrm{Im}(\mathbf{u})$
Output: $\mathbf{u} = L_p\mathbf{y}$

Algorithm 10: Multiply L_p times a vector

Input: $A_1, \ldots, A_T, P_1, \ldots, P_T, \mathbf{y}$
1 **for** $k = 1, \ldots, T$ **do**
2 $\quad \mathbf{u} \leftarrow \mathbf{u} +$ solve$(A_i^{|p|}, \mathbf{y})$ using CG preconditioned with P_i
3 **end**
Output: $\mathbf{u} = S_p\mathbf{y}$

Algorithm 11: Multiply S_p times a vector

Implementation details and computational complexity. The preconditioners P_i can be computed using an incomplete Cholesky factorization. In our test we observe that a 1e-4 threshold is enough to ensure the convergence of Alg.11 to 1e-8 precision in just 2 or 3 iterations. Since in our case the A_i are Laplacians, another excellent preconditioner can be obtained by using a Combinatorial Multi Grid method (CMG).

In our experiments, the CMG preconditioner performed similarly (but slightly worse) than the incomplete Cholesky.

A precise estimate of M in Alg.9 step 2 can be obtained using a Krylov eigensolver with Alg.11 as subroutine. As for m, since each A_i^p is positive definite and p is a negative integer, an estimate is obtained via Weyl's inequality (e.g. (Wilkinson, 1965))

$$m = \lambda_{\max}(A_1)^p + \cdots + \lambda_{\max}(A_T)^p \leq \lambda_{\min}(S_p)\,.$$

The number of contour points N can be chosen by using the geometric convergence of φ_N. In our experiments, we chose a precision $\tau =$1e-8 and we set

$$N = \left|(\ln(M/m) + 6)\ln(\tau)/2\pi^2\right|.$$

The contour points have been calculated using the code from (Driscoll, 2005).

On the computational cost of the method. Our analysis shows that it is proportional to the number of edges in each layer, i.e. Alg.9 scales to large sparse datasets. Let $c(A_i)$ be the cost of multiplying $c(A_i)$ times a vector (which is proportional to the number of nonzeros in A_i, i.e. the number of edges in the layer i when A_i is the normalized Laplacian of the i-th layer). Let K_1, K_2, K_3 be the number of iterations of GMRES,GMRES and PCG in lines 5, 3 and 2 of Algorithms 9, 10 and 11, respectively. Each instance of solve$(A_i^{|p|}, \mathbf{y})$ in Alg.11 requires $K_3 p\, c(A_i)$ operations per step. So the cost of Alg.11 is roughly $pK_3 \sum_{i=1}^{T} c(A_i)$. This implies that the cost of Alg.10 is $NK_2K_3p \sum_{i=1}^{T} c(A_i)$. Therefore, the cost of solving the linear system $(I + \lambda L_p)^{-1}\mathbf{y}$ with Alg.9 is

$$K_1 N K_2 K_3 p\big(c(A_1) + \cdots + c(A_T)\big)\,,$$

showing that the method scales as the number of nonzeros in each layer, as claimed. It is important to notice that the Algorithm allows for a high level of parallelism. In fact, the computation of the preconditioners P_i at step 1 of Alg.9, the **for** at step 2 of Alg.10 and the **for** at step 1 of Alg.11 can all be run in parallel. A time execution analysis is provided in Fig 5.5, where we can see that the time execution of our approach is competitive to the state of the art as TSS(Tsuda *et al.*, 2005), outperforming AGML(Nie *et al.*, 2016), SGMI(Karasuyama and Mamitsuka, 2013) and SMACD(Gujral and Papalexakis, 2018).

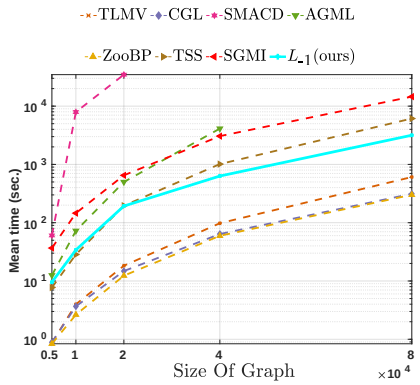


Figure 5.5: Mean execution time of 10 runs. L_{-1}(ours) stands for the power mean Laplacian regularizer together with our proposed matrix-free method. We generate multilayer graphs with two layers, with two classes of same size with parameters $p_{\text{in}} = 0.05$ and $p_{\text{in}} = 0.025$ and graph sizes $[0.5, 1, 2, 4, 8] \times 10^4$. Our matrix free approach (solid blue curve) outperforms AGML(Nie *et al.*, 2016), SGMI(Karasuyama and Mamitsuka, 2013) and SMACD(Gujral and Papalexakis, 2018).

5.6 EXPERIMENTS

In this section we present further experiments to evaluate the proposed approach. In Subsection 5.6.1 we present experiments on multilayer graphs from real-world settings, where we observe that our approach is competitive to the state of the art, and in Subsection 5.6.2 we present a numerical analysis of the effect of the regularization parameter, based on both synthetic multilayer graphs following the Multilayer Stochastic Block Model and on real datasets. We find that tuning the regularization parameter is consistent across synthetic and real datasets.

5.6.1 Experiments on Real Datasets

In this section we compare the performance of the proposed approach with state of the art methods on real-world datasets. We consider the following datasets: *3-sources* (Liu *et al.*, 2013), which consists of news articles that were covered by news sources BBC, Reuters and Guardian; *BBC* (Greene and Cunningham, 2005) and *BBC Sports* (Greene and Cunningham, 2009) news articles, a dataset of Wikipedia articles with ten different classes (Rasiwasia *et al.*, 2010), the hand written *UCI* digits dataset with six different set of features, and citations datasets *CiteSeer*(Lu and Getoor, 2003), *Cora*(McCallum *et al.*, 2000) and *WebKB*(Texas)(Craven *et al.*, 2011). For each dataset we build the corresponding layer adjacency matrices by taking the symmetric k-nearest neighbour graph using the Pearson linear correlation as similarity measure (i.e. we take the k neighbors with the highest correlation), and use the unweighted version of it. Datasets CiteSeer, Cora and WebKB have only two layers, where the first one is a fixed precomputed citation layer, and the second one the corresponding k-nearest neighbour graph built from document features.

Baseline methods: TSS (Tsuda *et al.*, 2005), which identifies an optimal linear combination of graph Laplacians, SGMI (Karasuyama and Mamitsuka, 2013), which performs label propagation by sparse integration, TLMV (Zhou and Burges, 2007), which is a weighted arithmetic mean of adjacency matrices, CGL (Argyriou *et al.*, 2006), which is a convex combination of the pseudo inverse Laplacian kernel, AGML (Nie *et al.*, 2016), which is a parameter-free method for optimal graph layer weights, ZooBP (Eswaran *et al.*, 2017), which is a fast approximation of Belief Propagation, and SMACD (Gujral and Papalexakis, 2018), which is a tensor factorization method designed for semi-supervised learning. Finally we set the parameters for TSS to ($c = 10, c_0 = 0.4$), SMACD ($\lambda = 0.01$)[2], TLMV ($\lambda = 1$), SGMI ($\lambda_1 = 1, \lambda_2 = 10^{-3}$) and $\lambda = 0.1$ for L_1 and $\lambda = 10$ for L_{-1} and L_{-10}. We do not perform cross validation in our experimental setting due to the large execution time in some of the methods considered here. Hence we fix the parameters for each method in all experiments.

We fix the nearest neighbourhood size to $k = 10$ and generate 10 samples of labeled nodes, where the percentage of labeled nodes per class is in the range $\{1\%, 5\%, 10\%, 15\%, 20\%, 25\%\}$. The average test errors are presented in table 5.1,

[2]this is the default value in the code released by the authors: github.com/egujr001/SMACD

where the **best** (resp. second best) performances are marked with bold fonts and gray background (resp. with only gray background). We can see that the first and second best positions are in general taken by the power mean Laplacian regularizers L_1, L_{-1}, L_{-10}, being clear for all datasets except with 3-sources. Moreover we can see that in 77% of all cases L_{-1} performs either best or second best, further verifying that our proposed approach based on the power mean Laplacian for semi-supervised learning in multilayer graphs is a competitive alternative to state of the art methods[3].

[3]Communications with the authors of (Gujral and Papalexakis, 2018) could not clarify the performance of SMACD.

3sources	1%	5%	10%	15%	20%	25%
TLMV	29.8	21.5	**20.8**	20.3	15.5	16.5
CGL	50.2	45.5	36.4	30.6	23.8	19.8
SMACD	91.5	91.1	91.2	90.9	90.7	91.3
AGML	**23.9**	26.3	33.9	33.3	26.1	22.0
ZooBP	31.0	21.9	21.3	19.8	15.0	15.3
TSS	29.8	23.9	33.1	34.6	34.8	35.0
SGMI	34.4	26.6	25.4	24.4	19.1	17.9
L_1	33.5	23.9	23.4	20.1	15.6	**14.6**
$L_{\text{-}1}$	28.4	**20.0**	21.8	22.0	17.2	17.9
$L_{\text{-}10}$	40.9	29.1	21.9	**19.3**	**14.8**	14.7

BBC	1%	5%	10%	15%	20%	25%
TLMV	**29.0**	19.3	13.2	11.1	9.3	8.8
CGL	72.5	52.3	36.1	27.4	22.0	17.1
SMACD	74.4	73.5	72.8	72.6	72.5	72.4
AGML	60.0	34.2	18.6	13.1	11.0	9.5
ZooBP	31.1	20.1	15.0	12.2	10.0	9.1
TSS	40.4	26.1	20.9	20.1	19.8	19.7
SGMI	37.6	28.9	24.9	22.8	20.7	19.3
L_1	31.3	22.8	17.4	13.5	10.2	8.9
$L_{\text{-}1}$	31.0	**17.0**	**11.5**	**10.5**	**9.2**	**8.7**
$L_{\text{-}10}$	51.6	26.9	16.6	12.8	10.3	9.5

BBCS	1%	5%	10%	15%	20%	25%
TLMV	25.6	12.6	10.5	7.5	6.4	5.4
CGL	79.2	51.6	34.9	23.4	16.5	12.7
SMACD	77.8	80.6	82.4	96.4	98.4	98.3
AGML	34.6	17.4	12.1	7.0	6.0	5.4
ZooBP	33.8	13.9	11.3	8.8	7.6	6.2
TSS	23.9	13.2	14.1	12.3	13.1	12.2
SGMI	31.9	19.6	16.6	15.5	14.8	12.1
L_1	29.9	15.0	13.5	10.6	8.7	7.2
$L_{\text{-}1}$	**23.8**	**11.6**	**8.7**	**6.3**	**5.8**	**5.1**
$L_{\text{-}10}$	48.7	22.5	14.2	9.1	7.8	6.1

Wikipedia	1%	5%	10%	15%	20%	25%
TLMV	65.7	56.8	46.4	43.1	40.8	39.2
CGL	87.3	83.0	82.5	82.2	83.0	83.0
SMACD	85.4	85.6	85.4	85.3	86.8	90.0
AGML	71.3	66.6	48.1	42.1	38.4	37.3
ZooBP	67.6	58.0	47.0	43.8	41.2	39.8
TSS	87.7	84.7	83.3	81.9	82.3	81.4
SGMI	69.3	84.8	84.5	83.8	83.2	82.8
L_1	68.2	61.1	53.6	48.3	44.1	42.3
$L_{\text{-}1}$	**59.1**	**52.3**	**40.2**	**36.3**	**35.1**	**34.1**
$L_{\text{-}10}$	66.9	57.2	43.2	38.7	36.3	34.9

UCI	1%	5%	10%	15%	20%	25%
TLMV	28.9	20.4	16.3	14.4	13.7	12.7
CGL	81.8	64.0	54.6	49.1	46.7	46.7
SMACD	73.6	81.0	90.0	90.0	86.2	81.9
AGML	25.3	17.2	15.2	13.2	12.5	12.0
ZooBP	30.8	21.7	17.6	15.1	14.1	13.0
TSS	**24.0**	17.6	16.6	15.9	15.8	15.6
SGMI	36.0	44.4	50.9	50.4	50.2	48.8
L_1	31.3	23.8	18.7	15.6	14.4	13.2
$L_{\text{-}1}$	30.5	**17.1**	**13.8**	**12.6**	**12.3**	**11.9**
$L_{\text{-}10}$	57.0	33.8	23.7	17.6	15.3	13.4

Citeseer	1%	5%	10%	15%	20%	25%
TLMV	51.5	39.4	36.5	33.7	31.6	30.3
CGL	89.3	71.8	58.0	49.8	44.5	40.9
SMACD	90.7	90.4	67.0	65.5	66.8	68.9
AGML	**47.3**	**32.3**	**29.6**	**28.2**	**27.5**	**27.0**
ZooBP	63.6	41.9	38.7	35.8	33.8	32.2
TSS	58.5	49.5	45.9	42.1	39.8	38.4
SGMI	59.4	46.8	44.0	42.3	40.5	39.2
L_1	56.3	44.1	41.2	38.5	36.1	34.7
$L_{\text{-}1}$	52.4	39.0	35.6	32.6	30.9	29.5
$L_{\text{-}10}$	68.6	54.6	48.5	43.0	39.7	37.2

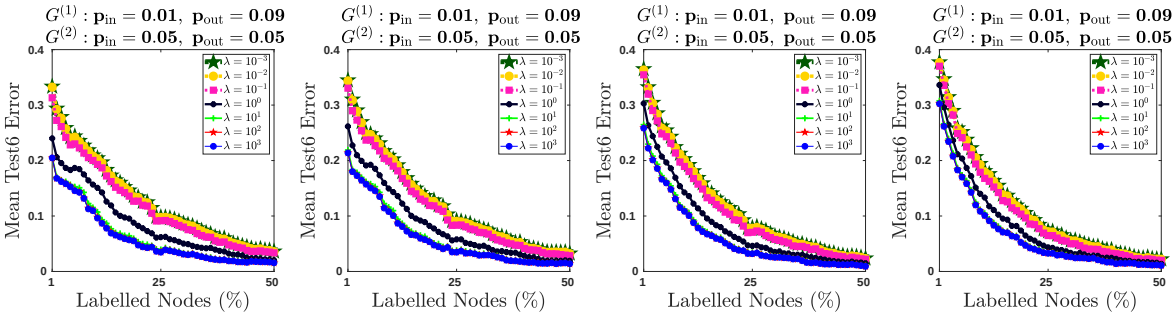
Cora	1%	5%	10%	15%	20%	25%
TLMV	46.0	34.1	28.8	25.8	22.5	20.6
CGL	85.5	70.1	56.5	49.1	44.2	40.0
SMACD	75.6	76.7	78.7	78.7	81.0	87.1
AGML	54.7	36.0	25.4	**20.7**	**18.1**	**16.5**
ZooBP	54.7	38.0	32.9	30.2	27.6	26.2
TSS	**38.8**	**27.7**	**24.1**	21.5	20.0	19.1
SGMI	57.3	47.7	43.0	41.8	40.1	38.5
L_1	50.7	38.2	33.4	31.2	28.2	25.6
$L_{\text{-}1}$	43.2	31.8	24.5	21.1	18.8	17.2
$L_{\text{-}10}$	62.0	46.3	35.4	29.4	25.2	22.3

WebKB	1%	5%	10%	15%	20%	25%
TLMV	58.6	49.4	45.6	47.2	47.6	48.2
CGL	80.4	82.4	84.4	86.9	82.7	89.2
SMACD	87.3	87.2	87.2	87.4	87.8	87.8
AGML	56.5	50.3	46.8	44.7	47.6	46.8
ZooBP	52.0	45.0	38.7	38.5	**36.4**	**33.5**
TSS	60.9	51.0	50.5	47.3	49.2	48.7
SGMI	**44.9**	**39.7**	41.9	**34.9**	40.3	52.5
L_1	58.5	49.0	44.8	44.3	44.5	44.4
$L_{\text{-}1}$	49.9	45.5	40.7	39.5	39.9	40.3
$L_{\text{-}10}$	52.3	41.9	**38.0**	38.1	36.8	39.5

Table 5.1: Experiments in real datasets. Notation: **best** performances are marked with bold fonts and gray background and second best performances with only gray background.

5.6.2 Analysis on the Effect of Regularization Parameter

Experiments under Multilayer Stochastic Block Model. We analyze the effect of the regularization parameter λ under the Multilayer Stochastic Block Model. The experimental setting is as follows: We fix the parameters of the first layer $G^{(1)}$ and second layer $G^{(2)}$ to $p_{\text{in}}^{(1)} = 0.09, p_{\text{out}}^{(1)} = 0.01, p_{\text{in}}^{(2)} = 0.05, p_{\text{out}}^{(2)} = 0.05$. We consider values of $\lambda \in \{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$, different amount of labeled nodes $\{1\%, \ldots, 50\%\}$. We sample five random multilayer graphs with the corresponding parameters and 5 random samples of labeled nodes with a fixed percentage, and present the average classification error. In Fig. 5.6 we can see that in general the larger the value of λ the smaller the classification error. In particular we can see that the performance does not present any relevant changes with $\lambda \leq 10^{-1}$.

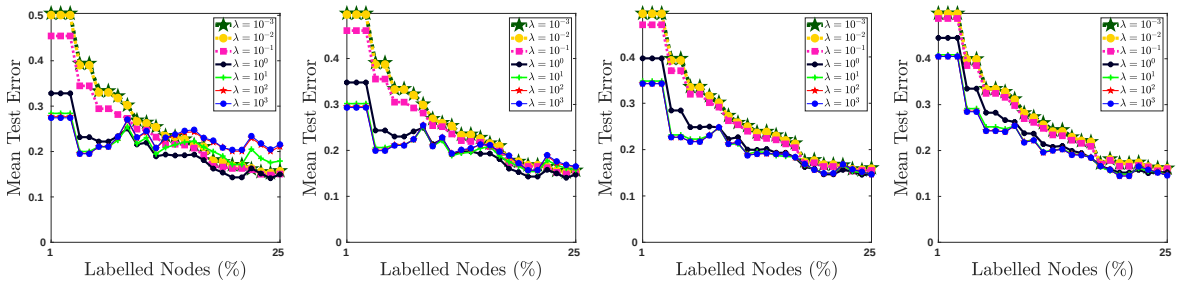


(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.6: Mean test classification error under MSBM for different values of λ. Details in Sec. 5.6.2.

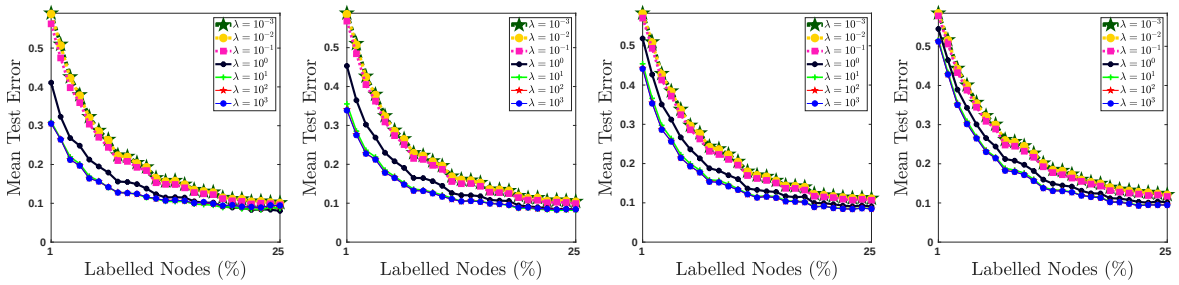
Experiments with real-world datasets. We analyze the effect of the regularization parameter λ with real-world datasets considered in Section 5.6. For each dataset we build the corresponding layer adjacency matrices by taking the symmetric k-nearest neighbour graph and take as similarity measure the Pearson linear correlation, (i.e. we take the k neighbours with highest correlation), and use the unweighted version of it.

We fix the nearest neighbourhood size to $k = 10$ and generate 10 samples of labeled nodes, where the percentage of labeled nodes per class is in the range $\{1\%, 2\%, \ldots, 25\%\}$. The average test errors are presented in Figs. 5.7–5.14, for the power mean Laplacian regularizers L_{-1}, L_{-2}, L_{-5}, and L_{-10}. We can see that in general the best performance, i.e. the smallest mean test classification error, corresponds to values of $\lambda = 10, 10^2, 10^3$, verifying the choice of $\lambda = 10$ presented in Section 5.6. Moreover, we can see that the mean test error in general decreases with larger amounts of labeled data, which verifies our previous experiments on multilayer graphs following the Multilayer Stochastic Block Model.



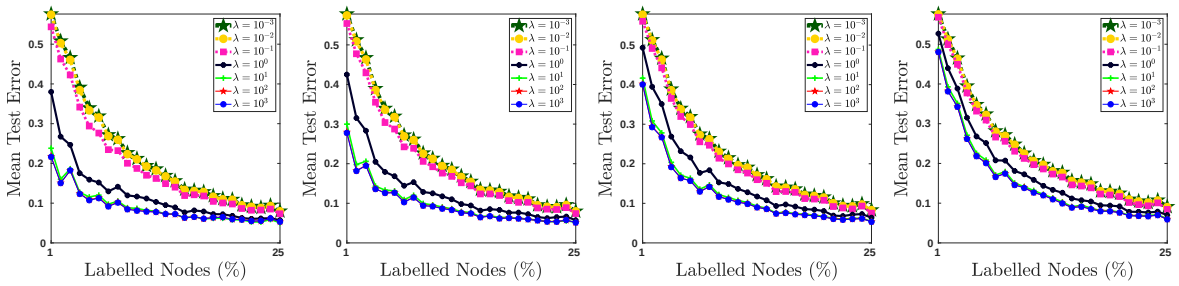
(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.7: Mean test classification error on 3sources for different values of λ.



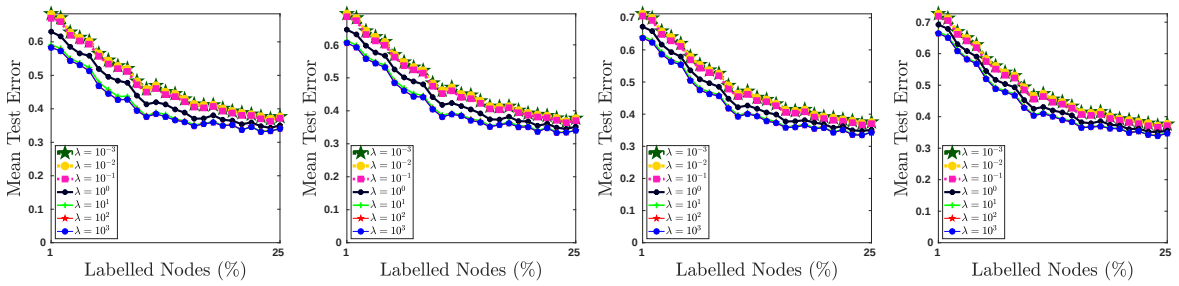
(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.8: Mean test classification error on BBC for different values of λ.



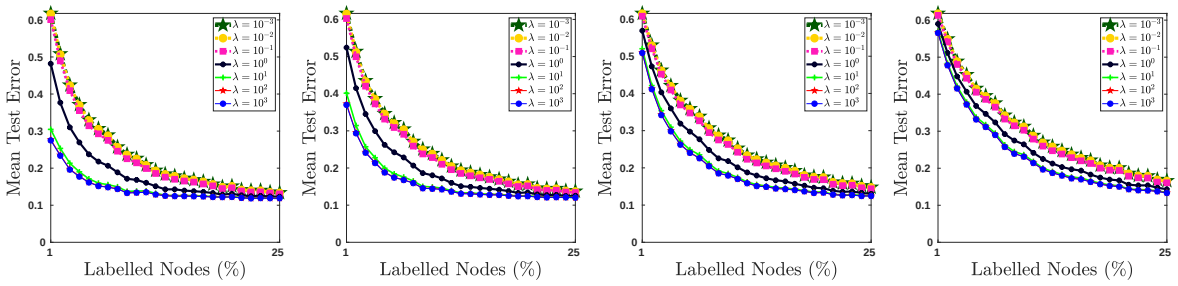
(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.9: Mean test classification error on BBCS for different values of λ.



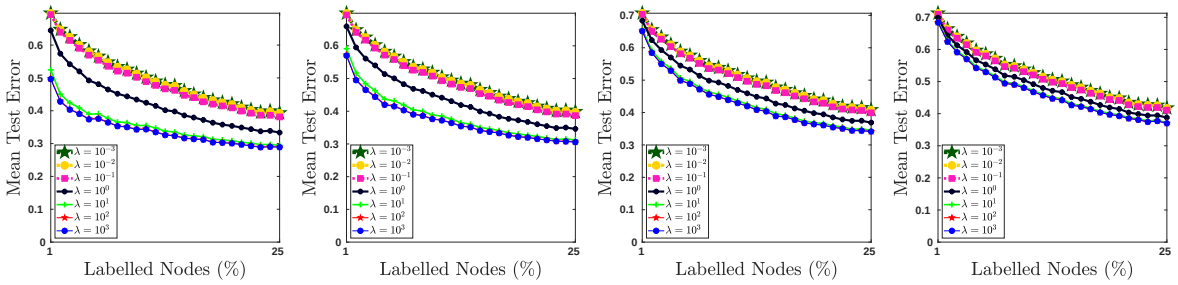
(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.10: Mean test classification error on Wikipedia for different values of λ.



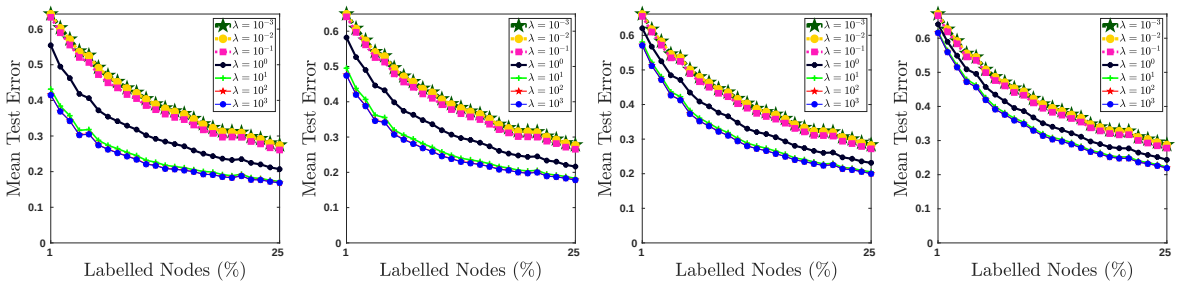
(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.11: Mean test classification error on UCI for different values of λ.



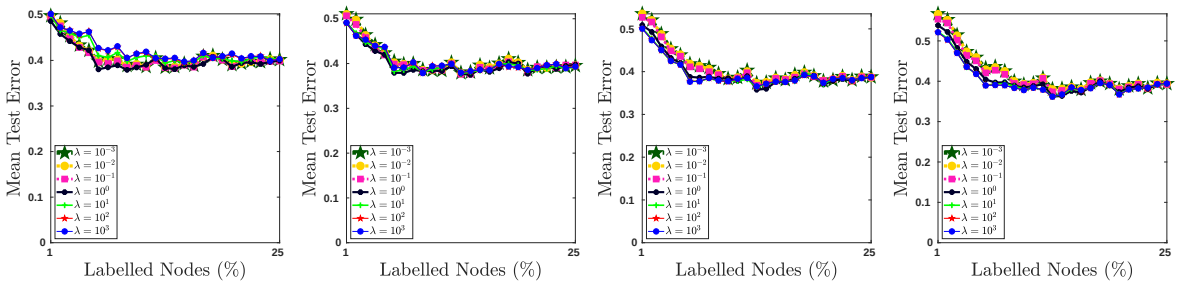
(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.12: Mean test classification error on Citeseer for different values of λ.



(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.13: Mean test classification error on Cora for different values of λ.



(a) L_{-1} (b) L_{-2} (c) L_{-5} (d) L_{-10}

Figure 5.14: Mean test classification error on WebKB for different values of λ.

5.7 CONCLUSION

In this chapter we have studied the task of semi-supervised learning on multilayer graphs. We have introduced the Power Mean Laplacian as a multilayer graph regularizer and analyzed it under a suitable stochastic block model in expectation. We have shown that our proposed approach obtains a good classification performance under suitable conditions, and provided verifications through extensive numerical experiments. Furthermore, our proposed approach does not perform worse than the state of the art on real datasets. Moreover, we presented a matrix-free numerical scheme, showing that our proposed approach is scalable to sparse multilayer graphs, outperforming the time execution of several state of the art approaches.

