Graph signal processing tailored for subgraph focus and community structure

Présentée le 11 mars 2022
Faculté des sciences et techniques de l'ingénieur
Laboratoire de traitement d'images médicales
Programme doctoral en génie électrique
pour l'obtention du grade de Docteur ès Sciences
par

Miljan PETROVIC

Acceptée sur proposition du jury
Prof. J.-Ph. Thiran, président du jury
Prof. D. N. A. Van De Ville, Dr R. P. Liégeois, directeurs de thèse
Prof. S. Barbarossa, rapporteur
Prof. S. Achard, rapporteuse
Prof. P. Frossard, rapporteur
"I suppose – I suppose I keep telling myself never to take anything for granted."

— Archibald Joseph Cronin, *The Citadel*

To my family, and groundhogs-the-pilgrims and other friends.
Acknowledgements

My quest through PhD studies wouldn’t have been possible if there hadn’t been for Professor Dimitri Van De Ville who graciously decided to give me a chance to come to Switzerland and join the EPFL research community under his supervision in the first place. Foremost, I owe him huge gratitude for the guidance throughout all the four years of work, and for always keeping me on track with the big picture of scientific research. No less is deserving for the same my thesis co-director Dr Raphaël Liégeois, whose help on the hands-on matters of research was invaluable.

I would also like to thank other coauthors on the publications that make this thesis – Maria Giulia Preti and Thomas Bolton – for their tremendous help during our collaboration. A great deal of gratitude also goes to all members of MIPLab who made the working experience at Campus Biotech impeccable, with special thanks to Ms. Manuela da Silva for handling the burden of administration. I am forever indebted to Younes Farouj, Serafeim Loukas, and Nicolas Gninenko. The brainstorming sessions with these friends shaped many aspects of my research over the years, and helped overcome many cul-de-sacs. With this in mind, credit also goes to the atmospheres of the great establishments that are the La Petite Reine, L’Apothicaire, and Le Phare.

The support of my family and friends in Serbia made both moving abroad and treading the path of many research challenges so much easier and enjoyable, even when homesickness sets in – especially the support of Nina Đenić. Thank you father Branišlav, mother Sonja, and brother Dušan for making this thesis possible, and for imbuing me with confidence and assertivness whenever I needed them.

Geneva, March, 2022

M. P.
Abstract

Community structure in graph-modeled data appears in a range of disciplines that comprise network science. Its importance relies on the influence it bears on other properties of graphs such as resilience, or prediction of missing connections. Nevertheless, research to date seems to overlook its effect on the properties of signals defined in the domain of graphs' vertices. Indeed, the framework of graph signal processing mainly focuses on local connectivity patterns reflected by the graph Laplacian, and the Laplacian's effect on signals as the graph equivalent of a differential operator. This dissertation investigates the aforementioned interplay between graph signals and the underlying community structure. We make an effort to answer questions like – do signal's values align and in what way with the communities?; does localization of signal's energy favors a community as the vertex support? Answers to these questions should provide a clearer perspective on the relation between graph signals and networks at the level of subgraphs.

This dissertation consists of two main parts. First, we investigate a particular approach – based on modularity matrix – of informing the graph Fourier transform about the communities without explicitly detecting them. Thereof, we show that the derived community-aware operators on signals, such as filtering or subsampling, provide a complementary view on the framework to the conventional one based on the Laplacian. Indeed, reduced signal variability within a community seems to be a valuable metric of important signal behavior, aside from its smoothness. Secondly, we explore the intricacies of a broader definition of a community – a subgraph of any special interest, possibly identified through metadata on vertices instead of the underlying edge connectivity. Within this context, the goal is to understand the benefits of processing signals in a subgraph-restricted way. We design a new, more computationally stable type of Slepians – bandlimited signals with energy concentrated on a subgraph. Consequently, we show that such bandlimited vectors can be successfully employed to identify a signal's oscillatory pattern localized in a subgraph of interest, by means of a modified filtering procedure. Findings from both lines of research confirm the need and benefits of a better understanding of the interaction between communities and graph signals.

Keywords: Networks, community structure, graph signal processing, graph Slepians, modularity matrix.
Résumé

La structure de la communauté dans les données modélisées par des graphes apparaît dans une gamme de disciplines qui comprennent la science des réseaux. Son importance repose sur l’influence qu’elle exerce sur d’autres propriétés des graphes telles que la résilience ou la prédiction des connexions manquantes. Néanmoins, les recherches à ce jour semblent ignorer son effet sur les propriétés des signaux définis dans le domaine des sommets des graphes. En effet, le traitement du signal sur graphes se concentre principalement sur des modèles de connectivité locale reflétés par le Laplacien graphique, et l’effet du Laplacien sur les signaux en tant qu’équivalent graphique d’un opérateur différentiel. Cette thèse étudie l’interaction susmentionnée entre les signaux graphiques et la structure de la communauté sous-jacente. Nous nous efforçons de répondre à des questions telles que : les valeurs du signal s’alignent-elles et de quelle manière avec les communautés ; la localisation de l’énergie du signal favorise-t-elle une communauté comme support de vertex ? Les réponses à ces questions devraient fournir une perspective plus claire sur la relation entre les signaux de graphes et les réseaux au niveau des sous-graphes.

Cette thèse se compose de deux parties principales. Premièrement, nous étudions une approche particulière - basée sur la matrice de modularité - pour informer la transformée de Fourier du graphe sur les communautés sans les détecter explicitement. A partir de celle-ci, nous montrons que les opérateurs communautaires dérivés sur les signaux, tels que le filtrage ou le sous-échantillonnage, fournissent une vue complémentaire dans ce cadre par rapport au cadre conventionnel basé sur le Laplacien. En effet, la variabilité réduite du signal au sein d’une communauté semble être une mesure précieuse du comportement important du signal, en dehors de sa régularité. Deuxièmement, nous explorons les subtilités d’une définition plus large d’une communauté - un sous-graphe d’un intérêt particulier, éventuellement identifié par des métadonnées sur des sommets au lieu de la connectivité de bord sous-jacente. Dans ce contexte, l’objectif est de comprendre les avantages du traitement des signaux de manière restreinte aux sous-graphes. Nous concevons un nouveau type de Slepians - des signaux à bande limitée avec une énergie concentrée sur un sous-graphe - plus stable sur le plan computationnel. Par conséquent, nous montrons que de tels vecteurs à bande limitée peuvent être utilisés avec succès pour identifier le modèle oscillatoire d’un signal localisé dans un sous-graphe d’intérêt, au moyen d’une procédure de filtrage modifiée. Les résultats des deux axes de recherche confirment le besoin et les avantages d’une meilleure compréhension de l’interaction entre les communautés et les signaux graphiques.
Résumé

Mots-clés: Les réseaux, la structure de la communauté, le traitement du signal sur graphes, Slepians sur graphes, la matrice de modularité.


## Contents

<table>
<thead>
<tr>
<th>Acknowledgements</th>
<th>i</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract (English &amp; Français)</td>
<td>iii</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Dissertation organization and main contributions</td>
<td>2</td>
</tr>
<tr>
<td>2 Background</td>
<td>9</td>
</tr>
<tr>
<td>2.1 Data and graphs</td>
<td>9</td>
</tr>
<tr>
<td>2.2 Networks and communities</td>
<td>10</td>
</tr>
<tr>
<td>2.2.1 What is a community?</td>
<td>11</td>
</tr>
<tr>
<td>2.2.2 Community detection</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Graph signal processing</td>
<td>13</td>
</tr>
<tr>
<td>2.4 Uncertainty principle and energy localization</td>
<td>15</td>
</tr>
<tr>
<td>3 Informing signal processing techniques with community structure</td>
<td>19</td>
</tr>
<tr>
<td>3.1 Journal article: Community-aware graph signal processing: modularity defines new ways of processing graph signals</td>
<td>20</td>
</tr>
<tr>
<td>3.1.1 Introduction</td>
<td>20</td>
</tr>
<tr>
<td>3.1.2 Graph signal processing</td>
<td>21</td>
</tr>
<tr>
<td>3.1.3 Community structure</td>
<td>23</td>
</tr>
<tr>
<td>3.1.4 Community-aware graph signal processing</td>
<td>24</td>
</tr>
<tr>
<td>3.1.5 Validation for neuroimaging</td>
<td>32</td>
</tr>
<tr>
<td>3.1.6 Conclusion &amp; outlook</td>
<td>32</td>
</tr>
<tr>
<td>3.2 Journal preprint: Graph uncertainty principle revisited: the importance of community structure</td>
<td>35</td>
</tr>
<tr>
<td>3.2.1 Introduction</td>
<td>35</td>
</tr>
<tr>
<td>3.2.2 A primer on graph signal processing and community structure</td>
<td>37</td>
</tr>
<tr>
<td>3.2.3 Localization of graph signals</td>
<td>40</td>
</tr>
<tr>
<td>3.2.4 Analysis of random graphs</td>
<td>48</td>
</tr>
<tr>
<td>3.2.5 Surrogate analysis of a publications graph</td>
<td>54</td>
</tr>
<tr>
<td>3.2.6 Conclusion</td>
<td>57</td>
</tr>
</tbody>
</table>
## Contents

4 Processing global signals in local environment 59  
4.1 Journal article: Guided graph spectral embedding: application to the *c. elegans* connectome .......................... 60  
4.1.1 Introduction ................................................. 60  
4.1.2 Methods .................................................... 62  
4.1.3 Mathematical considerations ............................. 67  
4.1.4 Results ..................................................... 70  
4.1.5 Discussion ............................................... 76  
4.2 Conference proceedings article: Slepian guided filtering of graph signals .......................... 82  
4.2.1 Introduction ................................................. 82  
4.2.2 Background ................................................. 83  
4.2.3 Slepian guided graph signal filtering ..................... 85  
4.2.4 Power spectrum density estimation on random graphs ........................................ 87  
4.2.5 Guided filtering of brain signals .......................... 90  
4.2.6 Conclusions and future work ............................. 96  

5 Summary and future perspectives 99  
5.1 Summary of findings .......................................... 99  
5.2 Perspectives of future work .................................. 100  

A Supplementary material for Chapter 3 103  
A.1 Supplementary material for Section 3.2 ...................... 103  

B Supplementary material for Chapter 4 105  
B.1 Supplementary material for Section 4.1 ...................... 105  

Bibliography 115  
Curriculum Vitae 131
1 Introduction

1.1 Motivation

Graph-modeled data is known to exhibit community structure – the presence of prominently more or less edges between nodes in a subgraph as opposed to those that run between nodes belonging to different subgraphs (Fortunato and Hric, 2016). Such a phenomenon appears in many disparate datasets like social networks (Jeub et al., 2015), transportation (Guimera et al., 2005), metabolic (Guimerà and Amaral, 2005) or neural networks (Meunier et al., 2010). The importance of communities arises from their relation to other graph properties explored in network science. In the presence of community structure – resilience of a complex system is improved by affecting the phase transition (Dong et al., 2018); immunization strategies are allowed to target a small number of influential spreaders in epidemics networks (Cherifi et al., 2019); prediction of missing edges in partly known graphs is possible, like in the cases of a network of associations between terrorists or a food web (Clauset et al., 2008). A large body of work is thus dedicated to the detection of communities (Mukerjee, 2021; Newman, 2006; Nicolini et al., 2017; Lu et al., 2018).

Opposite to the mesoscale graph topology reflected in community structure, stands the extremely local connectivity pattern (edge between two nodes) encoded globally over the whole graph in the Laplacian matrix (Chung, 1997). The Laplacian thus provides support for graph signal processing (GSP) framework (Shuman et al., 2013), by expressing the graph equivalent of a second-order differential operator for signals associated with nodes. Thereupon, concepts regarding processing discrete-time series get extended into the domain of signals on graphs. The framework encompasses the graph Fourier transform and allows filtering (Shuman et al., 2013; Ortega et al., 2018), stationarity analysis (Marques et al., 2017), wavelet transforms (Hammond et al., 2011), formulation of uncertainty principle (Tsitsvero et al., 2016), etc.

However, signal operators defined in this manner only account for the edge connectivity pattern from the perspective of graph Laplacian. To the authors’ knowledge, at the time of publication of results in this thesis, there was no attempt to incorporate the edge connectedness at
Chapter 1. Introduction

the level of community structure into the signal processing techniques. Given the importance and presence of community structure in graph-modeled data, there is the need to acquire such a perspective on the processing framework that would consider how the communities do or do not constrain signal behavior. This thesis acknowledges that perspective with two relevant lines of research.

The first line of inquiry builds upon a way to include information on communities into the graph Fourier transform and consequently other operators. The key goal here is to achieve this without performing community detection before signal processing, thus simplifying the approach. With this information accessible – here by means of modularity matrix (Newman, 2006) – there remains to be explored what can we learn about the signals and their relation to community structure, that is otherwise not attainable in the conventional GSP framework.

The other research direction deals with the case when a community or any subgraph of particular interest is included in the prior knowledge. We aim to understand the benefits of processing signals in a restricted way, by leveraging the concepts of energy concentration on the finite vertex support (a subgraph) and bandlimitedness. Graph Slepians (Van De Ville et al., 2017) are the main tool we use to confirm whether signals display relevant localized behavior, and to elucidate the effect a subgraph or a community can have on the properties of graph signals.

1.2 Dissertation organization and main contributions

The present doctoral thesis provides novel insights into ways how community structure of networks affects properties of graph signals, and how those signals can be processed by accounting for the underlying communities. Another objective of this dissertation is to dig into the relationship between the local connectivity patterns and the global topology of the graph, with an emphasis on how this relationship affects properties of signals, such as bandlimitedness and finite vertex support. The main body of the thesis represents a compilation of three published manuscripts, and one manuscript in preparation. Chapter 2 serves as the background for the processing framework that is extended and explored in the latter chapters. The manuscripts containing the main results of the work are encompassed by Chapters 3 and 4\(^1\). Finally, Chapter 65 gives a conclusion and outlines future perspectives of the performed research.

The remainder of this section gives a summary of the main research questions addressed in the chapters of the dissertation and the scientific publications that they include. Within the preparation of all included publications, I have been involved in the development of the mathematical framework, data processing, interpretation of the results, and writing and proof-reading the manuscript. I would also like to acknowledge the contribution of my coauthors to these publications since the research presented in the thesis came from a close collaboration.

\(^1\)Some of the chapters’ sections rephrase specific sentences and include corrections of typos, which renders them slightly different from the published versions of the manuscripts.
1.2. Dissertation organization and main contributions

Motivation

- Importance of community structure to network resilience, edge prediction, etc.
- Graph signal processing (GSP) conventionally does not consider the effect communities might have on the signal behaviour.
- Global connectivity patterns incorporated in GSP framework might not be apt for accessing local properties of graph signals.

Contributions

Chapter 3: Informing signal processing techniques with community structure


- Modularity matrix stands as the shift operator.
- Components of the Fourier domain reflect reduced signal variability within a community in contrast to global Laplacian smoothness.
- Generalization of filtering, denoising, surrogate generation, and optimal sampling and reconstruction to the community-aware framework.


- Signal energy localization on a community vs. random subset of nodes.
- Bandlimitedness of smooth vs. modular signals.
- Effect of type and strength of community structure on signal localization.

Chapter 4: Processing global signals in local environment


- Slepian criterion combining high energy concentration and low modified embedded distance.
- A basis of Slepian signals without degenerate subspaces.


- Capturing local signal changes in mean and variance by graph Slepians vs. eigenvectors of Laplacian.
- Filtering signals by altering contributions of localized spectral components.

Figure 1.1 – Thesis overview: Main contributions to the frameworks of graph signal processing and network science.
Chapter 1. Introduction

Chapter 3: Informing signal processing techniques with community structure

The shift operator, a matrix description of a graph conventionally set to be the adjacency or Laplacian, is the centerpiece of all operators on graph signals. Its eigendecomposition yields the spectral components of the graph Fourier domain. A straightforward way to incorporate the knowledge about graph community structure into GSP operators is to define spectral components as orthogonal signals that align with communities. Such signals are for example the eigenvectors of the modularity matrix, commonly used for community detection (Newman, 2006). Therefore, setting the modularity matrix as the shift operator provides all signal processing techniques with knowledge about communities without explicitly detecting them. Though modularity is not without limitations and there are numerous other ways of accessing community structure (Fortunato and Hric, 2016; Nicolini et al., 2017), it is the one available in an explicit matrix form needed for the GSP framework. Other criteria for community structure that do not include a notion of an expected edge between two nodes (given by an entry of the modularity matrix) would have to be incorporated into signal processing techniques in a different way than replacing the shift operator. These criteria go beyond the scope of this thesis and included publications, but are briefly discussed in Chapter 5. Two following publications explore different community-aware signal operators, and the uncertainty principle and signal localization for the case of the community-aware Fourier domain, respectively.

Section 3.1: Community-aware graph signal processing: modularity defines new ways of processing graph signals (Journal article)


• Q1 What information about graph signals can be accessed exclusively by the signal’s projection onto a spectral basis that is a proxy of community structure?

• Q2 What is the nature of oscillations in the eigenvectors of modularity matrix? What is the equivalent of Laplacian notion of smoothness if the Fourier spectral basis is the set of eigenvectors of modularity matrix?

• Q3 How do graph signal operators change under the assumption of the community structure-based Fourier domain? What, if any, are the computational challenges imposed by the modularity matrix?

In this article, the community structure of a graph is reflected by the eigenvectors of the modularity matrix. These are used to define the Fourier transform of graph signals, and thus inform the spectral domain of the present communities without explicitly detecting them. By considering modularity instead of the Laplacian matrix as the key graph description, the notion of signal’s global smoothness is replaced by the reduced signal variability within a community
1.2. Dissertation organization and main contributions

(Q2). The article provides frameworks of filtering, optimal sampling and reconstruction, surrogate data generation, and denoising. An intriguing finding in the context of human brain graphs is that the alignment of signals with the anti-modular structure (communities with fewer edges within the group than outside) is relevant for predicting certain behavioral measures (Q1). Indeed, the approach of community-aware signal processing gives a unique perspective on graph signals. For example, on a transportation network, the optimal set of nodes for subsampling globally smooth signals turns out to include mostly peripheral nodes – those whose signal value is hard to predict. On the contrary, if signals that align with the community structure are considered, the nodes with the greatest predictive power tend to be picked out (Q1). This work demonstrates how operators on signals change upon consideration of communities, and discusses the few easily circumvented computational challenges of the modularity matrix – the density of entries, and non-definiteness (Q3).

Section 3.2: Graph uncertainty principle revisited: the importance of community structure (Journal article in preparation)

Petrovic M, and Van De Ville D. Graph uncertainty principle revisited: the importance of community structure, in preparation.

- Q1 How is the set of possible vertex-spectrum limited signals on a graph, affected by the choice of bandlimitedness within smooth or modular spectral components, in graphs with community structure?

- Q2 How is the set of possible vertex-spectrum limited signals on a graph, affected by the choice of nodal support of signals being a random subset of nodes or a community?