6 CONCLUSIONS AND FUTURE WORK

In this thesis we have studied the task of extending spectral clustering and semi-supervised learning to signed and multilayer networks. We have seen that each setting studied here brings different challenges and opportunities.

We have observed that most of state of the art approaches are basically some sort of arithmetic mean of graph matrices. Based on this observation we proposed novel extensions based on a one-parameter family of means called power means. Particular cases of the power means are the harmonic, geometric and arithmetic means.

We analyze our contributions theoretically and provide extensive numerical evaluations. Our analysis is based on suitable extensions of the Stochastic Block Model. Some of our results under the Stochastic Block Model are as follows: First, we show robustness under noise for negative values of the power parameter of the matrix power means. We prove this under the Stochastic Block Model and identify the regimes where this holds. Second, we identify, for multilayer graphs, that the Power Mean Laplacian is able to effectively merge the information encoded by different layers, even in cases where global information is available only by taking all layers. Third, we provide numerical evidence that the Power Mean Laplacian, for multilayer graphs, is able to recover the right clusters in cases where the information between layers is not consistent. Fourth, we provide concentration bounds for the eigenvalues and eigenvectors of the (Signed) Power Mean Laplacian.

In our analysis we have included extensive numerical evaluation on real world datasets and show that our approach is competitive to the state of the art. For the case of spectral clustering in signed networks we have shown that our approach consistently identifies explicit clustering structure in a dataset where it was previously conjectured that there was no clustering structure.

Further, we have shown that all our approaches are scalable to large sparse graphs. We have proposed numerical schemes that are matrix-free, in the sense that they never explicitly compute the (Signed) Power Mean Laplacian, both for the case of clustering and semi-supervised learning. Our matrix-free numerical schemes are based on Krylov subspace methods and quadrature methods.

Due to the ill-posed nature of clustering, several opportunities remain naturally open for the multilayer network setting. In what follows we briefly share challenges that we consider relevant and interesting.

Future Work. The task of clustering and semi-supervised learning in multilayer networks provide a rich scenario for future projects and open questions. We briefly sketch some of them.

Spectral clustering provides a framework that nicely connects the discrete optimization problem of graph cuts, with a continuous relaxation related to the quadratic form associated to the graph Laplacian. Hence, it remains unclear what is the corresponding graph cut for the different Power Mean Laplacians. For instance, is it possible to say that the harmonic mean of Laplacians is related to the harmonic mean of graph cuts?

Clustering with contraints has received a relevent amount of attention, particularly in the context of fairness (Kleindessner *et al.*, 2019). A clustering is fair (Chierichetti *et al.*, 2017) if every demographic group is proportionally represented in every cluster. It is unclear how to extend the notion of fair clustering to multilayer graphs, where demographic groups have different kinds of interactions and dynamics per layer.

Modularity provides interesting connections between approaches like the Louvain method (Blondel *et al.*, 2008), and spectral methods. Hence, it is unclear if, for the case of multilayer graphs, there can be any connection between modularity and the Power Mean Laplacian as in the single layer case.

Currently there is an important amount of attention towards deep learning on graphs (Chen *et al.*, 2019), particularly on graph convolutional networks (Kipf and Welling, 2017). An interesting task is to extend the current approaches to the multilayer setting and explore if the matrix power means are advantageous in this setting.

Moreover, graph convolutional networks have been considered for the task of time series forecasting (Wu *et al.*, 2020). It remains as an open question if, for the case where a graph is available per time snapshot, the matrix power means provide any aid for this task. Further, this motivates the analysis of novel families of means that support the analysis on subsets of time windows where graph time series are available.

It remains an open question how matrix power means can be used for ranking and node centrality in multilayer graphs (Tudisco *et al.*, 2018). While there is already relevant work based on tensors, it remains unclear how this can be done with matrix power means.

Another attractive line of research focuses on optimal transport (Villani, 2008). It is unclear if Wasserstein Barycenters (Cuturi and Doucet, 2014) of multilayer graphs(Titouan *et al.*, 2019) are effective as tools for clustering, and if there is any connection to matrix power means. Moreover, in (Takatsu, 2011) it is shown that the

optimal transport map between multilvariate gaussians is expressed in terms of a matrix geometric mean. Hence, it remains open if different matrix means can be related in a similar way.

A related approach for standard graphs has been proposed in (Abbe *et al.*, 2020) and is based on a thresholded powered adjacency matrix, showing robustness under suitable generative block models. Whereas in our work we have shown that negative powers present an interesting behavior on multilayer graphs, they show that indeed positive powers of the adjacency matrix provide provable robustness. It remains as an open question how these two approaches can be related, and if the approach from (Abbe *et al.*, 2020) can be extended to multilayer graphs.

A PROOF OF THEOREMS 3.5 AND 3.6

A.1 PROOF OF THEOREM 3.5

For the proof of Theorem 3.5, we first present Theorem A.1 which is a general version that allows to choose different diagonal shifts of the Laplacians together with different edge probabilities.

Theorem A.1. *Let G^+ and G^- be random graphs with independent edges $\mathbb{P}(W^+_{i,j} = 1) = p^+_{ij}$ and $\mathbb{P}(W^-_{i,j} = 1) = p^-_{ij}$. Let δ^+, δ^- be the minimum expected degrees of G^+ and G^-, respectively. Let $C^+_p = p^{1/p}\beta^{1-1/p}$, and $C^-_p = |p|^{1/|p|}\alpha^{-(3+1/|p|)}$. Choose $\epsilon > 0$. Then there exist constants $k^+ = k^+(\epsilon/2)$ and $k^- = k^-(\epsilon/2)$ such that if $\delta^+ > k^+ \ln n$, and $\delta^- > k^- \ln n$ then with probability at least $1 - \epsilon$,*

$$\left\| L_p - \mathcal{L}_p \right\| \leq C^+_p m_p \left(2\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^+}}, 2\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^-}} \right)^{1/p}$$

for $p \geq 1$, with p integer and

$$\left\| L_p - \mathcal{L}_p \right\| \leq C^-_p m_{|p|} \left(2\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^+}}, 2\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^-}} \right)^{1/|p|}$$

for $p \leq -1$, with p integer, and where we have $L_p = M_p(L^+_{\mathrm{sym}} + \alpha I, Q^-_{\mathrm{sym}} + \alpha I)$, and $\mathcal{L}_p = M_p(\mathcal{L}^+_{\mathrm{sym}} + \alpha I, \mathcal{Q}^-_{\mathrm{sym}} + \alpha I)$.

Before starting the proof of Theorem A.1, we present an upper bound on the matrix power mean.

Theorem A.2. *Let $A_1, \ldots, A_T, B_1, \ldots, B_T$ be symmetric matrices where $\alpha \leq \lambda(A_i) \leq \beta$, $\alpha \leq \lambda(B_i) \leq \beta$ for $i = 1, \ldots T$ and $\alpha, \beta > 0$.*

Let $C^+_p = p^{1/p}\beta^{1-1/p}$ and $C^-_p = |p|^{1/|p|}\alpha^{-(3+1/|p|)}$. Then, for $p \geq 1$, with p integer

$$\begin{aligned} &\left\| M_p(A_1, \ldots, A_T) - M_p(B_1, \ldots, B_T) \right\| \\ &\quad \leq C^+_p m_p\left(\|A_1 - B_1\|, \ldots, \|A_T - B_T\| \right)^{1/p} \end{aligned}$$

and, for $p \leq -1$, with p integer

$$\begin{aligned} &\left\| M_p(A_1, \ldots, A_T) - M_p(B_1, \ldots, B_T) \right\| \\ &\quad \leq C^-_p m_{|p|}\left(\|A_1 - B_1\|, \ldots, \|A_T - B_T\| \right)^{1/|p|} \end{aligned}$$

Proof. The proof is contained in Section A.2. □

Observe that the upper bound in Theorem A.2 is general in the sense that it is suitable for symmetric definite matrices with bounded spectrum, and for an arbitrary number of matrices.

We are now ready to prove Theorem A.1.

Proof of Theorem A.1. Let

$$A_1 = L^+_{\text{sym}}, \qquad B_1 = \mathcal{L}^+_{\text{sym}}$$
$$A_2 = Q^-_{\text{sym}}, \qquad B_2 = \mathcal{Q}^-_{\text{sym}}$$

with the corresponding signed power mean Laplacian

$$L_p = M_p(L^+_{\text{sym}}, Q^-_{\text{sym}}), \qquad \mathcal{L}_p = M_p(\mathcal{L}^+_{\text{sym}}, \mathcal{Q}^-_{\text{sym}})$$

We start with the case $p \geq 1$, with p integer. Let $C^+_p = p^{1/p}\beta^{1-1/p}$. By Theorem A.2 we have

$$\left\|L_p - \mathcal{L}_p\right\| \leq C^+_p m_p(\left\|L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}\right\|, \left\|Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}}\right\|)^{1/p}$$

Let $\gamma = (\gamma_1, \gamma_2)$ where

$$\gamma_1 = 2\sqrt{\tfrac{3\ln(8n/\epsilon)}{\delta^+}} \qquad \gamma_2 = 2\sqrt{\tfrac{3\ln(8n/\epsilon)}{\delta^-}}$$

Define $a = c\, m_p(\gamma)$ and $c = C^+_p$. Then,

$$\begin{aligned}
\mathbb{P}(\left\|L_p - \mathcal{L}_p\right\| > a) \leq& \mathbb{P}\left(c\, m_p(\left\|L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}\right\|, \left\|Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}}\right\|) > a\right) \\
=& \mathbb{P}\left(m_p(\left\|L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}\right\|, \left\|Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}}\right\|) > \frac{a}{c}\right) \\
=& \mathbb{P}\left(\left\|L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}\right\|^p + \left\|Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}}\right\|^p > 2\left(\frac{a}{c}\right)^p\right) \\
=& \mathbb{P}\left(\left\|L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}\right\|^p + \left\|Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}}\right\|^p > \sum_{i=1}^{2}\gamma_i^p\right) \\
\leq& \mathbb{P}\left(\left\{\left\|L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}\right\|^p > \gamma_1^p\right\} \cup \right. \\
& \qquad \left. \left\{\left\|Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}}\right\|^p > \gamma_2^p\right\}\right) \\
\leq& \mathbb{P}\left(\left\|L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}}\right\|^p > \gamma_1^p\right) \\
& + \mathbb{P}\left(\left\|Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}}\right\|^p > \gamma_2^p\right) \qquad \text{(A.1)}
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{P}\left(\left\| L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}} \right\| > \gamma_1 \right) \\
&\quad + \mathbb{P}\left(\left\| Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}} \right\| > \gamma_2 \right) \\
&= \mathbb{P}\left(\left\| L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}} \right\| > 2\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^+}} \right) \\
&\quad + \mathbb{P}\left(\left\| Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}} \right\| > 2\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^-}} \right) \\
&= \mathbb{P}\left(\left\| L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}} \right\| > 2\sqrt{\frac{3\ln(4n/\hat{\epsilon})}{\delta^+}} \right) \\
&\quad + \mathbb{P}\left(\left\| Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}} \right\| > 2\sqrt{\frac{3\ln(4n/\hat{\epsilon})}{\delta^-}} \right) \\
&= \mathbb{P}\left(\left\| L^+_{\text{sym}} - \mathcal{L}^+_{\text{sym}} \right\| > 2\sqrt{\frac{3\ln(4n/\hat{\epsilon})}{\delta^+}} \right) \\
&\quad + \mathbb{P}\left(\left\| L^-_{\text{sym}} - \mathcal{L}^-_{\text{sym}} \right\| > 2\sqrt{\frac{3\ln(4n/\hat{\epsilon})}{\delta^-}} \right) \\
&\leq \hat{\epsilon} + \hat{\epsilon} \qquad \text{(A.2)} \\
&= \epsilon
\end{aligned}
$$

where $\hat{\epsilon} = \epsilon/2$. Inequality (A.1) follows from Boole's inequality. Inequality (A.2) comes from applying Theorem A.5 from (Chung and Radcliffe, 2011) to G^+ and G^-, with corresponding minimum expected degree δ^+, and δ^-, respectively, and $\hat{\epsilon}$, and

$$
\begin{aligned}
\left\| Q^-_{\text{sym}} - \mathcal{Q}^-_{\text{sym}} \right\| &= \|(I + T) - (I + \mathcal{T})\| \\
&= \|(I - T) - (I - \mathcal{T})\| \\
&= \left\| L^-_{\text{sym}} - \mathcal{L}^-_{\text{sym}} \right\|
\end{aligned}
$$

where

$$
\begin{aligned}
T &= (D^-)^{-1/2} W^- (D^-)^{-1/2} \\
\mathcal{T} &= (\mathcal{D}^-)^{-1/2} \mathcal{W}^- (\mathcal{D}^-)^{-1/2}
\end{aligned}
$$

Thus,

$$
\mathbb{P}\left(\left\| L_p - \mathcal{L}_p \right\| \geq a \right) < \epsilon
$$

and hence

$$
\mathbb{P}\left(\left\| L_p - \mathcal{L}_p \right\| \leq a \right) < 1 - \epsilon
$$

completing the proof for the case $p \geq 1$.

For the proof of the case $p \leq -1$ with p integer, let $c = |p|^{1/|p|} \alpha^{-(3+1/|p|)}$, and proceed as for the previous case with $|p|$. □

We now finally give the proof for Theorem 3.5.

Proof of Theorem 3.5. We will adapt to our particular case the general version presented in Theorem A.1. We do this by showing that our Stochastic Block Model approach together with the shift of our model are particular cases of Theorem A.1.

First, note that the spectrum of the normalized Laplacians L^+_{sym} and Q^-_{sym} is upper bounded by two, i.e. $\lambda(L^+_{\text{sym}}), \lambda(Q^-_{\text{sym}}) \in [0, 2]$. Hence, by adding a diagonal shift we get $\lambda(L^+_{\text{sym}} + \alpha I), \lambda(Q^-_{\text{sym}} + \alpha I) \in [\alpha, 2 + \alpha]$. Letting $\alpha = \varepsilon$ and $\beta = 2 + \alpha$ we get the shift corresponding to the particular case from Theorem 3.5.

Further, observe that our SBM model is obtained by setting $p^+_{ij} = p^+_{\text{in}}$ and $p^-_{ij} = p^-_{\text{in}}$ if v_i, v_j belong to the same cluster and $p^+_{ij} = p^+_{\text{out}}$ and $p^-_{ij} = p^-_{\text{out}}$ if v_i, v_j belong to different clusters.

Moreover, under the Stochastic Block Model here considered, the induced expected graphs are regular, and thus all nodes have the same degree. Hence, the minimum expected degrees of G^+ and G^- are

$$\delta^+ = \frac{n}{k}(p^+_{\text{in}} + (k-1)p^+_{\text{out}})$$
$$\delta^- = \frac{n}{k}(p^-_{\text{in}} + (k-1)p^-_{\text{out}})$$

Thus, taking these settings into Theorem A.1 we get the desired result, except that the condition on the minimum expected degrees is that there exists constants $k^+ = k^+(\epsilon/2)$, and $k^- = k^-(\epsilon/2)$ such that the desired concentration holds.

To overcome this, observe in the proof of Theorem A.5 (p.9) that the condition $\delta > k\ln(n)$ comes from the requirement

$$\sqrt{\frac{3\ln(4n/\epsilon)}{\delta}} < 1$$

Thus, by setting $\delta > 3\ln(4n/\epsilon)$ the condition is fulfilled. In our case, this yields to $\delta^+ > 3\ln(4n/\hat{\epsilon}) = 3\ln(8n/\epsilon)$ and $\delta^- > 3\ln(4n/\hat{\epsilon}) = 3\ln(8n/\epsilon)$, leading to the desired result.

□

A.2 PROOF OF THEOREM A.2

Before going into the proof, a set of preliminary results are necessary. In what follows, for Hermitian matrices A and B we mean by $A \preceq B$ that $B - A$ is positive semidefinite (see (Bhatia, 1997), Ch. 5, and (Tropp, 2015), Ch. 2.1.8 for more details). We now proceed with the definition of a operator monotone function:

Definition 1 ((Tropp, 2015) Ch. 8.4.2, (Bhatia, 1997) Ch. 5.)**.** *Let $f{:}I \to \mathbb{R}$ be a function on an interval I of the real line. The function f is operator monotone on I when $A \preceq B$ implies $f(A) \preceq f(B)$ for all Hermitian matrices A and B whose eigenvalues are contained in I.*

The following result states that the negative inverse is operator monotone.

Proposition A.1 ((Bhatia, 1997), Prop. V.1.6, (Tropp, 2015), Prop. 8.4.3)**.** *The function* $f(t) = -\frac{1}{t}$ *is operator monotone on* $(0, \infty)$*.*

The following result states that the effect of operator monotone functions can be upper bounded in a helpful way.

Theorem A.3 ((Bhatia, 1997), Theorem. X.3.8)**.** *Let* f *be an operator monotone function on* $(0, \infty)$ *and let* A, B *be two positive definite matrices that are bounded below by* a*; i.e.* $A \geq aI$ *and* $B \geq aI$ *for the positive number* a*. Then for every unitarily invariant norm*

$$|||f(A) - f(B)||| \leq f'(a)|||A - B|||$$

Applying this to the case of the negative inverse leads to the following Corollary.

Corollary A.1. *Let* A, B *be two positive definite matrices that are bounded below by* a*; i.e.* $A \geq aI$ *and* $B \geq aI$ *for the positive number* a*. Then for every unitarily invariant norm*

$$|||A^{-1} - B^{-1}||| \leq \frac{1}{a^2}|||A - B|||$$

Proof. Let $f(t) = -\frac{1}{t}$. Then, by Proposition A.1 we know that f is operator monotone. Since $f'(t) = 1/t^2$, it follows from Theorem A.3

$$|||A^{-1} - B^{-1}||| = |||f(A) - f(B)||| \leq f'(a)|||A - B||| = \frac{1}{a^2}|||A - B|||$$

□

The next results states a useful result on positive powers between zero and one.

Corollary A.2 ((Bhatia, 1997), Eq. X.2)**.** *Let* A, B *be two positive semidefinite matrices. Then, for* $0 \leq r \leq 1$

$$\|A^r - B^r\| \leq \|A - B\|^r$$

Its equivalent to positive integer powers is stated in the following result.

Proposition A.2 (See (Bhatia, 1997), Eq. IX.4)**.** *For any two matrices* X, Y*, and for* $m = 1, 2, \ldots,$

$$\|X^m - Y^m\| \leq mM^{m-1} \|X - Y\|$$

where $M = \max(\|X\|, \|Y\|)$*.*

Next we show that the spectrum of the matrix power mean is well bounded for positive powers larger than one.

Proposition A.3. *Let* $A_1, \ldots, A_T$ *be symmetric positive definite matrices that are bounded below and above by* α *and* β*; i.e.* $\alpha I \leq A_i \leq \beta I$ *for positive numbers* α *and* β*. Then, for* $p \geq 1$*, with* p *integer*

$$\alpha \leq \lambda(M_p(A_1, \ldots, A_T)) \leq \beta$$

Proof. Let $S_p(A_1, \ldots, A_T) = \frac{1}{T}\sum_{i=1}^{T} A_i^p$. Then

$$\begin{aligned}\langle x, S_p(A_1, \ldots, A_T)x\rangle &= \langle x, \left(\frac{1}{T}\sum_{i=1}^{T} A_i^p\right)x\rangle \\ &= \frac{1}{T}\sum_{i=1}^{T}\langle x, A_i^p x\rangle\end{aligned}$$

Thus, we obtain the following upper bound

$$\begin{aligned}\max_{\|x\|=1}\langle x, S_p(A_1, \ldots, A_T)x\rangle &= \max_{\|x\|=1}\frac{1}{T}\sum_{i=1}^{T}\langle x, A_i^p x\rangle \\ &\leq \frac{1}{T}\sum_{i=1}^{T}\max_{\|x\|=1}\langle x, A_i^p x\rangle \\ &\leq \beta^p\end{aligned}$$

Hence, $\lambda_{max}(S_p(A_1, \ldots, A_T)) \leq \beta^p$, and thus we obtain the corresponding upper bound $\lambda_{max}(M_p(A_1, \ldots, A_T)) \leq \beta$.

In a similar way we obtain the following lower bound,

$$\begin{aligned}\min_{\|x\|=1}\langle x, S_p(A_1, \ldots, A_T)x\rangle &= \min_{\|x\|=1}\frac{1}{T}\sum_{i=1}^{T}\langle x, A_i^p x\rangle \\ &\geq \frac{1}{T}\sum_{i=1}^{T}\min_{\|x\|=1}\langle x, A_i^p x\rangle \\ &\geq \alpha^p\end{aligned}$$

Hence, $\lambda_{min}(S_p(A_1, \ldots, A_T)) \geq \alpha^p$, and thus we obtain the corresponding lower bound $\lambda_{min}(M_p(A_1, \ldots, A_T)) \geq \alpha$.

Therefore, $\alpha \leq \lambda(M_p(A_1, \ldots, A_T)) \leq \beta$. □

We now present results for $p \geq 1$ of Theorem A.2.

Results for the case $p \geq 1$

The following two propositions are the main ingredients for the upper bound presented in Theorem A.2 for the case $p \geq 1$.

Proposition A.4. *Let $A_1, \ldots, A_T, B_1, \ldots, B_T$ be symmetric positive semidefinite matrices. Then, for $p \geq 1$ with p integer*

$$\begin{aligned}&\left\|M_p(A_1, \ldots, A_T) - M_p(B_1, \ldots, B_T)\right\| \\ &\quad\leq \left\|M_p^p(A_1, \ldots, A_T) - M_p^p(B_1, \ldots, B_T)\right\|^{\frac{1}{p}}\end{aligned}$$

Proof. Let $S_p(A_1, \ldots, A_T) = \frac{1}{T}\sum_{i=1}^{T} A_i^p$ and $r = 1/p$. Then,

$$\begin{aligned}
&\left\|M_p(A_1, \ldots, A_T) - M_p(B_1, \ldots, B_T)\right\| \\
&\quad = \left\|S_p^{1/p}(A_1, \ldots, A_T) - S_p^{1/p}(B_1, \ldots, B_T)\right\| \\
&\quad = \left\|S_p^{r}(A_1, \ldots, A_T) - S_p^{r}(B_1, \ldots, B_T)\right\| \\
&\quad \leq \left\|S_p(A_1, \ldots, A_T) - S_p(B_1, \ldots, B_T)\right\|^r \\
&\quad = \left\|S_p(A_1, \ldots, A_T) - S_p(B_1, \ldots, B_T)\right\|^{1/p} \\
&\quad = \left\|M_p^p(A_1, \ldots, A_T) - M_p^p(B_1, \ldots, B_T)\right\|^{1/p}
\end{aligned}$$

where the inequality comes from Corollary A.2, giving the desired result. □

Proposition A.5. *Let $A_1, \ldots, A_T, B_1, \ldots, B_T$ be symmetric positive semidefinite matrices such that $\lambda(A_i) \leq \beta$ and $\lambda(B_i) \leq \beta$ for $i = 1, \ldots T$. Then, for $p \geq 1$,*

$$\begin{aligned}
&\left\|M_p^p(A_1, \ldots, A_T) - M_p^p(B_1, \ldots, B_T)\right\| \\
&\qquad \leq p\beta^{p-1} m_p\big(\left\|A_1 - B_1\right\|, \ldots, \left\|A_T - B_T\right\|\big)
\end{aligned}$$

Proof. Let $\beta_i = \max(\|A_i\|, \|B_i\|)$. Then,

$$\begin{aligned}
&\left\|M_p^p(A_1, \ldots, A_T) - M_p^p(B_1, \ldots, B_T)\right\| \\
&\quad = \left\|\left(\frac{1}{T}\sum_{i=1}^{T} A_i^p\right) - \left(\frac{1}{T}\sum_{i=1}^{T} B_i^p\right)\right\| \\
&\quad = \left\|\frac{1}{T}\sum_{i=1}^{T} A_i^p - B_i^p\right\| \\
&\quad \leq \frac{1}{T}\sum_{i=1}^{T} \left\|A_i^p - B_i^p\right\| \\
&\quad \leq \frac{1}{T}\sum_{i=1}^{T} p(\beta_i)^{p-1} \left\|A_i - B_i\right\| \\
&\quad \leq \frac{1}{T}\sum_{i=1}^{T} p\beta^{p-1} \left\|A_i - B_i\right\| \\
&\quad = p\beta^{p-1}\left(\frac{1}{T}\sum_{i=1}^{T} \left\|A_i - B_i\right\|\right) \\
&\quad = p\beta^{p-1} m_1\big(\left\|A_1 - B_1\right\|, \ldots, \left\|A_T - B_T\right\|\big) \\
&\quad \leq p\beta^{p-1} m_p\big(\left\|A_1 - B_1\right\|, \ldots, \left\|A_T - B_T\right\|\big)
\end{aligned}$$

where: the first inequality follows from the triangular inequality, the second inequality follows from Proposition A.2, the third inequality follows as $\beta_i \leq \beta$, and the last inequality comes from the monotonicity of the scalar power means. □

The next Lemma contains the proof corresponding to the case of positive powers of Theorem A.2.

Lemma A.1 (Theorem A.2 for the case $p \geq 1$)**.** *Let $A_1, \dots, A_T, B_1, \dots, B_T$ be symmetric positive semidefinite matrices where $\lambda(A_i) \leq \beta$ and $\lambda(B_i) \leq \beta$ for $i = 1, \dots T$. Let $C_p^+ = p^{1/p}\beta^{1-1/p}$. Let $p \geq 1$, with p integer Then,*

$$\begin{aligned}&\left\|M_p(A_1, \dots, A_T) - M_p(B_1, \dots, B_T)\right\| \\ &\quad \leq C_p^+ m_p\big(\|A_1 - B_1\|, \dots, \|A_T - B_T\| \big)^{1/p}\end{aligned}$$

Proof. We can see that

$$\begin{aligned}&\left\|M_p(A_1, \dots, A_T) - M_p(B_1, \dots, B_T)\right\| \\ &\quad \leq \left\|M_p^p(A_1, \dots, A_T) - M_p^p(B_1, \dots, B_T)\right\|^{\frac{1}{p}} \\ &\quad \leq \left(p\beta^{p-1} m_p\big(\|A_1 - B_1\|, \dots, \|A_T - B_T\| \big) \right)^{\frac{1}{p}} \\ &\quad = C_p^+ m_p\big(\|A_1 - B_1\|, \dots, \|A_T - B_T\| \big)^{1/p}\end{aligned}$$

where the first inequality comes from Proposition A.4, and the second inequality comes from Proposition A.5. □

Results for the case $p \leq -1$

The following two propositions are the main ingredients for the upper bound presented in Theorem A.2 for the case $p \leq -1$.

Proposition A.6. *Let $A_1, \dots, A_T, B_1, \dots, B_T$ be symmetric positive definite matrices where $\alpha \leq \lambda(A_i)$ and $\alpha \leq \lambda(B_i)$ for $i = 1, \dots T$, and $\alpha > 0$. Then, for $p \leq -1$, with p integer*

$$\begin{aligned}&\left\|M_p(A_1, \dots, A_T) - M_p(B_1, \dots, B_T)\right\| \\ &\leq \frac{1}{\alpha^2} \left\|M_{|p|}^{|p|}(A_1^{-1}, \dots, A_T^{-1}) - M_{|p|}^{|p|}(B_1^{-1}, \dots, B_T^{-1})\right\|^{1/|p|}\end{aligned}$$

Proof. Let $S_p(A_1, \dots, A_T) = \frac{1}{T}\sum_{i=1}^T A_i^p$. Then,

$$\begin{aligned}&\left\|M_p(A_1, \dots, A_T) - M_p(B_1, \dots, B_T)\right\| \\ &= \left\|S_p^{1/p}(A_1, \dots, A_T) - S_p^{1/p}(B_1, \dots, B_T)\right\| \\ &= \left\|S_p^{1/|p|}(A_1, \dots, A_T)^{-1} - S_p^{1/|p|}(B_1, \dots, B_T)^{-1}\right\| \\ &= \left\|S_{|p|}^{1/|p|}(A_1^{-1}, \dots, A_T^{-1})^{-1} - S_{|p|}^{1/|p|}(B_1^{-1}, \dots, B_T^{-1})^{-1}\right\|\end{aligned}$$

$$
\begin{aligned}
&= \left\| M_{|p|}(A_1^{-1}, \dots, A_T^{-1})^{-1} - M_{|p|}(B_1^{-1}, \dots, B_T^{-1})^{-1} \right\| \\
&\leq \frac{1}{\alpha^2} \left\| M_{|p|}(A_1^{-1}, \dots, A_T^{-1}) - M_{|p|}(B_1^{-1}, \dots, B_T^{-1}) \right\| \\
&\leq \frac{1}{\alpha^2} \left\| M_{|p|}^{|p|}(A_1^{-1}, \dots, A_T^{-1}) - M_{|p|}^{|p|}(B_1^{-1}, \dots, B_T^{-1}) \right\|^{1/|p|}
\end{aligned}
$$

where the first inequality follows from Corollary A.1 and Proposition A.3, whereas the second inequality follows from Proposition A.4. □

Proposition A.7. *Let $A_1, \dots, A_T, B_1, \dots, B_T$ be symmetric positive definite matrices such that $\alpha \leq \lambda(A_i)$ and $\alpha \leq \lambda(B_i)$ for $i = 1, \dots T$. Then, for $p \leq -1$, with p integer*

$$
\begin{aligned}
&\left\| M_{|p|}^{|p|}(A_1^{-1}, \dots, A_T^{-1}) - M_{|p|}^{|p|}(B_1^{-1}, \dots, B_T^{-1}) \right\| \\
&\qquad \leq |p| \alpha^{-(1+|p|)} m_{|p|}\big(\|A_1 - B_1\|, \dots, \|A_T - B_T\| \big)
\end{aligned}
$$

Proof. Let $\alpha_i = \min(\|A_i\|, \|B_i\|)$, then it clearly follows that $\frac{1}{\alpha_i} = \max(\left\|A_i^{-1}\right\|, \left\|B_i^{-1}\right\|)$. Thus,

$$
\begin{aligned}
&\left\| M_{|p|}^{|p|}(A_1^{-1}, \dots, A_T^{-1}) - M_{|p|}^{|p|}(B_1^{-1}, \dots, B_T^{-1}) \right\| \\
&\qquad = \left\| \left(\frac{1}{T} \sum_{i=1}^{T} (A_i^{-1})^{|p|} \right) - \left(\frac{1}{T} \sum_{i=1}^{T} (B_i^{-1})^{|p|} \right) \right\| \\
&\qquad = \left\| \frac{1}{T} \sum_{i=1}^{T} (A_i^{-1})^{|p|} - (B_i^{-1})^{|p|} \right\| \\
&\qquad \leq \frac{1}{T} \sum_{i=1}^{T} \left\| (A_i^{-1})^{|p|} - (B_i^{-1})^{|p|} \right\| \\
&\qquad \leq \frac{1}{T} \sum_{i=1}^{T} |p| \Big(\frac{1}{\alpha_i} \Big)^{|p|-1} \left\| A_i^{-1} - B_i^{-1} \right\| \\
&\qquad \leq |p| \Big(\frac{1}{\alpha} \Big)^{|p|-1} \left(\frac{1}{T} \sum_{i=1}^{T} \left\| A_i^{-1} - B_i^{-1} \right\| \right) \\
&\qquad \leq |p| \Big(\frac{1}{\alpha} \Big)^{|p|-1} \left(\frac{1}{T\alpha^2} \sum_{i=1}^{T} \|A_i - B_i\| \right) \\
&\qquad = |p| \Big(\frac{1}{\alpha} \Big)^{|p|+1} \left(\frac{1}{T} \sum_{i=1}^{T} \|A_i - B_i\| \right) \\
&\qquad = |p| \Big(\frac{1}{\alpha} \Big)^{|p|+1} m_1\big(\|A_1 - B_1\|, \dots, \|A_T - B_T\| \big) \\
&\qquad \leq |p| \Big(\frac{1}{\alpha} \Big)^{|p|+1} m_{|p|}\big(\|A_1 - B_1\|, \dots, \|A_T - B_T\| \big) \\
&\qquad = |p| \alpha^{-(1+|p|)} m_{|p|}\big(\|A_1 - B_1\|, \dots, \|A_T - B_T\| \big)
\end{aligned}
$$

where: the first inequality follows from the triangular inequality, the second inequality follows from Proposition A.2, the third inequality follows as $\alpha \leq \alpha_i$, and the fourth inequality follows as Corollary A.1, and the last inequality comes from the monotonicity of the scalar power means. □

The next Lemma contains the proof corresponding to the case of negative powers of Theorem A.2.

Lemma A.2 (Theorem A.2 for the case $p \leq -1$)**.** *Let $A_1, \ldots, A_T, B_1, \ldots, B_T$ be symmetric positive definite matrices where $\alpha \leq \lambda(A_i)$ and $\alpha \leq \lambda(B_i)$ for $i = 1, \ldots T$. Let $C_p^- = |p|^{1/|p|}\alpha^{-(3+1/|p|)}$. Let $p \leq -1$ with p integer. Then,*

$$\begin{aligned}&\left\|M_p(A_1, \ldots, A_T) - M_p(B_1, \ldots, B_T)\right\| \\ &\qquad \leq C_p^- m_{|p|}\big(\, \|A_1 - B_1\|, \ldots, \|A_T - B_T\| \,\big)^{1/|p|}\end{aligned}$$

Proof.

$$\begin{aligned}&\left\|M_p(A_1, \ldots, A_T) - M_p(B_1, \ldots, B_T)\right\| \\ &\leq \frac{1}{\alpha^2}\left\|M_{|p|}^{|p|}(A_1^{-1}, \ldots, A_T^{-1}) - M_{|p|}^{|p|}(B_1^{-1}, \ldots, B_T^{-1})\right\|^{1/|p|} \\ &\leq \frac{1}{\alpha^2}\left(|p|\alpha^{-(1+|p|)} m_{|p|}\big(\, \|A_1 - B_1\|, \ldots, \|A_T - B_T\| \,\big)\right)^{1/|p|} \\ &= C_p^- m_{|p|}\big(\, \|A_1 - B_1\|, \ldots, \|A_T - B_T\| \,\big)^{1/|p|}\end{aligned}$$

where the first inequality comes from Proposition A.6, and the second inequality comes from Proposition A.7. □

We are now ready to prove the result of Theorem A.2.

Proof of Theorem A.2. For the case $p \geq 1$ see Lemma A.1. For the case $p \leq -1$ see Lemma A.2.

□

A.3 PROOF OF THEOREM 3.6

Before giving the proof of Theorem 3.6 we need to present two auxiliary results.

The following is an auxiliary technical result that extends an implicit result stated in (Rohe *et al.*, 2011)(p.1908-1909) for the Frobenius norm to the case of the operator norm.

Lemma A.3. *Let $X, \mathcal{X} \in \mathbb{R}^{n\times k}$ be matrices with orthonormal columns. Let U, V be orthonormal matrices and Σ a diagonal matrix such that*

$$\mathcal{X}^T X = U\Sigma V^T$$

where the diagonal entries of Σ are the cosines of the principal angles between the column space of X and the column space of $\mathcal{X}$. Let $O = UV^T$. Then,

$$\frac{1}{\sqrt{2}}\|X - \mathcal{X}O\| \leq \|\sin\Theta(\mathcal{X}, X)\|$$

Proof. For the proof we use the identity $X^T\mathcal{X}O = (V\Sigma U^T)UV^T = V\Sigma V^T$, and the fact that $\|A\| = \sqrt{\lambda_{\max}(A^TA)}$. That is,

$$\begin{aligned}
(X-&\mathcal{X}O)^T(X - \mathcal{X}O) \\
&= (X^T - O^T\mathcal{X}^T)(X - \mathcal{X}O) \\
&= X^TX - X^T\mathcal{X}O - O^T\mathcal{X}^TX + O^T\mathcal{X}^T\mathcal{X}O \\
&= I - X^T\mathcal{X}O - O^T\mathcal{X}^TX + O^TO \\
&= I - X^T\mathcal{X}O - O^T\mathcal{X}^TX + I \\
&= 2I - X^T\mathcal{X}O - O^T\mathcal{X}^TX \\
&= 2I - V\Sigma V^T - V\Sigma V^T \\
&= 2(I - V\Sigma V^T)
\end{aligned}$$

Thus,

$$\begin{aligned}
\|X - \mathcal{X}O\|^2 &= \lambda_{\max}\Big((X - \mathcal{X}O)^T(X - \mathcal{X}O)\Big) \\
&= 2\lambda_{\max}(I - V\Sigma V^T) \\
&= 2\max_i(1 - \cos\Theta_i) \\
&\leq 2\max_i(1 - \cos^2\Theta_i) \\
&= 2\max_i(\sin^2\Theta_i) \\
&= 2\|\sin\Theta\|^2
\end{aligned}$$

Hence, $\frac{1}{\sqrt{2}}\|X - \mathcal{X}O\| \leq \|\sin\Theta(\mathcal{X}, X)\|$ □

The next result is a useful representation of the Davis-Kahan theorem. It is a technical adaption from the Frobenius norm to the operator norm based on Lemma A.3 and Theorem A.7.

Theorem A.4. *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\mu_1 \geq \ldots \geq \mu_p$ and $\hat{\mu}_1 \geq \ldots \geq \hat{\mu}_p$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1}) > 0$, where $\mu_0 := \infty$ and $\mu_{p+1} := -\infty$. Let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \mu_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\mu}_j\hat{v}_j$ for $j = r, r+1, \ldots, s$. Then there exists an orthogonal matrix $O \in \mathbb{R}^{d\times d}$ such that*

$$\frac{1}{\sqrt{2}}\|V - \hat{V}O\| \leq \frac{2d^{1/2}\|\hat{\Sigma} - \Sigma\|}{\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1})}$$

Proof. By theorem A.7 we have

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{2\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1})}.$$

From lemma A.3 we can see that

$$\frac{1}{\sqrt{2}}\|V - \hat{V}O\| \leq \|\sin\Theta(\hat{V}, V)\|$$

Moreover, as $\sin\Theta(\hat{V}, V)$ is a diagonal matrix, it holds that

$$\begin{aligned}\|\sin\Theta(\hat{V}, V)\|^2 &= \max_i(\sin^2\Theta_i) \\ &\leq \sum_i^p \sin^2\Theta_i \\ &= \|\sin\Theta(\hat{V}, V)\|_F^2\end{aligned}$$

Thus,

$$\frac{1}{\sqrt{2}}\|V - \hat{V}O\| \leq \|\sin\Theta(\hat{V}, V)\| \leq \|\sin\Theta(\hat{V}, V)\|_F$$

Further, it is straightforward to see that

$$\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}}) \leq d^{1/2}\|\hat{\Sigma} - \Sigma\|$$

Thus, all in all, we have

$$\begin{aligned}\frac{1}{\sqrt{2}}\|V - \hat{V}O\| &\leq \|\sin\Theta(\hat{V}, V)\| \\ &\leq \|\sin\Theta(\hat{V}, V)\|_F \\ &\leq \frac{2d^{1/2}\|\hat{\Sigma} - \Sigma\|}{\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1})}\end{aligned}$$

which completes the proof. □

We are now ready to give the proof of Theorem 3.6.

Proof of Theorem 3.6. The proof is an application of the Davis-Kahan theorem as presented in Theorem A.4. Observe that in Theorem A.4 the eigenvalues are sorted in a decreasing way i.e. $\mu_1 \geq \cdots \geq \mu_n$, whereas in our case they are sorted in an increasing manner i.e. $\lambda_1 \leq \cdots \leq \lambda_n$.

Notationally, let the variables p, s, r from Theorem A.4 be defined as $p = s = n, r = p - k + 1$.

We first focus in the case for $p \leq -1$. For this case we are interested in the k-smallest eigenvalues, i.e. $\lambda_1, \ldots, \lambda_k$, which correspond to $\mu_p, \ldots, \mu_r$, where $\mu_p = \lambda_1$ and $\mu_r = \lambda_k$.

By definition, in Theorem A.4 we have that $\mu_{p+1} = -\infty$. Thus, $\mu_p - \mu_{p+1} = \infty$. Further, we can see $\mu_{r-1} - \mu_r = \lambda_{k+1} - \lambda_k = (1+\varepsilon) - m_p(1-\rho^+ + \varepsilon, 1+\rho^- + \varepsilon)$ and hence by Eq.C.7

$$\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1}) = (1+\varepsilon) - m_p(1-\rho^+ + \varepsilon, 1+\rho^- + \varepsilon)$$

which by Theorem A.4 leads to the following inequality

$$\|V_k - \mathcal{V}_k O_k\| \leq \frac{2^{3/2}k^{1/2}}{\gamma}\|L_p - \mathcal{L}_p\| = \frac{\sqrt{8k}}{\gamma}\|L_p - \mathcal{L}_p\|$$

By applying Theorem 3.5, we know that if

$$\delta^+ = \frac{n}{k}(p^+_{\text{in}} + (k-1)p^+_{\text{out}}) > 3\ln(8n/\epsilon), \text{ and}$$
$$\delta^- = \frac{n}{k}(p^-_{\text{in}} + (k-1)p^-_{\text{out}}) > 3\ln(8n/\epsilon)$$

then with probability at least $1-\epsilon$

$$\|V_k - \mathcal{V}_k O_k\| \leq \frac{\sqrt{8k}}{\gamma} C_p^- m_{|p|}^{1/|p|} \left(\sqrt{\frac{3\ln(8n/\epsilon)}{\delta^+}}, \sqrt{\frac{3\ln(8n/\epsilon)}{\delta^-}}\right)$$

yielding the desired result. The case for $p \geq 1$ is similar, where instead of k the value $k' = k-1$ is used.

□

A.4 MAIN BUILDING BLOCK FOR OUR RESULTS

In this section present two results from (Chung and Radcliffe, 2011) that are the main building blocks for our results.

Theorem A.5 ((Chung and Radcliffe, 2011)). *Let G be a random graph, where $pr(v_i \sim v_j) = p_{ij}$, and each edges is independent of each other edge. Let A be the adjacency matrix of G, so $A_{ij} = 1$ if $v_i \sim v_j$ and 0 otherwise, and $\bar{A} = E(A)$, so $\bar{A}_{ij} = p_{i,j}$. Let D be the diagonal matrix with $D_{ii} = deg(v_i)$, and $\bar{D} = E(D)$. Let δ be the minimum expected degree of G, and $L = I - D^{-1/2}AD^{-1/2}$ the (normalized) Laplacian matrix for G. Choose $\epsilon > 0$. Then there exists a constant $k = k(\epsilon)$ such that if $\delta > k\ln n$, then the probability at least $1-\epsilon$, the eigenvalues of L and $\bar{L}$ satisfy*

$$|\lambda_j(L) - \lambda_j(\bar{L})| \leq 2\sqrt{\frac{3\ln(4n/\epsilon)}{\delta}}$$

for all $1 \leq j \leq n$, where $\bar{L} = I - \bar{D}^{-1/2}\bar{A}\bar{D}^{-1/2}$.

Although this theorem is presented as the main result, one can see in the proof of theorem A.5 in (Chung and Radcliffe, 2011), that in deed what they proved was a concentration bound for $\|L - \bar{L}\|$.

Theorem A.6 ((Chung and Radcliffe, 2011)). *Assume that conditions of Theorem A.5 hold. Choose $\epsilon > 0$. Then there exists a constant $k = k(\epsilon)$ such that if $\delta > k \ln n$, then*

$$\mathbb{P}\left(\|L - \bar{L}\| \leq 2\sqrt{\frac{3\ln(4n/\epsilon)}{\delta}} \right) > 1 - \epsilon \tag{A.3}$$

Theorem A.7 (Yu *et al.*, 2015)). *Let $\Sigma, \hat{\Sigma} \in \mathbb{R}^{p\times p}$ be symmetric, with eigenvalues $\mu_1 \geq \ldots \geq \mu_p$ and $\hat{\mu}_1 \geq \ldots \geq \hat{\mu}_p$ respectively. Fix $1 \leq r \leq s \leq p$ and assume that $\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1}) > 0$, where $\mu_0 := \infty$ and $\mu_{p+1} := -\infty$. Let $d := s - r + 1$, and let $V = (v_r, v_{r+1}, \ldots, v_s) \in \mathbb{R}^{p\times d}$ and $\hat{V} = (\hat{v}_r, \hat{v}_{r+1}, \ldots, \hat{v}_s) \in \mathbb{R}^{p\times d}$ have orthonormal columns satisfying $\Sigma v_j = \mu_j v_j$ and $\hat{\Sigma}\hat{v}_j = \hat{\mu}_j \hat{v}_j$ for $j = r, r+1, \ldots, s$. Then*

$$\|\sin\Theta(\hat{V}, V)\|_{\mathrm{F}} \leq \frac{2\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1})}.$$

Moreover, there exists an orthogonal matrix $\hat{O} \in \mathbb{R}^{d\times d}$ such that

$$\|\hat{V}\hat{O} - V\|_{\mathrm{F}} \leq \frac{2^{3/2}\min(d^{1/2}\|\hat{\Sigma} - \Sigma\|, \|\hat{\Sigma} - \Sigma\|_{\mathrm{F}})}{\min(\mu_{r-1} - \mu_r, \mu_s - \mu_{s+1})}.$$

PROOF OF THEOREMS 4.2, AND 4.3

B

In this setting, we fix the number k of cluster to $k = 3$.
For convenience, we slightly overload the notation for the remaining of this section: we denote by n the size of each cluster $\mathcal{C}_1, \dots, \mathcal{C}_k$, i.e. $|\mathcal{C}_i| = |\mathcal{C}| = n$ for $i = 1, \dots, k$. Thus, the size of the graph is expressed in terms of the number and size of clusters, i.e. $|V| = nk$.

Furthermore, we suppose that for $t = 1, 2, 3$, the expected adjacency matrix $\mathcal{W}^{(t)} \in \mathbb{R}^{3n \times 3n}$ of $G^{(t)}$, are given, for all $i, j = 1, \dots, 3n$, as

$$\mathcal{W}_{ij}^{(t)} = \begin{cases} p_{in} & \text{if } v_i, v_j \in \mathcal{C}_t \text{ or } v_i, v_j \in \overline{\mathcal{C}_t} \\ p_{out} & \text{otherwise,} \end{cases}$$

where $0 < p_{out} \leq p_{in} \leq 1$. For $t = 1, 2, 3$ and $\epsilon \geq 0$, let $\mathcal{D}^{(t)} = \operatorname{diag}(\mathcal{W}^{(t)}\mathbf{1})$,

$$\mathcal{L}_{\text{sym}}^{(t)} = I - (\mathcal{D}^{(t)})^{-1/2} \mathcal{W}^{(t)} (\mathcal{D}^{(t)})^{-1/2} + \epsilon I,$$

and for a nonzero integer p let

$$\mathcal{L}_p = M_p(\mathcal{L}_{\text{sym}}^{(1)}, \mathcal{L}_{\text{sym}}^{(2)}, \mathcal{L}_{\text{sym}}^{(3)}),$$

where we assume that $\epsilon > 0$ if $p < 0$. Consider further $\boldsymbol{\chi}_1, \boldsymbol{\chi}_2, \boldsymbol{\chi}_3 \in \mathbb{R}^{3n}$ the vectors defined as

$$\boldsymbol{\chi}_1 = \mathbf{1}, \quad \boldsymbol{\chi}_2 = \mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_2}, \quad \boldsymbol{\chi}_3 = \mathbf{1}_{\mathcal{C}_1} - \mathbf{1}_{\mathcal{C}_3}.$$

In opposition to the previous model, it turns out that $\mathcal{L}_{\text{sym}}^{(1)}, \mathcal{L}_{\text{sym}}^{(2)}, \mathcal{L}_{\text{sym}}^{(3)}$ do not commute and thus do not share the same eigenvectors. Hence, we can not derive an explicit expression for $\mathcal{L}_p$. In particular this implies that we need to use different mathematical tools in order to study the eigenpairs of $\mathcal{L}_p$.

The first main result of this section, presented in Theorem B.1, shows that, in general, the ground truth clusters can not be reconstructed from the 3 smallest eigenvectors of $\mathcal{L}_{\text{sym}}^{(t)}$ for any $t = 1, 2, 3$.

Theorem B.1. *If $1 \geq p_{\text{in}}^{+} > p_{\text{out}}^{+} > 0$, then for any $t = 1, 2, 3$, there exist scalars $\alpha > 0$ and $\beta > 0$ such that the eigenvectors of $\mathcal{L}_{\text{sym}}^{(t)}$ corresponding to the two smallest eigenvalues are*

$$\boldsymbol{\varkappa}_1 = \alpha \mathbf{1}_{\mathcal{C}_t} + \mathbf{1}_{\overline{\mathcal{C}_t}} \quad \textit{and} \quad \boldsymbol{\varkappa}_2 = -\beta \mathbf{1}_{\mathcal{C}_t} + \mathbf{1}_{\overline{\mathcal{C}_t}}$$

whereas any vector orthogonal to both $\boldsymbol{\varkappa}_1$ and $\boldsymbol{\varkappa}_2$ is an eigenvector for the third smallest eigenvalue.

In fact, we prove even more by giving a full description of the eigenvectors of $L_{sym}^{(t)}$ as well as the ordering of their corresponding eigenvalues. These results can be found in Lemma B.11 below.

Our second main result is the following Theorem B.2. It shows that the ground truth clusters can always be recovered from the three smallest eigenvectors of $\mathcal{L}_p$.

Theorem B.2. *Let p be any nonzero integer and assume that $\epsilon > 0$ if $p < 0$. Furthermore, suppose that $0 < p_{out} < p_{in} \leq 1$. Then, there exists λ_i such that $\mathcal{L}_p \boldsymbol{\chi}_i = \lambda_i \boldsymbol{\chi}_i$ for $i = 1, 2, 3$ and $\lambda_1, \lambda_2, \lambda_3$ are the three smallest eigenvalues of $\mathcal{L}_p$.*

We actually prove more than just Theorem B.2. In fact, a full description of the eigenvectors of $\mathcal{L}_p$ and of the ordering of their corresponding eigenvalues is given in Lemma B.17 below.

For the proof of Theorem B.2, and the corresponding additional results, we proceed as follows. First we assume that $n = |\mathcal{C}_i| = 1$ and prove our claims. Then, we generalize these results to the case $n > 1$. For the sake of clarity, as we will need to refer to the case $n = 1$ for the proofs of the case $n > 1$, we put a tilde on the matrices in $\mathbb{R}^{3\times 3}$.

B.1 THE CASE $n = 1$

Suppose that $n = 1$, then $\tilde{\mathcal{L}}_{\text{sym}} = \mathcal{L}_{\text{sym}}^{(1)}$ is given by

$$\tilde{\mathcal{L}}_{\text{sym}} = \tau I_3 - \tilde{\mathcal{D}}^{-1/2} \tilde{\mathcal{W}} \tilde{\mathcal{D}}^{-1/2} = \tau I_3 - \tilde{\mathcal{M}},$$

where $\tau = 1 + \epsilon$, $\tilde{\mathcal{W}} = \mathcal{W}^{(1)}$, $\tilde{\mathcal{D}} = \text{diag}(\tilde{\mathcal{W}}\mathbf{1})$, $\tilde{\mathcal{M}} = \tilde{\mathcal{D}}^{-1/2} \tilde{\mathcal{W}} \tilde{\mathcal{D}}^{-1/2}$

$$\tilde{\mathcal{W}} = \begin{pmatrix} p_{in} & p_{out} & p_{out} \\ p_{out} & p_{in} & p_{in} \\ p_{out} & p_{in} & p_{in} \end{pmatrix}, \qquad \tilde{\mathcal{D}} = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & \beta \end{pmatrix}, \qquad \tilde{\mathcal{M}} = \begin{pmatrix} a & b & b \\ b & c & c \\ b & c & c \end{pmatrix}, \tag{B.1}$$

and $\alpha, \beta, a, b, c > 0$ are given by

$$\begin{aligned} \alpha &= p_{in} + 2p_{out}, \qquad \beta = 2p_{in} + p_{out}, \\ a &= \frac{p_{in}}{\alpha}, \qquad b = \frac{p_{out}}{\sqrt{\alpha\beta}}, \qquad c = \frac{p_{in}}{\beta}. \end{aligned}$$

Moreover, note that for any $(\lambda, \mathbf{v}) \in \mathbb{R} \times \mathbb{R}^3$ we have

$$\tilde{\mathcal{M}}\mathbf{v} = \lambda \mathbf{v} \qquad \Longleftrightarrow \qquad \tilde{\mathcal{L}}_{\text{sym}} \mathbf{v} = (\tau - \lambda)\mathbf{v}. \tag{B.2}$$

This implies that we can study the spectrum of $\tilde{\mathcal{M}}$ in order to obtain the spectrum of $\tilde{\mathcal{L}}_{\text{sym}}$. We have the following lemma:

Lemma B.1. *Suppose that $p_{out} > 0$ and let $\Delta > 0$ be defined as $\Delta = \sqrt{(a-2c)^2 + 8b^2}$. Then the eigenvalues of $\tilde{\mathcal{M}}$ are*

$$\tilde{\lambda}_1 = 0, \qquad \tilde{\lambda}_2 = \frac{a + 2c - \Delta}{2}, \qquad \tilde{\lambda}_3 = 1,$$

and it holds $\tilde{\lambda}_1 < \tilde{\lambda}_2 < \tilde{\lambda}_3$. Furthermore, the corresponding eigenvectors are given by

$$\mathbf{u}_1 = (0, -1, 1)^\top, \qquad \mathbf{u}_2 = \Big(\frac{a - 2c - \Delta}{2b}, 1, 1\Big)^\top, \qquad \mathbf{u}_3 = \big(\sqrt{\alpha}, \sqrt{\beta}, \sqrt{\beta}\big)^\top,$$

and it holds $\frac{a-2c-\Delta}{2b} < 0$.

Proof. The equality $\tilde{\mathcal{M}}\mathbf{u}_1 = 0$ follows from a direct computation. Furthermore, note that $\mathbf{u}_3 = \tilde{\mathcal{D}}^{1/2}\mathbf{1}$ and so

$$\tilde{\mathcal{M}}\mathbf{u}_3 = \tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2}\mathcal{D}^{1/2}\mathbf{1} = \tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\mathbf{1} = \mathbf{u}_3$$

implying $\tilde{\mathcal{M}}\mathbf{u}_3 = \mathbf{u}_3$. Now, let $s_\pm = \frac{a-2c\pm\Delta}{2b}$. Then s_+ and s_- are the solutions of the quadratic equation $bs^2 + (2c - a)s - 2b = 0$ which can be rearranged as $as + 2b = (bs + 2c)s$. The latter equation is equivalent to

$$\begin{cases} as + 2b = \lambda s \\ bs + 2c = \lambda \end{cases} \qquad \Longleftrightarrow \qquad \tilde{\mathcal{M}} \begin{pmatrix} s \\ 1 \\ 1 \end{pmatrix} = \lambda \begin{pmatrix} s \\ 1 \\ 1 \end{pmatrix}.$$

Hence, $\mathbf{u}_\pm = (s_\pm, 1, 1)$ are both eigenvectors of $\tilde{\mathcal{M}}$ corresponding to the eigenvalues

$$\lambda_\pm = b\, s_\pm + 2c = \frac{a - 2c \pm \Delta}{2} + 2c = \frac{a + 2c \pm \Delta}{2}.$$

Note in particular that we have $\mathbf{u}_2 = \mathbf{u}_-$ and $\tilde{\lambda}_2 = \lambda_-$. This concludes the proof that $(\lambda_i, \mathbf{u}_i)$ are eigenpairs of $\tilde{\mathcal{M}}$ for $i = 1, 2, 3$. We now show that $\tilde{\lambda}_1 < \tilde{\lambda}_2 < \tilde{\lambda}_3$ and $(a - 2c - \Delta)/2b < 0$.