This article builds upon the one in Section 3.1 and compares the Fourier domains defined by eigenvectors of Laplacian and modularity matrix. It explores the set of possible bandlimited signals with high energy concentration on a subgraph when community structure is present in the graph. We relate the breadth of the set with certain eigengaps of the graph description matrices under the stochastic block model of communities. The main finding results from theoretical considerations and an analysis of a large ensemble of random graphs. Bandlimitedness of smooth signals tends to allow for high energy concentration on a random subset of nodes, whereas the bandlimitedness of modular signals allows higher energy localization on a community (Q1, Q2). Furthermore, if the disassortative structure is present (subgraphs with weaker nodal intraconnectons than connections with the rest of the graph), the best localization is reached when limiting the signal's spectrum to components corresponding to eigenvalues of the modularity matrix with the highest absolute values (Q1).
Chapter 4: Processing global signals in local environment

Besides inferring communities from the edge connectivity pattern of a graph, it is often useful to consider subgraphs derived from metadata available for nodes (e.g. age in social networks, or sensory/motor function in a cellular network). These subgraphs may correspond to communities, which is why they are sometimes taken as the ground truth in the evaluation of community detection algorithms. The correspondence is present only in certain cases, depending on the specific graph topology (Peel et al., 2017), so one needs to assume this with caution. Nevertheless, even if the subgraphs do not represent communities in the same sense of edge connectivity, they are still of great interest, especially for investigating the localization properties of graph signals. Indeed, graph Slepians (Van De Ville et al., 2017) – bandlimited signals with localized energy on a subgraph – provide crucial information about the interplay between the local and global topology of the graph. Within this context, the following publications concern both the design and computation of Slepians for the purpose of analyzing the graph, and the adoption of Slepians as a tool to process signals residing on graph nodes. The first publication generalizes graph Slepians to signals that are both well concentrated and smooth in a given subgraph and consequently applies them to the embedding of the graph’s nodes. The second introduces filtering of graph signals through a process of modifying their local spectral properties as reflected by the Slepians.

Section 4.1: Guided graph spectral embedding: application to the c. elegans connectome (Journal article)


- Q1 Is there a way to circumvent the computational challenge of degenerate solutions to Slepian criteria?
- Q2 How can one combine the criteria of energy concentration and modified embedded distance for the purpose of finding signals both smooth and well-concentrated on a subgraph?
- Q3 What is the practical application of such signals beyond the reach of standard graph Slepians?

This article introduces $\zeta$-Slepians with the criterion that combines smoothness, bandlimitedness, and the high energy concentration in a given subgraph. It shows that starting from the criteria for energy concentration and modified embedded distance, one can reach the solution to the desired signals as the eigenvectors of a simple expression in terms of the adjacency matrix (Q2). The matrix criterion provides a stable eigendecomposition without degenerate eigenspaces in the desired range of solution. This is a consequence of the matrix criterion
which, as compared to those of energy concentration and modified embedded distance, does not have such a clear upper bound on the rank (Q1). In this work, we use smooth concentrated Slepian signals to embed the nodes of a graph representing the neural system of *Caenorhabditis elegans*. The findings confirm known observations on the worm’s neural network by embeddings that align with concepts such as somatic positions of cells. However, certain unexpected details appear in some of the embeddings. For example, clustering of a single interneuron with the sensory ones reflected the need for directing the experimental research towards a more detailed biological analysis of that interneuron (Q3).

**Section 4.2: Slepian guided filtering of graph signals (Conference proceedings article)**


- **Q1** What are the specific local – on a subgraph – behavior patterns of a signal that cannot be identified by the global spectral oscillations such as the Laplacian eigenvectors?
- **Q2** What are the challenges of filtering a graph signal by modifying the coefficients of its projection onto graph Slepians?
- **Q3** What is the importance of Slepian-based locally filtered signals in the context of human brain networks?

In this publication, the filtering of graph signals is reformulated to windowing the coefficients given by the signal’s projection onto Slepian vectors. Hence, instead of removing or preserving global oscillatory patterns reflected in graph Laplacian, we manipulate those localized on a subgraph of interest. Indeed, we show on an ensemble of random graphs that localized increase of variance or mean of the signal leads to prominent changes in the graph spectral domain defined by the Slepians, whereas the global Laplacian spectrum remains of the same shape (Q1). This principle can be exploited to locally affect a signal's energy or smoothness, depending on the choice of the Slepian criterion. However, the proposed filtering is limited by the Slepian criterion's degenerate eigenspaces. The spectral window of the filter would have to impose the same coefficient at all components in an eigenspace for the filter to have an equivalent form in the vertex domain, i.e. form of a polynomial of the shift operator (Q2). This might be resolved by the combined Slepian criterion (Petrovic et al., 2019), which is discussed in Chapter 5. Here, we provide a proof-of-concept using a human brain graph built from diffusion-weighted imaging, and signals derived from functional magnetic resonance imaging. We compare filtering based on different Slepian criteria and the Laplacian filtering. The only approach that could identify the expected negative correlation between neural activity in certain brain regions and the task paradigm was filtering based on modified embedded distance Slepians (Q3).
2 Background

This chapter introduces the foundations of signals and data processing framework explored in the following Chapters. It provides the background for graph models, analysis of networks, and processing signals in the domain of graph nodes.

2.1 Data and graphs

Modeling data as graphs allows sets foundation for an analysis through mathematical tools within graph theory (Joyner and Melles, 2017). Graph’s nodes\(^1\) and edges can be inferred by considering the source of data. For example, the analysis of electronic health records is possible by constructing a temporal graph where nodes are events (treatments, laboratory tests, etc.) and edges reflect the temporal sequence of those events. Such a graph allows for an alternative disease diagnosis or risk prediction (Liu et al., 2015).

Similarly, there may be datasets of temporal measurements available. Such multivariate signals emerge for example from a set of sensor measurements (Stankovic et al., 2019). If temperature is measured at different spatial locations, then a conventional ordering of the sensors’ samples in a row for a single time point is not intuitive. Rather, sensors are seen as nodes of a graph, and the geographical distances between them are constructed into edges. Then, a joint analysis of signals in both time and graph vertices domain is possible (Shuman et al., 2016; Stanković et al., 2020). The notion of distance is not restricted to geographical meaning. Indeed, an edge can reflect a generalized distance (or similarity) between data points represented by nodes, such as in the graph of patients, i.e. the aforementioned electronic health records (Rotmensch et al., 2017).

In certain cases, the edge connectivity between data points is inherently present or intuitive, so there is no need to choose a distance metric suitable for modeling (Chen et al., 2009). Such is the case of social networks\(^2\) where nodes are people, and edges denote friendships or other in-

\(^1\)Though the term node is somewhat more common in the network science community, and the term vertex within the mathematics community, this dissertation uses the two interchangeably.

\(^2\)Similarly to the previous comment regarding nodes and vertices, the terms graph and network are used
Chapter 2. Background

The analysis of such graphs bears significance both in sociology and other disciplines like epidemics research (Eubank et al., 2004). Further intuitive definitions of graphs emerge in the context of human brain networks (Sporns, 2011). Here, nodes of the graph are usually regions of the brain, and edges may represent the number of anatomical neural paths between regions (Maier-Hein et al., 2017), or a certain quantifiable relation between neural activities across regions (coherence, correlation, etc.) (Preti et al., 2017).

Regardless of the meaning of nodes and edges, graph theory combined with tools from signal processing and machine learning allows exploring the phenomena observed in all or many modeled datasets. Study of the percolation on graphs – removal of nodes or edges – can give insight into the bottlenecks of transportation systems (Hamedmoghadam et al., 2021), help design a reliable wireless network (Franceschetti et al., 2007), or it may facilitate the simulation of epidemiological scenarios (Piraveenan et al., 2013). Certain metrics of both nodes and the graph as a whole – e.g. clustering coefficient, centralities, path lengths, and others – are useful in the prediction of relevant quantities and traits in neuroscience (Mattar et al., 2018) and biochemistry (Yang et al., 2020). For the work presented in this thesis, the crucial methodology is the assessment of community structure (Cherifi et al., 2019), which is relevant in neuroscience (Sporns and Betzel, 2016), biology (Sah et al., 2014), analysis of transportation networks (Guimera et al., 2005), etc. Details on communities and their detection are given in the following section of the thesis.

From a formal mathematical perspective, a graph \( G \) of size \( N \) is defined as a set of nodes, edges and weights denoted together by \( (V, E, W) \). The node set \( V \) comprises of \( N \) nodes indexed from 1 to \( N \). The edge set \( E \) contains tuples \((i, j)\) that form a subset of \( V \times V \), with \( i, j \) being the appropriate node indices. The mapping \( W : (i, j) \rightarrow \mathbb{R} \) associates every edge from \( E \) to a real scalar weight \( w_{i,j} \). A compact representation of a graph is its adjacency matrix \( A \) whose entry \( A_{i,j} \) equals \( w_{i,j} \) if the edge exists \((i, j) \in E\), and equals 0 otherwise. A nonzero entry \( A_{i,j} \) reflects that there exists an edge running from node \( j \) to node \( i \). In the work of this dissertation, the weights are positive \((A_{i,j} \geq 0)\), or in a certain case binary \((A_{i,j} \in \{0, 1\})\). Also, we constrain the presented research to undirected graphs, i.e. those for which \( A \) is symmetric, i.e. \( A = A^T \). In such a graph, the degree \( k_i \) of node \( i \), sometimes referred to as its strength if the weights are nonbinary, is given by \( k_i = \sum_{j=1}^{N} A_{i,j} \). Other important matrices and graph-related concepts are formally introduced in the following sections of this Chapter and in the relevant sections of the publications in Chapters 3 and 4.

2.2 Networks and communities

This section gives a reminder of the nature of communities and how they can be identified in a graph. It serves to lay a foundation for the adoption of modularity matrix (Newman, 2006) as the centerpiece of the work presented in Chapter 3.
2.2. Networks and communities

2.2.1 What is a community?

As graph representations of complex data started to evolve through the field of network science, a great amount of effort has been addressed to exploring communities (Fortunato and Hric, 2016). Although there is, so far, no unique definition of community, researchers came to favor certain approaches and definitions more than others. At first, a community was considered to be a subset of nodes with more edges connecting them to each other than to nodes outside this subset. More specifically, in (Radicchi et al., 2004), a (strong) community is defined as the set of nodes that all have internal degrees (within the subset) greater than its external degree. A weak community takes into consideration only the average internal and external degrees of the subgraph’s nodes. Similarly, in (Hu et al., 2008), minimal internal and maximal external degrees are taken into account. However, these definitions suffer from not being able to distinguish communities of significantly different size Fortunato and Hric (2016). Therefore, other definitions of communities emerged that are intended to overcome this challenge.

The alternative approaches to defining communities rely on the comparison of the edge connectivity pattern to the connectivity of a certain random graph model. In this context, a set of nodes is regarded as a community if the number of its internal edges is higher than expected for the random graph with no communities. This definition is at the core of community detection algorithms such as the inference of a stochastic block model (Wasserman and Anderson, 1987) and the optimization of modularity (White and Harary, 2001; Newman, 2004). The stochastic block model (SBM) is a random graph that draws each edge from a Bernoulli distribution. Inferring the probabilities of edges between all pairs of nodes in a graph gives a structural model describing communities. This one is being implicitly compared to an Erdős–Rényi model (Erdős and Rényi, 1959), which is a specific case of SBM where all edge probabilities are equal. Alternatively, the approach involving the metric of modularity compares the observed graph with a configuration model (Newman, 2006). As in the case of Erdős–Rényi, this model places edges randomly across pairs of nodes, so no communities are present. However, it has the same degree sequence as the observed graph, whereas the degree sequence of an expected Erdős–Rényi is a vector of all equal entries.

Hitherto, only one type of community is mentioned, also known as the module or an assortative subgraph. There exist other types with slightly different definitions. A disassortative community, or an anti-module, denotes an opposite kind of structure to the assortative. These communities have less edges connecting nodes in the subset to each other than to those outside the subset. The extremal cases of such structure are bipartite and k-partite graphs (Newman, 2006), which emerge when modeling different types of objects as nodes, e.g. in a network of movies and actors who appeared in them (Newman et al., 2001). Finally, there is a core-periphery structure, found for example in human brain networks (Bassett et al., 2013). In this setting, a core is a subset of densely interconnected nodes, whereas a periphery is a subset of nodes scarcely connected to each other but moderately connected to nodes from the core. Fig. 2.1 illustrates all three types of community structure, together with their expected adjacency matrices under SBM models able to describe them. The work in Chapter 3 relies on
Chapter 2. Background

Figure 2.1 – Illustration of the expected adjacency matrix under different SBM models (darker shades denote higher matrix entries), and a representative graph instance of such model. (a) Model containing modules, or assortative communities. (b) Model containing anti-modules, or disassortative communities. (c) Model containing core-periphery structure. (d) Model containing no community structure – Erdős–Rényi. Figure reused (with adapted captions) with permission from Fortunato and Hric (2016). ©2016, Elsevier

the community assessment able to recognize both assortative and disassortative structure. Additional inclusion of core-periphery goes beyond the scope of this dissertation.

2.2.2 Community detection

The problem of identifying communities is quite challenging from both an accuracy and computational complexity point of view (Yang et al., 2016), which is why a variety of dedicated algorithms exist. Some of those are based on statistical inference (Shuo and Chai, 2016; Zhang and Peixoto, 2020), including the stochastic block model (Karrer and Newman, 2011). Others exploit dynamical processes on graphs such as different types of random walks (Pons and Latapy, 2005; de Guzzi Bagnato et al., 2018), linear and nonlinear diffusion (Jeub et al., 2015; Ibrahim and Gleich, 2019), or synchronization (Huang et al., 2019). Finally, there are algorithms that optimize a certain objective function like modularity (Blondel et al., 2008) or surprise (Aldecoa and Marin, 2013). Indeed, in the work of this dissertation, we are interested in an algorithm for maximizing the modularity metric, but more specifically, its variant (Newman, 2006) belonging to the group of spectral algorithms (Tang et al., 2019; Tiomoko Ali and Couillet, 2018). Approaching communities from a graph spectral domain allows us to straightforwardly make the connection to the graph Fourier transform and signal processing framework as
shown in Chapter 3.

Modularity is a metric that quantifies a partitioning of graph nodes into disjoint subsets, i.e. communities. If the partitioning assigns two nodes to the same community, then the difference between the observed and expected edge weight across them contributes to the positive value of the metric. Formally, modularity $q_P$ of partitioning $P$ into $K$ subsets $C_i, i = 1, 2, ..., K$ writes:

$$q_P = \frac{1}{2M} \sum_{i,j}^N (A_{ij} - \frac{k_i k_j}{2M}) \delta(C_i, C_j)$$

(2.1)

where $N$ is the number of nodes, $M$ is the sum of all edge weights in the graph $2M = \sum_{i=1}^N k_i$, and $\delta$ is the Kronecker delta function. If the value of $q_P$ is positive (negative), the subsets in the partitioning reflect an assortative (disassortative) community structure. An algorithm for community detection maximizes $q_P$ over all partitions (Blondel et al., 2008). Such optimization problem is NP-complete (Brandes et al., 2006, 2008) and different approaches are used for solving it (Clauset et al., 2004; Guimerà and Amaral, 2005). The spectral algorithm relevant for the work in Chapter 3 builds around the modularity matrix (Newman, 2006):

$$Q = A - P = A - \frac{1}{2M} kk^T$$

(2.2)

where matrix $P$ describes the expected graph of the random configuration model. For a bipartitioning indicator vector with entries in $\{1, -1\}$ denoting one of the two communities, the quadratic form of $Q$ is equivalent to the modularity in Eq. (2.1). Therefore, detection of assortative communities or modules is performed by maximizing this quadratic form. The optimization problem is relaxed to the eigendecomposition of $Q$, and communities are found by clustering the nodes through their embeddings by the entries of several eigenvectors with the highest eigenvalues. If the quadratic form is minimized, anti-modules are identified. The higher the relevant eigenvalue, the closer are the entries of an eigenvector that correspond to nodes in the same community. This fact is exploited in the framework of Chapter 3 where a similar phenomenon is observed regarding values of a graph signal.

### 2.3 Graph signal processing

Additional quantitative data associated with graph nodes are regarded as graph signals and processed in a manner analogous to the conventional processing of time series. Each sample of a graph signal is assigned to a single node, yielding a compact representation as a vector whose entries follow the same labeling as that of the columns of the graph's adjacency matrix (Fig. 2.2). Therefore, the framework of graph signal processing (GSP) takes advantage of linear algebra and matrix computations (Shuman et al., 2013; Sandryhaila and Moura, 2014a). A fundamental operator in the framework is the shift $\mathbf{S}$ that acts on a signal $\mathbf{s}$ through matrix-
vector multiplication:

$$\mathbf{s}_{\text{shift}} = \mathbf{S} \cdot \mathbf{s}_{\text{in}} \quad (2.3)$$

The shifted signal $\mathbf{s}_{\text{shift}}$ is supposed to replicate the effect of time-shifting in regular one-dimensional signals. To this purpose, the choice for $\mathbf{S}$ is often the adjacency $\mathbf{A}$ or the Laplacian matrix:

$$\mathbf{L} = \mathbf{K} - \mathbf{A} = \text{diag}(\mathbf{k}) - \mathbf{A} \quad (2.4)$$

though alternative choices are available (Gavili and Zhang, 2017; Dees et al., 2019; Narang and Ortega, 2012). The reason behind the use of Laplacian as shift operator lies in its interpretation as the second-order derivative operator. Furthermore, in the spectral domain, Laplacian provides an intuitive representation of graph frequencies.

Spectral analysis of graph signals focuses on the oscillatory patterns of signal values across nodes connected by an edge (Sandryhaila and Moura, 2013, 2014b). The crucial constructs are the graph Fourier transform and its inverse:

$$\hat{\mathbf{s}} = \mathbf{U}^{-1} \mathbf{s}, \quad \mathbf{s} = \mathbf{U} \hat{\mathbf{s}} \quad (2.5)$$

where $\hat{\mathbf{s}}$ is the vector of spectral coefficients, and matrix $\mathbf{U}$ comes from the eigendecomposition of the shift, i.e. Laplacian:

$$\mathbf{L} = \mathbf{U} \Lambda \mathbf{U}^T \quad (2.6)$$

Spectral components in the graph Fourier domain are the eigenvectors of $\mathbf{L}$ sorted as columns of $\mathbf{U}$. The oscillating frequency of an eigenvector is quantified by the corresponding eigenvalue from the diagonal matrix $\Lambda$. Indeed, a low eigenvalue denotes "slow" oscillations – when the
entries of eigenvector at positions of connected nodes have similar values. The opposite is true for high frequencies. By projecting a graph signal onto such spectral components, the oscillatory pattern of the signal itself is assessed (Fig. 2.3). Within the framework built around graph Laplacian, signals exhibiting low-frequency oscillations are referred to as smooth. Many signal operators exist that rely on definitions of graph shift and the graph Fourier transform. These include filtering (Shuman et al., 2016), wavelets (Hammond et al., 2011), Hilbert transform (Venkitaraman et al., 2019), etc.

2.4 Uncertainty principle and energy localization

Important elements of GSP prominent in the following chapters of this dissertation are the uncertainty principle and localization of signal energy. Before putting these elements into the context of graphs, I give a brief reminder within the conventional signal processing setting. For continuous square-integrable function $x(t)$, the time-limiting and spectrum-limiting
Chapter 2. Background

operators $\mathcal{D}$ and $\mathcal{B}$ write:

$$
\mathcal{D} x(t) = \begin{cases} 
  x(t) & |t| \leq T/2 \\
  0 & |t| > T/2 
\end{cases}
$$

(2.7)

$$
\mathcal{B} x(t) = \frac{1}{2\pi} \int_{-\Omega}^{\Omega} \hat{X}(\omega)e^{j\omega t} d\omega
$$

(2.8)

where $\hat{X}(\omega)$ is the Fourier transform of $x(t)$. Operator $\mathcal{D}$ forces a finite support upon the signal in the time domain, whereas $\mathcal{B}$ forces the finite support onto the signal’s Fourier representation. Further, parameters $\alpha^2$ and $\beta^2$ quantify the localization of signal’s energy in the range $[-T/2, T/2]$ and in the spectral subband $[-\Omega, \Omega]$, respectively:

$$
\alpha^2 = \frac{||\mathcal{D} x(t)||_2^2}{||x(t)||_2^2}
$$

$$
\beta^2 = \frac{||\mathcal{B} x(t)||_2^2}{||x(t)||_2^2}
$$

(2.9)

The uncertainty principle then states that there can exist a signal with particular values of localization parameters $\alpha^2$ and $\beta^2$ if and only if

$$
\cos^{-1} \alpha + \cos^{-1} \beta \geq \cos^{-1} \sqrt{\lambda_0}
$$

(2.10)

where $\lambda_0$ is the largest eigenvalue of the eigensystem of joint localization in both domains:

$$
\mathcal{B} \mathcal{D} \psi = \lambda \psi
$$

(2.11)

This eigenvector equation yields solutions $\psi$ that could correspond to high $\lambda$ (close to 1). These are the signals with the best joint-domain energy localization possible for the given time and spectral supports. They are commonly referred to as Slepian signals, in connection with the researcher who studied them under the name of prolate spheroidal wave functions (Slepian and Pollak, 1961).

As expected, the analogous equations that define uncertainty principle and Slepian signals on graphs follow remarkably similar forms as those in Eqs. (2.10, 2.11) (Agaskar and Lu, 2013; Tsitsvero et al., 2016). However, in this context, the finite support in time is replaced by the finite nodal support, i.e. signal energy localizes onto a subset of nodes. The vertex-limiting operator $\mathbf{D}$ is a diagonal matrix with elements $\mathbf{D}_{i,i} \in \{0,1\}$ reflecting the nodal support. The
2.4. Uncertainty principle and energy localization

spectrum-limiting operator $B$ writes:

$$ B = U \Sigma U^T $$

(2.12)

where columns of $U$ are eigenvectors of the shift operator, and $\Sigma$ is a binary diagonal matrix indicating finite support among Fourier basis components – similarly as $D$. Then, according to the graph uncertainty principle, a signal $x$ with energy localization parameters $\alpha^2$ and $\beta^2$:

$$ \alpha^2 = \frac{||Dx||^2}{||x||^2} $$

$$ \beta^2 = \frac{||Bx||^2}{||x||^2} $$

(2.13)

can exist on the given graph for specified supports if and only if (Tsitsvero et al., 2016):

$$ \cos^{-1} \alpha + \cos^{-1} \beta \geq \cos^{-1}(\sigma_{\text{max}}(BD)) $$

$$ \cos^{-1}(\sqrt{1-\alpha^2}) + \cos^{-1} \beta \geq \cos^{-1}(\sigma_{\text{max}}(BD)) $$

$$ \cos^{-1} \alpha + \cos^{-1}(\sqrt{1-\beta^2}) \geq \cos^{-1}(\sigma_{\text{max}}(BD)) $$

$$ \cos^{-1}(\sqrt{1-\alpha^2}) + \cos^{-1}(\sqrt{1-\beta^2}) \geq \cos^{-1}(\sigma_{\text{max}}(BD)) $$

(2.14)

where $\sigma_{\text{max}}$ denotes the highest singular value of a matrix. The area in the parameter space $(\alpha^2, \beta^2)$ bounded by the curves in Eq. (2.14) and corresponding to all possible graph signals is known as the admissible region (Fig. 2.4).

Finally, graph Slepian signals – those with the highest possible values of both localization parameters (Fig. 2.5a) – are the solutions to the eigenvector equation:

$$ BDB\psi = \lambda \psi $$

(2.15)

Rather than the high energy localization in the nodal domain, Slepians can be defined to exhibit maximally smooth oscillations within a finite nodal support (Van De Ville et al., 2017) (Fig. 2.5b). Moreover, they can emerge from a combination of the two criteria (Petrovic et al., 2019). Further details on these types of Slepian signals are given in the appropriate subsections of the publications in Chapter 4.
Figure 2.4 – Illustration of the uncertainty principle. Parameters $\alpha^2$ and $\beta^2$ denote energy localization on a subgraph and spectral subband, respectively. The admissible region corresponds to the blue shaded area. Here, $\sigma_{\text{max}}$ denotes the highest singular value of a matrix, and $D$ and $B$ are the vertex- and spectrum-limiting operators. Figure reused with permission from Tsitsvero et al. (2016). ©2016, IEEE

Figure 2.5 – Two types of Slepian signals on a mesh graph. The nodal support encompasses nodes of the head of the animal. a) Slepians found by maximization of energy localization/concentration $\mu$. b) Slepians found by maximizing Laplacian smoothness on the finite support. This smoothness is quantified by the so-called modified embedded distance $\xi$. Parameter $\lambda$ reflects the global Laplacian smoothness of signals. Figure reused with permission from (Van De Ville et al., 2017). ©2017, IEEE
This chapter deals with a particular way of informing the graph signal operators of the community structure in the given graph. The work relies on the definition of graph Fourier transform as the projection of a signal onto eigenvectors of modularity matrix (Newman, 2006; White and Smyth, 2005). The first section is a published article that sets the foundation of this framework and discusses different operators on signals. The second section is a journal manuscript in preparation that shares a perspective on the graph uncertainty principle and the signal localization in vertex and spectral domains within the same framework.
Chapter 3. Informing signal processing techniques with community structure

3.1 Journal article: Community-aware graph signal processing: modularity defines new ways of processing graph signals

(This article has been published in IEEE Signal Processing Magazine, DOI: 10.1109/MSP.2020.3018087)

Miljan Petrović\textsuperscript{a,b,*}, Raphaël Liégeois\textsuperscript{a,b,*}, Thomas A. W. Bolton\textsuperscript{a,b,c}, Dimitri Van De Ville\textsuperscript{a,b}

\textsuperscript{a} Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, Campus Biotech, Geneva, Switzerland
\textsuperscript{b} Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland
\textsuperscript{c} Department of Decoded Neurofeedback, ATR Computational Neuroscience Laboratories, Soraku-gun, Kyoto, Japan

* These authors contributed equally and are considered joint-first authors.

Abstract

The emerging field of graph signal processing (GSP) allows to transpose classical signal processing operations (e.g., filtering) to signals on graphs. The GSP framework is generally built upon the graph Laplacian, which plays a crucial role to study graph properties and measure graph signal smoothness. Here instead, we propose the graph modularity matrix as the centerpiece of GSP, in order to incorporate knowledge about graph community structure when processing signals on the graph, but without the need for community detection. We study this approach in several generic settings such as filtering, optimal sampling and reconstruction, surrogate data generation, and denoising. Feasibility is illustrated by a small-scale example and a transportation network dataset, as well as one application in human neuroimaging where community-aware GSP reveals relationships between behavior and brain features that are not shown by Laplacian-based GSP. This work demonstrates how concepts from network science can lead to new meaningful operations on graph signals.

Keywords: Graph Signal Processing, Laplacian, Modularity, Community Structure, Networks, Graph Fourier Transform

3.1.1 Introduction

Network science is a multidisciplinary field that accounts for complex structure of data, providing new interpretations of datasets in diverse scientific disciplines ranging from humanities to physics and biomedicine. Naturally, analysis of network data relies on methods from graph theory, but also from statistical mechanics, statistical inference, advanced visualization, and domain knowledge from applied fields. More recently, graph signal processing (GSP) emerged as a new research theme at the intersection between signal processing and graph theory, with a
3.1. Journal article: Community-aware graph signal processing: modularity defines new ways of processing graph signals

particular focus on processing graph signals that associate values to the nodes of the graph. In many cases, the graph Fourier transform was defined by the eigendecomposition of the graph Laplacian; i.e., the eigenvectors of the Laplacian are considered as graph Fourier basis vectors, and the associated eigenvalues are graph frequencies (Shuman et al., 2013). Such graph Fourier transform can then generalize various classical signal processing tools to graphs (Shuman et al., 2013; Sandryhaila and Moura, 2013), such as the wavelet transform (Hammond et al., 2011), as well as theoretical considerations about graph uncertainty principles (Tsitsvero et al., 2016).

The graph Laplacian defines the second-order derivative on the graph and is therefore linked to smoothness, but alternative operators can explore other properties of graphs and graph signals. For example, community structure is a particularly interesting concept from network science where nodes inside a community are more strongly interconnected than with the rest of the graph (Fortunato and Hric, 2016; Newman, 2004). Community structure turned out to be present and relevant for a broad range of applications in sociology (Jeub et al., 2015), transportation (Guimera et al., 2005), biology (Guimerà and Amaral, 2005) or neuroscience (Meunier et al., 2010). In practice, communities can be found by maximizing the modularity index that evaluates the density of connections within clusters against a degree-matched graph where no cluster preference exists (Newman, 2006). Similar to Laplacian-based spectral clustering, where the Laplacian eigenvectors with smallest non-zero eigenvalues are considered since they optimize the convex relaxation of the graph cut criterion (von Luxburg, 2007), one approach for community detection is to compute the eigendecomposition of the modularity operator and consider the eigenvectors with largest eigenvalues (Fortunato and Hric, 2016).

In this paper, we set the foundations for community-aware GSP by introducing the modularity operator at the heart of the framework. This allows to define GSP operations that are aware of the graph community structure, but without the need of explicit community detection. After recalling basic GSP notions (Section 3.1.2), we define the modularity index and corresponding operator, highlighting the differences with the Laplacian (Section 3.1.3). We then detail how GSP operations such as filtering, sampling, and denoising, can be rendered community aware (Section 3.1.4). Using the OpenFlights and a functional magnetic resonance imaging (fMRI) datasets, we illustrate the benefits of community-aware GSP over its Laplacian-based counterpart (Sections 3.1.4 and 3.1.5).

3.1.2 Graph signal processing

We consider an undirected graph \( G = (\mathcal{V}, \mathcal{E}) \) with node set \( \mathcal{V} \) of cardinality \( N \) and edge set \( \mathcal{E} \). \( G \) can also be represented by the \( N \times N \) weighted adjacency matrix \( A \), whose entry \( a_{i,j} \) is non-zero and indicates the edge weight for an edge \( (i, j) \in \mathcal{E} \) that runs from node \( i \) to node \( j \). For an undirected graph, \( A \) is symmetric; i.e., it holds that \( a_{i,j} = a_{j,i} \) and \( A = A^T \). We will refer to a subgraph \( G_S = (\mathcal{V}_S, \mathcal{E}_S) \) by its node set \( \mathcal{V}_S \subset \mathcal{V} \) and assume \( \mathcal{E}_S \) containing all edges \( (i, j) \)
Chapter 3. Informing signal processing techniques with community structure

between nodes in $\mathcal{G}$. A graph signal associated to $\mathcal{G}$ is a vector $x \in \mathbb{R}^N$ that attributes values $x_i$ to the nodes $i = 1, 2, \ldots, N$. The neighborhood of a node $i$ is defined as the set of nodes $\mathcal{N}_i$ connected to it. A graph shift operator is defined as a linear operator on the space of signals, such that each entry of the shifted graph signal is a linear combination of input signal values, which often only involves neighboring entries to the one at hand (Sandryhaila and Moura, 2013). Therefore, the shift operator can be represented by a symmetric matrix $S \in \mathbb{R}^{N \times N}$ that associates weights $s_{ij}$ to edges $(i, j)$ such that $x_{\text{shift}} = Sx$. We will consider graph operators $\mathcal{H}$ that are shift-invariant under $S$ and thus satisfy $\mathcal{H}Sx = S\mathcal{H}x$ and can be represented as a matrix polynomial of $S$ (Ortega et al., 2018); i.e., $\mathcal{H} = p(S) = \sum_{k=0}^{K} h_k S^k$, with maximum degree of $N - 1$ due to the Cayley-Hamilton theorem.

The eigendecomposition of the shift operator provides the factorization

$$S = U \Lambda U^T,$$  \hspace{1cm} (3.1)

where $U = [u_1, \ldots, u_N]$ contains the $N$ eigenvectors and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_N)$ is a diagonal matrix with the corresponding eigenvalues. This allows to write the graph operator $\mathcal{H}$ alternatively as $\mathcal{H} = p(S) = U \text{diag}(\tilde{h}) U^T = \mathcal{U} \tilde{\Lambda} \mathcal{U}^T$, where the entries $\tilde{h}_i = p(\lambda_i) = \sum_{k=0}^{K} h_k \lambda_i^k$ of $\tilde{H}$ yield the spectral characterization of the graph operator. For the perspective of GSP, a given shift operator $S$ defines the Graph Fourier transform (GFT) of the graph signal $x$ as (Shuman et al., 2013):

$$\hat{x} = U^T x, \quad \text{and} \quad x = U \hat{x},$$  \hspace{1cm} (3.2)

where $U$ is defined as in Eq. (3.1) and $\hat{x}$ contains the spectral coefficients of the GFT. The graph operator $\mathcal{H}$ can then be implemented elegantly in the graph Fourier domain as

$$x_{\text{out}} = \mathcal{H}x = p(S)x = U p(\Lambda) U^T x = U \tilde{\Lambda} U^T x,$$  \hspace{1cm} (3.3)

which allows to directly specify $\tilde{H}$ in terms of a spectral window (e.g., low-pass, band-pass, high-pass) for graph filtering operations (Ortega et al., 2018). Beyond filtering, other operations have been extended to the graph domain, such as stationarity analysis (Marques et al., 2017), wavelet transforms (Hammond et al., 2011), or convolutional neural networks (Defferrard et al., 2016).

One common choice for $S$ is the weighted graph adjacency matrix $A$ (Sandryhaila and Moura, 2013; Huang et al., 2018). Another one is the Laplacian matrix $L = D - A$, where $D = \text{diag}(k_1, k_2, \ldots, k_N)$ is the degree matrix with $k_i = \sum_{j=1}^{N} a_{i,j}$ the weighted degree (Hammond et al., 2011; Leonardi and Van De Ville, 2013). For the latter, the eigenvalues are sometimes referred to as graph frequencies and reflect smoothness in terms of the signal variation norm of the corresponding eigenvectors (Ortega et al., 2018). For a graph signal $x$, its smoothness is measured by the quadratic form

$$q_L(x) = \sum_{i \neq j} a_{i,j} (x_i - x_j)^2 = x^T Lx = \sum_{i=1}^{N} \lambda_i x_i^2,$$  \hspace{1cm} (3.4)

which shows that measuring smoothness in the spectral domain can be done by weighting with
3.1. Journal article: Community-aware graph signal processing: modularity defines new ways of processing graph signals

the graph frequencies. The example in Fig. 3.1 illustrates the Laplacian eigendecomposition for a simple graph and will be discussed in more details later.

3.1.3 Community structure

Communities refer to dense subgraphs $P_c \subset N$, $c = 1, \ldots, C$, that are well separated from each other, and manifested at the “mesoscale” level between local nodal and global graph properties (Fortunato and Hric, 2016). A large number of measures have been proposed with the purpose to discover community structure of an observed graph. For our aim, it is insightful to first revisit the graph Laplacian as it relates to one aspect of community structure, which is quantifying the separation between subgraphs. Specifically, the splitting of a graph into two mutually exclusive subgraphs $P_1$ and $P_2$ can be encoded by a vector $s$ whose entries $s_i = +1$ or $-1$ indicate whether a node $i$ belongs to the first or second subgraph, respectively. The graph cut size—number of connections running between the two subgraphs—can then be related to the Laplacian as

$$R = \frac{1}{2} \sum_{i,j \atop s_i \neq s_j} a_{i,j} = \frac{1}{2} \sum_{i,j} \left( \frac{1 - s_i s_j}{2} \right) a_{i,j} = \frac{1}{4} s^\top L s. \quad (3.5)$$

Optimizing $R$ by convex relaxation of $s$ (i.e., allowing the entries to take any value) leads to the well-known spectral clustering (von Luxburg, 2007). The eigenvector of $L$ with smallest non-zero eigenvalue (a.k.a. Fiedler vector) provides the solution to the bipartition problem. Recent work has also used graph wavelets to enable multiscale subgraph discovery (Tremblay and Borgnat, 2014).

The network-science view on community structure considers the adjacency matrix as a realization of an underlying stochastic model that defines edge probabilities within and between subgraphs. Stochastic block models (SBMs) (Lee and Wilkinson, 2019) are the best known generative models that can express assortativity (preferential connectivity within a node’s subgraph, leading to community structure), but also disassortativity (preferential connectivity to a subgraph to which the node does not belong) and core-periphery structure (densely interconnected core and periphery to the core). SBMs can be fitted by statistical inference to an observed graph, or can generate random graphs with predefined structure. Modularity, denoted by $Q$, is a specific graph measure derived from stochastic considerations that quantifies density of subgraphs by comparison against a null model:

$$Q = \frac{1}{2} \sum_{i,j \atop s_i = s_j} (a_{i,j} - z_{i,j}) = \frac{1}{2} \sum_{i} \left( \frac{1 + s_i}{2} \right) (a_{i,i} - z_{i,i}) = \frac{1}{4} s^\top \left( A - \frac{kk^\top}{2M} \right) s, \quad (3.6)$$

where $s$ is a vector encoding the graph partition into two communities, $z_{i,j} = \frac{k_i k_j}{2M}$ is the edge probability between nodes $i$ and $j$ according to the null model, $M = \sum_{i=1}^N k_i$ is the total edge weight, and $Q$ is the modularity matrix. Choosing this null model allows for comparisons
against a reference that preserves the graph degree distribution (i.e., \( \sum_{j=1}^{N} a_{i,j} = \sum_{j=1}^{N} z_{i,j} \) for \( i = 1, 2, \ldots, N \)), with edges placed evenly (Newman, 2006). Hence \( Q \) encodes the difference between edge densities in the original graph and in a degree-matched null model. This model is known as the configuration model and is commonly used to define \( Q \), but other null models can be considered (Newman, 2006; Massen and Doye, 2005; Newman, 2002).