As $\Delta > 0$, we have $\lambda_- < \lambda_+$. We prove $\lambda_- > 0$. As $p_{in} > p_{out}$ by assumption, the definition of $a, b, c > 0$ implies that

$$\begin{aligned} b^2 &= \frac{p_{out}^2}{(2p_{in} + p_{out})(p_{in} + 2p_{out})} \\ &< \frac{p_{in}^2}{(2p_{in} + p_{out})(p_{in} + 2p_{out})} = ac. \end{aligned}$$

And from $ac > b^2$ it follows that $a^2 + 4ac + 4c^2 > a^2 - 4ac + 4c^2 + 8b^2$ which implies that $(a + 2c)^2 > (a - 2c)^2 + 8b^2 = \Delta^2$. Hence, $a + 2c - \Delta > 0$ and thus $\lambda_- > 0$. Thus we have $0 < \lambda_- < \lambda_+$. Now, as $\tilde{\mathcal{M}}$ has strictly positive entries, the Perron-Frobenius theorem (see for instance Theorem 1.1 in (Tudisco *et al.*, 2015)) implies that $\tilde{\mathcal{M}}$ has a unique nonnegative eigenvector $\mathbf{u}$. Furthermore, $\mathbf{u}$ has positive entries and its corresponding eigenvalue is the spectral radius of $\tilde{\mathcal{M}}$. As $\mathbf{u}_3 = \tilde{\mathcal{D}}^{1/2}\mathbf{1}$ has positive entries and is an eigenvector of $\tilde{\mathcal{M}}$, we have $\mathbf{u} = \mathbf{u}_3$. It follows that $\rho(\tilde{\mathcal{M}}) = \lambda_+ = \tilde{\lambda}_3$. Furthermore, $\mathbf{u}_2$ must have a strictly negative entry and thus it holds $s_- < 0$. □

Combining the results of Lemma B.1 and Equation (B.2) we directly obtain the following corollary which fully describes the eigenvectors of $\tilde{\mathcal{L}}_{\text{sym}}$ as well as the ordering of the corresponding eigenvalues:

Corollary B.1. *There exists $\tilde{\lambda} \in (0,1)$ and $s_- < 0 < s_+ < 1$ such that*

$$\big(\tau - 1, (s_+, 1, 1)^\top\big), \qquad \big(\tau - \tilde{\lambda}, (s_-, 1, 1)^\top\big), \qquad \big(\tau, (0, -1, 1)^\top\big)$$

are the eigenpairs of $\tilde{\mathcal{L}}_{\text{sym}}$.

Proof. The only thing which is not directly implied by Lemma B.1 and Equation (B.2) is that $s_+ < 1$. But this follows again from Lemma B.1. Indeed, by the Perron-Frobenius theorem the nonnegative eigenvector is unique, i.e. $(s_+, 1, 1)$ and $(\sqrt{\alpha}, \sqrt{\beta}, \sqrt{\beta})$ must span the same line. Hence we have

$$s_+ = \sqrt{\frac{\alpha}{\beta}} = \sqrt{\frac{p_{in} + 2p_{out}}{2p_{in} + p_{out}}}.$$

As $p_{out} < p_{in}$, we get $0 < s_+ < 1$. □

Now, we study the spectral properties of $\tilde{\mathcal{L}}_p = \mathcal{L}_p \in \mathbb{R}^{3\times 3}$. To this end, for $t = 1, 2, 3$ let $\tilde{\mathcal{W}}^{(t)} = \mathcal{W}^{(t)}, \tilde{\mathcal{L}}^{(t)}_{\text{sym}} = \mathcal{L}^{(t)}_{\text{sym}} \in \mathbb{R}^{3\times 3}$. Furthermore, consider the permutation matrices $\tilde{P}_1, \tilde{P}_2, \tilde{P}_3 \in \mathbb{R}^{3\times 3}$ defined as

$$\tilde{P}_1 = I_3, \quad \tilde{P}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \tilde{P}_3 = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, we have $\tilde{\mathcal{W}}^{(t)} = \tilde{P}_t \tilde{\mathcal{W}} \tilde{P}_t$ for $t = 1, 2, 3$. The following lemma relates $\tilde{\mathcal{L}}^{(t)}_{\text{sym}}$ and $\tilde{\mathcal{L}}_{\text{sym}}$.

Lemma B.2. *For $t = 1, 2, 3$, we have $\tilde{P}_t = \tilde{P}_t^{-1} = \tilde{P}_t^\top$ and $\tilde{\mathcal{L}}^{(t)}_{\text{sym}} = \tilde{P}_t \tilde{\mathcal{L}}_{\text{sym}} \tilde{P}_t$.*

Proof. The identity $\tilde{P}_t = \tilde{P}_t^{-1} = \tilde{P}_t^\top$ follows by a direct computation. Now, as $\tilde{P}_t \mathbf{1} = \mathbf{1}$, we have $\tilde{P}_t \tilde{\mathcal{W}} \tilde{P}_t \mathbf{1} = \tilde{P}_t \tilde{\mathcal{W}} \mathbf{1}$. Assuming the exponents on the vector in the following expressions are taken component wise, we have $\operatorname{diag}(\tilde{\mathcal{W}}\mathbf{1})^{-1/2} = \operatorname{diag}\big((\tilde{\mathcal{W}}\mathbf{1})^{-1/2}\big)$ and thus

$$\begin{aligned} \operatorname{diag}(\tilde{P}_t \tilde{\mathcal{W}} \tilde{P}_t \mathbf{1})^{-1/2} &= \operatorname{diag}\big((\tilde{P}_t \tilde{\mathcal{W}} \tilde{P}_t \mathbf{1})^{-1/2}\big) \\ &= \operatorname{diag}\big(\tilde{P}_t (\tilde{\mathcal{W}} \mathbf{1})^{-1/2}\big) \\ &= \tilde{P}_t \operatorname{diag}\big((\tilde{\mathcal{W}} \mathbf{1})^{-1/2}\big) \tilde{P}_t \\ &= \tilde{P}_t \operatorname{diag}(\tilde{\mathcal{W}} \mathbf{1})^{-1/2} \tilde{P}_t \\ &= \tilde{P}_t \tilde{\mathcal{D}}^{-1/2} \tilde{P}_t. \end{aligned}$$

It follows that

$$\begin{aligned}\tilde{\mathcal{L}}_{\text{sym}}^{(t)} &= \tau\tilde{P}_t\tilde{P}_t - \tilde{P}_t\tilde{\mathcal{D}}^{-1/2}\tilde{P}_t\tilde{P}_t\tilde{\mathcal{W}}\tilde{P}_t\tilde{P}_t\tilde{\mathcal{D}}^{-1/2}\tilde{P}_t \\ &= \tau\tilde{P}_t\tilde{P}_t - \tilde{P}_t\tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2}\tilde{P}_t \\ &= \tilde{P}_t(\tau I_3 - \tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2})\tilde{P}_t \\ &= \tilde{P}_t\tilde{\mathcal{L}}_{\text{sym}}\tilde{P}_t,\end{aligned}$$

which concludes our proof. □

Combining Corollary B.1 with Lemma B.2, we directly obtain the following

Corollary B.2. *There exists $\tilde{\lambda} \in (0,1)$ and $s_- < 0 < s_+$ such that*

$$(\tau - 1, \tilde{P}_t(s_+, 1, 1)^\top), \qquad (\tau - \lambda, \tilde{P}_t(s_-, 1, 1)^\top), \qquad (\tau, \tilde{P}_t(0, -1, 1)^\top)$$

are the eigenpairs of $\tilde{\mathcal{L}}_{\text{sym}}^{(t)}$ for $t = 1, 2, 3$.

A similar argument as in the proof of Lemma 2.1 implies that the eigenvectors of $\tilde{\mathcal{L}}_p$ coincide with those of the matrix $\tilde{\mathsf{L}}_p \in \mathbb{R}^{3\times 3}$ defined as

$$\tilde{\mathsf{L}}_p = (\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p + (\tilde{\mathcal{L}}_{\text{sym}}^{(2)})^p + (\tilde{\mathcal{L}}_{\text{sym}}^{(3)})^p = 3\tilde{\mathcal{L}}_p^p.$$

We study the spectral properties of $\tilde{\mathsf{L}}_p$. To this end, we consider the following subspaces of matrices:

$$\mathcal{U}_3 = \left\{ \begin{pmatrix} s_1 & s_2 & s_2 \\ s_3 & s_5 & s_4 \\ s_3 & s_4 & s_5 \end{pmatrix} \;\middle|\; s_1, \dots, s_5 \in \mathbb{R} \right\},$$

$$\mathcal{Z}_3 = \left\{ \begin{pmatrix} t_1 & t_2 & t_2 \\ t_2 & t_1 & t_2 \\ t_2 & t_2 & t_1 \end{pmatrix} \;\middle|\; t_1, t_2 \in \mathbb{R} \right\}.$$

We prove that for every p, it holds $(\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p \in \mathcal{U}_3$ and $\tilde{\mathsf{L}}_p \in \mathcal{Z}_3$. We need the following lemma:

Lemma B.3. *The following holds:*

1. *For all $\tilde{A}, \tilde{B} \in \mathcal{U}_3$ we have $\tilde{A}\tilde{B} \in \mathcal{U}_3$.*
2. *If $\tilde{A} \in \mathcal{U}_3$ and $det(\tilde{A}) \neq 0$, then $\tilde{A}^{-1} \in \mathcal{U}_3$.*
3. *$\mathcal{Z}_3 = \tilde{P}_1\mathcal{U}_3\tilde{P}_1 + \tilde{P}_2\mathcal{U}_3\tilde{P}_2 + \tilde{P}_3\mathcal{U}_3\tilde{P}_3$.*

Proof. Let $\tilde{A} \in \mathcal{U}_3, \tilde{C} \in \mathcal{Z}_3$ be respectively defined as

$$\tilde{A} = \begin{pmatrix} s_1 & s_2 & s_2 \\ s_3 & s_5 & s_4 \\ s_3 & s_4 & s_5 \end{pmatrix}, \quad \tilde{C} = \begin{pmatrix} t_1 & t_2 & t_2 \\ t_2 & t_1 & t_2 \\ t_2 & t_2 & t_1 \end{pmatrix}$$

1. Follows from a direct computation.

2. If $det(\tilde{A}) \neq 0$, then $\tilde{A}$ is invertible and

$$det(\tilde{A})\tilde{A}^{-1} = \begin{pmatrix} s_5^2 - s_4^2 & s_2(s_4 - s_5) & s_2(s_4 - s_5) \\ s_3(s_4 - s_5) & s_1 s_5 - s_2 s_3 & s_2 s_3 - s_1 s_4 \\ s_3(s_4 - s_5) & s_2 s_3 - s_1 s_4 & s_1 s_5 - s_2 s_3 \end{pmatrix}.$$

   It follows that $\tilde{A}^{-1} \in \mathcal{U}_3$.

3. We have

$$\sum_{i=1}^{3} \tilde{P}_i \tilde{A} \tilde{P}_i = \begin{pmatrix} s_1 + 2s_5 & s_2 + s_3 + s_4 & s_2 + s_3 + s_4 \\ s_2 + s_3 + s_4 & s_1 + 2s_5 & s_2 + s_3 + s_4 \\ s_2 + s_3 + s_4 & s_2 + s_3 + s_4 & s_1 + 2s_5 \end{pmatrix} \tag{B.3}$$

   and conversely, there clearly exists $s_1, \ldots, s_4$ such that $s_1 + 2s_5 = t_1$ and $s_2 + s_3 + s_4 = t_2$, so we have $\sum_{i=1}^{3} \tilde{P}_i \tilde{A} \tilde{P}_i = \tilde{C}$ implying the reverse inclusion.

□

Now, we show that $\tilde{\mathsf{L}}_p \in \mathcal{Z}_3$ for all nonzero integer p.

Lemma B.4. *For every integer $p \neq 0$ we have $\tilde{\mathsf{L}}_p \in \mathcal{Z}_3$.*

Proof. From (B.1), we know that $\tilde{\mathcal{L}}_{\text{sym}}^{(1)} \in \mathcal{U}_3$. By point 2 in Lemma B.3, this implies that $(\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^{\text{sign}(p)} \in \mathcal{U}_3$. Now point 1 of Lemma B.3 implies that $(\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p = \left((\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^{\text{sign}(p)}\right)^{|p|} \in \mathcal{U}_3$. Finally, by Lemma B.2 and point 3 in Lemma B.3, we have

$$\tilde{\mathsf{L}}_p = \sum_{t=1}^{3} (\tilde{\mathcal{L}}_{\text{sym}}^{(t)})^p = \sum_{t=1}^{3} \tilde{P}_t (\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p \tilde{P}_t \in \mathcal{Z}_3,$$

which concludes the proof. □

Matrices in $\mathcal{Z}_3$ have the interesting property that they have a simple spectrum and they all share the same eigenvectors. Indeed we have the following:

Lemma B.5. *Let $\tilde{C} \in \mathcal{Z}_3$ and t_1, t_2 be such that $\tilde{C} = (t_1 - t_2)I_3 + t_2\tilde{E}$ where $\tilde{E} \in \mathbb{R}^{3\times 3}$ is the matrix of all ones. Then the eigenpairs of $\tilde{C}$ are given by:*

$$\left(t_1 - t_2, (-1, 0, 1)^\top\right), \quad \left(t_1 - t_2, (-1, 1, 0)^\top\right), \qquad \left(t_1 + 2t_2, (1, 1, 1)^\top\right).$$

Proof. Follows from a direct computation. □

So, the last thing we need to discuss is the order of the eigenvalues of $\tilde{\mathsf{L}}_p$. To this end, we study the sign pattern of the powers of this matrix.

Lemma B.6. *For every positive integer $p > 0$ we have $(\tilde{\mathcal{L}}_{\text{sym}}^p)_{i,j} < 0 < (\tilde{\mathcal{L}}_{\text{sym}}^p)_{i,i} < \tau^p$ for all $i, j = 1, 2, 3$ with $i \neq j$. For every negative integer $p < 0$ we have $(\tilde{\mathcal{L}}_{\text{sym}}^p)_{i,j} > 0$ for all $i, j = 1, 2, 3$.*

Proof. First, assume that $p > 0$ and let $\tilde{S} = \tilde{\mathcal{D}}^{-1}\tilde{\mathcal{W}}$. We have

$$\begin{aligned}
\tilde{\mathcal{L}}_{\text{sym}}^{p} &= (\tau I_3 - \tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2})^p \\
&= \sum_{r=0}^{p} \binom{p}{r} \tau^{p-r}(-1)^r (\tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2})^r \\
&= \tilde{\mathcal{D}}^{1/2}\Big(\sum_{r=0}^{p} \binom{p}{r} \tau^{p-r}(-1)^r (\tilde{\mathcal{D}}^{-1}\tilde{\mathcal{W}})^r\Big)\tilde{\mathcal{D}}^{-1/2} \\
&= \tilde{\mathcal{D}}^{1/2}(\tau I_3 - \tilde{S})^p \tilde{\mathcal{D}}^{-1/2}.
\end{aligned}$$

As $\tilde{\mathcal{D}}^{1/2}$ and $\tilde{\mathcal{D}}^{-1/2}$ are diagonal with positive diagonal entries, the sign of the entries of $\tilde{\mathcal{L}}_{\text{sym}}^{p}$ coincide with those of $(\tau I_3 - \tilde{S})^p$. Furthermore, we have $(\tilde{\mathcal{L}}_{\text{sym}}^{p})_{i,i} = ((\tau I_3 - \tilde{S})^p)_{i,i}$ for all i. Now the matrix $\tilde{S}$ is row stochastic, that is $\tilde{S}\mathbf{1} = \mathbf{1}$ and has the following form

$$\tilde{S} = \begin{pmatrix} 1-2\hat{a} & \hat{a} & \hat{a} \\ 1-2\hat{b} & \hat{b} & \hat{b} \\ 1-2\hat{b} & \hat{b} & \hat{b} \end{pmatrix} \quad \hat{a} = \frac{\hat{\alpha}}{1+2\hat{\alpha}}, \quad \hat{b} = \frac{1}{2+\hat{\alpha}}$$

where $\hat{\alpha} = p_{out}/p_{in} \in (0,1)$. Let

$$\gamma = (\hat{a} - \hat{b}) = \frac{p_{out}^2 - p_{in}^2}{(2p_{in} + p_{out})(2p_{out} + p_{in})} < 0,$$

$$\mu = (1 - 2\hat{b}) = \frac{p_{in}}{2p_{out} + p_{in}} > 0$$

We have the following result about $(\tau I_3 - \tilde{S})^p$.

Lemma B.7. *For all positive integer p, we have*

$$(\tau I_3 - \tilde{S})^p = \frac{1}{2\gamma + 1}\begin{pmatrix} q_p & r_p & r_p \\ s_p & t_p & u_p \\ s_p & u_p & t_p \end{pmatrix}$$

where q_p, r_p, s_p, t_p, u_p are given by

$$\begin{aligned}
q_p &= \mu(\tau - 1)^p + 2\hat{a}(2\gamma + \tau)^p, \\
r_p &= \hat{a}\big[(\tau - 1)^p - (2\gamma + \tau)^p\big], \\
s_p &= \frac{\mu}{\hat{a}} r_p, \\
t_p &= \hat{a}\big[(\tau - 1)^p + \tau^p\big] + \frac{\mu}{2}\big[\tau^p + (2\gamma + \tau)^p\big], \\
u_p &= \hat{a}\big[(\tau - 1)^p - \tau^p\big] - \frac{\mu}{2}\big[\tau^p - (2\gamma + \tau)^p\big].
\end{aligned} \tag{B.4}$$

Proof. Please see Section B.3 □

Note that as $p_{in} > p_{out} > 0$, we have

$$\begin{aligned}\delta = 2\gamma + 1 = 2(\hat{a} - \hat{b}) + 1 = 2\hat{a} + \mu \\ = \frac{5p_{in}p_{out} + 4p_{out}^2}{2p_{in}^2 + 5p_{in}p_{out} + 2p_{out}^2} \in (0,1),\end{aligned}$$

Furthermore, as $\tau \geq 1$ and $\gamma < 0$, we have $\delta \leq (2\gamma + \tau) < \tau$. It follows that

$$0 < \mu(\tau-1)^p + 2\hat{a}\delta^p \leq q_p < \mu(\tau-1)^p + 2\hat{a}\tau^p \leq \delta\tau^p < \tau^p, \text{and}$$
$$0 < \hat{a}\big[(\tau-1)^p + \tau^p\big] + \frac{\mu}{2}(\tau^p + \delta^p) \leq t_p < 2\hat{a}\tau^p + \mu\tau^p = \delta\tau^p < \tau^p.$$

Finally, we have

$$r_p = \hat{a}\big[(\tau-1)^p - \big(\delta + (\tau-1)\big)^p\big] < 0, \quad s_p = \frac{\mu}{\hat{a}} r_p < 0.$$

Now, suppose that $p < 0$, then we have $\tau > 1$ and

$$(\tau I_3 - \tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2})^{-1} = \sum_{k=0}^{\infty} (\tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2})^k.$$

As $\tilde{\mathcal{M}} = \tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2}$ is a matrix with strictly positive entries, this implies that $\tilde{\mathcal{L}}_{\text{sym}}^{-1}$ has positive entries as well. Furthermore, it also implies that $\tilde{\mathcal{L}}_{\text{sym}}^{p} = (\tilde{\mathcal{L}}_{\text{sym}}^{-1})^{|p|}$ is positive for every $p < 0$. □

We can now use Lemma B.6 to determine the ordering of the eigenvalues of $\tilde{\mathsf{L}}_p$.

Lemma B.8. *Let $t_1, t_2 \in \mathbb{R}$ be such that it holds $\tilde{\mathsf{L}}_p = (t_1 - t_2)I_3 + t_2\tilde{E}$. Furthermore, for any nonzero integer p, it holds $0 < t_1 - t_2 < t_1 + 2t_2$ if $p < 0$ and $t_1 - t_2 > t_1 + 2t_2$ otherwise.*

Proof. If $p < 0$, then we must have $\tau > 1$ for $\tilde{\mathsf{L}}_p$ to be well defined. By Lemma B.6, $(\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p$ has strictly positive entries. Hence, $\tilde{\mathsf{L}}_p = \sum_{t=1}^{3} \tilde{P}_t(\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p\tilde{P}_t$ is also a matrix with positive entries. It follows that $t_1 - t_2 > 0$ and $t_2 > 0$ so that $0 < t_1 - t_2 < t_1 + 2t_2$. Now assume that $p > 0$, Lemma B.6 implies that $(\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p$ with positive diagonal elements and negative off-diagonal. It follows from (B.3) that $\tilde{\mathsf{L}}_p$ also has positive diagonal elements and negative off-diagonal. Hence, we have $t_2 < 0 < t_1$ and thus $t_1 - t_2 > t_1 + 2t_2$ which concludes the proof. □

We have the following corollary on the spectral properties of the Laplacian p-mean.

Corollary B.3. *Let p be a nonzero integer and let $\epsilon \geq 0$ if $p > 0$ and $\epsilon > 0$ if $p < 0$. Define*

$$\tilde{\mathcal{L}}_p = \Big(\frac{(\tilde{\mathcal{L}}_{\text{sym}}^{(1)})^p + (\tilde{\mathcal{L}}_{\text{sym}}^{(2)})^p + (\tilde{\mathcal{L}}_{\text{sym}}^{(3)})^p}{3}\Big)^{1/p},$$

then there exists $0 \leq \tilde{\lambda}_1 < \tilde{\lambda}_2$ such that the eigenpairs of $\tilde{\mathcal{L}}_p$ are given by

$$\big(\tilde{\lambda}_1, (-1,0,1)^\top\big), \qquad \big(\tilde{\lambda}_1, (-1,1,0)^\top\big), \qquad \big(\tilde{\lambda}_2, (1,1,1)^\top\big).$$

Proof. First, note that $\tilde{\mathcal{L}}_p = \left(\frac{1}{3}\tilde{\mathsf{L}}_p\right)^{1/p}$ hence as they are positive semi-definite matrices, $\tilde{\mathcal{L}}_p$ and $\tilde{\mathsf{L}}_p$ share the same eigenvectors. Precisely, we have $\tilde{\mathsf{L}}_p\mathbf{v} = \lambda\mathbf{v}$ if and only if $\tilde{\mathcal{L}}_p\mathbf{v} = f(\lambda)\mathbf{v}$ where $f(t) = (t/3)^{1/p}$. Now, by Lemmas B.4 and B.5 we know all eigenvectors of $\tilde{\mathsf{L}}_p$ and the corresponding eigenvalues are $\theta_1 = t_1 - t_2$ and $\theta_2 = t_1 + 2t_2$. Finally, using Lemma B.8 and the fact that f is increasing if $p > 0$ and decreasing if $p < 0$ we deduce the ordering of $\tilde{\lambda}_i = f(\theta_i)$. □

B.2 THE CASE $n > 1$

We now generalize the previous results to the case $n > 1$. To this end, we use mainly the properties of the Kronecker product $\otimes$ which we recall is defined for matrices $A \in \mathbb{R}^{m_1\times m_2}, B \in \mathbb{R}^{m_3\times m_4}$ as the block matrix $A \otimes B \in \mathbb{R}^{m_1m_3\times m_2m_4}$ with m_1m_2 blocks of the form $A_{i,j}B \in \mathbb{R}^{m_3\times m_4}$ for all i, j. In particular, for $n > 1$, if E denotes the matrix of all ones in $\mathbb{R}^{n\times n}$, we have then $\mathcal{W}^{(t)} = \tilde{\mathcal{W}}^{(t)} \otimes E$ for every $t = 1, 2, 3$. Furthermore, let us define $\mathcal{W} = \tilde{\mathcal{W}} \otimes E$ and $P_t = \tilde{P}_t \otimes I_n$ for $t = 1, 2, 3$ so that $\mathcal{W}^{(t)} = P_t\mathcal{W}P_t$ for $t = 1, 2, 3$. Finally, let $\mathcal{L}_{\text{sym}} = \tau I_{3n} - \mathcal{D}^{-1/2}\mathcal{W}\mathcal{D}^{-1/2}$ where we recall that $\tau = 1 + \epsilon$ and $\mathcal{D} = \text{diag}(\mathcal{W}\mathbf{1})$. The normalized Laplacians of $\mathcal{W}$ and $\tilde{\mathcal{W}}$ are related in the following lemma:

Lemma B.9. *It holds*

$$\mathcal{L}_{\text{sym}} = \tau I_{3n} - \left[\tfrac{1}{n}\tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2}\right] \otimes E.$$

Proof. First, note that $\mathcal{D} = n\tilde{\mathcal{D}} \otimes I_n$, as $(A_1 \otimes B_1)(A_2 \otimes B_2) = (A_1A_2 \otimes B_1B_2)$ for any compatible matrices A_1, A_2, B_1, B_2. We have

$$\begin{aligned}\mathcal{D}^{-1/2}\mathcal{W}\mathcal{D}^{-1/2} &= \frac{(\tilde{\mathcal{D}}^{-1/2} \otimes I_n)(\tilde{\mathcal{W}} \otimes E)(\tilde{\mathcal{D}}^{-1/2} \otimes I_n)}{n} \\ &= \tfrac{1}{n}\tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2} \otimes E,\end{aligned}$$

which concludes the proof. □

In order to study the eigenpairs of $\mathcal{L}_{\text{sym}}$, we combine Lemma B.1 with the following theorem from (Horn and Johnson, 1991) which implies that eigenpairs of Kronecker products are Kronecker products of the eigenpairs:

Theorem B.3 (Theorem 4.2.12, (Horn and Johnson, 1991))**.** *Let $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times n}$. Let (λ, x) and (μ, y) be eigenpairs of A and B respectively. Then $(\lambda\mu, x \otimes y)$ is an eigenpair of $A \otimes B$.*

Indeed, the above theorem implies that the eigenpairs of $\mathcal{D}^{-1/2}\mathcal{W}\mathcal{D}^{-1/2}$ are Kronecker products of the eigenpairs of $\tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2}$ and E. As we already know those of $\tilde{\mathcal{D}}^{-1/2}\tilde{\mathcal{W}}\tilde{\mathcal{D}}^{-1/2}$, we briefly describe those of E:

Lemma B.10. *Let $E \in \mathbb{R}^{n\times n}$, $n \geq 2$ be the matrix of all ones, then the eigenpairs of E are given by $(n, \mathbf{1})$ and $(0, \mathbf{v}_1), \ldots, (0, \mathbf{v}_{n-1})$ where $\mathbf{v}_k \in \mathbb{R}^n$ is given as*

$$(\mathbf{v}_k)_j = \begin{cases} 1 & \text{if } j \leq k, \\ -k & \text{if } j = k+1, \\ 0 & \text{otherwise.} \end{cases} \tag{B.5}$$

Proof. As $E = \mathbf{1}\mathbf{1}^\top$, it is clear that $(n, \mathbf{1})$ is an eigenpair of E. Now, for every i we have $E\mathbf{v}_i = (\mathbf{1}^\top v_i)\mathbf{1}$ and $\mathbf{1}^\top \mathbf{v}_i = i - i = 0$. □

We can now describe the spectral properties of $\mathcal{L}_{\text{sym}}^{(t)}$ for $t = 1, 2, 3$.

Lemma B.11. *There exists $\lambda \in (0, 1)$ and $s_- < 0 < s_+ < 1$ such that, for $t = 1, 2, 3$, the eigenpairs of $\mathcal{L}_{\text{sym}}^{(t)}$ are given by*

$$\begin{aligned} &\big(\tau - 1, P_t(s_+, 1, 1)^\top \otimes \mathbf{1}\big), \quad \big(\tau, P_t(s_+, 1, 1)^\top \otimes \mathbf{v}_k\big), \\ &\big(\tau, P_t(0, -1, 1)^\top \otimes \mathbf{1}\big), \quad \big(\tau, P_t(0, -1, 1)^\top \otimes \mathbf{v}_k\big), \\ &\big(\tau - \lambda, P_t(s_-, 1, 1)^\top \otimes \mathbf{1}\big), \quad \big(\tau, P_t(s_-, 1, 1)^\top \otimes \mathbf{v}_k\big), \end{aligned}$$

for $k = 1, \ldots, n-1$, where $\mathbf{v}_k$ is defined as in (B.5).

Proof. Follows from Lemmas B.1, B.10 and Theorem B.3. □

Similarly to the case $n = 1$, let us consider $\mathsf{L}_p \in \mathbb{R}^{3n\times 3n}$ defined as

$$\mathsf{L}_p = (\mathcal{L}_{\text{sym}}^{(1)})^p + (\mathcal{L}_{\text{sym}}^{(2)})^p + (\mathcal{L}_{\text{sym}}^{(3)})^p = 3\mathcal{L}_p^p.$$

Again, we note that the eigenvectors of L_p and $3\mathcal{L}_p^p$ are the same. Now, let us consider the sets $\mathcal{U}_{3n} \subset \mathbb{R}^{3n\times 3n}$ and $\mathcal{Z}_{3n} \subset \mathbb{R}^{3n\times 3n}$ defined as

$$\begin{aligned} \mathcal{U}_{3n} &= \big\{ s_0 I_{3n} - \tilde{A} \otimes E \;\big|\; \tilde{A} \in \mathcal{U}_3, s_0 \in \mathbb{R} \big\}, \\ \mathcal{Z}_{3n} &= \big\{ t_0 I_{3n} - \tilde{C} \otimes E \;\big|\; \tilde{C} \in \mathcal{Z}_3, s_0 \in \mathbb{R} \big\}. \end{aligned}$$

Note that, as $s_0 I_3 + \mathcal{U}_3 = \mathcal{U}_3$ and $s_0 I_3 + \mathcal{Z}_3 = \mathcal{Z}_3$ for all $s_0 \in \mathbb{R}$, the definitions of $\mathcal{U}_{3n}$ and $\mathcal{Z}_{3n}$ reduce to that of $\mathcal{U}_3$ and $\mathcal{Z}_3$ when $n = 1$. We prove that $\mathsf{L}_p \in \mathcal{Z}_{3n}$ for all nonzero integer p. To this end, we first prove the following lemma which generalizes Lemma B.3.

Lemma B.12. *The following holds:*

1. *$\mathcal{U}_{3n}$ is closed under multiplication, i.e. for all $A, B \in \mathcal{U}_{3n}$ we have $AB \in \mathcal{U}_{3n}$.*
2. *If $A \in \mathcal{U}_{3n}$ satisfies $\det(A) \neq 0$, then $A^{-1} \in \mathcal{U}_{3n}$.*
3. *$\mathcal{Z}_3 = P_1 \mathcal{U}_{3n} P_1 + P_2 \mathcal{U}_{3n} P_2 + P_3 \mathcal{U}_{3n} P_3$.*

Proof. Let $A, B \in \mathcal{U}_{3n}, C \in \mathcal{Z}_{3n}$ and $s_0, r_0, t_0 \in \mathbb{R}$, $\tilde{A}, \tilde{B} \in \mathcal{U}_3, \tilde{C} \in \mathcal{Z}_3$ such that $A = s_0 I_{3n} - \tilde{A} \otimes E$, $B = r_0 I_{3n} - \tilde{B} \otimes E$ and $C = t_0 I_{3n} - \tilde{C} \otimes E$.

1. We have
$$A\,B = s_0 r_0 I_{3n} + (n\tilde{A}\tilde{B} - s_0\tilde{B} - r_0\tilde{A}) \otimes E$$
As $\tilde{A}\tilde{B} \in \mathcal{U}_3$ by point 1 in Lemma B.3, we have $(n\tilde{A}\tilde{B} - s_0\tilde{B} - r_0\tilde{A}) \in \mathcal{U}_3$ and so $AB \in \mathcal{U}_{3n}$.

2. First note that as A is invertible, it holds $s_0 \neq 0$. Furthermore, using von Neumann series, we have
$$\begin{aligned}(s_0 I_{3n} - \tilde{A} \otimes E)^{-1} &= s_0^{-1}(I_{3n} - s_0^{-1}\tilde{A} \otimes E)^{-1} \\ &= s_0^{-1} \sum_{k=0}^{\infty} s_0^{-k} (\tilde{A} \otimes E)^k \\ &= s_0^{-1} \sum_{k=0}^{\infty} s_0^{-k} n^{k-1} (\tilde{A}^k \otimes E).\end{aligned}$$
As $\tilde{A}^k \in \mathcal{U}_3$ for all k by point 1 in Lemma B.3 we have that $S_\nu = s_0^{-1} \sum_{k=0}^{\nu} s_0^{-k} n^{k-1} (\tilde{A}^k \otimes E) \in \mathcal{U}_{3n}$ for all $\nu = 0, 1, \ldots$ As $\lim_{\nu\to\infty} S_\nu = A^{-1}$ and $\mathcal{U}_{3n}$ is closed, it follows that $A^{-1} \in \mathcal{U}_{3n}$.

3. Note that for $i = 1, 2, 3$ it holds
$$P_i\,A\,P_i = s_0 I_{3n} - (\tilde{P}_i \tilde{A} \tilde{P}_i \otimes E).$$
Hence, we have
$$\sum_{i=1}^{3} P_i\,A\,P_i = 3 s_0 I_{3n} - \Big(\sum_{i=1}^{3} \tilde{P}_i \tilde{A} \tilde{P}_i \Big) \otimes E.$$
We know from point 3 in Lemma B.3 that $\sum_{i=1}^{3} \tilde{P}_i \tilde{A} \tilde{P}_i \in \mathcal{U}_3$ and thus $\sum_{i=1}^{3} \tilde{P}_i\,\tilde{A}\,\tilde{P}_i \in \mathcal{Z}_3$. Finally, note that by choosing the coefficients in $\tilde{A}$ in the same way as in the proof of point 3 in Lemma B.3, we have $A = C$ with $s_0 = t_0$. This concludes the proof.

□

We can now prove that $\mathsf{L}_p \in \mathcal{Z}_{3n}$.

Lemma B.13. *For every nonzero integer p, we have $\mathsf{L}_p \in \mathcal{Z}_{3n}$.*

Proof. As $\mathcal{L}_{\text{sym}} = \mathcal{L}_{\text{sym}}^{(1)} \in \mathcal{U}_{3n}$, we have $\mathcal{L}_{\text{sym}}^{p} \in \mathcal{U}_{3n}$ by points 1 and 2 in Lemma B.12. We prove that $\mathsf{L}_p = \sum_{t=1}^{3} P_t \mathcal{L}_{\text{sym}}^{p} P_t$. To this end, note that, with the convention that powers on vectors are considered component wise, for $t = 1, 2, 3$, we have
$$\begin{aligned}\operatorname{diag}(P_t \mathcal{W} P_t \mathbf{1})^{-1/2} &= \operatorname{diag}\big(P_t (\mathcal{W}\mathbf{1})^{-1/2}\big) \\ &= P_t \operatorname{diag}\big((\mathcal{W}\mathbf{1})^{-1/2}\big) P_t \\ &= P_t \mathcal{D}^{-1/2} P_t.\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\operatorname{diag}(P_t\mathcal{W}P_t\mathbf{1})^{-1/2}P_t\mathcal{W}P_t\operatorname{diag}(P_t\mathcal{W}P_t\mathbf{1})^{-1/2}\\
&\quad= P_t\mathcal{D}^{-1/2}P_t^2\mathcal{W}P_t^2\mathcal{D}^{-1/2}P_t\\
&\quad= P_t\mathcal{D}^{-1/2}\mathcal{W}\mathcal{D}^{-1/2}P_t.
\end{aligned}$$

This implies that $\mathcal{L}_{\text{sym}}^{(t)} = P_t\mathcal{L}_{\text{sym}}P_t$ for $t = 1,2,3$ and thus we obtain the desired expression for L_p. Point 3 in Lemma B.12 finally imply that $\mathsf{L}_p \in \mathcal{Z}_{3n}$. □

We combine Theorem B.3 and Lemmas B.5, B.10 to obtain the following:

Lemma B.14. *Let $C \in \mathcal{Z}_{3n}$ and t_0, t_1, t_2 such that $C = t_0 I_{3n} - ((t_1 - t_2)I_3 + t_2\tilde{E}) \otimes E$. Then, the eigenpairs of C are given by*

$$\begin{aligned}
&\big(t_0 - n(t_1 - t_2), (-1,0,1)^\top \otimes \mathbf{1}\big),\\
&\big(t_0 - n(t_1 - t_2), (-1,1,0)^\top \otimes \mathbf{1}\big),\\
&\big(t_0 - n(t_1 + 2t_2), (1,1,1)^\top \otimes \mathbf{1}\big).
\end{aligned}$$

and, with $\mathbf{v}_i$ defined as in (B.5)*,*

$$\begin{aligned}
&\big(t_0, (-1,0,1)^\top \otimes \mathbf{v}_i\big), \qquad \big(t_0, (-1,1,0)^\top \otimes \mathbf{v}_i\big),\\
&\big(t_0, (1,1,1)^\top \otimes \mathbf{v}_i\big), \qquad i = 1,\dots,n-1.
\end{aligned}$$

Similar to Lemma B.8, we have following the lemma for deciding the order of the eigenvectors of L_p.

Lemma B.15. *For every positive $p > 0$ we have $(\mathcal{L}_{\text{sym}}^p)_{i,j} < 0 < (\mathcal{L}_{\text{sym}}^p)_{i,i} < \tau^p$ for all $i, j = 1,\dots,3n$ with $i \neq j$. For every negative $p < 0$ we have $(\mathcal{L}_{\text{sym}}^p)_{i,j} > 0$ for all $i, j = 1,\dots,3n$.*

Proof. Let $\mathcal{M} = \mathcal{D}^{-1/2}\mathcal{W}\mathcal{D}^{-1/2}$, then by Lemma B.9, we have $\mathcal{M} = \frac{1}{n}(\tilde{\mathcal{M}} \otimes E)$. Now, for $p > 0$, it holds:

$$\begin{aligned}
\mathcal{L}_{\text{sym}}^p &= (\tau I_{3n} - \mathcal{M})^p\\
&= (\tau I_{3n} - \tfrac{1}{n}(\tilde{\mathcal{M}} \otimes E))^p\\
&= \sum_{k=0}^{p} \binom{p}{k} \tau^{p-k}(-1)^k n^{-k}(\tilde{\mathcal{M}}^k \otimes E^k)\\
&= \tau^p I_{3n} + \sum_{k=1}^{p} \binom{p}{k} \tau^{p-k}(-1)^k n^{-k}(\tilde{\mathcal{M}}^k \otimes E^k)\\
&= \tau^p I_{3n} + \Big(\sum_{k=1}^{p} \binom{p}{k} \tau^{p-k}(-1)^k \tilde{\mathcal{M}}^k\Big) \otimes E\\
&= \tau^p I_{3n} + \big(\tilde{\mathcal{L}}_{\text{sym}}^p - \tau^p I_3\big) \otimes E. \qquad \text{(B.6)}
\end{aligned}$$

By Lemma B.6, we know that $(\tilde{\mathcal{L}}^p_{\text{sym}})_{i,j} < 0$ if $i \neq j$ and $(\tilde{\mathcal{L}}^p_{\text{sym}})_{i,i} - \tau^p < 0$ for all i. Hence, the matrix $\tilde{Q} = \tilde{\mathcal{L}}^p_{\text{sym}} - \tau^p I_3$ has strictly negative entries. Thus, all the off-diagonal elements of $\mathcal{L}^p_{\text{sym}}$ are strictly negative. Finally, note that

$$(\mathcal{L}^p_{\text{sym}})_{i,i} = \tau^p + (\tilde{\mathcal{L}}^p_{\text{sym}} \otimes E)_{i,i} - \tau^p = (\tilde{\mathcal{L}}^p_{\text{sym}} \otimes E)_{i,i} > 0.$$

This concludes the proof for the case $p > 0$. The case $p < 0$ can be proved in the same way as for the case $n = 1$ (see Lemma B.6). □

Observation B.1. *We note that Equation* (B.6) *implies the following relation between* L_p *and* $\tilde{\mathsf{L}}_p$*:*

$$\mathsf{L}_p = 3\tau^p I_{3n} + \big(\tilde{\mathsf{L}}_p - \tau^p I_3\big) \otimes E. \tag{B.7}$$

Lemma B.16. *Let* $t_0, t_1, t_2 \in \mathbb{R}$ *be such that* $\mathsf{L}_p = t_0 I_{3n} - ((t_1 - t_2)I_3 + t_2\tilde{E}) \otimes E$*. Furthermore, for any integer* $p \neq 0$*, it holds* $t_0 < t_0 - n(t_1 - t_2) < t_0 - n(t_1 + 2t_2)$ *if* $p < 0$ *and* $t_0 > t_0 - n(t_1 - t_2) > t_0 - n(t_1 + 2t_2)$ *otherwise.*

Proof. The proof is essentially the same as that of Lemma B.8. Indeed, if $p < 0$, then L_p is strictly positive and thus $t_2 < 0$ as $(\mathsf{L}_p)_{1,3n} > 0$, $t_1 - t_2 < 0$ as $(\mathsf{L}_p)_{1,n} > 0$ and $t_0 - t_1 > 0$ as $(\mathsf{L}_p)_{1,1} > 0$. This means that $t_1 - t_2 > t_1 + 2t_2$ and so $t_0 - n(t_1 - t_2) < t_0 - n(t_1 + 2t_2)$. Furthermore, this shows that $t_0 - n(t_1 - t_2) > t_0$. Now, if $p > 0$, by Lemma B.15 we have $t_2 > 0$ as $(\mathsf{L}_p)_{1,3n} < 0$, $t_1 - t_2 > 0$ as $(\mathsf{L}_p)_{1,n} < 0$ and $t_0 - t_1 > 0$ as $(\mathsf{L}_p)_{1,1} > 0$. It follows that $t_1 - t_2 < t_1 + 2t_2$ and thus $t_0 - n(t_1 - t_2) > t_0 - n(t_1 + 2t_2)$. Finally, as $t_1 - t_2 > 0$, we have $t_0 > t_0 - n(t_1 - t_2)$ which concludes the proof. □

We conclude by giving a description of the spectral properties of $\mathcal{L}_p$.

Lemma B.17. *Let* p *be any nonzero integer and assume that* $\epsilon > 0$ *if* $p < 0$*. Define*

$$\mathcal{L}_p = \Big(\frac{(\mathcal{L}^{(1)}_{\text{sym}})^p + (\mathcal{L}^{(2)}_{\text{sym}})^p + (\mathcal{L}^{(3)}_{\text{sym}})^p}{3}\Big)^{1/p},$$

then there exists $0 \leq \lambda_1, \lambda_2 < \lambda_3$ *such that all the eigenpairs of* $\mathcal{L}_p$ *are given by*

$$\begin{aligned}
&\big(\lambda_1, (-1,0,1)^\top \otimes \mathbf{1}\big), \qquad &&\big(\lambda_3, (-1,0,1)^\top \otimes \mathbf{v}_i\big)\\
&\big(\lambda_1, (-1,1,0)^\top \otimes \mathbf{1}\big), \qquad &&\big(\lambda_3, (-1,1,0)^\top \otimes \mathbf{v}_i\big)\\
&\big(\lambda_2, (1,1,1)^\top \otimes \mathbf{1}\big), \qquad &&\big(\lambda_3, (1,1,1)^\top \otimes \mathbf{v}_i\big),
\end{aligned}$$

and $i = 1, \ldots, n-1$*, where* $\mathbf{v}_i$ *is defined in* (B.5)*.*

Proof. The proof is the same as that of Corollary B.3 where one uses Lemmas B.13, B.14, B.16 instead of Lemmas B.4, B.5, B.8. □

B.3 PROOF OF LEMMA B.7

The proof is by induction. We first verify the base case $p = 1$ to later consider the inductive step where we prove the general case for $p + 1$.

Please recall that

$$\tilde{S} = \begin{pmatrix} 1-2\hat{a} & \hat{a} & \hat{a} \\ 1-2\hat{b} & \hat{b} & \hat{b} \\ 1-2\hat{b} & \hat{b} & \hat{b} \end{pmatrix} \quad \hat{a} = \frac{\hat{\alpha}}{1+2\hat{\alpha}}, \quad \hat{b} = \frac{1}{2+\hat{\alpha}}$$

where $\hat{\alpha} = p_{out}/p_{in} \in (0, 1)$, and

$$\gamma = (\hat{a} - \hat{b}) = \frac{p_{out}^2 - p_{in}^2}{(2p_{in} + p_{out})(2p_{out} + p_{in})} < 0,$$

$$\mu = (1 - 2\hat{b}) = \frac{p_{in}}{2p_{out} + p_{in}} > 0$$

Base Case: We verify the conditions for the case $p = 1$. The corresponding verifications are contained in: Lemma B.18 for q_1, Lemma B.19 for r_1, Lemma B.20 for s_1, Lemma B.21 for t_1, and Lemma B.22 for u_1.

Lemma B.18.

$$\frac{q_1}{(2\gamma+1)} = (\tau I_3 - \tilde{S})_{11}$$

Proof. We can see that,

$$\begin{aligned} q_1 &= \mu(\tau - 1) + 2\hat{a}(2\gamma + \tau) \\ &= \mu(\tau - 1) + 2\hat{a}(2\gamma + \tau) + (2\hat{a} - 2\hat{a}) \\ &= (1 - 2\hat{b})(\tau - 1) + 2\hat{a}(2(\hat{a} - \hat{b}) + \tau) + (2\hat{a} - 2\hat{a}) \\ &= \tau(2\hat{a} - 2\hat{b} + 1) + 2\hat{a}(2\hat{a} - 2\hat{b} + 1) - (2\hat{a} - 2\hat{b} + 1) \\ &= (\tau + 2\hat{a} - 1)(2\hat{a} - 2\hat{b} + 1) \\ &= (\tau + 2\hat{a} - 1)(2\gamma + 1) \end{aligned}$$

Hence, $\frac{q_1}{(2\gamma+1)} = \tau + 2\hat{a} - 1 = (\tau I_3 - \tilde{S})_{11}$ □

Lemma B.19.

$$\frac{r_1}{(2\gamma+1)} = (\tau I_3 - \tilde{S})_{12}$$

Proof. We can see that,

$$r_1 = \hat{a}\big[(\tau - 1) - (2\gamma + \tau)\big] = -\hat{a}(2\gamma + 1)$$

Hence, $\frac{r_1}{(2\gamma+1)} = -\hat{a} = (\tau I_3 - \tilde{S})_{12}$ □

Lemma B.20.

$$\frac{s_1}{(2\gamma+1)} = (\tau I_3 - \tilde{S})_{21}$$

Proof. We can see that,

$$s_1 = \frac{\mu}{\hat{a}} r_p = \frac{\mu}{\hat{a}} \hat{a}\big[(\tau-1) - (2\gamma+\tau)\big] = -\mu(2\gamma+1) = -(1-2\hat{b})(2\gamma+1)$$

Hence, $\frac{s_1}{(2\gamma+1)} = -(1-2\hat{b}) = (\tau I_3 - \tilde{S})_{21}$ □

Lemma B.21.

$$\frac{t_1}{(2\gamma+1)} = (\tau I_3 - \tilde{S})_{22}$$

Proof. We can see that,

$$\begin{aligned}
t_1 &= \hat{a}\big[(\tau-1)+\tau\big] + \frac{\mu}{2}\big[\tau + (2\gamma+\tau)\big] \\
&= \hat{a}(2\tau-1) + \mu(\tau+\gamma) \\
&= \tau(2\hat{a}+\mu) + (\mu\gamma - \hat{a}) \\
&= \tau(2\hat{a}+1-2\hat{b}) + \big[(1-2\hat{b})\gamma - \hat{a}\big] \\
&= \tau(2\gamma+1) - 2\gamma\hat{b} + \gamma - \hat{a} \\
&= \tau(2\gamma+1) - 2\gamma\hat{b} - \hat{b} \\
&= \tau(2\gamma+1) - \hat{b}(2\gamma+1) \\
&= (\tau - \hat{b})(2\gamma+1)
\end{aligned}$$

Hence, $\frac{t_1}{(2\gamma+1)} = (\tau - \hat{b}) = (\tau I_3 - \tilde{S})_{22}$ □

Lemma B.22.

$$\frac{u_1}{(2\gamma+1)} = (\tau I_3 - \tilde{S})_{23}$$

Proof. We can see that,

$$\begin{aligned}
u_1 &= \hat{a}\big[(\tau-1)-\tau\big] - \frac{\mu}{2}\big[\tau - (2\gamma+\tau)\big] \\
&= -\hat{a} + \mu\gamma \\
&= -\hat{a} + (1-2\hat{b})\gamma \\
&= -2\hat{b}\gamma - \hat{a} + \gamma \\
&= -2\hat{b}\gamma - \hat{b} \\
&= -\hat{b}(2\gamma+1)
\end{aligned}$$

Hence, $\frac{u_1}{(2\gamma+1)} = -\hat{b} = (\tau I_3 - \tilde{S})_{23}$ □

Inductive step: We verify the conditions for the case $p+1$. The corresponding verifications are contained in: Lemma B.23 for q_{p+1}, Lemma B.24 for r_{p+1}, Lemma B.25 for s_{p+1}, Lemma B.26 for t_{p+1}, and Lemma B.27 for u_{p+1}.

As a reference for our verifications, one can easily see that:

$$\begin{aligned}(\tau I_3 - \tilde{S})^{p+1} &= (\tau I_3 - \tilde{S})^p(\tau I_3 - \tilde{S}) \\ &= \Big(\frac{1}{2\gamma+1}\Big)^2 \begin{pmatrix} q_p & r_p & r_p \\ s_p & t_p & u_p \\ s_p & u_p & t_p \end{pmatrix} \begin{pmatrix} q_1 & r_1 & r_1 \\ s_1 & t_1 & u_1 \\ s_1 & u_1 & t_1 \end{pmatrix} \\ &= \Big(\frac{1}{2\gamma+1}\Big)^2 \begin{pmatrix} q_pq_1 + r_ps_1 + r_ps_1 & q_pr_1 + r_pt_1 + r_pu_1 & q_pr_1 + r_pu_1 + r_pt_1 \\ s_pq_1 + t_ps_1 + u_ps_1 & s_pr_1 + t_pt_1 + u_pu_1 & s_pr_1 + t_pu_1 + u_pt_1 \\ s_pq_1 + u_ps_1 + t_ps_1 & s_pr_1 + u_pt_1 + t_pu_1 & s_pr_1 + u_pu_1 + t_pt_1 \end{pmatrix}\end{aligned}$$

Lemma B.23.

$$\frac{q_{p+1}}{(2\gamma+1)} = (\tau I_3 - \tilde{S})^{p+1}_{11}$$

Proof. Since $(\tau I_3 - \tilde{S})^{p+1}_{11} = \frac{q_pq_1+2r_ps_1}{(2\gamma+1)^2}$ we can see that

$$\begin{aligned} q_p\frac{q_1}{2\gamma+1} &= \big[\mu(\tau-1)^p + 2\hat{a}(2\gamma+\tau)^p\big](\tau + 2\hat{a} - 1)\,, \text{and} \\ r_p\frac{s_1}{2\gamma+1} &= \hat{a}\big[(\tau-1)^p - (2\gamma+\tau)^p\big](-1+2\hat{b}) = -\hat{a}\mu\big[(\tau-1)^p - (2\gamma+\tau)^p\big]\end{aligned}$$

and hence

$$\begin{aligned}\frac{q_pq_1 + 2r_ps_1}{2\gamma+1} &= \big[\mu(\tau-1)^p + 2\hat{a}(2\gamma+\tau)^p\big](\tau+2\hat{a}-1) - 2\hat{a}\mu\big[(\tau-1)^p - (2\gamma+\tau)^p\big] \\ &= (\tau-1)^p\big[\mu(\tau+2\hat{a}-1) - 2\hat{a}\mu\big] + (2\gamma+\tau)^p\big[2\hat{a}(\tau+2\hat{a}-1) + 2\hat{a}\mu\big] \\ &= (\tau-1)^p\big[\mu(\tau-1)\big] + (2\gamma+\tau)^p\big[2\hat{a}(\tau+2\hat{a}-1) + 2\hat{a}\mu\big] \\ &= (\tau-1)^p\big[\mu(\tau-1)\big] + 2\hat{a}(2\gamma+\tau)^p\big[(\tau+2\hat{a}-1) + \mu\big] \\ &= (\tau-1)^p\big[\mu(\tau-1)\big] + 2\hat{a}(2\gamma+\tau)^p\big[(\tau+2\hat{a}-1) + 1 - 2\hat{b}\big] \\ &= (\tau-1)^p\big[\mu(\tau-1)\big] + 2\hat{a}(2\gamma+\tau)^p\big[2\gamma+\tau\big] \\ &= \mu(\tau-1)^{p+1} + 2\hat{a}(2\gamma+\tau)^{p+1}\end{aligned}$$

Hence,

$$(\tau I_3 - \tilde{S})^{p+1}_{11} = \frac{\mu(\tau-1)^{p+1} + 2\hat{a}(2\gamma+\tau)^{p+1}}{2\gamma+1} = \frac{q_{p+1}}{2\gamma+1}$$

□

Lemma B.24.

$$\frac{r_{p+1}}{(2\gamma+1)} = (\tau I_3 - \tilde{S})^{p+1}_{12}$$

Proof. Since $(\tau I_3 - \tilde{S})^{p+1}_{12} = \frac{q_p r_1 + r_p t_1 + r_p u_1}{(2\gamma+1)^2}$ we can see that

$$\begin{aligned}
q_p \frac{r_1}{2\gamma+1} &= \big[\mu(\tau-1)^p + 2\hat{a}(2\gamma+\tau)^p\big](-\hat{a}), \\
r_p \frac{t_1}{2\gamma+1} &= \hat{a}\big[(\tau-1)^p - (2\gamma+\tau)^p\big](\tau-\hat{b}), \text{and} \\
r_p \frac{u_1}{2\gamma+1} &= \hat{a}\big[(\tau-1)^p - (2\gamma+\tau)^p\big](-\hat{b})
\end{aligned}$$

and hence

$$r_p \frac{t_1+u_1}{2\gamma+1} = \hat{a}\big[(\tau-1)^p - (2\gamma+\tau)^p\big](\tau-2\hat{b})$$

therefore:

$$\begin{aligned}
\frac{q_p r_1 + r_p(t_1+u_1)}{2\gamma+1} &= -\hat{a}\big[\mu(\tau-1)^p + 2\hat{a}(2\gamma+\tau)^p\big] + \hat{a}\big[(\tau-1)^p - (2\gamma+\tau)^p\big](\tau-2\hat{b}) \\
&= \hat{a}\big\{\big[-\mu(\tau-1)^p - 2\hat{a}(2\gamma+\tau)^p\big] + \big[(\tau-1)^p - (2\gamma+\tau)^p\big](\tau-2\hat{b})\big\} \\
&= \hat{a}\big\{(\tau-1)^p\big[-\mu + (\tau-2\hat{b})\big] + (2\gamma+\tau)^p\big[-2\hat{a} - (\tau-2\hat{b})\big]\big\} \\
&= \hat{a}\big\{(\tau-1)^p\big[-\mu + (\tau-2\hat{b})\big] + (2\gamma+\tau)^p\big[-(\tau+2\gamma)\big]\big\} \\
&= \hat{a}\big\{(\tau-1)^p\big[-(1-2\hat{b}) + (\tau-2\hat{b})\big] + (2\gamma+\tau)^p\big[-(\tau+2\gamma)\big]\big\} \\
&= \hat{a}\big\{(\tau-1)^p\big[\tau-1\big] + (2\gamma+\tau)^p\big[-(\tau+2\gamma)\big]\big\} \\
&= \hat{a}\big\{(\tau-1)^{p+1} - (2\gamma+\tau)^{p+1}\big\}
\end{aligned}$$

Hence,

$$(\tau I_3 - \tilde{S})^{p+1}_{12} = \frac{\hat{a}\big\{(\tau-1)^{p+1} - (2\gamma+\tau)^{p+1}\big\}}{2\gamma+1} = \frac{r_{p+1}}{2\gamma+1}$$

□

Lemma B.25.

$$\frac{s_{p+1}}{(2\gamma+1)} = (\tau I_3 - \tilde{S})^{p+1}_{21}$$

Proof. Since $(\tau I_3 - \tilde{S})^{p+1}_{21} = \frac{s_p q_1 + s_1(t_p+u_p)}{(2\gamma+1)^2}$ we can see that

$$\begin{aligned}
s_p \frac{q_1}{2\gamma+1} &= \frac{\mu}{\hat{a}} r_p\big[(\tau-1+2\hat{a})\big] = \mu\big[(\tau-1)^p - (2\gamma+\tau)^p\big]\big[(\tau-1+2\hat{a})\big] \\
t_p \frac{s_1}{2\gamma+1} &= \big\{\hat{a}\big[(\tau-1)^p + \tau^p\big] + \frac{\mu}{2}\big[\tau^p + (2\gamma+\tau)^p\big]\big\}(-1+2\hat{b}), \text{and} \\
u_p \frac{s_1}{2\gamma+1} &= \big\{\hat{a}\big[(\tau-1)^p - \tau^p\big] - \frac{\mu}{2}\big[\tau^p - (2\gamma+\tau)^p\big]\big\}(-1+2\hat{b})
\end{aligned}$$

Observe that

$$\frac{s_1(t_p+u_p)}{2\gamma+1} = \big[2\hat{a}(\tau-1)^p + \mu(2\gamma+\tau)^p\big]\big[-1+2\hat{b}\big]$$