The solution to maximizing \( Q \) is found by spectral clustering using the eigenvectors of \( Q \) with largest eigenvalues, identifying “modules” with high assortativity. Similarly, “anti-modules” with high dissortativity can be found by minimizing modularity. In fact, \( Q \) is a rank-one perturbation of \( A \), and consequently, Weyl’s inequality informs us that eigenvalues \( \lambda_i(Q) \) and \( \lambda_i(A) \) of \( Q \) and \( A \) are interleaved, i.e., \( \lambda_1(A) \geq \lambda_1(Q) \geq \lambda_2(A) \geq \lambda_2(Q) \geq \ldots \geq \lambda_N(A) \geq \lambda_N(Q) \), where the sequences of eigenvalues are in descending order. This result confirms that the modularity matrix of any simple undirected graph (that is, without self-loops) has both positive and negative eigenvalues (Bolla et al., 2015). The existence of both positive and negative eigenvalues implies that such a graph can be analyzed in terms of modular and anti-modular spectral components of \( Q \). Eigenvectors with zero eigenvalues are modularity-neutral; e.g., the constant vector \( 1 \) is in the kernel of \( Q \) due to \( Q \cdot 1 = 0 \).

Further illustration of the differences between spectral properties of \( L \) and \( Q \) is provided in Fig. 3.1 for a toy graph with \( N = 10 \) nodes, 5 of which form a strong (fully connected) community and the others are weakly connected. Fig. 3.1a shows the graph and a plot of eigenvalues \( \lambda_i(L) \) (blue solid line) and \( \lambda_i(Q) \) (red solid line) in the conventional ascending and descending order, respectively. To better highlight the differences between the corresponding eigenvectors \( u_i(L) \) and \( u_i(Q) \), respectively, we also plot smoothness \( u_i(Q)^\top L u_i(Q) \) of modularity eigenvectors (blue dotted line), and modularity\(^1 \) \( u_i(L)^\top Qu_i(L) \) of Laplacian eigenvectors (red dotted line). Fig. 3.1b shows the degree-matched null model graph \( k k^\top 2M \). Several eigenvectors \( u_i(Q) \) and \( u_i(L) \) \((i = 1, 2, 3, 10)\) are shown in Fig. 3.1c. Although \( u_i(L) \) are optimized for smoothness, they show high values at specific nodes (except the constant eigenvector \( u_1(L) \) with zero eigenvalue). Therefore, the Fiedler vector \( u_2(L) \), with lowest graph cut size, does not provide a correct partitioning. The eigenvectors of \( Q \) though are optimized for the modularity index and \( u_1(Q) \) provides a conspicuous split between the communities. \( Q \) also has a constant eigenvector \( u_4(Q) \) with zero smoothness. Curiously, the modularity of \( u_3(L) \) is actually the highest among the Laplacian eigenvectors, but still does not provide a convincing partitioning. Eigenvectors of \( Q \) with negative eigenvalues, such as \( u_{10}(Q) \), are driven by smoothness across modules and signal variability within modules.

### 3.1.4 Community-aware graph signal processing

The Laplacian operator \( L \) is the common choice of shift operator in GSP (Shuman et al., 2013; Defferrard et al., 2016; Huang et al., 2018) from which the GFT and all operations are derived. Instead, we propose to use the modularity matrix \( Q \) as shift operator. Interestingly, the

\(^1\)The quadratic form associated to the modularity matrix will be formally introduced in Sect. IV.
3.1. Journal article: Community-aware graph signal processing: modularity defines new ways of processing graph signals

Figure 3.1 – Eigenvalues and eigenvectors of the graph Laplacian (L) and modularity matrices (Q). (a) Underlying graph structure (top) and corresponding L (blue) and Q (red) eigenspectra (solid lines) and quadratic forms of smoothness and modularity (dotted lines) and (b) corresponding degree-matched null model used to compute Q (Eq. (3.6)). (c) Selected eigenvectors of L (blue) and Q (red) matrices. Value and sign of eigenvectors’ entries are reflected by the height and up-down direction, respectively, of the vertical bars.

Modularity matrix is a non-local operator since the second term that originates from the null model “spreads out” the signal over the whole graph according to the degree distribution—and not only the local neighborhood. Based on this generalized notion of shift operator, we will obtain GSP operations that become aware of the graph community structure, but without the need of explicit community detection. Given a graph signal \( x \), its modularity is computed by the quadratic form

\[
q_Q(x) = \sum_{i,j} a_{i,j} x_i x_j - \frac{\left( \sum_{i,j} a_{i,j} x_i \right)^2}{\sum_{i,j} a_{i,j}} = x^T Q x. \tag{3.7}
\]

Since \( Q \) is not positive semi-definite, \( q_Q(x) \) can take positive and negative values, depending whether signal variations follow modular or anti-modular organization (Newman, 2006). Thus, the quadratic form \( q_Q(x) \) is not a variation norm of the graph signal \( x \), which is needed for some GSP operations such as regularization. We overcome this limitation by introducing

\[
q_{Q^+}(x) = x^T Q^+ x, \tag{3.8}
\]

based on \( Q^+ = \lambda_{\text{max}}^{(Q)} I - Q \) where \( \lambda_{\text{max}}^{(Q)} \) is the largest eigenvalue of \( Q \) and \( I \) is the identity matrix. Since \( Q^+ \) is positive semi-definite, \( q_{Q^+}(x) \) is a non-negative function of \( x \). A low value of \( q_{Q^+}(x) \)
reflects that the graph signal \( x \) follows modular organization of the graph. On the contrary, a high value of \( q_{Q^+} (x) \) is obtained for graph signals reflecting the anti-modular organization. In other words, \( q_{Q^+} (x) \) can be interpreted as the modularity-based graph signal variation of \( x \). Minimization of this metric is achieved by the eigenvectors of \( Q^+ \) that define the spectral basis of a GFT exploiting modularity of graph signals. Since \( Q^+ \) and \( Q \) have the same eigenvectors with eigenvalues that are reversed and shifted, the eigenvectors of \( Q \) define a proper GFT basis that is built up according to modularity/anti-modularity. Similarly, denoting \( Q^- = Q - \lambda_{\min}^{(Q)} \cdot I \) allows to define a variation norm \( q_{Q^-} (x) \) that encodes anti-modular organization in low values, while exploiting the same spectral basis since \( Q^- \) also has the same eigenvectors as \( Q \).

We now illustrate utility of community-aware GSP tools using data from the OpenFlights Airports Database (https://openflights.org/data.html) that consists of 3281 airports and 67202 routes (Fig. 3.2a). Graph nodes denote airports that are connected by an undirected binary edge if there exists an airline route between them. Node colors reflect a graph signal computed as the sum of both departing and incoming flights at each airport, which was then demeaned and scaled to unit variance. We considered a ground truth community structure based on the continent to which each airport belongs, resulting in a partition of the nodes into the six following communities: Europe, Africa, Asia, Oceania, North America, and South America (Guimera et al., 2005). The inset of Fig. 3.2a shows the total number of flights leaving from or arriving to the airports in the eastern part of the North American continent. It can be seen that Atlanta airport has more traffic than JFK airport in New-York and that the vast majority of airports have very low traffic.

### 3.1.4.1 Filtering

From the general definition of GSP filtering proposed in Eq. (3.3), community-aware filtering uses the modularity-based spectral domain with a spectral window \( \tilde{h} \):

\[
x_{out} = U \underbrace{\text{diag}(\tilde{h})}_{=: \tilde{h}} U^T x,
\]

where \( U \) contains the eigenvectors of \( Q \). While low- and high-pass filtering are natural operations when using the Laplacian GFT, modularity-based GFT allows to define a modular filter (i.e., \( \tilde{h} \) has non-zero weights on spectral components with positive eigenvalues) or an anti-modular filter (i.e., \( \tilde{h} \) only has non-zero weights on spectral components with negative eigenvalues).

The community-aware filtering was applied on the graph signal of Fig. 3.2a and was compared to a Laplacian-based filtering. The passband, that is, the range of eigenvalue indices with non-zero filter weights, of the modular (anti-modular) filter includes all 1125 (1159) strictly positive (negative) eigenvalues of \( Q \), and the smooth (non-smooth) filter was matched so as to capture the same number of spectral Laplacian components (Fig. 3.2b). Within a passband \([N_1, N_2]\), the \( i^{th} \) entry of \( \tilde{h} \) was set to \( |\lambda_i| / \sum_{k=N_1}^{N_2} |\lambda_k| \) for modular, anti-modular and non-smooth filterings.
and to $1 - \frac{|\lambda_i|}{\sum_{k=N_1}^{N_2} |\lambda_k|}$ for the smooth filtering, which accounts for strength of modularity or smoothness of the components. Finally, for each filtered signal we computed a measure of within-community variability, denoted $\Delta_C$, and defined as the standard deviation of the filtered signal values within a ground-truth community, averaged over the 6 communities.

Figure 3.2 – Application of the GSP framework on the OpenFlights. (a) Graph nodes correspond to airports and an edge connects two nodes when at least one flight connects the two corresponding airports. Graph signal (number of flights at each airport) is reflected in nodes’ colors. The inset shows a zoom on New-York (JFK) and Atlanta (ATL) airports. (b) Eigenvalues of the graph Laplacian (blue) and modularity matrix (red). Dashed gray lines and arrows represent limits of the filtering passbands, left for modular and smooth, right for anti-modular and non-smooth. (c) Laplacian filtering of the graph signal shown in panel (a) yields smooth and non-smooth signals, and (d) community-aware filtering yields modular and anti-modular signals. The value of within-community variability ($\Delta_C$) is shown for the four filtered signals.
Chapter 3. Informing signal processing techniques with community structure

As shown in Fig. 3.2c, the Laplacian-based filtering extracts smooth and non-smooth parts of the graph signal. The smooth signal tends to capture widespread fluctuations over the graph whereas the non-smooth signal contains rather localized peaks that partially correspond to the extreme values in the original signal (e.g., ATL and JFK airports), suggesting that the underlying community structure is not a predominant feature encoded in Laplacian-filtered signals. In contrast, the modular signal shown in Fig. 3.2d reflects the community structure of the underlying graph by the clusters of high values in North America, Europe and Asia. This effect is further supported by within-community variability ($\Delta C$) that is lower in the modular signal than in the smooth one. In other words, modular-based filtering can be seen as promoting smoothness within communities. On the contrary, the anti-modular signal promotes variability within communities, as this signal shows higher $\Delta C$ compared to the non-smooth signal.

In order to further explore the roles of particular nodes in the different filtering operations, we focus on two airports: ATL and JFK. While these two airports are both highly connected, as seen from Fig. 3.2a, they play different roles in the graph community structure. Indeed, ATL has a within-community z-score degree ($Z_{in}$) (Guimerà and Amaral, 2005) of 8.98 and an outside-community z-score degree ($Z_{out}$) of 5.97, while for JFK, we have $Z_{in} = 4.22$ and $Z_{out} = 15.29$. Therefore, ATL has stronger connections within its community than between communities, and vice versa for JFK. For Laplacian filtering, the signal values of both JFK and ATL are evened out in the smooth signal, and stand out in the non-smooth signal (insets of Fig. 3.2c). However, community-aware filtering picks up differences between these airports by a relatively stronger value of JFK in the modular signal and of ATL in the anti-modular signal (insets of Fig. 3.2d). Since modularity-promoting filtering favors smoothing within the communities, the value of the strongly within-community connected ATL will be more reduced than for JFK. The large value of ATL is captured in the anti-modular signal as it stands out with respect to values of its within-community neighbors. This suggests that modular/anti-modular signal identifies nodes with high values and strong inter-modular/intra-modular connectivity. Overall, the results reveal that community-aware filtering can attenuate or enhance values of nodes according to their connectivity within or between communities.

Finally, the need for the eigendecomposition of $Q$ in Eq. (3.9) can be circumvented by implementing the filtering operation in the vertex domain by a polynomial matrix function $p(Q)$ as suggested in Eq. (3.3), which is equivalent to applying the spectral window $H = p(\Lambda)$. In order to further improve computational efficiency of filtering for large-scale but sparse graphs, one can break down the operation $Qx$ into $Qx = Ax - (1/2M)kk^T x$, where the first term is a sparse matrix-vector multiplication, and the second term can be evaluated by consecutively computing $k^T x$ and then multiplying the resulting scalar with $k/2M$. Therefore, the dense matrix $Q$ does never need to be stored explicitly. For an undirected graph with $M'$ edges and $N$ nodes, computing $Ax$ takes $\Theta(2M')$ time, and $Lx$ takes $\Theta(2M' + N)$ (in big O notation). The term $(1/2M)kk^T x$ has complexity of $\Theta(N)$. Consequently, $Qx$ takes $\Theta(2M' + N)$, identical to $Lx$. For a polynomial filter of order $K$ applied to a large sparse graph, this reverts to $\Theta(KM')$. 

28
3.1. Journal article: Community-aware graph signal processing: modularity defines new ways of processing graph signals

3.1.4.2 Optimal sampling & reconstruction

Finding the subset of nodes from which a signal can be optimally reconstructed has been extended to the graph domain in the context of bandlimited graph signals \( \mathbf{x} = \mathbf{Bx} = \mathbf{U\Sigma U^T x} \) (Tsitsvero et al., 2016), where \( \mathbf{U} \) contains the eigenvectors of the shift operator, and \( \mathbf{\Sigma} \) is a diagonal matrix indicating the passband. The noisy graph signal \( \mathbf{y} = \mathbf{x} + \mathbf{n} \), with \( \mathbf{n} \) additive independent and identically distributed (i.i.d.) noise, is sampled into \( \mathbf{x}_s = \mathbf{Ry} \) where the diagonal matrix \( \mathbf{R} \) indicates with 0’s and 1’s the sampled nodes. Reconstruction denotes the procedure of finding \( \mathbf{x}_{\text{rec}} \) from \( \mathbf{x}_s \), such that the mean squared error \( \mathbb{E}[||\mathbf{x}_{\text{rec}} - \mathbf{x}||^2_2] \) is minimized (Tsitsvero et al., 2016). The minimization condition further extends to the choice of optimal sampling procedure since sampling at specific nodes can limit the potential performance of the subsequent reconstruction. One of the solutions (Tsitsvero et al., 2016) to the problem of finding (sub)optimal sampling and reconstruction defines sampling as finding \( \mathbf{R}^* \) via:

\[
\mathbf{R}^* = \arg\max_{\mathbf{R}} || \mathbf{\Sigma U^T R} ||_F. \tag{3.10}
\]

Solving Eq. (3.10) amounts to selecting nodes for the optimal sampling subset for which the columns of \( \mathbf{\Sigma U^T} \) have the highest \( l_2 \)-norm. Given the graph signal \( \mathbf{x}_s \) sampled at nodes defined in \( \mathbf{R}^* \), the reconstruction follows:

\[
\mathbf{x}_{\text{rec}} = \mathbf{V\Psi^{-1}V^Tx}_s, \tag{3.11}
\]

where \( \mathbf{V} \) and \( \mathbf{\Psi} \) contain the eigenvectors and eigenvalues of \( \mathbf{BR^*B^T} \).

We explore how well the graph signal presented in Fig. 3.2a can be reconstructed using the above framework and considering either \( \mathbf{L} \) or \( \mathbf{Q} \) as shift operator. We set to 500 the number of nodes to be sampled and use a spectral band including 200 components with lowest (\( \mathbf{L} \)), or highest positive (\( \mathbf{Q} \)) eigenvalues. The set of optimal nodes in these two cases is given in Fig. 3.3.

![Figure 3.3](image-url)

**Figure 3.3** – Optimal subset of 500 nodes for subsampling smooth (blue) or modular (red) signals of bandwidth 200.
Chapter 3. Informing signal processing techniques with community structure

The Laplacian-based sampling subset is composed of more peripheral nodes with low degree (1.78 ± 1.01) whereas the modularity-based sampling contains nodes with high degree (53.12 ± 43.61), which are important for inter- and intra-community connectivity. Only two nodes were found to belong to both subsets. An interpretation for this is that while the Laplacian framework focuses on preserving values on nodes where that value is hard to predict due to their low connectivity, the modularity framework maximizes predictability of all nodes by selecting nodes with high degree. This is in accordance with the assumption that traffic at a well connected airport could be a good predictor of the traffic at airports connected to it. Finally, the average reconstruction error is found to be significantly lower ($p < 0.01$, paired $t$-test over nodes) for the modular-based framework than for the Laplacian-based framework (0.0001 ± 0.0002 vs. 0.0006 ± 0.0037). This result supports the relevance of the modularity matrix as shift operator rather than the Laplacian in applications where community structure is pertinent.

3.1.4.3 Surrogate data generation

Surrogates play an essential role in non-parametric statistical testing to provide data under the null hypothesis; i.e., randomizing measurements while also preserving some properties. For instance, phase randomization preserves the moduli of the Fourier coefficients while their phases are randomized, leading to surrogate data with the same autocorrelation properties as the original data. This framework was extended to graph signals using the Laplacian, yielding surrogate data that preserve smoothness of the original graph signal (Pirondini et al., 2016). We propose to transpose this method to community-aware representations in order to preserve modular organization of a given graph signal. In particular, a community-aware surrogate signal $x_{\text{surr}}$ of the graph signal $x$ is given by

$$x_{\text{surr}} = UC\hat{x},$$

where $\hat{x}$ is the modularity-based GFT of $x$ and $C$ a diagonal matrix with random entries 1 or $-1$, thereby preserving the modularity index of the original signal. The null distribution of any test statistic can then be obtained from multiple realizations of $x_{\text{surr}}$ and compared against its value for the empirical signal $x$. This could be refined to a more specific null model by only changing the signs of (anti-)modular components; i.e., entries in $C$ corresponding to positive (negative) eigenvalues of $Q$.

We applied this framework to the signal of Fig. 3.2a by permuting signs of (i) all Laplacian-based, (ii) all modularity-based, (iii) only modular, and (iv) only anti-modular spectral coefficients. For each case, we generated 10000 surrogate samples that were used to test whether the original signal value is higher than expected under the null hypothesis. The test used an $\alpha$ level of 0.05 Bonferroni-corrected for multiple comparisons. In (i) and (ii), no nodes were found with values significantly different from their surrogates, but when only randomizing modular (anti-modular) components, 16 (2) airports revealed higher values than expected.
These airports had lower values of \( Z_{in} \) \((-0.35 \pm 0.16)\) than \( Z_{out} \) \((-0.18 \pm 0.03)\), indicating these nodes have stronger connectivity with other communities. Considering the results of the filtered signal values of JFK and ATL (Fig. 3.2d), one could expect that high signal values at these nodes can be explained by the underlying community structure. However, surrogate testing showed they cannot be explained by community structure alone. Similarly, two ‘outlier’ airports are identified when only randomizing signs of anti-modular spectral coefficients. The results illustrate the complementary roles of modular and anti-modular parts to describe a graph signal. In the context of OpenFlights, this corroborates the assumption of relevant community structure being present in the graph that is only accounted for by modularity-based GSP, and can then be used to assess to what extent graph signals follow this underlying graph structure.

### 3.1.4.4 Denoising

Another generic GSP operation is the recovery of the graph signal \( x \) from its noisy observation \( y = x + n \). The variational formulation puts forward a data-fitting term and a regularization term:

\[
\arg\min_x ||x - y||^2 + \mu \cdot x^\top P x, \tag{3.13}
\]

where \( \mu \) is the regularization tuning parameter and the quadratic form of \( P \) reflects prior knowledge about \( x \). A classical choice is \( P = L \), which corresponds to assuming that the graph signal \( x \) should be smooth on the graph. Since \( L \) is positive semi-definite, the cost function in Eq. (3.13) is convex and has a unique optimal solution. The same is true if \( P \) is chosen as \( Q^+ \) and \( Q^- \) (cf. Eq. (3.8)) in order to favor modular or anti-modular organization of \( x \), respectively.

Performance of these reconstruction approaches is illustrated using the original signal of Fig. 3.2a. This signal was normalized to unit norm, corrupted with additive Gaussian noise of different variance \( \sigma^2 \) ranging between 0.01 and 1, and the optimal value of \( \mu \) was determined using an oracle approach. For small to intermediate noise levels, we found that imposing a modular structure on \( x \) (i.e., \( P = Q^+ \)) yielded the best performance (RMS error is 0.0048 for \( \sigma^2 = 0.01 \), and 0.0083 for \( \sigma^2 = 0.25 \)). The error increases by an order of magnitude (0.0106 for \( \sigma^2 = 0.01 \), and 0.0168 for \( \sigma^2 = 0.25 \)) when using a Laplacian regularizer (i.e., \( P = L \)), and whereas for anti-modular regularization (i.e., \( P = Q^- \)) similar values of RMS are reached (0.0096 for \( \sigma^2 = 0.01 \), and 0.0145 for \( \sigma^2 = 0.25 \)). The advantage of modular regularization decreases for larger noise and the reconstruction errors become comparable when \( \sigma^2 = 1 \) (all errors above 0.15).

The assumptions of the different regularizers can be summarized as follows: Laplacian (L) favors smoothness of the graph signal by measuring differences between adjacent nodes; modularity (\( Q^+ \)) favors smoothness of the graph signal between nodes weighted by their closeness community-wise (assortative mixing); anti-modularity (\( Q^- \)) favors smoothness of the graph signal between nodes weighted inversely by this closeness (dissortative mixing). An explanation of the superior performance of modularity-based regularization in the present
example is that similarly high air traffic is more bound to modular organization than to neighborhood relationships; e.g., low traffic of a small island airport connected to several high-traffic mainland hubs of different communities, would lead to a high signal variation through the Laplacian lens, but not through the modularity lens since the island is not close to the large communities. In the end, it is the nature of the graph signal and how it relates to the underlying graph structure that will motivate the use of one or the other regularizer.

3.1.5 Validation for neuroimaging

The results presented above are built from a flight network with known community structure. Likewise, many real-world networks exhibit community structure and we therefore expect the proposed framework to provide a more appropriate way to analyse the corresponding graph signals. We further illustrate the benefits of community-aware GSP in a validation experiment using brain anatomical and functional data from the Human Connectome Project (Essen et al., 2013) using a parcellation of the cerebral cortex ($N = 360$). The graph structure was defined by counting the number of fiber tracts in diffusion-weighted MRI (Preti and Ville, 2019), and the graph signals are the activity patterns at different timepoints obtained from functional MRI time series reflecting activity in each brain region (Essen et al., 2013). For each region, the timecourse was z-scored (centered and unit variance).

The experiment consisted in exploring the link between brain imaging data and 62 behavioral scores for 181 healthy volunteers. To that aim, functional time series were filtered using either the anatomical graph Laplacian or modularity matrix following the procedure described in Section 3.1.4.1, and the filtered time series were averaged to yield, for each subject, a metric reflecting (non-)smoothness or (anti-)modular structure of brain function in each brain region (Medaglia et al., 2017). The link between these measures and the 62 behavioral scores was then computed using a nested cross-validation scheme and the $R^2$ coefficient of determination was used to quantify the strength of the association between brain function and behavior.

Fig. 3.4 shows the values of $R^2$ for the 62 behavioral measures and different bandwidths. It can first be seen that in most cases using a narrow bandwidth yields stronger $R^2$ which suggests that the information of interest is captured in the very first (non-)smooth or (anti-)modular eigenvectors. Then, we also observe that community-aware filtering reveals links between brain function and behavior that are not captured by Laplacian-based filterings. For example, the anti-modular functional signal shows strong links with various personal character traits while the modular signal mainly captures information about cognitive features.

3.1.6 Conclusion & outlook

Measures of community structure have been extensively used in network science to probe the organization of complex networks. Importantly, the tools that have been developed for
3.1. **Journal article: Community-aware graph signal processing: modularity defines new ways of processing graph signals**


processing graph signals expressed on these networks are based on the graph Laplacian and thus blind to underlying community structure. We proposed to make GSP community-aware, not by detecting communities, but by defining operations based on the modularity matrix. This provides a natural interpretation of the modularity spectrum in terms of modular and anti-modular contributions, though it requires adaptation when a variation metric is needed. We showed, using several examples, that community-aware GSP acts meaningfully differently compared to classical GSP. Considering the variety of datasets with community structure, the proposed framework will find its use in a wide range of fields and applications.

One extension of Laplacian- and modularity-based GSP is to account for directed graphs. We believe this is beyond the scope of the present paper but the interested reader is referred to (Sardellitti et al., 2017) for defining Laplacian-based spectral bases of directed graphs, or modularity matrices for directed graphs (Malliaros and Vazirgiannis, 2013) using in- and out-degrees of nodes. Another extension of the community-aware GSP framework could include different null models in the modularity criterion. Specifically, the Bernoulli model preserves the average degree (Newman, 2006), whereas the configuration model can be modified to
Chapter 3. Informing signal processing techniques with community structure

exclude self-loops (Massen and Doye, 2005), or to consider possible correlation between
degrees of nodes (Newman, 2002). Finally, communities could also be defined at the level of
edges instead of nodes to deal with overlapping communities (Ahn et al., 2010), or even trian-
gles and higher-order simplicial complexes as in higher-order Laplacian-based topological
GSP (Barbarossa and Sardellitti, 2020).

Acknowledgment

This work was supported in part by the Swiss National Science Foundation (200021_17 5506),
the CHIST-ERA IVAN project (20CH21_174081), the Japan JST ERATO Grant (JPMJER1801),
and the Center for Biomedical Imaging (CIBM).
3.2 Journal preprint: Graph uncertainty principle revisited: the importance of community structure

(This manuscript is in preparation for submittal to a journal.)

Miljan Petrović a,b, Dimitri Van De Ville a,b

a Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, Campus Biotech, Geneva, Switzerland
b Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

Abstract

Conventional uncertainty principles assess joint localization in time and frequency and play an important role in signal processing and physics. Their transposition into the graph setting comes with several challenges as the graph structure itself alters the definition of the spectral domain, commonly defined through the eigenvectors of the graph Laplacian. Motivated by the ubiquitous presence of community structure in real-world networks, we explored whether an alternative definition of the graph spectral domain, specifically the one based on the graph modularity operator, is able to improve uncertainty bounds obtained with the graph Laplacian. Using eigengap analysis of the spectrum-defining matrices and numerical results on random graphs under the stochastic block model, we demonstrate that such is the case under certain conditions depending on the trade-off between modularity versus smoothness.

Keywords: Graph uncertainty principle, Graph Laplacian, Community structure, Modularity, Signal localization

3.2.1 Introduction

Graph signal processing (GSP) sets a framework to define meaningful operations on signals defined on graphs, thus non-Cartesian domains (Shuman et al., 2013; Sandryhaila and Moura, 2013; Ortega et al., 2018; Girault et al., 2018) present in disciplines such as neuroscience (Huang et al., 2016), image processing (Cheung et al., 2018), big sensor data (Sandryhaila and Moura, 2014a). The extension of classical harmonic analysis on graphs is pursued by defining the graph Fourier transform (GFT) using the eigenvectors of the graph Laplacians such that the spectral representation of graph signals can be obtained by projecting on these basis vectors. Graph uncertainty principles (Agaskar and Lu, 2013; Tsitsvero et al., 2016) then come into play to characterize the interplay between localization in the vertex and spectral domains, respectively. These theoretical insights spurred further research into concepts such as energy vertex-frequency distributions (Daković et al., 2019) and orthogonal bases for signal
decomposition into vertex-frequency localized signals (Van De Ville et al., 2017; Petrovic et al., 2019), i.e. graph generalizations of Slepian functions (Slepian and Pollak, 1961; Demesmaeker et al., 2018).

The graph Laplacian is the fundamental operator to characterize graph structure; i.e., it can be used for optimizing graph cuts in partitioning problems von Luxburg (2007) or to measure the smoothness of graph signals. While the GFT is conventionally defined through the eigendecomposition of the graph Laplacian, recent work shows the possibility of using alternative spectral domains Petrovic et al. (2020), which can be defined by other concepts of graph organization such as community structure Newman (2006); Fortunato and Hric (2016). Communities are subsets of nodes that are more strongly connected between them than with nodes of other communities, a phenomenon that is observed in many different research areas; e.g., transportation Guimera et al. (2005), biology Guimerà and Amaral (2005), neuroscience Betzel et al. (2019), and sociology Girvan and Newman (2002). One approach to measuring the strength of community structure is based on the modularity index Newman (2006). The framework of community-aware GSP Petrovic et al. (2020) proposes the eigendecomposition of the modularity matrix Newman (2006); White and Smyth (2005) to define the GFT and subsequently generalize filtering, optimal sampling, noisy reconstruction, and randomization of graph signals. The potential gain of such tailored GSP lies in the assessment of how much graph signals respect the underlying graph structure, which can further be quantified by vertex-spectral uncertainty analysis.

A fundamental justification for considering alternative definitions of graph spectral domains can be provided by graph uncertainty principles. In the conventional setting, uncertainty is basically the trade-off between localization properties in the original and spectral domains, where the spectral domain is defined in a “universal” way (e.g., Fourier transform on a uniform grid). In the graph setting, the specification of the spectral domain plays an active role as it can depend on the graph in different ways. Following the seminal work on graph uncertainty principles for the Laplacian spectral domain (Agaskar and Lu, 2013; Tsitsvero et al., 2016), we will investigate whether modularity-derived spectral representations can be advantageous, and quantify this potential w.r.t. the strength and type of community structure.

The structure of the article is as follows. First, in Sect. 3.2.2, we start with a primer on GSP and community structure. We will discuss the modularity matrix as an alternative graph operator w.r.t. the Laplacian one. The stochastic block model (SBM) Lee and Wilkinson (2019) will be considered to reproduce prototypical adjacency matrices of graphs with assortative structure (i.e., stronger connectivity within communities than between) or disassortative structure (i.e., stronger connectivity between communities than within). Then, in Sect. 3.2.3, we present our theoretical derivations that link certain eigengaps of the spectrum-defining operator with the localization properties of the graph signal. In Sect. 3.2.4, we present in detail how these insights can be made operational for different types of SBM graphs, with varying graph community size and graph signal concentration, allowing us to identify under which conditions modularity-based spectral representation is advantageous compared to classical
3.2. Journal preprint: Graph uncertainty principle revisited: the importance of community structure

Finally, in Sect. 3.2.5, graph signal localization is explored for a real-world network representing scientific publications.

3.2.2 A primer on graph signal processing and community structure

We start with an introduction to the main principles of GSP that are most important for our purposes i.e. definitions of graphs and signals, graph Fourier transform, Laplacian and modularity operators, and the stochastic block model.

3.2.2.1 Graph signal processing and the Laplacian

Graphs are defined by a set $\mathcal{N}$ of $N$ nodes, and a set $\mathcal{E}$ of edges represented as ordered pairs $(i, j)$ where $i, j \in \mathcal{N}$. The edges are associated with positive weights that are conveniently arranged in the adjacency matrix $A$ of size $N \times N$. The entry $A_{i,j}$ is the weight of the edge running from node $j$ to $i$. We constrain further analysis to undirected graphs for which $A = A^\top$. A graph signal is represented by an $N$-dimensional vector $x \in \mathbb{R}^N$ that associates a scalar $x_i$ value to each node $i$.

Many graph properties can be assessed by the graph Laplacian operator, which is characterized by the matrix

$$L = \text{Diag}(k) - A,$$

where $k$ is the vector of nodal degrees with $k_i = \sum_{j=1}^{N} A_{i,j}$, and Diag makes a diagonal matrix with the entries of a vector. The Laplacian acts on the space of graph signals, such that $x_{\text{out}} = Lx$ is the second-order derivative of $x$. The eigendecomposition of $L$,

$$LU = U\Lambda,$$

identifies the eigenvectors as the columns of $U$ and the eigenvalues $\lambda_i$ on the diagonal of $\Lambda$. By construction, $L$ is positive-semidefinite and its eigenvalues are non-negative. The eigenvectors and eigenvalues have different interpretations, one of them being the solutions to the convex relaxation of the graph cut minimization problem (von Luxburg, 2007; Fiedler, 1973) and thus providing the basis of several spectral clustering algorithms (Gunnemann et al., 2013; Li et al., 2019). For GSP, the Laplacian eigenvectors are often taken as graph Fourier basis vectors, and the eigenvalues as squared frequency values. Consequently, the GFT of a graph signal is defined as

$$\hat{x} = U^\top x,$$

$$x = U\hat{x},$$

where $\hat{x}$ contains the spectral coefficients of $x$. The quadratic form of the Laplacian provides
an estimate of the smoothness of the graph signal:

\[ q_L(x) = x^\top L x = x^\top U \Lambda U^\top x = x^\top \Lambda \hat{x}^2 = \sum_{j=1}^{N} \lambda_k \hat{x}_k^2. \] (3.17)

Graph signals that strongly load on “low-frequency” Laplacian eigenvectors will have low \( q_L \) and vice versa. Spectral considerations using the Laplacian have led various GSP operations and concepts such as filtering and convolution (Shuman et al., 2013; Ortega et al., 2018; Sandryhaila and Moura, 2013), stationarity of signals (Perraudin and Vandergheynst, 2017), graph sampling and learning (Tanaka et al., 2020; Dong et al., 2015), and wavelets (Hammond et al., 2011; Leonardi and Van De Ville, 2013).

### 3.2.2.2 Community structure and modularity

Though the Laplacian plays a central role in graph clustering (von Luxburg, 2007), the observation of community structure (Newman, 2006; Fortunato, 2010) in many real-world networks has led to alternative graph operators. Community detection is based on accounting for a graph null model that has no preferential connectivity within or between communities; i.e., the modularity matrix (Newman, 2006; Fortunato and Hric, 2016) is defined as

\[ Q = A - \frac{1}{m} k k^\top \] (3.18)

where \( m = \sum_{i=1}^{N} k_i \) is the total edge weight, and the second term is the expected (rank-one) adjacency matrix of the configuration model, a random graph that has the same degree sequence as \( A \), but without any preferential connectivity. Hence, \( Q \) is the difference between the observed graph and the expectation of a reference random graph.

The quadratic form of the modularity matrix is associated with the modularity index

\[ q_Q(x) = x^\top Q x, \] (3.19)

which is why the eigendecomposition of the modularity matrix also serves as the spectral approach to community detection, where the eigenvector indicates the modularity index. Because in general, \( Q \) is a non-definite matrix, both positive and negative eigenvalues can occur. Eigenvectors corresponding to positive eigenvalues indicate community partitioning, while those with negative eigenvalues reveal k-partite structure (i.e., subsets of nodes with inferred edge set smaller than expected) (Newman, 2006). Since real-world networks have both positive and negative eigenvalues, it is suggested that both modular and anti-modular structures of the graph carry important information, so one mustn’t neglect the anti-modular in favor of the more commonly explored modular (Fasino and Tudisco, 2018b; Newman, 2006). A high positive (low negative) eigenvalues of \( Q \) denote the modularity (anti-modularity) of their corresponding eigenvectors, i.e. the entries of eigenvectors will vary less (more) within a community.
Community-aware GSP is exploiting the spectral domain defined by the modularity matrix $Q$ Petrovic et al. (2020). In this case, the eigenvalues do not measure smoothness or frequency, but the modularity of spectral components. When signals on graphs with community structure tend to follow this structure, modularity-based GSP leads to new meaningful approaches to filtering and processing graph signals.

### 3.2.2.3 Stochastic block model

The pattern of edge connectivity in a graph where a probability of an edge between two nodes depends on the membership of those nodes to certain subsets or blocks is described by the stochastic block model (SBM) Lee and Wilkinson (2019). The model represents a group of mathematical rules for the construction of random graph instances with predefined block structure, but it can also be used to infer the underlying edge probabilities of a given graph for the purpose of identifying its blocks.

Formally, SBM describes a graph with $N$ nodes and $K$ blocks by the $N \times K$ membership matrix $\Theta$ for which $\Theta_{i,j}$ equals 1 if $i^{th}$ node belongs to $j^{th}$ block and 0 otherwise, and by the $K \times K$ connectivity matrix $T$ for which $T_{i,j}$ gives the probability of an edge between nodes in $i^{th}$ and $j^{th}$ block, respectively. The expected adjacency matrix (Fasino and Tudisco, 2018a) of such graph is:

$$A = \Theta T \Theta^T. \quad (3.20)$$

Three distinct types of relationships between $i^{th}$ and $j^{th}$ block can be identified by $T$, more specifically, by the two diagonal entries corresponding to blocks $T_{i,i}$ and $T_{j,j}$ and one off-diagonal entry $T_{i,j}$. If both diagonal entries are high and off-diagonal low, the blocks represent communities or assortative structure. If the opposite is true, diagonal entries are both low, and off-diagonal high, then the graph contains anti-communities or disassortative structure. Finally, if $T_{i,i} \gg T_{i,j} \gg T_{j,j}$ then the core-periphery structure is present, where one densely intraconnected $i^{th}$ block - the core - is moderately connected to the $j^{th}$ block of nodes with very low intraconnectivity - the periphery.

Since communities and anti-communities fall into the broader definition of blocks, inference of SBM on a graph is sometimes used for community detection Peng et al. (2017), as it is demonstrated in Sect. 3.2.5. However, within the exploration of random graphs in Sect. 3.2.4, we make use of the SBM model of community structure to generate random graph instances. Such model is simple in the sense that all communities have the same edge probability $a$ within the blocks, and the edge probability $b$ between nodes in different communities is also fixed, thus making the block connectivity matrix write:

$$T = b 1_K 1_K^T + (a - b) I_{K \times K}, \quad (3.21)$$

where $1_K$ and $I_{K \times K}$ denote a length $K$ vector of all ones and an identity matrix of size $K \times K$, respectively. The degree sequence of the expected graph under this model, with zeros replacing
the diagonal entries of $A$ to exclude self-loops, is a constant vector:

$$A1_N = c1_N$$

$$c = \left(\frac{N}{K} - 1\right)a + \left(\frac{N}{K} - \frac{N}{K}\right)b$$  \hspace{1cm} (3.22)

### 3.2.3 Localization of graph signals

Two essential operators are needed to characterize the localization of signals in the vertex and spectral domains. Possible combinations of localization in the vertex and spectrum domains are characterized by the admissibility region.