Therefore,

$$
\begin{aligned}
&\frac{s_p q_1 + s_1(t_p + u_p)}{2\gamma+1} = \\
&\mu\big[(\tau-1)^p - (2\gamma+\tau)^p\big]\big[(\tau-1+2\hat{a})\big] + \big[2\hat{a}(\tau-1)^p + \mu(2\gamma+\tau)^p\big]\big[-1+2\hat{b}\big] = \\
&\mu\big[(\tau-1)^p - (2\gamma+\tau)^p\big]\big[(\tau-1+2\hat{a})\big] + \big[2\hat{a}(\tau-1)^p + \mu(2\gamma+\tau)^p\big](-\mu) = \\
&\mu\big\{\big[(\tau-1)^p - (2\gamma+\tau)^p\big]\big[(\tau-1+2\hat{a})\big] - \big[2\hat{a}(\tau-1)^p + \mu(2\gamma+\tau)^p\big]\big\} = \\
&\mu\big\{(\tau-1)^p\big[(\tau-1+2\hat{a}) - 2\hat{a}\big] + (2\gamma+\tau)^p\big[-(\tau-1+2\hat{a}) - \mu)\big]\big\} = \\
&\mu\big\{(\tau-1)^p(\tau-1) - (2\gamma+\tau)^p\big[(\tau-1+2\hat{a}) + \mu)\big]\big\} = \\
&\mu\big\{(\tau-1)^p(\tau-1) - (2\gamma+\tau)^p\big[(\tau-1+2\hat{a}) + (1-2\hat{b})\big]\big\} = \\
&\mu\big\{(\tau-1)^p(\tau-1) - (2\gamma+\tau)^p(2\gamma+1)\big\} = \\
&\mu\big\{(\tau-1)^{p+1} - (2\gamma+\tau)^{p+1}\big\}
\end{aligned}
$$

Hence,

$$
(\tau I_3 - \tilde{S})^{p+1}_{21} = \frac{\mu\big\{(\tau-1)^{p+1} - (2\gamma+\tau)^{p+1}\big\}}{2\gamma+1} = \frac{s_{p+1}}{2\gamma+1}
$$

□

Lemma B.26.

$$
\frac{t_{p+1}}{(2\gamma+1)} = (\tau I_3 - \tilde{S})^{p+1}_{22}
$$

Proof. Since $(\tau I_3 - \tilde{S})^{p+1}_{22} = \frac{s_p r_1 + t_p t_1 + u_p u_1}{(2\gamma+1)^2}$ we can see that

$$
\begin{aligned}
s_p\frac{r_1}{2\gamma+1} &= \frac{\mu}{\hat{a}} r_p \frac{r_1}{2\gamma+1} = \frac{\mu}{\hat{a}} r_p\big(-\hat{a}\big) = -\mu r_p = -\mu\hat{a}\big[(\tau-1)^p - (2\gamma+\tau)^p\big], \\
t_p\frac{t_1}{2\gamma+1} &= \big\{\hat{a}\big[(\tau-1)^p + \tau^p\big] + \frac{\mu}{2}\big[\tau^p + (2\gamma+\tau)^p\big]\big\}\big[\tau - \hat{b}\big], \text{and} \\
u_p\frac{u_1}{2\gamma+1} &= \big\{\hat{a}\big[(\tau-1)^p - \tau^p\big] - \frac{\mu}{2}\big[\tau^p - (2\gamma+\tau)^p\big]\big\}\big[-\hat{b}\big]
\end{aligned}
$$

and hence

$$
\begin{aligned}
\frac{s_p r_1 + t_p t_1 + u_p u_1}{2\gamma+1} &= \hat{a}(\tau-1)^p\big\{-\mu - \hat{b} + (\tau-\hat{b})\big\} \\
&\quad + \frac{\mu}{2}(2\gamma+1)^p\big\{2\hat{a} - \hat{b} + \tau - \hat{b}\big\} \\
&\quad + \tau^p\big\{\hat{a}\hat{b} + \frac{\mu\hat{b}}{2} + \hat{a}(\tau-\hat{b}) + \frac{\mu(\tau-\hat{b})}{2}\big\} \\
&= \hat{a}(\tau-1)^p\big\{-(1-2\hat{b}) + (\tau-2\hat{b})\big\} \\
&\quad + \frac{\mu}{2}(2\gamma+1)^p\big\{2(\hat{a}-\hat{b}) + \tau\big\} \\
&\quad + \tau^p\big\{\hat{a}\tau + \frac{\mu\tau}{2}\big\}
\end{aligned}
$$

$$
\begin{aligned}
&= \hat{a}(\tau-1)^p(\tau-1) \\
&+ \frac{\mu}{2}(2\gamma+1)^p(2\gamma+\tau) \\
&+ \hat{a}\tau^{p+1} + \frac{\mu}{2}\tau^{p+1} \\
&= \hat{a}\big[(\tau-1)^p + \tau^p\big] + \frac{\mu}{2}\big[\tau^p + (2\gamma+\tau)^p\big]
\end{aligned}
$$

Hence,

$$
(\tau I_3 - \tilde{S})^{p+1}_{22} = \frac{\hat{a}\big[(\tau-1)^p + \tau^p\big] + \frac{\mu}{2}\big[\tau^p + (2\gamma+\tau)^p\big]}{2\gamma+1} = \frac{t_{p+1}}{2\gamma+1}
$$

□

Lemma B.27.

$$
\frac{u_{p+1}}{(2\gamma+1)} = (\tau I_3 - \tilde{S})^{p+1}_{23}
$$

Proof. Since $(\tau I_3 - \tilde{S})^{p+1}_{23} = \frac{s_p r_1 + t_p u_1 + u_p t_1}{(2\gamma+1)^2}$ we can see that

$$
\begin{aligned}
s_p \frac{r_1}{2\gamma+1} &= \frac{\mu}{\hat{a}} r_p \frac{r_1}{2\gamma+1} = \frac{\mu}{\hat{a}} r_p\big(-\hat{a}\big) = -\mu r_p = -\mu\hat{a}\big[(\tau-1)^p - (2\gamma+\tau)^p\big], \\
t_p \frac{u_1}{2\gamma+1} &= \big\{\hat{a}\big[(\tau-1)^p + \tau^p\big] + \frac{\mu}{2}\big[\tau^p + (2\gamma+\tau)^p\big]\big\}\big[-\hat{b}\big] \\
u_p \frac{t_1}{2\gamma+1} &= \big\{\hat{a}\big[(\tau-1)^p - \tau^p\big] - \frac{\mu}{2}\big[\tau^p - (2\gamma+\tau)^p\big]\big\}\big[\tau-\hat{b}\big]
\end{aligned}
$$

and hence

$$
\begin{aligned}
\frac{s_p r_1 + t_p u_1 + u_p t_1}{2\gamma+1} &= \hat{a}(\tau-1)^p\big[-\mu-\hat{b}+(\tau-\hat{b})\big] \\
&+ \frac{\mu}{2}(2\gamma+\tau)^p\big[2\hat{a}-\hat{b}+(\tau-\hat{b})\big] \\
&+ \tau^p\big[-\hat{a}\hat{b} - \frac{\mu\hat{b}}{2} - \hat{a}(\tau-\hat{b}) - \frac{\mu}{2}(\tau-\hat{b})\big] \\
&= \hat{a}(\tau-1)^p\big[-(1-2\hat{b})+(\tau-2\hat{b})\big] \\
&+ \frac{\mu}{2}(2\gamma+\tau)^p\big[2(\hat{a}-\hat{b})+\tau\big] \\
&+ \tau^p\big[-\hat{a}\tau - \frac{\mu\tau}{2}\big] \\
&= \hat{a}(\tau-1)^p(\tau-1) \\
&+ \frac{\mu}{2}(2\gamma+\tau)^p(2\gamma+\tau) \\
&- \hat{a}\tau^{p+1} - \frac{\mu}{2}\tau^{p+1} \\
&= \hat{a}\big[(\tau-1)^{p+1} - \tau^{p+1}\big] - \frac{\mu}{2}\big[\tau^{p+1} - (2\gamma+\tau)^{p+1}\big]
\end{aligned}
$$

Hence,

$$(\tau I_3 - \tilde{S})^{p+1}_{23} = \frac{\hat{a}\left[(\tau-1)^{p+1} - \tau^{p+1}\right] - \frac{\mu}{2}\left[\tau^{p+1} - (2\gamma+\tau)^{p+1}\right]}{2\gamma+1} = \frac{u_{p+1}}{2\gamma+1}$$

□

Therefore, with the base case ($p = 1$) and the induction step ($p + 1$) for all entries of $(\tau I_3 - \tilde{S})^p$ the proof of Lemma B.7is finished.

C PROOF OF THEOREMS 5.1 AND 5.2

C.1 PROOF OF THEOREM 5.1

For the proof of Theorem 5.1 we first present some results that are necessary. The following Lemma states the eigenvalues and eigenvectors of expected adjacency matrices according to the Stochastic Block Model here considered.

Lemma C.1. *Let $\mathcal{C}_1, \ldots, \mathcal{C}_k$ be clusters of equal size $|\mathcal{C}| = n/k$. Let $\mathcal{W} \in \mathbb{R}^{n \times n}$ be defined as*

$$\mathcal{W} = (p_{\text{in}} - p_{\text{out}}) \sum_{i=1}^{k} \mathbf{1}_{\mathcal{C}_i} \mathbf{1}_{\mathcal{C}_i}^T + p_{\text{out}} \mathbf{1}\mathbf{1}^T \tag{C.1}$$

and let $\chi_1, \ldots, \chi_k \in \mathbb{R}^n$ be defined as

$$\chi_1 = \mathbf{1}, \qquad \chi_r = \sum_{j=1}^{r} \mathbf{1}_{\mathcal{C}_j} - r\mathbf{1}_{\mathcal{C}_r} \tag{C.2}$$

for $r = 2, \ldots, k$. Then, $\chi_1, \ldots, \chi_k$ are orthogonal eigenvectors of $\mathcal{W}$, with eigenvalues

$$\lambda_1 = |\mathcal{C}|(p_{\text{in}} + (k-1)p_{\text{out}}), \qquad \lambda_r = |\mathcal{C}|(p_{\text{in}} - p_{\text{out}}) \tag{C.3}$$

Proof. Please note that from the definition that the matrix $\mathcal{W}$ is equal to p_{in} in the block diagonal and p_{out} elsewhere. We first consider the following matrix vector products that can be easily verified:

$$\mathcal{W}\mathbf{1} = |\mathcal{C}|(p_{\text{in}} + (k-1)p_{\text{out}})\mathbf{1} \tag{C.4}$$
$$\mathcal{W}\mathbf{1}_{\mathcal{C}_i} = |\mathcal{C}|(p_{\text{in}}\mathbf{1}_{\mathcal{C}_i} + p_{\text{out}}\mathbf{1}_{\overline{\mathcal{C}}_i}) \tag{C.5}$$

Moreover, we can see that

$$\begin{aligned}
\mathcal{W}\left(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_i}\right) &= |\mathcal{C}|\left(\left(p_{\text{in}}\mathbf{1}_{\mathcal{C}_j} + p_{\text{out}}\mathbf{1}_{\overline{\mathcal{C}}_j}\right) - \left(p_{\text{in}}\mathbf{1}_{\mathcal{C}_i} + p_{\text{out}}\mathbf{1}_{\overline{\mathcal{C}}_i}\right)\right) \\
&= |\mathcal{C}|\left(p_{\text{in}}\left(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_i}\right) + p_{\text{out}}\left(\mathbf{1}_{\overline{\mathcal{C}}_j} - \mathbf{1}_{\overline{\mathcal{C}}_i}\right)\right) \\
&= |\mathcal{C}|\left(p_{\text{in}}\left(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_i}\right) - p_{\text{out}}\left(\mathbf{1}_{\mathcal{C}_i} - \mathbf{1}_{\mathcal{C}_j}\right)\right) \\
&= |\mathcal{C}|(p_{\text{in}} - p_{\text{out}})\left(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_i}\right)
\end{aligned}$$

Now we show that $\chi_2, \ldots, \chi_k$ are eigenvectors of $\mathcal{W}$.

$$\begin{aligned}
\mathcal{W}\chi_r &= \mathcal{W}\left(\sum_{j=1}^{r} \mathbf{1}_{\mathcal{C}_j} - r\mathbf{1}_{\mathcal{C}_r}\right) \\
&= \mathcal{W}\sum_{j=1}^{r}(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_r}) \\
&= \sum_{j=1}^{r}\mathcal{W}\left(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_r}\right) \\
&= \sum_{j=1}^{r}|\mathcal{C}|(p_{\text{in}} - p_{\text{out}})\left(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_r}\right) \\
&= |\mathcal{C}|(p_{\text{in}} - p_{\text{out}})\sum_{j=1}^{r}\left(\mathbf{1}_{\mathcal{C}_j} - \mathbf{1}_{\mathcal{C}_r}\right) \\
&= |\mathcal{C}|(p_{\text{in}} - p_{\text{out}})\left(\sum_{j=1}^{r}\mathbf{1}_{\mathcal{C}_j} - r\mathbf{1}_{\mathcal{C}_r}\right) \\
&= |\mathcal{C}|(p_{\text{in}} - p_{\text{out}})\chi_r \\
&= \lambda_r\chi_r
\end{aligned}$$

Furthermore, we can see that eigenvectors $\chi_2, \ldots, \chi_k$ are orthogonal.

Let $2 \leq r < s \leq k$, then

$$\begin{aligned}
\chi_r^T\chi_s &= \left(\sum_{j_1=1}^{r}\mathbf{1}_{\mathcal{C}_{j_1}} - r\mathbf{1}_{\mathcal{C}_r}\right)^T\left(\sum_{j_2=1}^{s}\mathbf{1}_{\mathcal{C}_{j_2}} - s\mathbf{1}_{\mathcal{C}_s}\right) \\
&= \sum_{j_1=1}^{r}\sum_{j_2=1}^{s}\mathbf{1}_{\mathcal{C}_{j_1}}^T\mathbf{1}_{\mathcal{C}_{j_2}} - s\sum_{j_1=1}^{r}\mathbf{1}_{\mathcal{C}_{j_1}}^T\mathbf{1}_{\mathcal{C}_s} - r\sum_{j_2=1}^{s}\mathbf{1}_{\mathcal{C}_{j_2}}^T\mathbf{1}_{\mathcal{C}_r} + rs\mathbf{1}_{\mathcal{C}_r}^T\mathbf{1}_{\mathcal{C}_s} \\
&= \sum_{j_1=1}^{r}\sum_{j_2=1}^{s}\mathbf{1}_{\mathcal{C}_{j_1}}^T\mathbf{1}_{\mathcal{C}_{j_2}} - r\sum_{j_2=1}^{s}\mathbf{1}_{\mathcal{C}_{j_2}}^T\mathbf{1}_{\mathcal{C}_r} \\
&= \sum_{j_1=1}^{r}\sum_{j_2=1}^{s}\mathbf{1}_{\mathcal{C}_{j_1}}^T\mathbf{1}_{\mathcal{C}_{j_2}} - r\mathbf{1}_{\mathcal{C}_r}^T\mathbf{1}_{\mathcal{C}_r} \\
&= \sum_{j_1=1}^{r}\sum_{j_2=1}^{s}\left(\mathbf{1}_{\mathcal{C}_{j_1}}^T\mathbf{1}_{\mathcal{C}_{j_2}}\right) - r|\mathcal{C}| \\
&= \sum_{j_1=1}^{r}\left(\mathbf{1}_{\mathcal{C}_{j_1}}^T\mathbf{1}_{\mathcal{C}_{j_1}}\right) - r|\mathcal{C}| \\
&= \sum_{j_1=1}^{r}|\mathcal{C}| - r|\mathcal{C}| \\
&= r|\mathcal{C}| - r|\mathcal{C}|
\end{aligned}$$

$$= 0$$

where in the third step we used that fact that $\mathbf{1}_{\mathcal{C}_r}^T \mathbf{1}_{\mathcal{C}_s} = 0$ as $r < s$, and $\mathbf{1}_{\mathcal{C}_{j_1}}^T \mathbf{1}_{\mathcal{C}_s} = 0$ as $j_1 < s$. Finally, we can see that for $2 \leq r \leq k$

$$\chi_1^T \chi_r = \mathbf{1}^T \left(\sum_{j=1}^{r} \mathbf{1}_{\mathcal{C}_j} - r\mathbf{1}_{\mathcal{C}_r} \right) = \sum_{j=1}^{r} \left(\mathbf{1}^T \mathbf{1}_{\mathcal{C}_j} \right) - r\mathbf{1}^T \mathbf{1}_{\mathcal{C}_r} = r|\mathcal{C}| - r|\mathcal{C}| = 0$$

and hence $\chi_1, \ldots, \chi_k$ are orthogonal eigenvectors of the matrix $\mathcal{W}$.

□

The following Lemma shows the eigenvectors and eigenvalues of the power mean Laplacian in expectation under the considered Stochastic Block Model.

Lemma C.2. *Let $E(\mathbb{G})$ be the expected multilayer graph with T layers following the multilayer SBM with k classes $\mathcal{C}_1, \ldots, \mathcal{C}_k$ of equal size and parameters $\left(p_{\text{in}}^{(t)}, p_{\text{out}}^{(t)} \right)_{t=1}^{T}$. Then the eigenvalues of the power mean Laplacian $\mathcal{L}_p$ are*

$$\lambda_1(\mathcal{L}_p) = \varepsilon, \qquad \lambda_i(\mathcal{L}_p) = m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}), \qquad \lambda_j(\mathcal{L}_p) = 1 + \varepsilon \tag{C.6}$$

with eigenvectors

$$\chi_1 = \mathbf{1}, \qquad \chi_i = \sum_{j=1}^{i} \mathbf{1}_{\mathcal{C}_j} - i\mathbf{1}_{\mathcal{C}_i}$$

where $(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}})_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})/(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \varepsilon$, $t = 1, \ldots, T$, $i = 2, \ldots, k$, and $j = k+1, \ldots, |V|$,

Proof. From Lemma C.1 we know that $\chi_1, \ldots, \chi_k$ are eigenvectors of $\mathcal{W}^{(1)}, \ldots, \mathcal{W}^{(T)}$. In particular, we have seen that

$$\lambda_1^{(t)} = |\mathcal{C}|(p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}), \quad \lambda_i^{(t)} = |\mathcal{C}|(p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)})$$

for $i = 2, \ldots, k$. Further, as matrices $\mathcal{W}^{(1)}, \ldots, \mathcal{W}^{(T)}$ share all their eigenvectors, they are simultaneously diagonalizable, i.e. there exists a non-singular matrix Σ such that $\Sigma^{-1}\mathcal{W}^{(t)}\Sigma = \Lambda^{(t)}$, where $\Lambda^{(t)}$ are diagonal matrices $\Lambda^{(t)} = \text{diag}(\lambda_1^{(t)}, \ldots, \lambda_k^{(t)}, 0, \ldots, 0)$.

As we assume that all clusters are of the same size $|\mathcal{C}|$, the expected layer graphs are regular graphs with degrees $d^{(1)}, \ldots, d^{(T)}$. Hence, the normalized Laplacians of the expected layer graphs can be expressed as

$$\mathcal{L}_{\text{sym}}^{(t)} = \Sigma(I - \frac{1}{d^{(t)}}\Lambda^{(t)})\Sigma^{-1}$$

Thus, we can observe that

$$\lambda_1^{(t)}(\mathcal{L}_{\text{sym}}^{(t)}) = 0, \qquad \lambda_i^{(t)}(\mathcal{L}_{\text{sym}}^{(t)}) = 1 - \rho^{(t)}, \qquad \lambda_j^{(t)}(\mathcal{L}_{\text{sym}}^{(t)}) = 1,$$

for $i = 2, \dots, k$, and $j = k+1, \dots, |V|$, where

$$\rho^{(t)} = (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)}) / (p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)})$$

for $t = 1, \dots, T$. By obtaining the power mean Laplacian on diagonally shifted matrices,

$$\mathcal{L}_p = M_p(\mathcal{L}_{\text{sym}}^{(1)} + \varepsilon I, \dots, \mathcal{L}_{\text{sym}}^{(1)} + \varepsilon I)$$

we have by Lemma 2.1

$$\begin{aligned} \lambda_1(\mathcal{L}_p) &= m_p(\lambda_1^{(1)} + \varepsilon, \dots, \lambda_1^{(T)} + \varepsilon) = \varepsilon \\ \lambda_i(\mathcal{L}_p) &= m_p(1 - \rho^{(1)} + \varepsilon, \dots, 1 - \rho^{(T)} + \varepsilon) = m_p(\boldsymbol{\rho}_\varepsilon) \\ \lambda_j(\mathcal{L}_p) &= m_p(\lambda_j^{(1)} + \varepsilon, \dots, \lambda_j^{(T)} + \varepsilon) = 1 + \varepsilon \end{aligned} \tag{C.7}$$

where $(\boldsymbol{\rho}_\varepsilon)_t = 1 - (p_{\text{in}}^{(t)} - p_{\text{out}}^{(t)}) / (p_{\text{in}}^{(t)} + (k-1)p_{\text{out}}^{(t)}) + \varepsilon$, and $t = 1, \dots, T$, $i = 2, \dots, k$, and $j = k+1, \dots, |V|$, □

The following Lemma describes the general form of the solution matrix

$$F = (f^{(1)}, \dots, f^{(k)})$$

where the columns of F are obtained from the following optimization problem

$$f^{(r)} = \arg\min_{f \in \mathbb{R}^n} \|f - CY^{(r)}\|^2 + \lambda f^T L_p f$$

Observe that this setting contains as a particular case the problem described in Eq. (5.1).

Lemma C.3. *Let $E(\mathbb{G})$ be the expected multilayer graph with T layers following the multilayer SBM with k classes $\mathcal{C}_1, \dots, \mathcal{C}_k$ of equal size and parameters $\left(p_{\text{in}}^{(t)}, p_{\text{out}}^{(t)}\right)_{t=1}^{T}$. Let $\boldsymbol{\rho}_\varepsilon$ be defined as in Lemma C.2. Let $n_1, \dots, n_k$ be the number of labeled nodes per class. Let $C \in \mathbb{R}^{n \times n}$ be a diagonal matrix with $C_{ii} = c_r$ for $v_i \in \mathcal{C}_r$. Let $l(v_i)$ be the label of node v_i, i.e. $l(v_i) = r$ if and only if $v_i \in \mathcal{C}_r$. Let the solution matrix $F = (f^{(1)}, \dots, f^{(k)})$ where*

$$f^{(r)} = \arg\min_{f \in \mathbb{R}^n} \|f - CY^{(r)}\|^2 + \mu f^T \mathcal{L}_p f$$

Then the solution matrix F is such that:

- *If $r < l(v_i)$, then*

$$f_i^{(r)} = c_r \frac{n_r}{n} \alpha + c_r n_r \beta \left((1 - l(v_i)) \frac{1}{\left\| \boldsymbol{\chi}_{l(v_i)} \right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{1}{\left\| \boldsymbol{\chi}_j \right\|^2} \right)$$

- *If* $r > l(v_i)$, *then*

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1-r)\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right)$$

- *If* $r = l(v_i)$, *then*

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1-r)^2\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right)$$

where $\alpha = \frac{1}{1+\mu\varepsilon} - \frac{1}{1+\mu(1+\varepsilon)}$, *and* $\beta = \frac{1}{1+\mu m_p(\rho_\varepsilon)} - \frac{1}{1+\mu(1+\varepsilon)}$.

Proof. Let $U \in \mathbb{R}^{n\times n}$ be an orthonormal matrix such that $U = (u_1, u_2, \ldots, u_n)$, with $u_i = \chi_i / \|\chi_i\|$ for $i = 1, \ldots, k$, where $\chi_1, \ldots, \chi_k$ are eigenvectors of the power mean Laplacian as described in Lemma C.2.

The power mean Laplacian $\mathcal{L}_p$ is a symmetric positive semidefinite matrix (see Lemma C.2) and hence we can express $\mathcal{L}_p$ as $U\Lambda U^T$ where Λ is a diagonal matrix with entries $\Lambda_{ii} = \lambda_i(\mathcal{L}_p)$, with $i = 1, \ldots, n$. Hence, we can see that

$$(I + \mu\mathcal{L}_p)^{-1} = (I + U\Lambda U^T)^{-1} = (U(I+\Lambda)U^T)^{-1} = U(I+\Lambda)^{-1}U^T = U\Omega U^T$$

where Ω is a diagonal matrix with entries $\Omega_{ii} = \frac{1}{1+\mu\lambda_i}$, with $i = 1, \ldots, n$.

From Lemma C.2 we know that $\lambda_{k+1} = \cdots = \lambda_n = 1 + \varepsilon =: \widehat{\omega}$, and hence it follows that $\Omega_{ii} = \frac{1}{1+\mu\widehat{\omega}}$ for $i = k+1, \ldots, n$. Moreover, we can express Ω as the sum of two diagonal matrices, i.e.

$$\Omega = \omega I + \Theta$$

where $\omega = \frac{1}{1+\mu\widehat{\omega}}$ and $\Theta = \operatorname{diag}(\Omega_{11} - \omega, \ldots, \Omega_{kk} - \omega, 0, \ldots, 0)$. Observe that $\Theta_{11} = \Omega_{11} - \omega = \frac{1}{1+\mu\varepsilon} - \frac{1}{1+\mu(1+\varepsilon)} =: \alpha$ and $\Theta_{jj} = \Omega_{jj} - \omega = \frac{1}{1+\mu m_p(\rho_\varepsilon)} - \frac{1}{1+\mu(1+\varepsilon)} =: \beta$, for $j = 2, \ldots, k$.
Recall that we are interested in the equation

$$F = (I + \mu\mathcal{L}_p)^{-1}CY = U\Omega U^T CY \in \mathbb{R}^{n\times k},$$

where each column of $Y = [y^{(1)}, \ldots, y^{(k)}]$ is a class indicator of labeled nodes, i.e.

$$y_i^{(j)} = \begin{cases} 1 & \text{if } \ l(v_i) = j \\ 0 & \text{else} \end{cases} \tag{C.8}$$

Hence, each column of Y can be expressed as

$$y^{(j)} = \sum_{v_i \in V | l(v_i) = j} e_i \tag{C.9}$$

where $e_i \in \mathbb{R}^n$ and $(e_i)_i = 1$ and zero else. With this in mind, we now study the matrix-vector product $U\Omega U^T e_i$. Recall that $U\Theta U^T$ is a k-rank matrix. Hence we have

$$\begin{aligned}
U\Omega U^T e_i &= U(\omega I + \Theta)U^T e_i \\
&= \omega e_i + U\Theta U^T e_i \\
&= \omega e_i + \left(\sum_{j=1}^{k} \Theta_{jj} u_j u_j^T\right) e_i \\
&= \omega e_i + \left(\sum_{j=1}^{k} \frac{1}{\left\|\chi_j\right\|^2} \Theta_{jj} \chi_j \chi_j^T\right) e_i \\
&= \omega e_i + \frac{1}{n}\Theta_{11}\chi_1 + \left(\sum_{j=2}^{k} \frac{1}{\left\|\chi_j\right\|^2} \Theta_{jj} \chi_j \chi_j^T\right) e_i \\
&= \omega e_i + \frac{1}{n}\alpha\chi_1 + \beta\left(\sum_{j=2}^{k} \frac{1}{\left\|\chi_j\right\|^2} \chi_j \chi_j^T\right) e_i
\end{aligned}$$

where in the last steps we used the fact that $\chi_1^T e_i = \mathbf{1}^T e_i = 1$, and define $\alpha = \Theta_{11}$ and $\beta = \Theta_{jj}$ due to the fact that Θ_{jj} are all equal for $j = 2, \ldots, k$.

The remaining terms $\chi_j \chi_j^T e_i$ depend on the cluster to which the corresponding node v_i belongs to.