**Definition 1** (Vertex-limiting operator). The vertex-limiting operator $D$ is a diagonal matrix with $D_{i,i} = 1$ when node $i$ is included in the set $\mathcal{D}$, and $D_{i,i} = 0$ otherwise. The set $\mathcal{D}$ is referred to as the focus set, containing the nodes with nonzero entries in $D$. For a given focus, the set of completely vertex-localized graph signals $x$ satisfies $Dx = x$.

**Definition 2** (Spectrum-limiting operator). The spectrum-limiting operator $B$ is defined by

$$B = U\Sigma U^\top,$$  \hspace{1cm} (3.23)

where $U$ contains the graph spectral components in its columns, and $\Sigma$ is a diagonal matrix with $\Sigma_{i,i} = 1$ when the $i$-th spectral component is included in the set of components $\mathcal{B}$, and $\Sigma_{i,i} = 0$ otherwise. The operator’s bandwidth refers to the number of nonzero entries of $\Sigma$, i.e. the cardinality of $\mathcal{B}$. For a given spectral support $\mathcal{B}$, the set of completely spectrum-localized graph signals $x$ satisfies $Bx = x$.

**Definition 3** (Admissible region). The admissible region $\Gamma$ is the set of pairs $(\alpha, \beta)$ that characterize signals that could exist on a given graph. The two parameters $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$ denote the vertex- and spectrum-localized energy, respectively:

$$\alpha^2 = \frac{||Dx||_2^2}{||x||_2^2}$$

$$\beta^2 = \frac{||Bx||_2^2}{||x||_2^2}$$  \hspace{1cm} (3.24)

where $D$ and $B$ are the vertex- and spectrum-limiting operators.

In Sect. 3.2.3.1, we will build upon known bounds of the region $\Gamma$ that consider the Laplacian spectrum Tsitsvero et al. (2016). A link between the area of the admissible region and certain eigengaps of the graph matrix is recognized. In relation to this, we will propose the bilateral spectrum of the modularity matrix as the support of $B$, since spectral components with a strongly negative modularity index are different from Laplacian eigenvectors with high eigenvalues typically associated with noise Shuman et al. (2013). Differences between the Laplacian and modularity matrix relevant to the eigengaps are discussed in Sect. 3.2.3.2.
3.2. Journal preprint: Graph uncertainty principle revisited: the importance of community structure

3.2.3.1 Bounds of the admissible region

We explore how the choice of eigenvectors \( U \) to preserve in \( B \) and of the focus set in \( D \) affects \( \Gamma \). The admissible region is bounded by the curves Tsitsvero et al. (2016):

\[
\begin{align*}
\cos^{-1} \alpha + \cos^{-1} \beta & \geq \cos^{-1}(\sigma_{\text{max}}(BD)) \\
\cos^{-1}(\sqrt{1 - \alpha^2}) + \cos^{-1} \beta & \geq \cos^{-1}(\sigma_{\text{max}}(BD)) \\
\cos^{-1} \alpha + \cos^{-1}(\sqrt{1 - \beta^2}) & \geq \cos^{-1}(\sigma_{\text{max}}(BD)) \\
\cos^{-1}(\sqrt{1 - \alpha^2}) + \cos^{-1}(\sqrt{1 - \beta^2}) & \geq \cos^{-1}(\sigma_{\text{max}}(BD))
\end{align*}
\] (3.26)

where \( \sigma_{\text{max}} \) is the highest singular value, and the overline denotes a matrix complement as \( B = I - B, D = I - D \), with \( I \) being the identity matrix. Expanding the area of admissible region relates to decreasing the bound \( \cos^{-1}(\sigma_{\text{max}}(BD)) \); i.e., increasing \( \sigma_{\text{max}}(BD) \), whose maximal value is 1. The following derivations are also applicable for the bounds in the three other inequalities, but we focus only on the first one as it governs the part of \( \Gamma \) where signals are jointly well-localized in vertex and spectral domains; i.e., both \( \alpha \) and \( \beta \) are close to 1.

We first investigate the effect of spectrum-limiting operator \( B \) on the upper bound of \( \sigma_{\text{max}}(BD) \). For eigenvalues \( 1 \geq \mu_1 \geq \mu_2 \geq \ldots \geq \mu_N \geq 0 \) of \( B \) it holds:

\[
\sigma_{\text{max}}(BD) \leq \sigma_{\text{max}}(B) \cdot 1 = \mu_1 \tag{3.27}
\]

We show that if \( B \) preserves relevant information about eigenvectors of Laplacian, and there is a large eigengap between successive eigenvectors \( i \) and \( j \) for which \( \Sigma_{i,i} = 1 \) and \( \Sigma_{i,j} = 0 \), then the bound \( \mu_1 \) approaches its maximum 1. Formally, we use Davis–Kahan theorem to represent the preserving of relevant information about eigenvectors.

**Theorem 1** (Davis-Kahan theorem). For given eigenvectors \( x_r \) and \( z_r \) of matrices \( X \) and \( Z \), the expression \( \sin(z_r^\top x_r) \) reflecting their closeness, is upper bounded by:

\[
\sin(z_r^\top x_r) \leq \frac{||Z - X||_F}{\min(||\zeta_{r-1} - \xi_r||, ||\zeta_{r+1} - \xi_r||)} \tag{3.28}
\]

where \( \zeta_i \) and \( \xi_i, i = 1,2,\ldots,N \) are the eigenvalues of \( X \) and \( Z \).

The closeness of eigenvectors or its tight upper bound given by the Davis-Kahan theorem reflects similar information present in both vectors. If this is true for one full-band and one band-limited signal, then we can verify that almost all relevant information is contained within the chosen spectral subband. Therefore, we assume this for eigenvectors of \( L \) (or \( Q \)) and \( B \) in our formal findings.

**Proposition 1.** Let \( \Sigma_{i,i} = 1 \) for \( i = 1,2,\ldots,r \) and \( \Sigma_{i,i} = 0 \) for \( i > r \) and let \( \sin v_j^\top u_j \) between all pairs \( j = 1,2,\ldots,N \) of eigenvectors \( v_j \) of \( B \) and eigenvectors \( u_j \) of \( L \) be close to 0. If there is a large eigengap in Laplacian eigenvalues such that \( \lambda_r \approx 0 \) and \( \lambda_{r+1} >> 1 \), then there is a large
eigengap $\mu_r - \mu_{r+1}$ between eigenvalues $\mu_j$ of $\mathbf{B}$. (Note that $\lambda_j$ and $\mu_j$ are in an increasing and decreasing order w.r.t. index $j$, respectively.)

**Proof.** Eigenvalues of $\mathbf{B}$ can be found by:

$$\begin{align*}
\mu_r &= \lambda_{r+1} \pm |\lambda_{r+1} - \mu_r| \\
\mu_{r+1} &= \lambda_r + |\lambda_r - \mu_{r+1}|. 
\end{align*} \tag{3.29}$$

Expressions within absolute value brackets are high according to Th. 1 applied to $\sin(\mathbf{v}_r^T \mathbf{u}_r)$ and $\sin(\mathbf{v}_{r+1}^T \mathbf{u}_{r+1})$. Also, we choose a positive sign in the second expression so as to avoid negative value of $\mu_{r+1}$ due to low $\lambda_r$. The eigengap $\mu_r - \mu_{r+1}$ becomes:

$$\begin{align*}
\mu_r - \mu_{r+1} &= (\lambda_{r+1} - \lambda_r) \pm |\lambda_{r+1} - \mu_r| - |\lambda_r - \mu_{r+1}|. 
\end{align*} \tag{3.31}$$

The three terms are large under given assumptions, so assuming a negative sign in front of the second term might lead to a negative overall value of the expression, which is not in line with the assumed ordering of $\mu_j$. However, taking a positive sign in front of the second term yields an eigengap $\mu_r - \mu_{r+1}$ with a positive value in the order of magnitude of one large term. □

**Remark.** If for some $r$ there is a large eigengap $\mu_r - \mu_{r+1}$, since $\mu_i < 1, i = 1, 2, \ldots, N$, then the highest eigenvalue $\mu_1$ approaches 1. Therefore, the bound of $\sigma_{\max}(\mathbf{BD})$ in Eq. (3.27) also approaches 1, rendering large admissible regions possible (though depending on $\mathbf{D}$ too).

In brief, a large eigengap between spectral components that are or are not preserved by the spectrum-limiting operator works favorably for the admissible region. The same conclusion is reached whether $\lambda_j$ and $\mathbf{u}_i$ are considered eigenvalues and eigenvectors of the Laplacian or the modularity matrix. The differences in the proof for the case of modularity matrix only relate to the order of eigenvalues, which is conventionally decreasing for $\mathbf{Q}$.

Moreover, distances between eigenvalues appearing both in the Davis-Kahan theorem and the derived expression of eigengap of $\mathbf{B}$ are absolute-value distances. This suggests that a large admissible region might be reached if both positive and negative eigenvalues of $\mathbf{Q}$ are considered. Therefore, the absolute-value ordering of modularity’s eigenvalues could be beneficial for the admissible region when choosing $\Sigma$ and $\mathbf{B}$. In other words, eigenvectors relevant for localization of signals might include both those reflecting modular and anti-modular graph structure; i.e., those corresponding to highly positive and highly negative eigenvalues. We refer to the approach of including these into the definition of $\mathbf{B}$ as the localization in the bilateral modular spectrum. In Sect. 3.2.4, we compare the effect of this kind of spectrum-limiting operator on the admissible region against conventional ones. However, in the case of Laplacian, eigenvectors with high eigenvalues are often overlocalized and perceived as noise Hata and Nakao (2017); Shuman et al. (2013). Thus, it would only make sense to include the side of the spectrum which contains smooth spectral components (low
3.2. Journal preprint: Graph uncertainty principle revisited: the importance of community structure

eigenvalues). In conclusion, the modularity matrix uniquely provides an interpretable basis for \( B \) whether it includes one or both sides of the spectrum.

We move to discuss the effect of the focus set \( \mathcal{D} \) on the bound of the admissible region, in the presence of community structure. The effect is aptly illustrated with the approximation of \( B \) formed by spectral components that define partitioning into communities. Indeed, entries of eigenvectors of \( Q \) tend to have similar values, and more specifically the same sign if the entries correspond to nodes in the same community Newman (2006). In fact, we can formalize this for the SBM model.

**Lemma 1.** For an expected graph under an SBM model with \( N \) nodes and \( K \) communities (cf. Eq. (3.21) and Prop. 2), the spectral-limiting operator \( B \) that preserves all smooth eigenvectors \( L \) or modular eigenvectors of \( Q \) has a modified block-diagonal form. All entries within the \( K \) blocks of size \( N/K \) equal \((K−1)/N \) and the remaining matrix entries are \(-1/N \).

**Proof.** See Appendix A.1.

For simplicity, we can assume large \( N \) and \( K \) and approximate \( B = U \Sigma U^T \) with:

\[
B = I_K \otimes 1_{\frac{N}{K} \times \frac{N}{K}}
\]

Then, the joint localization operator \( BDB \) writes:

\[
BDB = \text{diag}(p) \otimes 1_{\frac{N}{K} \times \frac{N}{K}}
\]

where \( p \) is a length \( K \) vector with \( p_i \) equal to the number of nodes in \( \mathcal{D} \) which belong to \( i \)-th community.

Due to the block diagonal form of \( BDB \), this matrix has as many eigenvectors corresponding to non-zero eigenvalues as there are non-zero values of \( p \), that is, the number of communities over which the focus set spans. These eigenvectors \( \delta \) have entries equal to 1 for nodes belonging to the corresponding community, and 0 otherwise. This can be formalized for the eigenvector corresponding to the \( i \)-th community as \( \delta_i(j) = \delta_{C(j),i} \), where \( \delta \) is the Kronecker delta, and \( C(j) \) gives the community to which node \( j \) belongs. Then, the eigenvector equation writes:

\[
BDB\delta_i = (\text{diag}(p) \otimes 1_{\frac{N}{K} \times \frac{N}{K}})\delta_i = p_is_i\delta_i
\]

where \( s_i \) is the number of nodes belonging to the \( i \)-th community. In the case of all equal-sized communities, it is clear that the highest eigenvalue is \( \max_i(p_i) \), meaning that the bound of the admissible region is highest when a whole community is encompassed by \( \mathcal{D} \). Note that in previous equations, this value would be \( \sqrt{\frac{N}{K}} \), but for a proper unit-normalization of \( \delta \), it approaches the value of 1. For a graph with communities of different sizes, the highest eigenvalue of \( BDB \) becomes \( \max_i(p_is_i) \), which means the effect of community size is also
non-negligible. The largest admissible region is reached when the vertex-limiting operator focuses on all nodes of the largest community.

### 3.2.3.2 Eigengaps of Laplacian and modularity matrix

In the case of a graph with a strong community structure, Laplacian eigenvectors corresponding to the smallest non-zero eigenvalues will relate to the clustering properties of the graph; i.e., those minimizing the cut size von Luxburg (2007). Specifically, since the constant eigenvector (all entries equal to 1) has the smallest eigenvalue 0, and assuming a connected graph, the clustering partitions are reflected in the eigenvectors with the second to $K$th smallest eigenvalues, in a graph with $K$ communities. However, the eigenvectors of modularity matrix with the $K-1$ highest eigenvalues determine the community partitions in greater accuracy than the Laplacian eigenvectors Newman (2006). The set of these eigenvectors can be extracted as $U_r$ to explore the admissible region $\Gamma$ of modular or smooth signals on a community. Since the area of $\Gamma$ increases with the relevant eigengaps according to Th. 1, we are interested in the particular eigengaps $\eta$ of $L$ and $Q$ (Note that $\lambda^L_i$ and $\lambda^Q_i$ are in the ascending and descending order, respectively):

$$\eta(L) = \lambda^L_{K+1} - \lambda^L_K,$$  \hspace{1cm} (3.35)  

$$\eta(Q) = \lambda^Q_{K-1} - \lambda^Q_K. \hspace{1cm} (3.36)$$

We proceed to evaluate $\eta$ for a graph with $K$ communities in order to distinguish properties of different Fourier spectra that affect the admissible region in the presence of communities. We leverage the SBM model of random graphs to reflect communities. First, we find that the eigengaps $\eta(L)$ and $\eta(Q)$ are equal for the expected random graph (Prop. 2). Next, we compare the eigengaps of an instance of the random graph to the expected graph in the limit of a large number of nodes (Props. 3 and 4).

Following (3.21), the expected Laplacian matrix writes:

$$L = \text{diag}(A1_N) - A \hspace{1cm} (3.37)$$

and the aforementioned eigengap equals Ramakrishna et al. (2020):

$$\eta(L) = \frac{N}{K} (a + (K-1)b) - Nb = \frac{N}{K} (a - b) \hspace{1cm} (3.38)$$

where the smooth and non-smooth eigenvectors concentrate at the subspaces with eigenvalues $Nb$ and $\frac{N}{K} (a + (K-1)b)$, respectively. A similar result can be derived for the expected modular and anti-modular spectral components.

**Proposition 2.** We consider an SBM graph with communities of the same size and strength as in (3.21). The expected adjacency matrix is defined by replacing diagonal entries $A_{i,i}$ of the matrix in (3.20) with zeros, in order to account for the absence of self-loops. Then, the expected
modularity matrix:

\[
Q = A - \frac{A_1 N 1_N^T A}{1_N A 1_N^T}
\]  

(3.39)

has modular eigenvectors concentrated at the positive (assuming \(a > b\) and \(N \gg K\)) eigenvalue \(\frac{N}{K} (a - b) - a\), and the anti-modular at the negative \(-a\).

Proof. For the expected SBM graph defined according to Eqs. (3.20) and (3.21), the degree sequence is constant (cf. Eq. (3.22)). Therefore, the Laplacian corresponds to the negative adjacency matrix shifted by the identity scaled with \(c\) (cf. Eq. 3.37). Further, modularity matrix writes:

\[
Q = A - \frac{A_1 N 1_N^T A}{1_N A 1_N^T} = A - \frac{c}{N} 1_N 1_N^T
\]  

(3.40)

and corresponds to the adjacency rank-one modified by its constant eigenvector’s outer product. Therefore, the eigenvectors of \(L\) and \(Q\) are the same but have different eigenvalues. In fact, except for the one matching the eigenvector \(1_N\), the eigenvalue of a modular (anti-modular) eigenvector equals the difference between \(c\) and the eigenvalue of the corresponding smooth (non-smooth) eigenvector. This leads to values of \(\frac{N}{K} (a - b) - a\) for modular, and \(-a\) for the anti-modular spectral components. (The constant vector of ones corresponds to eigenvalue 0 for both matrices.)

Remark. Given the eigenvalues of the modularity matrix of the expected random graph, the eigengap \(\eta(Q)\) writes:

\[
\eta(Q) = \frac{N}{K} (a - b) - a - (-a) = \frac{N}{K} (a - b).
\]  

(3.41)

The eigengap of modularity matrix equals the one of the Laplacian (\(\eta(L) = \eta(Q)\)). This result points out that the differences between Laplacian and modularity eigengaps could become apparent only in the case of an instance of a random graph. An instance does not necessarily have a constant vector as the degree sequence as the expected graph does (Eq. (3.22)). It is important to account for these non-constant degrees in the computation of eigengaps. Therefore, in the following results, we modify the SBM model of random graphs.

Definition 4 (Model \(\mathcal{G}\) of random graphs). A random graph \(\mathcal{G}\) with \(N\) nodes, \(K\) communities and no self-loops, has the same adjacency matrix \(A\) of the expected graph \(E(\mathcal{G})\) as the SBM model in Prop. 2 (Eq. 3.21). The degree sequence of an instance of \(\mathcal{G}\) is determined by \(A_1 N + h\), where \(h\) is a length \(N\) zero-mean realization of a random variable from a log-normal distribution, known to be the degree distribution of many networks Broido and Clauset (2019). Though a graph with degrees \(A_1 N + h\) does not precisely correspond to an adjacency \(A\), it is assumed that the adjacency of the graph instance is also \(A\). Indeed, at the limit of large \(N\) the differences in the entries of \(A\) are negligible as compared to the order of magnitude of entries of \(h\). Furthermore, the variance of the distribution of \(h\) is considered low enough so that the graph instances do not exhibit negative degrees.
Proposition 3. In the limit of large graphs, the eigengap \( \eta(Q(G)) \) of the random graph instance approaches the eigengap of the expected graph \( \eta(Q(E(G))) \), or:

\[
\lim_{N \to \infty} \eta(Q(G)) = \eta(Q(E(G)))
\]  
(3.42)

Proof. Given the properties of \( G \), the modularity matrix of a random graph instance writes:

\[
Q(G) = A - \frac{(A1_N + h)(1^\top_N A + h^\top)}{1^\top_N A1_N + h^\top 1_N}
\]  
(3.43)

which boils down to:

\[
Q(G) = Q(E(G)) - \frac{hh^\top}{cN}
\]  
(3.44)

where \( Q(E(G)) \) is given by (3.39).