We first study the vector product $\chi_r^T e_i$. Observe that

$$\chi_r^T e_i = \left(\sum_{j=1}^{r} \mathbf{1}_{\mathcal{C}_j} - r\mathbf{1}_{\mathcal{C}_r}\right)^T e_i = \sum_{j=1}^{r} \left(\mathbf{1}_{\mathcal{C}_j}^T e_i\right) - r\mathbf{1}_{\mathcal{C}_r}^T e_i$$

Recall that $l(v_i)$ is the label of node v_i, i.e. $l(v_i) = r$ if and only if $v_i \in \mathcal{C}_r$. Since the nodes are ordered by class, we have that $l(v_i) = r$ for $i = (r-1)|\mathcal{C}| + 1, \ldots, r|\mathcal{C}|$. Then, we have

$$\sum_{j=1}^{r} \left(\mathbf{1}_{\mathcal{C}_j}^T e_i\right) - r\mathbf{1}_{\mathcal{C}_r}^T e_i = \begin{cases} 0 & \text{for} \quad r < l(v_i) \\ 1 - l(v_i) & \text{for} \quad r = l(v_i) \\ 1 & \text{for} \quad r > l(v_i) \end{cases} \tag{C.10}$$

Therefore,

$$\left(\sum_{j=2}^{k} \frac{1}{\left\|\chi_j\right\|^2} \chi_j \chi_j^T\right) e_i = (1 - l(v_i)) \frac{\chi_{l(v_i)}}{\left\|\chi_{l(v_i)}\right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{\chi_j}{\left\|\chi_j\right\|^2}$$

All in all we have

$$U\Omega U^T e_i = \omega e_i + \frac{1}{n}\alpha\chi_1 + \beta\left((1 - l(v_i)) \frac{\chi_{l(v_i)}}{\left\|\chi_{l(v_i)}\right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{\chi_j}{\left\|\chi_j\right\|^2}\right)$$

Moreover, the solution matrix F can now be described column-wise as follows

$$\begin{aligned}
f^{(r)} &= (I + \mu\mathcal{L}_p)^{-1} C y^{(r)} \\
&= c_r \left(\sum_{v_i \in V | l(v_i) = r} U \Omega U^T e_i \right) \\
&= c_r \left(\sum_{v_i \in V | l(v_i) = r} \omega e_i \right) + \frac{1}{n} c_r n_r \alpha \chi_1 + c_r n_r \beta \left((1 - l(v_i)) \frac{\chi_{l(v_i)}}{\left\| \chi_{l(v_i)} \right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{\chi_j}{\left\| \chi_j \right\|^2} \right) \\
&= \omega c_r y^{(r)} + c_r n_r \left(\frac{1}{n} \alpha \chi_1 + \beta \left((1 - r) \frac{\chi_r}{\left\| \chi_r \right\|^2} + \sum_{j=r+1}^{k} \frac{\chi_j}{\left\| \chi_j \right\|^2} \right) \right)
\end{aligned}$$

We now study the columns of matrix F. For this, observe that the i^{th} entry of the column corresponding to the class r, is obtained by $f_i^{(r)} = \langle e_i, f^{(r)} \rangle$, and hence have

$$\begin{aligned}
\langle e_i, f^{(r)} \rangle =& \langle e_i, \omega c_r y^{(r)} + c_r n_r \left(\frac{1}{n} \alpha \chi_1 + \beta \left((1 - r) \frac{\chi_r}{\left\| \chi_r \right\|^2} + \sum_{j=r+1}^{k} \frac{\chi_j}{\left\| \chi_j \right\|^2} \right) \right) \rangle \\
=& c_r \frac{n_r}{n} \alpha + c_r n_r \beta \langle e_i, \left((1 - r) \frac{\chi_r}{\left\| \chi_r \right\|^2} c_r + \sum_{j=r+1}^{k} \frac{\chi_j}{\left\| \chi_j \right\|^2} \right) \rangle
\end{aligned}$$

where $\langle e_i, \omega c_r y^{(r)} \rangle = 0$ for unlabeled nodes. Having this, we now proceed to study three different cases of the remaining inner product. We do this by considering the following cases and making use of Eq. (C.10):

First case: $f_i^{(r)}$ with $r < l(v_i)$. We first analyze the following term

$$\begin{aligned}
\langle e_i, \left((1 - r) \frac{\chi_r}{\left\| \chi_r \right\|^2} + \sum_{j=r+1}^{k} \frac{\chi_j}{\left\| \chi_j \right\|^2} \right) \rangle &= \langle e_i, (1 - r) \frac{\chi_r}{\left\| \chi_r \right\|^2} \rangle + \langle e_i, \sum_{j=r+1}^{k} \frac{\chi_j}{\left\| \chi_j \right\|^2} \rangle \\
&\overset{\text{(by first case of Eq.C.10)}}{=} \langle e_i, \sum_{j=r+1}^{k} \frac{\chi_j}{\left\| \chi_j \right\|^2} \rangle \\
&\overset{\text{(by cases of Eq.C.10)}}{=} (1 - l(v_i)) \frac{1}{\left\| \chi_{l(v_i)} \right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{1}{\left\| \chi_j \right\|^2}
\end{aligned}$$

Thus, we have

$$f_i^{(r)} = c_r \frac{n_r}{n} \alpha + c_r n_r \beta \left((1 - l(v_i)) \frac{1}{\left\| \chi_{l(v_i)} \right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{1}{\left\| \chi_j \right\|^2} \right)$$

Second case: $f_i^{(r)}$ **with** $r > l(v_i)$. We first analyze the following term

$$\langle e_i, \left((1-r)\frac{\chi_r}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{\chi_j}{\|\chi_j\|^2} \right) \rangle = \langle e_i, (1-r)\frac{\chi_r}{\|\chi_r\|^2} \rangle + \langle e_i, \sum_{j=r+1}^{k} \frac{\chi_j}{\|\chi_j\|^2} \rangle$$

$$\text{(by third case of Eq.C.10)} \qquad = (1-r)\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2}$$

Thus, we have

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1-r)\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right)$$

Third case: $f_i^{(r)}$ **with** $r = l(v_i)$. We first analyze the following term

$$\langle e_i, \left((1-r)\frac{\chi_r}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{\chi_j}{\|\chi_j\|^2} \right) \rangle = \langle e_i, (1-r)\frac{\chi_r}{\|\chi_r\|^2} \rangle + \langle e_i, \sum_{j=r+1}^{k} \frac{\chi_j}{\|\chi_j\|^2} \rangle$$

$$\text{(by second case of Eq.C.10)} \qquad = (1-r)^2\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2}$$

Thus, we have

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1-r)^2\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right)$$

These three cases are the desired conditions.

□

We now finally provide the proof for Theorem 5.1.

Proof of Theorem 5.1. The proof of this theorem builds on top of Lemma C.3, where the entries of the solution matrix $F = (f^{(1)}, \ldots, f^{(k)})$ are described, where

$$f^{(r)} = \arg\min_{f \in \mathbb{R}^n} \|f - CY^{(r)}\|^2 + \mu f^T \mathcal{L}_p f$$

Let $l(v_i)$ be the label of node v_i, i.e. $l(v_i) = r$ if and only if $v_i \in \mathcal{C}_r$. According to Lemma C.3 the entries of matrix F for unlabeled nodes are such that

- If $r < l(v_i)$, then

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1-l(v_i))\frac{1}{\left\|\chi_{l(v_i)}\right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{1}{\|\chi_j\|^2} \right)$$

- If $r > l(v_i)$, then

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1-r)\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right)$$

- If $r = l(v_i)$, then

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1-r)^2\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right)$$

where $\alpha = \frac{1}{1+\mu\varepsilon} - \frac{1}{1+\mu(1+\varepsilon)}$, and $\beta = \frac{1}{1+\mu m_p(\rho_\varepsilon)} - \frac{1}{1+\mu(1+\varepsilon)}$.

Observe that the case here considered corresponds to the case where the amount of labeled data per class is the same, i.e. $n_1 = \cdots = n_k$, and where the matrix C is the identity, i.e. $c_1 = \cdots c_r = 1$.

Moreover, the estimated label assignment for unlabeled nodes goes by the following rule

$$\hat{l}(v_i) = \arg\max\{f_i^{(1)}, \ldots, f_i^{(k)}\}$$

Hence, we need to find conditions so that the following inequality holds

$$f_i^{(j)} < f_i^{(l(v_i))} \qquad \forall\ j \neq l(v_i)$$

Hence, we consider the following two cases:

Case 1: $f_i^{(r)} < f_i^{(l(v_i))}$ **for** $r > l(v_i)$**.**

Let $r^* = l(v_i)$, and $r = r^* + \Delta$. Then, we have

$$\begin{aligned}
&f_i^{(r)} < f_i^{(l(v_i))} \Leftrightarrow \\
&f_i^{(r)} < f_i^{(r^*)} \Leftrightarrow \\
&\beta\left((1-r)\frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right) < \beta\left((1-r^*)^2\frac{1}{\|\chi_{r^*}\|^2} + \sum_{j=r^*+1}^{k} \frac{1}{\|\chi_j\|^2} \right) \Leftrightarrow \\
&0 < \beta\left((1-r^*)^2\frac{1}{\|\chi_{r^*}\|^2} - (1-r)\frac{1}{\|\chi_r\|^2} + \sum_{j=r^*+1}^{k} \frac{1}{\|\chi_j\|^2} - \sum_{j=r+1}^{k} \frac{1}{\|\chi_j\|^2} \right) \Leftrightarrow \\
&0 < \beta\left((1-r^*)^2\frac{1}{\|\chi_{r^*}\|^2} + (r-1)\frac{1}{\|\chi_r\|^2} + \sum_{j=r^*+1}^{k} \frac{1}{\|\chi_j\|^2} - \sum_{j=r^*+\Delta+1}^{k} \frac{1}{\|\chi_j\|^2} \right) \Leftrightarrow \\
&0 < \beta\left((1-r^*)^2\frac{1}{\|\chi_{r^*}\|^2} + (r-1)\frac{1}{\|\chi_r\|^2} + \sum_{j=r^*+1}^{r^*+\Delta} \frac{1}{\|\chi_j\|^2} \right) \Leftrightarrow \\
&0 < \beta
\end{aligned}$$

Case 2: $f_i^{(r)} < f_i^{(l(v_i))}$ for $r < l(v_i)$.
Let $r^* = l(v_i)$, and $r^* = r + \Delta$. Then, we have

$$
\begin{aligned}
& f_i^{(r)} < f_i^{(l(v_i))} \Leftrightarrow \\
& f_i^{(r)} < f_i^{(r^*)} \Leftrightarrow \\
& \beta \left((1-r^*) \frac{1}{\|\boldsymbol{\chi}_{r^*}\|^2} + \sum_{j=r^*+1}^{k} \frac{1}{\|\boldsymbol{\chi}_j\|^2} \right) < \beta \left((1-r^*)^2 \frac{1}{\|\boldsymbol{\chi}_{r^*}\|^2} + \sum_{j=r^*+1}^{k} \frac{1}{\|\boldsymbol{\chi}_j\|^2} \right) \Leftrightarrow \\
& 0 < \beta \left((1-r^*)^2 \frac{1}{\|\boldsymbol{\chi}_{r^*}\|^2} - (1-r^*) \frac{1}{\|\boldsymbol{\chi}_{r^*}\|^2} \right) \\
& 0 < \beta \left((1-r^*)^2 \frac{1}{\|\boldsymbol{\chi}_{r^*}\|^2} + (r^*-1) \frac{1}{\|\boldsymbol{\chi}_{r^*}\|^2} \right) \Leftrightarrow \\
& 0 < \beta
\end{aligned}
$$

All in all, from the two considered cases we can see that

$$f_i^{(j)} < f_i^{(l(v_i))} \qquad \forall\, j \neq l(v_i) \Longleftrightarrow 0 < \beta$$

In fact,

$$
\begin{aligned}
0 < \beta \Leftrightarrow \\
0 < \frac{1}{1+\mu m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}})} - \frac{1}{1+\mu(1+\varepsilon)} \Leftrightarrow \\
\frac{1}{1+\mu(1+\varepsilon)} < \frac{1}{1+\mu m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}})} \Leftrightarrow \\
1+\mu m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) < 1+\mu(1+\varepsilon) \Leftrightarrow \\
m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) < 1+\varepsilon
\end{aligned}
$$

which is the desired condition. □

C.2 PROOF OF THEOREM 5.2

We first give a general version of Theorem 5.2.

Theorem C.1. *Let $E(\mathbb{G})$ be the expected multilayer graph with T layers following the multilayer SBM with two classes $\mathcal{C}_1, \mathcal{C}_2$ of equal size and parameters $\left(p_{\text{in}}^{(t)}, p_{\text{out}}^{(t)}\right)_{t=1}^{T}$. Let n_1, n_2 nodes from classes $\mathcal{C}_1, \mathcal{C}_2$ be labeled, respectively. Let $\mu = 1$. Then, a zero test classification error is achieved if and only if*

$$m_p(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}}) < \min \left\{ \frac{(n_1+n_2)((1+\varepsilon)^2+1) - 2n_2}{2n_2 + (n_1+n_2)\varepsilon}, \frac{(n_1+n_2)((1+\varepsilon)^2+1) - 2n_1}{2n_1 + (n_1+n_2)\varepsilon} \right\}$$

where $(\boldsymbol{\rho}_{\boldsymbol{\varepsilon}})_l = 1 - (p_{\text{in}}^{(l)} - p_{\text{out}}^{(l)}) / (p_{\text{in}}^{(l)} + (k-1)p_{\text{out}}^{(l)}) + \varepsilon$, and $l = 1, \ldots, T$.

Proof. The proof of this theorem builds on top of Lemma C.3, where the entries of the solution matrix $F = (f^{(1)}, \ldots, f^{(k)})$ are described, where

$$f^{(r)} = \arg\min_{f \in \mathbb{R}^n} \|f - CY^{(r)}\|^2 + \mu f^T \mathcal{L}_p f$$

Let $l(v_i)$ be the label of node v_i, i.e. $l(v_i) = r$ if and only if $v_i \in \mathcal{C}_r$. According to Lemma C.3 the entries of matrix F for unlabeled nodes are such that

- If $r < l(v_i)$, then

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1 - l(v_i)) \frac{1}{\left\|\chi_{l(v_i)}\right\|^2} + \sum_{j=l(v_i)+1}^{k} \frac{1}{\left\|\chi_j\right\|^2} \right)$$

- If $r > l(v_i)$, then

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1 - r) \frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\left\|\chi_j\right\|^2} \right)$$

- If $r = l(v_i)$, then

$$f_i^{(r)} = c_r \frac{n_r}{n}\alpha + c_r n_r \beta \left((1 - r)^2 \frac{1}{\|\chi_r\|^2} + \sum_{j=r+1}^{k} \frac{1}{\left\|\chi_j\right\|^2} \right)$$

where $\alpha = \frac{1}{1+\mu\varepsilon} - \frac{1}{1+\mu(1+\varepsilon)}$, and $\beta = \frac{1}{1+\mu m_p(\rho_\varepsilon)} - \frac{1}{1+\mu(1+\varepsilon)}$.

Observe that the case here considered corresponds to the case with two classes, i.e. $k = 2$ with equal size classes $\mathcal{C}_1$ and $\mathcal{C}_2$ where the amount of labeled data per class is n_1 and n_2, respectively, with the matrix C as the identity, i.e. $c_1 = c_2 = 1$, and regularization parameter $\mu = 1$.

Moreover, the estimated label assignment for unlabeled nodes goes by the following rule

$$\hat{l}(v_i) = \arg\max\{f_i^{(1)}, f_i^{(2)}\}$$

Hence, we need to find conditions so that the following inequality holds

$$f_i^{(j)} < f_i^{(l(v_i))} \qquad \forall\ j \neq l(v_i)$$

Let $l(v_i) = 1 \Leftrightarrow v_i \in \mathcal{C}_1$, and $l(v_i) = 2 \Leftrightarrow v_i \in \mathcal{C}_2$. A quick computation following Lemma C.3 yields

- $f_i^{(1)} = \frac{n_1}{n}\alpha + n_1\beta(\frac{1}{\|\chi_2\|^2})$ for $v_i \in \mathcal{C}_1$, i.e. $l(v_i) = 1$
- $f_i^{(1)} = \frac{n_1}{n}\alpha - n_1\beta(\frac{1}{\|\chi_2\|^2})$ for $v_i \in \mathcal{C}_2$, i.e. $l(v_i) = 2$

- $f_i^{(2)} = \frac{n_2}{n}\alpha - n_2\beta(\frac{1}{\|\chi_2\|^2})$ for $v_i \in \mathcal{C}_1$, i.e. $l(v_i) = 1$
- $f_i^{(2)} = \frac{n_2}{n}\alpha + n_2\beta(\frac{1}{\|\chi_2\|^2})$ for $v_i \in \mathcal{C}_2$, i.e. $l(v_i) = 2$

Observing that $\|\chi_2\|^2 = n$ these conditions can be rephrased as follows

$$f^{(1)} = \frac{n_1}{n}\left((\alpha+\beta)\,\mathbf{1}_{\mathcal{C}} + (\alpha-\beta)\,\mathbf{1}_{\overline{\mathcal{C}}}\right)$$
$$f^{(2)} = \frac{n_2}{n}\left((\alpha-\beta)\,\mathbf{1}_{\mathcal{C}} + (\alpha+\beta)\,\mathbf{1}_{\overline{\mathcal{C}}}\right)$$

Hence, the conditions for correct label assignment of unlabeled nodes are

$$n_1\,(\alpha+\beta) > n_2\,(\alpha-\beta) \text{ and } n_2\,(\alpha+\beta) > n_1\,(\alpha-\beta)$$

Let $\Omega_{11} = \frac{1}{1+\varepsilon}, \Omega_{22} = \frac{1}{1+m_p(\boldsymbol{\rho}_\varepsilon)}$, and $\omega = \frac{1}{1+(1+\varepsilon)}$. Then, $\alpha = \Omega_{11} - \omega$, and $\beta = \Omega_{22} - \omega$.

By studying the first condition we observe

$$\begin{aligned}
n_1\,(\alpha+\beta) &> n_2\,(\alpha-\beta) \Leftrightarrow \\
n_1\,(\Omega_{11}-\omega+\Omega_{22}-\omega) &> n_2\,(\Omega_{11}-\omega-(\Omega_{22}-\omega)) \Leftrightarrow \\
n_1\,(\Omega_{11}+\Omega_{22}-2\omega) &> n_2(\Omega_{11}-\Omega_{22}) \Leftrightarrow \\
(n_1-n_2)\Omega_{11} + (n_1+n_2)\Omega_{22} &> 2n_1\omega \Leftrightarrow \\
\Omega_{22} &> \frac{1}{n_1+n_2}\left(2n_1\omega - (n_1-n_2)\Omega_{11}\right) \Leftrightarrow \\
\frac{1}{1+m_p(\boldsymbol{\rho}_\varepsilon)} &> \frac{1}{n_1+n_2}\left(2n_1\frac{1}{1+(1+\varepsilon)} - (n_1-n_2)\Omega_{11}\right) \Leftrightarrow \\
\frac{1}{1+m_p(\boldsymbol{\rho}_\varepsilon)} &> \frac{1}{n_1+n_2}\left(2n_1\frac{1}{2+\varepsilon} - (n_1-n_2)\frac{1}{1+\varepsilon}\right) \Leftrightarrow \\
\frac{1}{1+m_p(\boldsymbol{\rho}_\varepsilon)} &> \frac{1}{n_1+n_2}\left(\frac{2n_2+(n_1+n_2)\varepsilon}{(2+\varepsilon)(1+\varepsilon)}\right) \Leftrightarrow \\
1+m_p(\boldsymbol{\rho}_\varepsilon) &< (n_1+n_2)\left(\frac{(2+\varepsilon)(1+\varepsilon)}{2n_2+(n_1+n_2)\varepsilon}\right) \Leftrightarrow \\
m_p(\boldsymbol{\rho}_\varepsilon) &< (n_1+n_2)\left(\frac{(2+\varepsilon)(1+\varepsilon)}{2n_2+(n_1+n_2)\varepsilon}\right) - 1 \Leftrightarrow \\
m_p(\boldsymbol{\rho}_\varepsilon) &< \frac{(n_1+n_2)(2+\varepsilon)(1+\varepsilon) - (2n_2+(n_1+n_2)\varepsilon)}{2n_2+(n_1+n_2)\varepsilon} \Leftrightarrow \\
m_p(\boldsymbol{\rho}_\varepsilon) &< \frac{(n_1+n_2)((2+\varepsilon)(1+\varepsilon)-\varepsilon) - 2n_2}{2n_2+(n_1+n_2)\varepsilon} \Leftrightarrow \\
m_p(\boldsymbol{\rho}_\varepsilon) &< \frac{(n_1+n_2)((1+\varepsilon)^2+1) - 2n_2}{2n_2+(n_1+n_2)\varepsilon} \Leftrightarrow
\end{aligned}$$

The corresponding condition for $\mathcal{C}_2$ can be obtained in a similar way, yielding

$$m_p(\boldsymbol{\rho}_\varepsilon) < \frac{(n_1+n_2)((1+\varepsilon)^2+1)-2n_1}{2n_1+(n_1+n_2)\varepsilon}$$

Hence, both conditions hold if and only if

$$m_p(\boldsymbol{\rho}_\varepsilon) = m_p(\boldsymbol{\rho}_\varepsilon) < \min\left\{\frac{(n_1+n_2)((1+\varepsilon)^2+1)-2n_2}{2n_2+(n_1+n_2)\varepsilon}, \frac{(n_1+n_2)((1+\varepsilon)^2+1)-2n_1}{2n_1+(n_1+n_2)\varepsilon}\right\}$$

□

We are now ready to give the proof of Theorem 5.2.

Proof of Theorem 5.2. We first analyze the first condition of the right hand side of Theorem C.1. Let $g(\varepsilon) = \frac{(n_1+n_2)((1+\varepsilon)^2+1)-2n_2}{2n_2+(n_1+n_2)\varepsilon}$. Then,

$$g(0) = \frac{2(n_1+n_2)-2n_2}{2n_2} = \frac{n_1}{n_2}$$

Moreover, it is clear that g is monotone, as it is quadratic on ε on the numerator and linear on the denominator, and hence $g(0) < g(\varepsilon)$.

A similar procedure with the second condition of the right hand side of Theorem C.1 leads to the condition $\frac{n_2}{n_1}$, leading to the desired result. □

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

E. Abbe (2018). Community Detection and Stochastic Block Models: Recent Developments, *Journal of Machine Learning Research*, vol. 18(177), pp. 1–86. 22

E. Abbe, A. S. Bandeira, A. Bracher, and A. Singer (2014). Decoding Binary Node Labels from Censored Edge Measurements: Phase Transition and Efficient Recovery, *IEEE Transactions on Network Science and Engineering*, vol. 1(1), pp. 10–22. 23

E. Abbe, E. Boix-Adsera, P. Ralli, and C. Sandon (2020). Graph Powering and Spectral Robustness, *SIAM Journal on Mathematics of Data Science*, vol. 2(1), pp. 132–157. 84

P. Agrrawal, R. Borgman, J. M. Clark, and R. Strong (2010). Using the price-to-earnings harmonic mean to improve firm valuation estimates, *Journal of Financial Education*, pp. 98–110. 15

A. Aleta and Y. Moreno (2019). Multilayer Networks in a Nutshell, *Annual Review of Condensed Matter Physics*, vol. 10(1), pp. 45–62. 5

D. Aloise, A. Deshpande, P. Hansen, and P. Popat (2009). NP-Hardness of Euclidean Sum-of-Squares Clustering, *Machine Learning*, vol. 75, p. 245–248. 14

A. Argyriou, M. Herbster, and M. Pontil (2006). Combining Graph Laplacians for Semi–Supervised Learning, in *Neural Information Processing Systems (NeurIPS) 2006*. 7, 46, 61, 63, 75

V. Arsigny, P. Fillard, X. Pennec, and N. Ayache (2006). Log-Euclidean metrics for fast and simple calculus on diffusion tensors, *Magnetic resonance in medicine*, vol. 56, pp. 411–421. 17

V. Arsigny, P. Fillard, X. Pennec, and N. Ayache (2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices, *SIAM Journal on Matrix Analysis and Applications*, vol. 29, pp. 328–347. 16

D. Arthur and S. Vassilvitskii (2007). K-Means++: The Advantages of Careful Seeding, in *Symposium on Discrete Algorithms (SODA) 2007*. 15

O. Bachem, M. Lucic, and S. Lattanzi (2018). One-shot Coresets: The Case of k-Clustering, in *Artificial Intelligence and Statistic (AISTATS) 2018*. 15

N. Bansal, A. Blum, and S. Chawla (2004). Correlation Clustering, *Machine Learning*, vol. 56(1), pp. 89–113. 5, 20

D. S. Bassett, M. A. Porter, N. F. Wymbs, S. T. Grafton, J. M. Carlson, and P. J. Mucha (2013). Robust detection of dynamic community structure in networks, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 23(1), p. 013142. 6

F. Battiston, V. Nicosia, and V. Latora (2014). Structural measures for multiplex networks, *Physical Review E*, vol. 89, p. 032804. 5

M. Bazzi, L. G. S. Jeub, A. Arenas, S. D. Howison, and M. A. Porter (2016). A Framework for the Construction of Generative Models for Mesoscale Structure in Multilayer Networks, *arXiv:1608.06196*. 45, 49, 55, 56

M. Belkin, I. Matveeva, and P. Niyogi (2004). Regularization and Semi-supervised Learning on Large Graphs, in *Conference on Learning Theory (COLT) 2004*. 7, 61

K. V. Bhagwat and R. Subramanian (1978). Inequalities between means of positive operators, *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 83(3), p. 393–401. 16, 64

R. Bhatia (1997). *Matrix Analysis*, Springer New York. 88, 89

R. Bhatia (2009). *Positive definite matrices*, Princeton University Press. 16, 17

D. Bini and B. Ianazzo (2015). *The Matrix Means Toolbox*, *http://bezout.dm.unipi.it/software/mmtoolbox/*. 37

D. A. Bini and B. Iannazzo (2011). A Note on Computing Matrix Geometric Means, *Advances in Computational Mathematics*, vol. 35(2–4), p. 175–192. 17

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008). Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008(10), p. P10008. 83

S. Boccaletti, G. Bianconi, R. Criado, C. del Genio, J. G.-G. nes, M. Romance, I. S. na Nadal, Z. Wang, and M. Zanin (2014). The structure and dynamics of multilayer networks, *Physics Reports*, vol. 544(1), pp. 1 – 122. 5, 45

P. S. Bullen (2013). *Handbook of means and their inequalities*, vol. 560, Springer Science & Business Media. 16

X. Cao, C. Zhang, C. Zhou, H. Fu, and H. Foroosh (2015). Constrained Multi-View Video Face Clustering, *IEEE Transactions on Image Processing*, vol. 24(11), pp. 4381–4393. 45

D. Cartwright and F. Harary (1956). Structural balance: a generalization of Heider's theory., *Psychological Review*, vol. 63(5), pp. 277–293. 3, 18, 20, 45

I. Chami, Z. Ying, C. Ré, and J. Leskovec (2019). Hyperbolic Graph Convolutional Neural Networks, in *Neural Information Processing Systems (NeurIPS) 2019*. 7

O. Chapelle, B. Schölkopf, and A. Zien (2010). *Semi-Supervised Learning*, The MIT Press. 61

K. Chaudhuri, F. Chung, and A. Tsiatas (2012). Spectral Clustering of Graphs with General Degrees in the Extended Planted Partition Model, in *Conference on Learning Theory (COLT) 2012*. 35

P. Y. Chen and A. O. Hero (2017). Multilayer Spectral Graph Clustering via Convex Layer Aggregation: Theory and Algorithms, *IEEE Transactions on Signal and Information Processing over Networks*, vol. 3(3), pp. 553–567. 6, 7, 46, 47

Z. Chen, L. Li, and J. Bruna (2019). Supervised Community Detection with Line Graph Neural Networks, in *International Conference on Learning Representations (ICLR) 2019*. 83

K. Chiang, J. Whang, and I. Dhillon (2012). Scalable clustering of signed networks using balance normalized cut, in *Conference on Information and Knowledge Management (CIKM) 2012*. 4, 19, 21, 39

K.-Y. Chiang, N. Natarajan, A. Tewari, and I. S. Dhillon (2011). Exploiting Longer Cycles for Link Prediction in Signed Networks, in *Conference on Information and Knowledge Management (CIKM) 2011*. 4, 19

F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii (2017). Fair Clustering Through Fairlets, in *Neural Information Processing Systems (NeurIPS) 2017*. 83

F. Chung and M. Radcliffe (2011). On the spectra of general random graphs, *The electronic journal of combinatorics*, vol. 18(1). 35, 87, 97, 98

F. Chung, A. Tsiatas, and W. Xu (2013). Dirichlet PageRank and Ranking Algorithms Based on Trust and Distrust, *Internet Mathematics*, vol. 9(1), pp. 113–134. 19

M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery (2011). Learning to Extract Symbolic Knowledge from the World Wide Web, in *AAAI Conference on Artificial Intelligence 2011*. 59, 75

M. Cucuringu, P. Davies, A. Glielmo, and H. Tyagi (2019). SPONGE: A generalized eigenproblem for clustering signed networks, in *Artificial Intelligence and Statistics (AISTATS) 2019*. 4, 5, 19, 39

M. Cucuringu, A. Pizzoferrato, and Y. van Gennip (2018). An MBO scheme for clustering and semi-supervised clustering of signed networks, *arXiv:1901.03091*. 5, 19

M. Cuturi and A. Doucet (2014). Fast Computation of Wasserstein Barycenters, in *International Conference on Machine Learning (ICML) 2014*. 83

E. Davis and S. Sethuraman (2018). Consistency of modularity clustering on random geometric graphs, *Annals of Applied Probability*, vol. 28(4), pp. 2003–2062. 35

J. A. Davis (1967). Clustering and structural balance in graphs, *Human Relations*, vol. 20, pp. 181–187. 3, 18, 20

C. De Bacco, E. A. Power, D. B. Larremore, and C. Moore (2017). Community detection, link prediction, and layer interdependence in multilayer networks, *Physical Review E*, vol. 95, p. 042317. 5, 6, 45, 47, 59

M. De Domenico, V. Nicosia, A. Arenas, and V. Latora (2015). Structural reducibility of multilayer networks, *Nature Communications*, vol. 6, p. 6864. 45

T. Derr, Y. Ma, and J. Tang (2018). Signed Graph Convolutional Network, in *International Conference on Data Mining (ICDM) 2018*. 4, 19

M. Desai and V. Rao (1994). A characterization of the smallest eigenvalue of a graph, *Journal of Graph Theory*, vol. 18(2), pp. 181–194. 4, 20

X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov (2012). Clustering With Multi-Layer Graphs: A Spectral Perspective, *IEEE Transactions on Signal Processing*, vol. 60(11), pp. 5820–5831. 6, 45, 47

X. Dong, P. Frossard, P. Vandergheynst, and N. Nefedov (2014). Clustering on Multi-Layer Graphs via Subspace Analysis on Grassmann Manifolds, *IEEE Transactions on Signal Processing*, vol. 62(4), pp. 905–918. 6, 45

P. Doreian and A. Mrvar (2009). Partitioning signed social networks, *Social Networks*, vol. 31(1), pp. 1–11. 5, 19

T. A. Driscoll (2005). Algorithm 843: improvements to the Schwarz-Christoffel toolbox for MATLAB, *ACM Transactions on Mathematical Software (TOMS)*, vol. 31(2), pp. 239–251. 74

B. Ermis, E. Acar, and A. T. Cemgil (2015). Link prediction in heterogeneous data via generalized coupled tensor factorization, *Data Mining and Knowledge Discovery*, vol. 29(1), pp. 203–236. 5

D. Eswaran, S. Günnemann, C. Faloutsos, D. Makhija, and M. Kumar (2017). ZooBP: Belief Propagation for Heterogeneous Networks, in *International Conference on Very Large Data Bases (VLDB) 2017*. 8, 63, 75

G. L. Falher, N. Cesa-Bianchi, C. Gentile, and F. Vitale (2017). On the Troll-Trust Model for Edge Sign Prediction in Social Networks, in *Artificial Intelligence and Statistics (AISTATS) 2017*. 3, 18

M. Fasi and B. Iannazzo (2018). Computing the Weighted Geometric Mean of Two Large-Scale Matrices and Its Inverse Times a Vector, *SIAM Journal on Matrix Analysis and Applications*, vol. 39(1), pp. 178–203. 17

D. Fasino and F. Tudisco (2018). A modularity based spectral method for simultaneous community and anti-community detection, *Linear Algebra and its Applications*, vol. 542, pp. 605–623. 35

W. F. Ferger (1931). The Nature and Use of the Harmonic Mean, *Journal of the American Statistical Association*, vol. 26(173), pp. 36–40. 16

S. Fortunato (2010). Community detection in graphs, *Physics Reports*, vol. 486(3), pp. 75 – 174. 6

A. Fujita, P. Severino, K. Kojima, J. R. Sato, A. G. Patriota, and S. Miyano (2012). Functional clustering of time series gene expression data by Granger causality, *BMC systems biology*, vol. 6(1), p. 137. 18

J. Gallier (2016). Spectral theory of unsigned and signed graphs. Applications to graph clustering: a survey, *arXiv:1601.04692*. 5, 19

R. Gallotti and M. Barthelemy (2015). The multilayer temporal network of public transport in Great Britain, *Scientific Data*, vol. 2. 45

I. Giotis and V. Guruswami (2006). Correlation Clustering with a Fixed Number of Clusters, in *Symposium on Discrete Algorithms (SODA) 2006*. 5, 20

A. B. Goldberg, X. Zhu, and S. Wright (2007). Dissimilarity in Graph-Based Semi-Supervised Classification, in *Artificial Intelligence and Statistics (AISTATS) 2007*. 4

D. Greene and P. Cunningham (2005). Producing Accurate Interpretable Clusters from High-Dimensional Data, in *Knowledge Discovery in Databases (PKDD) 2005*. 59, 75

D. Greene and P. Cunningham (2009). A matrix factorization approach for integrating multiple data views, *Machine Learning and Knowledge Discovery in Databases*, pp. 423–438. 59, 75

E. Gujral and E. E. Papalexakis (2018). SMACD: Semi-supervised Multi-Aspect Community Detection, in *SIAM International Conference on Data Mining (SDM) 2018*. 8, 63, 74, 75, 76

A. Haans (2008). What does it mean to be average? The miles per gallon versus gallons per mile paradox revisited, *Practical Assessment, Research, and Evaluation*, vol. 13(1). 16

N. Hale, N. J. Higham, and L. N. Trefethen (2008). Computing A^α,\log(A), and related matrix functions by contour integrals, *SIAM Journal on Numerical Analysis*, vol. 46(5), pp. 2505–2523. 72, 73

Q. Han, K. S. Xu, and E. M. Airoldi (2015). Consistent Estimation of Dynamic and Multi-layer Block Models, in *International Conference on Machine Learning (ICML) 2015*. 35, 49

D. J. Hand (1994). Deconstructing Statistical Questions, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 157(3), pp. 317–356. 16

F. Harary (1953). On the notion of balance of a signed graph, *Michigan Mathematical Journal*, vol. 2, pp. 143–146. 3, 18, 20

G. Hardy, J. Littlewood, and G. Pólya (1934). *Inequalities*, Cambridge University Press. 16

X. He, L. Li, D. Roqueiro, and K. Borgwardt (2017). Multi-view Spectral Clustering on Conflicting Views, in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD) 2017*. 59

S. Heimlicher, M. Lelarge, and L. Massoulié (2012). Community detection in the labelled stochastic block model, *arXiv:1209.2910*. 23, 35, 49, 64

N. J. Higham (2008). *Functions of Matrices: Theory and Computation*, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA. 57

P. W. Holland, K. B. Laskey, and S. Leinhardt (1983). Stochastic blockmodels: First steps, *Social Networks*, vol. 5(2), pp. 109 – 137. 22

R. Horn and C. Johnson (1991). *Topics in Matrix Analysis*, Cambridge University Press. 107

H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen (2012). Affinity aggregation for spectral clustering, in *Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. 6, 46

R. Jenatton, N. L. Roux, A. Bordes, and G. R. Obozinski (2012). A latent factor model for highly multi-relational data, in *Neural Information Processing Systems (NeurIPS) 2012*. 6, 47

V. Jog and P.-L. Loh (2015). Information-theoretic bounds for exact recovery in weighted stochastic block models using the Renyi divergence, *arXiv:1509.06418*. 35, 49

A. Joseph and B. Yu (2016). Impact of regularization on spectral clustering, *Annals of Statistics*, vol. 44(4), pp. 1765–1791. 35

V. Kanade, E. Mossel, and T. Schramm (2016). Global and Local Information in Clustering Labeled Block Models, *IEEE Transactions on Information Theory*, vol. 62(10), pp. 5906–5917. 65

M. Karasuyama and H. Mamitsuka (2013). Multiple Graph Label Propagation by Sparse Integration, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 24(12), pp. 1999–2012. 7, 61, 63, 74, 75

T. Kato, H. Kashima, and M. Sugiyama (2009). Robust Label Propagation on Multiple Networks, *Transactions on Neural Networks*, vol. 20(1), pp. 35–44. 7, 61, 63

J. Kim and J.-G. Lee (2015). Community Detection in Multi-Layer Graphs: A Survey, *SIGMOD Record*. 7, 45

J. Kim, H. Park, J.-E. Lee, and U. Kang (2018). SIDE: Representation Learning in Signed Directed Networks, in *International World Wide Web Conference (WWW) 2018*. 4, 19

T. N. Kipf and M. Welling (2017). Semi-Supervised Classification with Graph Convolutional Networks, in *International Conference on Learning Representations (ICLR) 2017*. 7, 61, 83

A. Kirkley, G. T. Cantwell, and M. E. J. Newman (2019). Balance in signed networks, *Physical Review E*, vol. 99, p. 012320. 5, 19

N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila (2010). Towards rich mobile phone datasets: Lausanne data collection campaign, *International Conference on Pervasive Services (ICPS)*. 45

M. Kivelä, A. Arenas, M. Barthelemy, J. P. Gleeson, Y. Moreno, and M. A. Porter (2014). Multilayer networks, *Journal of Complex Networks*, vol. 2(3), pp. 203–271. 5, 6, 45

M. Kleindessner, S. Samadi, P. Awasthi, and J. Morgenstern (2019). Guarantees for Spectral Clustering with Fairness Constraints, in *International Conference on Machine Learning (ICML) 2019*. 83

A. Knyazev (2018). On spectral partitioning of signed graphs, in *SIAM Workshop on Combinatorial Scientific Computing 2018*. 5, 19

M. Koptelov, A. Zimmermann, B. Crémilleux, and L. Soualmia (2020). Link Prediction via Community Detection in Bipartite Multi-Layer Graphs, in *Proceedings of the 35th Annual ACM Symposium on Applied Computing 2020*. 5

D. Koutra, T.-Y. Ke, U. Kang, D. H. P. Chau, H.-K. K. Pao, and C. Faloutsos (2011). Unifying Guilt-by-Association Approaches: Theorems and Fast Algorithms, in *Machine Learning and Knowledge Discovery in Databases 2011*. 7, 63

A. Kumar and H. D. III (2011). A Co-training Approach for Multi-view Spectral Clustering, in *International Conference on Machine Learning (ICML) 2011*. 6, 45, 46

A. Kumar, P. Rai, and H. Daume (2011). Co-regularized Multi-view Spectral Clustering, in *Neural Information Processing Systems (NeurIPS) 2011*. 6, 45, 46, 59

S. Kumar, F. Spezzano, V. Subrahmanian, and C. Faloutsos (2016). Edge Weight Prediction in Weighted Signed Networks, in *International Conference on Data Mining (ICDM) 2016*. 3, 18

J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E. Luca, and S. Albayrak (2010). Spectral analysis of signed graphs for clustering, prediction and visualization, in *International Conference on Data Mining (ICDM) 2010*. 4, 19, 21, 46

C. M. Le, E. Levina, and R. Vershynin (2017). Concentration and regularization of random graphs, *Random Structures & Algorithms*, vol. 51(3), pp. 538–561. 35

J. Lei and A. Rinaldo (2015). Consistency of spectral clustering in stochastic block models, *Annals of Statistics*, vol. 43(1), pp. 215–237. 35

J. Leskovec, D. Huttenlocher, and J. Kleinberg (2010a). Predicting positive and negative links in online social networks, in *International World Wide Web Conference (WWW) 2010*. 4, 18

J. Leskovec, D. Huttenlocher, and J. Kleinberg (2010b). Signed Networks in Social Media, in *Conference on Human Factors in Computing Systems (CHI) 2010*. 18

J. Leskovec and A. Krevl (2014). *SNAP Datasets: Stanford Large Network Dataset Collection*, *http://snap.stanford.edu/data*. 38

Y. Lim and M. Pálfia (2012). Matrix power means and the Karcher mean, *Journal of Functional Analysis*, vol. 262, pp. 1498–1514. 17

J. Liu, C. Wang, J. Gao, and J. Han (2013). Multi-view clustering via joint nonnegative matrix factorization, in *SIAM International Conference on Data Mining (SDM) 2013*. 59, 75

S. Liu (2015). Multi-way dual Cheeger constants and spectral bounds of graphs, *Advances in Mathematics*, vol. 268, pp. 306 – 338. 20

S. P. Lloyd (1982). Least squares quantization in pcm, *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137. 14

Y. D. Lockerman, B. Sauvage, R. Allègre, J.-M. Dischler, J. Dorsey, and H. Rushmeier (2016). Multi-Scale Label-Map Extraction for Texture Synthesis, *ACM Transactions on Graphics*, vol. 35(4). 5

M. Lovric (Ed.) (2011). *International Encyclopedia of Statistical Science*, Springer Berlin Heidelberg, Berlin, Heidelberg. 15

Q. Lu and L. Getoor (2003). Link-based Classification, in *International Conference on Machine Learning (ICML) 2003*. 59, 75

H. Lütkepohl (1996). *Handbook of Matrices*, Wiley. 12

J. MacQueen (1967). Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics 1967*. 13

M. Mahajan, P. Nimbhorkar, and K. Varadarajan (2012). The planar k-means problem is NP-hard, *Theoretical Computer Science*, vol. 442, pp. 13 – 21. 14

A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore (2000). Automating the construction of internet portals with machine learning, *Information Retrieval*, vol. 3(2), pp. 127–163. 59, 75

P. Mercado, J. Bosch, and M. Stoll (2019a). Node Classification for Signed Social Networks Using Diffuse Interface Methods, in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD) 2019*. 4, 18

P. Mercado, A. Gautier, F. Tudisco, and M. Hein (2018). The Power Mean Laplacian for Multilayer Graph Clustering, in *Artificial Intelligence and Statistics (AISTATS) 2018*. 10, 22, 35

P. Mercado, F. Tudisco, and M. Hein (2016). Clustering Signed Networks with the Geometric Mean of Laplacians, in *Neural Information Processing Systems (NeurIPS) 2016*. 4, 17, 19, 20, 21, 22, 28, 29, 39

P. Mercado, F. Tudisco, and M. Hein (2019b). Generalized Matrix Means for Semi-Supervised Learning with Multilayer Graphs, in *Neural Information Processing Systems (NeurIPS) 2019*. 10

P. Mercado, F. Tudisco, and M. Hein (2019c). Spectral Clustering of Signed Graphs via Matrix Power Means, in *International Conference on Machine Learning (ICML) 2019*. 10

E. Mossel and J. Xu (2016). Local Algorithms for Block Models with Side Information, in *ACM Conference on Innovations in Theoretical Computer Science (ITCS) 2016*. 65

S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. D. Morris (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function, *Genome Biology*, vol. 9, pp. S4 – S4. 7

P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela (2010). Community Structure in Time-Dependent, Multiscale, and Multiplex Networks, *Science*, vol. 328(5980), pp. 876–878. 6, 45, 47

J. Newling and F. Fleuret (2017). K-Medoids For K-Means Seeding, in *Neural Information Processing Systems (NeurIPS) 2017*. 15

M. E. J. Newman (2006). Modularity and community structure in networks, *Proceedings of the National Academy of Sciences*, vol. 103(23), pp. 8577–8582. 6, 45

M. E. J. Newman and M. Girvan (2004). Finding and evaluating community structure in networks, *Physical Review E*, vol. 69, p. 026113. 6

F. Nie, J. Li, and X. Li (2016). Parameter-free Auto-weighted Multiple Graph Learning: A Framework for Multiview Clustering and Semi-supervised Classification, in *International Joint Conferences on Artificial Intelligence (IJCAI) 2016*. 7, 61, 63, 74, 75

S. Paul and Y. Chen (2016a). Consistent community detection in multi-relational data through restricted multi-layer stochastic blockmodel, *Electronic Journal of Statistics*, vol. 10(2), pp. 3807–3870. 6, 45

S. Paul and Y. Chen (2016b). Null Models and Modularity Based Community Detection in Multi-Layer Networks, *arXiv:1608.00623*. 6, 45, 47

S. Paul and Y. Chen (2020). Spectral and matrix factorization methods for consistent community detection in multi-layer networks, *Annals of Statistics*, vol. 48(1), pp. 230–250. 7, 35, 47

N. G. Pavlidis, V. P. Plagianakos, D. K. Tasoulis, and M. N. Vrahatis (2006). Financial forecasting through unsupervised clustering and neural networks, *Operational Research*, vol. 6(2), pp. 103–127. 18

T. P. Peixoto (2015). Inferring the mesoscale structure of layered, edge-valued, and time-varying networks, *Physical Review E*, vol. 92, p. 042807. 6, 45, 47

T. P. Peixoto (2019). *Bayesian Stochastic Blockmodeling*, chapter 11, pp. 289–332, John Wiley & Sons, Ltd. 6, 47

D. Petz (2007). *Quantum Information Theory and Quantum Statistics*, Springer Berlin Heidelberg. 17

T. Qin and K. Rohe (2013). Regularized Spectral Clustering under the Degree-Corrected Stochastic Blockmodel, in *Neural Information Processing Systems (NeurIPS) 2013*. 35

N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos (2010). A new approach to cross-modal multimedia retrieval, in *ACM Multimedia 2010*. 59, 75

M. Rocklin and A. Pinar (2011). Latent Clustering on Graphs with Multiple Edge Types, in *Algorithms and Models for the Web Graph 2011*. 6

K. Rohe, S. Chatterjee, B. Yu, *et al.* (2011). Spectral clustering and the high-dimensional stochastic blockmodel, *The Annals of Statistics*, vol. 39(4), pp. 1878–1915. 22, 35, 49, 94

Y. Saad (2011). *Numerical Methods for Large Eigenvalue Problems*, SIAM. 58

Y. Saad and M. H. Schultz (1986). GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM Journal on Scientific and Statistical Computing*, vol. 7(3), pp. 856–869. 72

A. Saade, F. Krzakala, M. Lelarge, and L. Zdeborová (2018). Fast Randomized Semi-Supervised Clustering, *Journal of Physics: Conference Series*, vol. 1036, p. 012015. 65

A. Saade, F. Krzakala, and L. Zdeborová (2014). Spectral Clustering of graphs with the Bethe Hessian, in *Neural Information Processing Systems (NeurIPS) 2014*. 5, 21

A. Saade, M. Lelarge, F. Krzakala, and L. Zdeborová (2015). Spectral detection in the censored block model, in *2015 IEEE International Symposium on Information Theory (ISIT) 2015*. 5, 21, 33, 34, 38

R. J. Sánchez-García, E. Cozzo, and Y. Moreno (2014). Dimensionality reduction and spectral properties of multilayer networks, *Physical Review E*, vol. 89, p. 052815. 5

P. Sarkar and P. J. Bickel (2015). Role of normalization in spectral clustering for stochastic blockmodels, *The Annals of Statistics*, vol. 43(3), pp. 962–990. 35

A. Schein, J. Paisley, D. M. Blei, and H. Wallach (2015). Bayesian Poisson Tensor Factorization for Inferring Multilateral Relations from Sparse Dyadic Event Counts, in *Conference on Knowledge Discovery and Data Mining (KDD) 2015*. 6, 45, 47

A. Schein, M. Zhou, D. Blei, and H. Wallach (2016). Bayesian Poisson Tucker Decomposition for Learning the Structure of International Relations, in *International Conference on Machine Learning (ICML) 2016*. 6, 45, 47

J. Sedoc, J. Gallier, D. Foster, and L. Ungar (2017). Semantic Word Clusters Using Signed Spectral Clustering, in *Association for Computational Linguistics (ACL) 2017*. 5, 19, 45

M. Shahriari and M. Jalili (2014). Ranking Nodes in Signed Social Networks, *Social Network Analysis and Mining*, vol. 4(1), p. 172. 19

L. Solá, M. Romance, R. Criado, J. Flores, A. García del Amo, and S. Boccaletti (2013). Eigenvector centrality of nodes in multiplex networks, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 23(3), p. 033131. 5

S. Sra and R. Hosseini (2016). Geometric optimization in machine learning, in *Algorithmic Advances in Riemannian Geometry and Applications 2016*, pp. 73–91, Springer. 17

N. Stanley, S. Shai, D. Taylor, and P. J. Mucha (2016). Clustering Network Layers with the Strata Multilayer Stochastic Block Model, *IEEE Transactions on Network Science and Engineering*, vol. 3(2), pp. 95–105. 45

A. Subramanya and P. P. Talukdar (2014). *Graph-Based Semi-Supervised Learning*, Morgan & Claypool Publishers. 61

S. Sun (2013). A survey of multi-view machine learning, *Neural Computing and Applications*, vol. 23(7), pp. 2031–2038. 3, 7, 45

A. Takatsu (2011). Wasserstein geometry of Gaussian measures, *Osaka Journal of Mathematics*, vol. 48(4), pp. 1005–1026. 83

J. Tang, C. Aggarwal, and H. Liu (2016a). Node Classification in Signed Social Networks, in *SIAM International Conference on Data Mining (SDM) 2016*. 4, 18

J. Tang, Y. Chang, C. Aggarwal, and H. Liu (2016b). A Survey of Signed Network Mining in Social Media, *ACM Computing Surveys*, vol. 49(3), pp. 42:1–42:37. 5, 19

W. Tang, Z. Lu, and I. S. Dhillon (2009). Clustering with Multiple Graphs, in *International Conference on Data Mining (ICDM) 2009*. 6, 45, 47

D. Taylor, R. S. Caceres, and P. J. Mucha (2017). Super-Resolution Community Detection for Layer-Aggregated Multilayer Networks, *Physical Review X*, vol. 7, p. 031056. 5, 6, 45, 46

D. Taylor, S. Shai, N. Stanley, and P. J. Mucha (2016). Enhanced Detectability of Community Structure in Multilayer Networks through Layer Aggregation, *Physical Review Letters*, vol. 116, p. 228301. 5, 45

V. Titouan, N. Courty, R. Tavenard, C. Laetitia, and R. Flamary (2019). Optimal Transport for structured data with application on graphs, in *International Conference on Machine Learning (ICML) 2019*. 83

J. A. Tropp (2015). An Introduction to Matrix Concentration Inequalities, *Foundations and Trends® in Machine Learning*, vol. 8(1-2), pp. 1–230. 88, 89

K. Tsuda, H. Shin, and B. Schölkopf (2005). Fast Protein Classification with Multiple Networks, *Bioinformatics*, vol. 21(2), pp. 59–65. 7, 61, 62, 63, 74, 75

F. Tudisco, F. Arrigo, and A. Gautier (2018). Node and Layer Eigenvector Centralities for Multiplex Networks, *SIAM Journal on Applied Mathematics*, vol. 78(2), pp. 853–876. 5, 83

F. Tudisco, V. Cardinali, and C. Fiore (2015). On complex power nonnegative matrices, *Linear Algebra and its Applications*, vol. 471, pp. 449–468. 101

United Nations Development Programme (1997). *Human Development Report 1997*, Human Development Report, Oxford University Press. 15, 16

C. Villani (2008). *Optimal Transport: Old and New*, Grundlehren der mathematischen Wissenschaften, Springer Berlin Heidelberg. 83

K. Viswanathan, S. Sachdeva, A. Tomkins, and S. Ravi (2019). Improved Semi-Supervised Learning with Multiple Graphs, in *Artificial Intelligence and Statistics (AISTATS) 2019*. 7, 61, 63

U. von Luxburg (2007). A tutorial on spectral clustering, *Statistics and Computing*, vol. 17(4), pp. 395–416. 11, 48

S. Wang, J. Tang, C. Aggarwal, Y. Chang, and H. Liu (2017). Signed Network Embedding in Social Media, in *SIAM International Conference on Data Mining (SDM) 2017*. 4, 19

J. H. Wilkinson (1965). *The algebraic eigenvalue problem*, vol. 662, Oxford Clarendon. 74

J. D. Wilson, J. Palowitch, S. Bhamidi, and A. B. Nobel (2017). Community Extraction in Multilayer Networks with Heterogeneous Community Structure, *Journal of Machine Learning Research*, vol. 18(149), pp. 1–49. 6, 45, 47

F. Wu, A. Souza, T. Zhang, C. Fifty, T. Yu, and K. Weinberger (2019). Simplifying Graph Convolutional Networks, in *International Conference on Machine Learning (ICML) 2019*. 7

Z. Wu, S. Pan, G. Long, J. Jiang, X. Chang, and C. Zhang (2020). Connecting the Dots: Multivariate Time Series Forecasting with Graph Neural Networks, *Conference on Knowledge Discovery and Data Mining (KDD)*. 83

R. Xia, Y. Pan, L. Du, and J. Yin (2014). Robust Multi-View Spectral Clustering via Low-Rank and Sparse Decomposition., in *AAAI Conference on Artificial Intelligence 2014*. 6, 59

C. Xu, D. Tao, and C. Xu (2013). A Survey on Multi-view Learning, *arXiv:1304.5634*. 3, 7, 45

J. Xu, L. Massoulié, and M. Lelarge (2014). Edge Label Inference in Generalized Stochastic Block Models: from Spectral Theory to Impossibility Results, in *Conference on Learning Theory (COLT) 2014*. 35, 49

M. Xu, V. Jog, and P.-L. Loh (2020). Optimal rates for community estimation in the weighted stochastic block model, *Annals of Statistics*, vol. 48(1), pp. 183–204. 35, 49

C. Yang, Y. Xiao, Y. Zhang, Y. Sun, and J. Han (2020). Heterogeneous Network Representation Learning: Survey, Benchmark, Evaluation, and Beyond, *arXiv:2004.00216*. 5

Z. Yang, W. W. Cohen, and R. Salakhutdinov (2016). Revisiting Semi-supervised Learning with Graph Embeddings, in *International Conference on Machine Learning (ICML) 2016*. 7, 61

J. Ye and L. Akoglu (2018). Robust Semi-Supervised Learning on Multiple Networks with Noise, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2018*. 61

Y. Yu, T. Wang, and R. J. Samworth (2015). A useful variant of the Davis-Kahan theorem for statisticians, *Biometrika*, vol. 102(2), pp. 315–323. 98

S. Yuan, X. Wu, and Y. Xiang (2017). SNE: Signed Network Embedding, in *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD) 2017*. 19

S.-Y. Yun and A. Proutiere (2016). Optimal Cluster Recovery in the Labeled Stochastic Block Model, in *Neural Information Processing Systems (NeurIPS) 2016*. 35, 49

P. Zadeh, R. Hosseini, and S. Sra (2016). Geometric mean metric learning, in *International Conference on Machine Learning (ICML) 2016*. 17

H. Zhang, L. Qiu, L. Yi, and Y. Song (2018). Scalable Multiplex Network Embedding, in *International Joint Conference on Artificial Intelligence (IJCAI) 2018*. 5

H. Zhao, Z. Ding, and Y. Fu (2017a). Multi-View Clustering via Deep Matrix Factorization., in *AAAI Conference on Artificial Intelligence 2017*. 6, 45

J. Zhao, X. Xie, X. Xu, and S. Sun (2017b). Multi-view learning overview: Recent progress and new challenges, *Information Fusion*, vol. 38, pp. 43 – 54. 7, 45

D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf (2003). Learning with Local and Global Consistency, in *Neural Information Processing Systems (NeurIPS) 2003*. 7, 61, 62, 64

D. Zhou and C. J. Burges (2007). Spectral clustering and transductive learning with multiple views, in *International Conference on Machine Learning (ICML) 2007*. 6, 7, 46, 47, 59, 61, 63, 75

X. Zhu and Z. Ghahramani (2002). Learning from labeled and unlabeled data with label propagation, Technical report. 68

X. Zhu, Z. Ghahramani, and J. Lafferty (2003). Semi-supervised Learning Using Gaussian Fields and Harmonic Functions, in *International Conference on Machine Learning (ICML) 2003*. 7, 61, 62, 68

H. Ziegler, M. Jenny, T. Gruse, and D. A. Keim (2010). Visual market sector analysis for financial time series data, in *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on 2010*. 18

PUBLICATIONS

6. P. Mercado, F. Tudisco, and M. Hein. Generalized Matrix Means for Semi-Supervised Learning with Multilayer Graphs . In Advances in Neural Information Processing Systems (**NeurIPS**) (2019)

5. P. Mercado, J. Bosch, and M. Stoll. Node Classification for Signed Social Networks Using Diffuse Interface Methods. In European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (**ECML-PKDD**) (2019)

4. P. Mercado, F. Tudisco, and M. Hein. Spectral Clustering of Signed Graphs via Matrix Power Means. In Proceedings of the 36th International Conference on Machine Learning (**ICML**) (2019)

3. F. Tudisco, P. Mercado, and M. Hein. Community Detection in Networks via Nonlinear Modularity Eigenvectors. In SIAM Journal on Applied Mathematics, 78:2393–2419 (2018)

2. P. Mercado, A. Gautier, F. Tudisco, and M. Hein. The Power Mean Laplacian for Multilayer Graph Clustering. In Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics (**AISTATS**) (2018)

1. P. Mercado, F. Tudisco, and M. Hein. Clustering Signed Networks with the Geometric Mean of Laplacians. In Advances in Neural Information Processing Systems (**NeurIPS**) (2016)