Under the Weyl’s theorem Weyl (1912) of perturbed matrices, the difference between eigenvalues of \( Q(G) \) and \( Q(E(G)) \) (assuming the same ordering) is upper bounded by the highest eigenvalue in absolute value (spectral norm) of the difference matrix, or formally:

\[
|\lambda_i(Q(G)) - \lambda_i(Q(E(G)))| \leq \left\| \frac{hh^\top}{cN} \right\|,
\]  
(3.45)

where \( \| \cdot \| \) denotes the spectral norm. Since this norm is always less than or equal to the Frobenius norm of a matrix, we give the limit of this upper bound at large \( N \):

\[
\lim_{N \to \infty} \|Q(E(G)) - Q(G)\|_F = \lim_{N \to \infty} \sqrt{\sum_{i,j=1,...,N} \left( \frac{h_i h_j}{cN} \right)^2} = 0.
\]  
(3.46)

Since the upper bound of the difference between eigenvalues approaches 0, so does the difference between any of the eigengaps, leading to the claim in (3.42).

The eigengaps in (3.35) of both modularity and Laplacian matrix tend to be larger if the graph exhibits more prominent communities Budel and Mieghem (2020). The result from Prop. 3 confirms that in the presence of communities, the modularity matrix of a graph preserves that large eigengap of the expected graph. Therefore, according to Prop. 1, eigenvectors of modularity are a relevant choice of spectral components in contexts where a large admissible region is desirable. Though the Laplacian approach could still yield a large \( \Gamma \) in certain cases, a graph instance’s eigengap of Laplacian matrix does not follow tightly the one of the expected graph.

Proposition 4. Under the same conditions as in Prop. 3, the Frobenius upper bound of the difference between eigenvalues of a large random graph and the expected one diverges to infinity.
3.2. Journal preprint: Graph uncertainty principle revisited: the importance of community structure

in the case of Laplacian matrix, or:

\[
\lim_{N \to \infty} ||L(G) - L(E(G))||_F = \infty
\]  

(3.47)

Proof. The Laplacian matrix \(L(G)\) of a random graph writes:

\[
L(G) = \text{diag}(h) + \text{diag}(A1_N) - A = L(E(G)) + \text{diag}(h)
\]  

(3.48)

where \(L(E(G))\) stands for the expected Laplacian from (3.37). Then, considering Weyl’s theorem, the bound on the difference between eigenvalues writes:

\[
|\lambda_i^{L(G)} - \lambda_i^{L(E(G))}| \leq ||\text{diag}(h)|| = \max_j |h_j|.
\]  

(3.49)

The bound \(\max_j |h_j|\) corresponds to the difference between the highest and the mean degree, or between the mean and the lowest, depending on which difference is greater. Since this value might be quite high, there is no assurance that the eigengap \(\eta(L(G))\) will not be significantly smaller than \(\eta(L(E(G)))\), no matter the graph size \(N\). Furthermore, we can leverage the fact that \(h\) is drawn from a log-normal distribution. Indeed, \(\max_j |h_j|\) follows the generalized extreme value (GEV) distribution, specifically Gumbel distribution, under the Fisher–Tippett–Gnedenko theorem. Therefore, it can be evaluated as the median of the corresponding GEV distribution. Let’s recall that if a random variable follows a normal distribution with variance \(\sigma^2\), then the maximal value of its sample follows the Gumbel distribution with median growing as \(\sim \sigma \sqrt{2\log N}\). Also, the exponential of that variable follows the log-normal distribution, so we can approximate the growth of the bound in (3.49) with Kinnison (1983):

\[
\max_j |h_j| \sim e^{\sigma \sqrt{2\log N}}.
\]  

(3.50)

This quantity increases with \(N\) and diverges for \(N \to \infty\).

Practical implication of (3.50) is that for large graphs, the extremal value \(\max_j |h_j|\) is finite, and very high. Consequently, the upper bound of the difference between eigenvalues of a graph instance and the expected graph is not tight. Therefore, the relevant eigengap that should ensure a large admissible region might be unacceptably small.

Remark (Alternative perspective on Props. 3 and 4). The eigenvalues \(\lambda_i^{L(E(G))}\) and \(\lambda_i^{Q(E(G))}\) of the expected graph can be considered to be the true mean value of some random variable estimated by the averages of eigenvalues \(\lambda_i^{L(G)}\) and \(\lambda_i^{Q(G)}\) over random graph instances. Then, it is only in the case of modularity matrix that this estimator approaches minimum (zero-valued) variance estimator for large \(N\), suggesting a graph description consistent with the SBM. As a consequence, \(\sigma_{\text{maxBD}}\) tends to have lower variance across graph instances as shown in the empirical results in the following section (Figs. 3.6 and 3.7).

In summary, large graphs exhibiting communities are more likely to have a large admissible
Chapter 3. Informing signal processing techniques with community structure

region if the spectrum-limiting operator focuses on the modular eigenvectors rather than on the smooth Laplacian eigenvectors, due to eigengap preservation. However, note that the results about eigengaps above did not take into consideration the effect of the vertex-limiting operator $D$. In the following section, we explore how $D$ can align with this spectral preference, or change it in favor of the Laplacian. We corroborate the findings on the admissible region and the eigengaps by empirical results on synthetic graphs generated under SBM.

3.2.4 Analysis of random graphs

3.2.4.1 Main findings

The following analysis is dedicated to verifying the theoretical results by considering generated graph instances. Several key points are being examined and corroborated:

- In finite-size graphs with community structure, the eigengaps relevant for admissible region tend to be higher for the modularity matrix than for the Laplacian.
- When the focus set is a community, the size of the admissible region is proportionally linked to the eigengaps of the graph matrix defining the spectral domain.
- Effect of the choice of modularity or Laplacian spectrum to the admissible regions is more prominent for the extremal cases of the vertex-limiting operator, i.e. when the focus set is either exactly a community or a random subset of nodes with no connection to the community structure. When the focus set is a community, the modularity matrix provides a larger admissible region in the presence of weak modular, or bilateral modular structure; otherwise Laplacian spectrum is the preferred choice.
- Preserving bilateral modular spectrum by the spectrum-limiting operator allows for better vertex-spectrum localization of graph signals as opposed to focusing only on the modular spectrum - when the graph contains anti-communities.

3.2.4.2 Data and methods

In order to explore the effect of vertex-limiting operator on the admissible region, we generated 50 graph instances with 10 communities, each with 50 nodes, under four different SBM models: $(a = 0.9, b = 0.1)$, $(a = 0.8, b = 0.2)$, $(a = 0.7, b = 0.3)$, and $(a = 0.3, b = 0.7)$. The value of $\sigma_{\text{max}}(BD)$ was calculated for all graphs, with $D$ preserving different percentage of nodes in one fixed community, such that the value of 2% corresponds to a random focus set, and 100% corresponds to the focus on the community. The operator $B$ preserves the first $K − 1$ eigenvectors that are most smooth (lowest eigenvalues of $L$), most modular (highest eigenvalues of $Q$), or most modular or anti-modular (eigenvalues of $Q$ with highest absolute value). Results are reported in Fig.3.5.

The rest of the results in this section (Figs. 3.6 to 3.9) came from the analysis of extremal
cases of the focus set being either a community or a random subset. The admissible regions of random SBM graphs have been evaluated within the space of parameters encompassing community strength \((a\) and \(b\)) and community size \(N/K\). We generated 50 graph instances for each point in this parameter space. All graphs had \(K = 10\) communities whose size was chosen as one of the 10 points in the logarithmic range from 10 to 100 nodes. Edge probabilities \(a\) and \(b\) were shifted from 0.1 to 0.9 with linear step 0.05. Bounds \(\sigma_{\text{max}}(\mathbf{BD})\) were calculated for every graph, and for the three types of spectra with fixed bandwidth \(K - 1\). Furthermore, eigengaps \((3.35)\) were computed for all generated graphs in order to explore their connection to the admissible region.

### 3.2.4.3 Overlap between focus set and a community

The importance of vertex-limiting operator for the admissible region is conveyed through the dependence of the region’s bound on the overlap between the focus set and a community (Fig. 3.5). The choice of a particular matrix spectrum seems to have little effect on this dependence for graphs with strong modular structure \((a = 0.9, b = 0.1)\). However, already at middle-strength communities \((a = 0.8, b = 0.2)\) we observe lower variance of \(\sigma_{\text{max}}(\mathbf{BD})\) for modular and bilateral modular approach compared to the Laplacian, with the difference between variances becoming significantly greater for the case of weak communities \((a = 0.7, b = 0.3)\). In fact, in the presence of weak communities, the Laplacian-based bound shows a horizontal trend of its mean value (across graph instances with the same model parameters), suggesting no relevant connection to the choice of \(\mathbf{D}\). On the other hand, the modularity matrix still exhibits low variance and an increasing trend w.r.t. the overlap between focus set and the community, across the weaker-community graphs. Furthermore, the increase of the bound of \(\Gamma\) as the focus becomes a community is preserved by the bilateral modular approach regardless of the type (assortative or disassortative) and the strength of community structure. The other two types of spectra cannot capture such dependence if anti-modular communities are present in the graph \((a = 0.3, b = 0.7)\).

According to the trends of \(\sigma_{\text{max}}(\mathbf{BD})\) for \((a = 0.7, b = 0.3)\) and \((a \leq 0.7, b \geq 0.3)\), the discriminating effect of the type of spectrum on the admissible regions becomes most apparent for the cases when \(\mathcal{S}\) is either a full community or a random subset of nodes. (Note the difference between the mean values.) This is the reason why the following results are only provided for these two cases.

### 3.2.4.4 Connection between eigengaps and admissible region

Probability density functions (PDF) of eigengaps \(\eta\) from \((3.35)\) are estimated since according to Prop. 1, the eigengaps are relevant for the size of \(\Gamma\). Thus, we are interested in comparing them for the Laplacian and modularity matrix (Fig. 3.6) under different SBM models: \((a = 0.9, b = 0.1)\), \((a = 0.8, b = 0.2)\), and \((a = 0.7, b = 0.3)\). As shown by \((3.38)\) and Prop. 2, the expected eigengaps of both matrices are \((N/K)(a - b)\). The PDFs are located below these for all cases,
Chapter 3. Informing signal processing techniques with community structure

Figure 3.5 – Size of $\Gamma$ as reflected by $\sigma_{\text{max}}(BD)$ over percentage of the focus set preserved by $D$ that falls into the same community for cases ($a = 0.9, b = 0.1$), ($a = 0.8, b = 0.2$), ($a = 0.7, b = 0.3$), and ($a = 0.3, b = 0.7$). Blue, orange and red plots reflect the cases when $B$ extracts smooth, modular and bilateral modular spectral components, respectively. Shaded areas correspond to $\pm$ standard deviation around the mean over graph instances of the same model. The graph has 10 communities, each comprising of 50 nodes.

but they do follow their increasing trend as the community size $N/K$ increases. The plots demonstrate that $\eta(Q)$ is higher, and closer to the expected gap than $\eta(L)$, with the difference $\eta(Q) - \eta(L)$ increasing as the communities get weaker. This is a consequence of the graph’s community structure, since the definition of the modularity matrix is in fact tailored to access communities. Furthermore, Props. 3 and 4 are corroborated by PDFs of $\eta(Q)$ being narrower, with the variance much lower than that of PDFs of $\eta(L)$.

Next, we examine (for the same three SBM models) the connection between $\eta$ and the bound of the admissible region $\sigma_{\text{max}}(BD)$. In Fig. 3.7, we compare the cases of Laplacian (blue) and modularity matrix (orange) through the scatter plots of $\sigma_{\text{max}}(BD)$ over $\eta$. When the vertex-limiting operator preserves a community (Fig. 3.7(a)), the bound does increase with the eigengap of modularity matrix. Though this is also somewhat true for the Laplacian in the cases of ($a = 0.9, b = 0.1$) and ($a = 0.8, b = 0.2$), the scatter of $\sigma_{\text{max}}(BD)$ at a single eigengap is much greater than for the modularity matrix, especially for low eigengaps (smaller graphs). Since the variance of eigengaps of $Q$ are lower (Fig. 3.6), so is the variance of $\sigma_{\text{max}}(BD)$. This makes the effect of the eigengaps to the bound, and their proportionality, much clearer. For the graphs with weaker communities ($a = 0.7, b = 0.3$), an increasing trend is not there for the Laplacian, but modularity still recognizes the eigengaps as relevant for $\sigma_{\text{max}}(BD)$. 

50
Figure 3.6 – Violin plots of eigengaps $\eta$ in Eq. (3.35) over different graph sizes $N$ for cases $(a = 0.9, b = 0.1)$, $(a = 0.8, b = 0.2)$, and $(a = 0.7, b = 0.3)$. The plots include probability density functions (PDF) of $\eta$ given by the Gaussian kernel estimator, with their means denoted by horizontal lines. The blue plots correspond to the case of Laplacian matrix, and orange ones to the modularity matrix. Values of eigengaps $(N/K)(a - b)$ expected under the SBM model are represented by asterisks.

The scatter patterns are very different under the condition that the focus is on a random subset of nodes. In this case (Fig. 3.7(b)), the same observation remains on the less variability of $\sigma_{\text{max}}$(BD) for modularity matrix, but no increasing trend is observed. The Laplacian bounds scatter widely, whereas modularity-based bound tends to fix around a certain value ($\sim 0.4$) regardless of $\eta$. In other words, the effect that eigengaps, being the spectral proxies of the presence of community structure Budel and Mieghem (2020), have on the joint vertex-spectrum localization becomes apparent in the spectrum only if the vertex-limiting operator recognizes a community. This relates to the intrinsic connection between the two seemingly-different domains, and the importance of the community structure to that connection.

3.2.4.5 Effect of the type of community structure

We further wish to compare the size of the admissible region across the communities’ strength and type. These are reflected by the SBM parameters $a$ and $b$. If $a > b$, the communities are modular and stronger as the difference $a - b$ increases, whereas they become more anti-modular as $b$ becomes greater than $a$. Fig. 3.8 shows surf plots of the differences between modular- and smooth-based $\sigma_{\text{max}}$(BD), as well as the differences between bilateral modular- and smooth-based $\sigma_{\text{max}}$(BD) across the combinations of $a$ and $b$ in the range from 0.1 to 0.9.
Chapter 3. Informing signal processing techniques with community structure

Figure 3.7 – Connection between the bound of the admissible region $\sigma_{\text{max}}(\mathbf{BD})$ and eigengaps $\eta$ for cases ($a = 0.9, b = 0.1$), ($a = 0.8, b = 0.2$), and ($a = 0.7, b = 0.3$). Scatter plots consist of blue and orange points, for which $K-1$ most smooth eigenvectors of $\mathbf{L}$ and most modular eigenvectors of $\mathbf{Q}$ are preserved by $\mathbf{B}$, respectively. The vertex-limiting operator $\mathbf{D}$ extracts (a) single community or (b) a random set of nodes of the same size as a community.

When $\mathbf{D}$ preserves a community (Fig. 3.8(a), left), the modular approach outperforms the smooth with higher $\sigma_{\text{max}}(\mathbf{BD})$ in almost all cases where $a > b$, i.e. where the nature of the community is modular. In fact, size of this region only approaches $a - b > 0$ for large $N$. Fig. 3.8 shows the case of graph size $N = 600$, which settles the area where modularity outperforms Laplacian at around $a - b > 0.2$. The boundaries (red curves in Fig. 3.8) between regimes where either approach is more useful are interpolated using zero-crossings. On the other hand, the negative values in the right upper half of the plot, suggest that the Laplacian is the preferred choice when there is an anti-modular structure present ($a - b < -0.2$) or no community
3.2. Journal preprint: Graph uncertainty principle revisited: the importance of community structure

Figure 3.8 – Differences between $\sigma_{\text{max}}(\text{BD})$ for the cases of localization in modular and smooth (left), and bilateral modular and smooth (right) spectral components (mean over graph instances for graph size $N = 600$) over various SBM models ($a$ and $b$ both in range from 0.1 to 0.9). The vertex-limiting operator $D$ extracts (a) a single community or (b) a random set of nodes of the same size as a community. Red curves denote the boundaries between graph structure modes, as estimated by zero-crossings.

structure at all ($a - b \approx 0$). A second boundary is observed at approximately $a - b > 0.6$, meaning that for graphs with strong communities, both types of spectra lead to similar results, corroborated by values close to 0 in the lower-left corner. When the bilateral modular spectrum is preserved (Fig. 3.8(a), right), the benefits of modularity spectrum over Laplacian become apparent for approximately $a - b < -0.2$ which is the case when the communities have fewer edges within itself than outside. This means that whereas smooth eigenvectors capture this type of structure better than modular ones, the best signal localization is reached in the case of bilateral spectrum, because it includes the anti-modular components which exactly reflect the described community type. Slightly higher $\sigma_{\text{max}}(\text{BD})$ is also observed even within $a - b > 0.2$, in the area denoting the weaker (modular) communities.

When $D$ focuses on a random set of nodes, accessing modular and bilateral modular spectrum leads to $\sigma_{\text{max}}(\text{BD})$ at around 0.4 across all values of $a$ and $b$. This is why the differences of $\sigma_{\text{max}}(\text{BD})$ with those in the case of Laplacian spectrum are almost identical (Fig. 3.8(b)).
All values of the bound are strictly negative, so the curves denoting the boundary between regimes are relying on the conspicuous crossing of the value $-0.1$, instead of the regular zero-crossing. This means that when the focus set is not a community, the Laplacian provides a larger admissible region for all underlying models of structure. Only in the case of strong communities (approximately $a - b > 0.6$), the effect of modularity spectrum becomes clearer, but even then only comparable to that of the Laplacian. For the modularity spectra to provide information not available to Laplacian, the focus has to be on the community, as it was previously suggested by the eigengaps (Fig. 3.6).

### 3.2.4.6 Effect of finite graph size

Since the results in Props. 3 and 4 consider the limit of large graphs ($N \to \infty$), it is important to investigate the effect of a finite graph size on the admissible region. We report the dependence of $\sigma_{\text{max}}(\text{BD})$ on $N$ in Fig. 3.9 for SBM models $(a = 0.9, b = 0.1)$, $(a = 0.8, b = 0.2)$, $(a = 0.7, b = 0.3)$, and $(a = 0.3, b = 0.7)$. When the focus set is a community (Fig. 3.9(a)), the bounds increase with $N$. The modular approaches reach a larger admissible region than the Laplacian-based as the community strength declines, similar to the results concerning the effect of $D$ and the eigengaps (Figs. 3.5 and 3.7). Moreover, once the nature of a community shifts to anti-modular as in $(a = 0.3, b = 0.7)$, the bilateral modular spectrum leads to significantly higher $\sigma_{\text{max}}(\text{BD})$ than either of the two other types of spectra, as in Fig. 3.5. This supports the reasoning behind absolute value based ordering of eigenvalues suggested previously in relation to Prop. (1). For large graphs, all bounds saturate at around 0.9 in the case of stronger communities, and so does the bound for bilateral modular spectrum in the case of anti-modular communities. This corroborates that a stable limit of large graphs (reached already for graphs with few hundreds of nodes) exists in line with Prop. 3, whereas high variability of the Laplacian bound at $(a \leq 0.7, b \geq 0.3)$ supports Prop. 4.

Note that in the case of the focus being a random set (Fig. 3.8(b)), the bounds decrease with $N$. The Laplacian approach renders higher $\sigma_{\text{max}}(\text{BD})$ as the community strength decreases but with much higher variance. These observations follow the previous findings that induce the need to adjust the spectrum-defining matrix to the vertex-limiting operator and the type of community structure present in the graph. Indeed, the finite graph size does not change the theoretical findings regarding the expected size of the admissible region (Sec. 3.2.3).

### 3.2.5 Surrogate analysis of a publications graph

Beyond random graphs, the sizes of admissible regions under the modularity and Laplacian spectra are explored for a graph of scientific publications. A presented use-case scenario should illuminate the practical side of inspection of $\Gamma$ in the analysis of real-world networks. By means of surrogate testing, we determined, for different spectral and vertex localization operators, whether the networks’ admissible regions can be explained by the inferred SBM model, and further, how this depends on the choice of the spectrum.
3.2. Journal preprint: Graph uncertainty principle revisited: the importance of community structure

Figure 3.9 – Size of $\Gamma$ as reflected by $\sigma_{\text{max}}(\mathbf{BD})$ over graph size $N$ (mean across graph instances of the same size) for cases ($a = 0.9$, $b = 0.1$), ($a = 0.8$, $b = 0.2$), ($a = 0.7$, $b = 0.3$), and ($a = 0.3$, $b = 0.7$). Blue, orange and red plots correspond to $\mathbf{B}$ preserving smooth, modular and bilateral modular spectrum, respectively. Shaded areas denote $\pm$ standard deviations around the means over graph instances. The vertex-limiting operator $\mathbf{D}$ focuses on (a) a single community or (b) a random set of nodes of the same size as a community.

The nodes of the explored network were extracted as publications from Google Scholar Cholewiak et al. (2020), where 100 of them were taken as the results of the search query "brain mri", and 100 of the "brain pet". The queries were chosen so as to expect two communities of publications, whose presented work in brain research relates to two different scan modalities. A graph's edge with weight 1 was set to connect two nodes if the underlying publications' titles overlap in 5 or more words. Finally, the largest connected component was extracted resulting
in an undirected binary graph with 156 nodes and 749 edges. The inferred SBM model Peng et al. (2017) recognized two communities with sizes of 84 and 72 nodes, and intracommunity edge probabilities of 0.1719 and 0.1618. The intercommunity edge probability was estimated to be approximately 11 times lower at 0.0156.

Fifty thousand surrogate graphs were generated under the inferred model, and the bound $\sigma_{\max}(\mathbf{B}\mathbf{D})$ was calculated for both the given graph and its surrogates. Three types of spectral components (modular, smooth, and bilateral modular) were extracted by $\mathbf{B}$ with bandwidths from 1 to 30, and operator $\mathbf{D}$ preserved 84 nodes out of which 1 to 84 belonged to the first community. Finally, for every pair of parameters (bandwidth and $\mathbf{D}$), a test was conducted by comparing $\sigma_{\max}(\mathbf{B}\mathbf{D})$ of the publications graph to the 95th percentile of the surrogates’ $\sigma_{\max}(\mathbf{B}\mathbf{D})$ (Fig. 3.10). The test’s null hypothesis is that the size of the admissible region is as expected by a random graph under an SBM model of its communities.

In the cases of modular and bilateral modular spectral components, the bound increases with the focus set approaching either of the two communities, but not in the case of Laplacian (Fig. 3.10(a); sides at which the percentage equals 0% or 100%). This aligns with the discussion revolving (3.34) since a focus set spreading across different communities is not expected to yield a large admissible region. However, for random graphs, this remains true for those with weaker communities (Fig. 3.5), whereas the publications graph can be considered to have strong communities ($a/b \approx 11$). This points to the fact that remarks on the admissible region given in previous sections extend to graphs with strong community structure probably due
3.2. Journal preprint: Graph uncertainty principle revisited: the importance of community structure

to specific traits of real-world networks (unequal sizes of communities, non-constant degree sequence, etc). Furthermore, there is an unclear preference of higher $\sigma_{\text{max}}(\text{BD})$ when one community is the focus than when the other is - in the case of Laplacian spectrum. The almost symmetric plots of $\sigma_{\text{max}}(\text{BD})$ (around the axis where percentage equals 50%) suggests equal treatment of both communities by the modularity matrix.

Results of the surrogate testing (Fig. 3.10(b)) show that across different spectra, the hypothesis is rejected within different subspaces of the two parameters. For modular spectrum, rejection occurs mainly when the focus set is almost completely overlapping with any of the communities. This supports the previous finding that modular spectral components combined with the focus being a community yield a large $\Gamma$, by emphasizing that its size is larger than in a random graph. In the case of the Laplacian, the null hypothesis is rejected when the bandwidth is higher than around 10 regardless of the focus set. This suggests that the graph Laplacian does indeed contain information relevant for a large admissible region, but remains blind to the community structure. Further, it supports the previous finding on its preference of a random subset of nodes as the focus (Fig. 3.8(b)). In the case of the bilateral modular spectrum, rejection occurs when the bandwidth is high and the focus is on one particular community. Similar to the values of $\sigma_{\text{max}}(\text{BD})$ for the Laplacian, there is an obscure asymmetry between focuses on the two communities. Possibly, the anti-modular motifs at different size scales of the graph topology are unequally dispersed between the modules, thus unevenly changing the overall nature of communities. This stresses out an important role of anti-modular components in the spectral content of localized signals, even when the graph has a conspicuous modular structure (as shown by the inferred SBM).

3.2.6 Conclusion

As the theoretical and random graph analysis showed, the community structure of a graph, including specific features such as communities’ type, size, and strength, greatly affect how wide is the range of jointly localized signals that can exist on a given graph. Depending on the aforementioned parameters, certain structural modes tend to be more appropriately accessed by either Laplacian or modularity matrix, thus determining the right choice of the spectrum in which a signal localizes. The main result recognizes (bilateral) modular or smooth spectral components as more convenient depending on whether vertex localization presents in a random subset of nodes or a community (and the nature of that community). An explanation is offered by the fact that the spectra access different features of signals, or specifically, the modular (anti-modular) components explore the hypothesis of expecting similar (different) signal values across nodes in the same community, whereas the smooth ones allow for similar values across connected nodes. In summary, in order to go beyond the possible limitations of the uncertainty principle, the spectral domain of signals should be chosen such that it accounts for the underlying structure.

The results in this article relate to the simple SBM model of a graph with several communities
of the same sizes and edge probabilities (similar edge densities). A more extensive analysis encompassing different structural graph models is needed to provide insight into the full extent of the structure’s effect on signal localization. Some of the models could include communities of very different edge densities, overlapping communities Shen et al. (2009), or core-periphery and hierarchical structure Schaub and Peel (2020). Note that community detection using modularity matrix is limited by community size Lancichinetti and Fortunato (2011). This calls for a potential need to modify $Q$ by scaling the null model matrix Reichardt and Bornholdt (2006) before setting its eigenvectors as the spectral components.

Further work on the topic can be directed to signal decomposition into orthogonal constituents optimally localized in both vertex and spectral domain, or the so-called graph Slepians. The Slepian basis has proved beneficial for multi-taper spectral analysis, compressive sensing Dav- enport and Wakin (2012), geophysics Simons et al. (2009), etc. Besides the graph Slepians based on energy concentration principle, found as eigenvectors of $BDB$, useful orthogonal bases can be found by the inclusion of eigenvalue scaling of spectral components as in Van De Ville et al. (2017), or by combining the two principles as in Petrovic et al. (2019), but adapted to the community structure modeled by modularity matrix. However, any Slepian condition other than $BDB$ leads to different bounds of the admissible region than the one explored here, thus requiring once again to revisit the uncertainty principle. A notable alternative is the vertex-spectrum localization operator $\cos(\theta)D + \sin(\theta)B$ for a certain angle $\theta$ Erb (2021). Whatever the chosen basis is, decomposition of graph signals into it presents a way to assess if a signal that localizes in a community does indeed tend to do so because of its spectral features imposed by the community structure.

Acknowledgments

This work was supported in part by the Swiss National Science Foundation (200021 17 5506).
4 Processing global signals in local environment

The work presented in this chapter makes use of graph Slepians (Van De Ville et al., 2017) – bandlimited signals with localized energy on the subset of vertices – in order to guide the signal processing towards a subgraph of interest. The first section is a published journal article where a new type of graph Slepians was designed. The criteria of local energy and local smoothness are being merged into a single computationally stable Slepian criterion. The second section is a conference proceeding article that introduces a Fourier domain built from graph Slepians and discusses filtering of graph signals within such domain.
4.1 Journal article: Guided graph spectral embedding: application to the *c. elegans* connectome

(This article has been published in Network Neuroscience, DOI: 10.1162/netn_a_00084)

Miljan Petrovića,b, Thomas A. W. Boltona,b, Maria Giulia Pretia,b, Raphaël Liégeois,a,b, Dimitri Van De Ville a,b

a Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, Campus Biotech, Geneva, Switzerland
b Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

Abstract

Graph spectral analysis can yield meaningful embeddings of graphs by providing insight into distributed features not directly accessible in nodal domain. Recent efforts in graph signal processing have proposed new decompositions—e.g., based on wavelets and Slepians—that can be applied to filter signals defined on the graph. In this work, we take inspiration from these constructions to define a new guided spectral embedding that combines maximizing energy concentration with minimizing modified embedded distance for a given importance weighting of the nodes. We show that these optimization goals are intrinsically opposite, leading to a well-defined and stable spectral decomposition. The importance weighting allows to put the focus on particular nodes and tune the trade-off between global and local effects. Following the derivation of our new optimization criterion, we exemplify the methodology on the *C. elegans* structural connectome. The results of our analyses confirm known observations on the nematode’s neural network in terms of functionality and importance of cells. Compared to Laplacian embedding, the guided approach, focused on a certain class of cells (sensory neurons, interneurons or motoneurons), provides more biological insights, such as the distinction between somatic positions of cells, and their involvement in low or high order processing functions.

*Keywords*: spectral graph domain, graph embedding, low-dimensional space, focused connectomics

4.1.1 Introduction

Many aspects of network science relate to graph partitioning—the grouping of nodes in subgraphs—and graph embedding—their representation in a low-dimensional space that accounts for graph topology (von Luxburg, 2007). Spectral graph theory motivates analytical methods based on the eigenvectors of fundamental graph operators, such as the adjacency
4.1. Journal article: Guided graph spectral embedding: application to the *c. elegans* connectome

and the Laplacian operators (Chung, 1997). For instance, the well-known graph cut problem can be convexly relaxed and solved by thresholding of the Laplacian eigenvector with the smallest non-zero eigenvalue, known as the Fiedler vector (Fiedler, 1989). More recently, new approaches in graph signal processing have taken advantage of the Laplacian eigenvectors to define the graph Fourier transform, which can then be used to process (i.e., filter) graph signals in the spectral domain (Shuman et al., 2013; Ortega et al., 2018); the spectral graph wavelet transform by (Hammond et al., 2011) is one such example.

The Laplacian eigenvectors also provide a meaningful embedding by mapping nodes onto a line, or higher-dimensional representation, that minimizes distances between connected nodes (Belkin and Niyogi, 2003). Other well-known embedding techniques use different metrics for distance in order to assess local graph properties, ranging from simple Euclidean distance in locally linear embedding (Roweis, 2000), to shortest path in Isomap (Tenenbaum, 2000), transition probability (Shen and Meyer, 2008), or conditional probability of an edge in t-distributed stochastic neighbor embedding (van der Maaten and Hinton, 2008). A time-dependent dynamical similarity measure has also been introduced (Schaub et al., 2018).

In addition, efforts have been made to employ global properties of the graph, such as in Sammon mapping (Sammon, 1969), where a cost function including all pairwise distances is optimized. In this manner, embedding is performed while taking in consideration both local (neighborhood) and global (distant nodes) properties of the graph. However, these techniques are not aware of the network at the mesoscale: one cannot guide the embedding by giving a certain subgraph more importance while still preserving local features and global topology characteristics.

In essence, the most powerful feature of graph spectral embedding is to effectively summarize local structure across the graph into low-dimensional global patterns. This is achieved, for instance, with the recently introduced concept of graph Slepians; i.e., graph signals that are bandlimited and take into account a subset of selected nodes. Specifically, two types of Slepian designs that respectively optimize for energy concentration and modified embedded distance have been introduced (Van De Ville, 2016; Van De Ville et al., 2017).

In this work, we further build on this framework by providing a simple way to guide analyses with additional flexibility. Guidance includes the selection of a given subgraph or group of nodes to study, and the ability to specify the intensity of the focus set on these selected nodes. With respect to graph Slepians, we hereby provide several extensions. First, we allow the selection process to be weighted, so that the importance of a node can be gradually changed. Second, we propose a new criterion that meaningfully combines the two existing ones; i.e., we want to maximize energy concentration and minimize modified embedded distance at the same time. Third, as we detail below, these two criteria are counteracting, and hence, we obtain stable solutions even at full bandwidth, where the original Slepian designs degenerate numerically. Fourth, we show how this criterion can be rewritten as an eigenvalue problem of an easy modification of the adjacency matrix, which can be interpreted as reweighting paths in the graph, and thus significantly simplifies the whole Slepian concept. The solution of the
eigendecomposition then defines the guided spectral domain, spanned by its eigenvectors. We illustrate the proposed approach with a proof-of-concept on the *Caenorhabditis elegans* (*C. elegans*) connectome. Through spectral embedding-based visualization, we observe the effects of focusing on a specific cellular population made of sensory neurons, interneurons or motoneurons, and we reveal trajectories of these neurons as a function of focus strength.

### 4.1.2 Methods

#### 4.1.2.1 Essential graph concepts

We consider an undirected graph with $N$ nodes, labeled $1, 2, \ldots, N$. The edge weights are contained in the symmetric weighted adjacency matrix $\tilde{A}$ with non-negative real-valued elements $\tilde{a}_{i,j}, i, j = 1, \ldots, N$. We also assume that the graph contains no self-loops; i.e., all diagonal elements $\tilde{a}_{i,i}$ are zero. The degree matrix $D$ is the diagonal matrix with elements $d_{i,i} = \sum_{j=1}^{N} \tilde{a}_{i,j}$. The graph Laplacian is defined as $\tilde{L} = D - \tilde{A}$ and can be interpreted as a second-order derivative operator on the graph. Here, we use the symmetrically normalized variants of the adjacency $\tilde{A}$ and graph Laplacian $\tilde{L}$ defined as $A = D^{-1/2} \tilde{A} D^{-1/2}$ and $L = I - A$. This normalization is often used in applications to emphasize the changes in topology and not in nodal degree (De Lange et al., 2014).

Let us define a graph signal as a vector of length $N$ that associates a value to each node (Shuman et al., 2013). One way to recognize the importance of the Laplacian and its eigendecomposition is to consider the smoothness of a graph signal $x$ as

$$ x^T L x = \sum_{i,j=1}^{N} a_{i,j} (x_i - x_j)^2, \quad (4.1) $$

which sums squared differences between signal values on nodes that are connected, proportionally to their link strength $a_{i,j}$. The eigenvectors of $L$ minimize this distance that is reflected by the eigenvalues, sorted by convention increasingly as $\lambda_1 = 0 \leq \lambda_2 \leq \ldots \leq \lambda_N$. Therefore, considering the eigenvectors associated to the smallest non-zero eigenvalues provides the Laplacian embedding of the nodes that minimizes distance in a lower-dimensional space (Belkin and Niyogi, 2003). The eigenvector with the smallest non-zero eigenvalue is also known as the Fiedler vector (Fiedler, 1989), which relates to the solution of the convex relaxation of the graph cut problem (von Luxburg, 2007).

Therefore, the eigendecomposition $L = U \Lambda U^T$ of the graph Laplacian is the cornerstone of spectral methods for graphs, as the eigenvectors $\{u_k\}, k = 1, \ldots, N$ (columns of $U$) play the role of graph Fourier components, and the associated eigenvalues $\{\lambda_k\}, k = 1, \ldots, N$, of frequencies (Chung, 1997). The graph Fourier transform (GFT) then provides the link between a graph signal $x$ and its spectral coefficients given by vector $\hat{x}$:
Figure 4.1 – Spectral embedding of the *C. elegans* connectome according to the eigenvectors of the Laplacian matrix with second and third smallest eigenvalues. The purpose of this work is to introduce guided spectral analysis; that is, to indicate direction by selecting a subset of nodes, and to adjust the strength of the focus set on this subset. Each colored circle in the figure depicts one *C. elegans* neuron. Light gray strokes link the cells that are connected by gap junctions or chemical synapses. Labels and connectivity were retrieved from (Varshney et al., 2011).
\( x = U \hat{x}, \) and \( \hat{x} = U^\top x. \)

### 4.1.2.2 Graph Slepians

In earlier work, the combination of the concepts of selectivity and bandwidth for graph signals has been used to define “graph Slepians” (Tsitsvero et al., 2016; Van De Ville, 2016; Van De Ville et al., 2017); i.e., bandlimited graph signals with maximal energy concentration in the subset of nodes \( S \)—a generalization of prolate spheroidal wave functions that were proposed fifty years ago on regular domains to find a trade-off between temporal and spectral energy concentrations (Slepian and Pollak, 1961; Slepian, 1978). The presence or absence of a node in \( S \) is encoded by the diagonal elements of the selection matrix \( S \); that is, we have \( S_{i,i} = \delta_{i \in S}, \) \( i = 1, \ldots, N, \) where \( \delta \) is the Kronecker delta. The Slepian design then boils down to finding the linear combination of Laplacian eigenvectors, encoded by spectral coefficients \( \hat{g} \), within the bandlimit \( W \) with maximal energy in \( S \), reverting to the Rayleigh quotient

\[
\mu = \frac{\hat{g}^\top W^\top U^\top S U W \hat{g}}{\hat{g}^\top \hat{g}},
\]

where \( W \) is a spectral selection matrix that has \( W \) ones on its diagonal followed by \( N - W \) zeros. This problem can be solved by the eigendecomposition of the concentration matrix \( C = W^\top U^\top S U W \) as \( C \hat{g}_k = \mu_k \hat{g}_k, \) \( k = 1, \ldots, W \). The graph Slepians \( g_k = U \hat{g}_k, \) \( k = 1, \ldots, W, \) are orthonormal over the entire graph as well as orthogonal over the subset \( S \); i.e., we have \( g_k^\top g_l = \delta_{k-l} \) as well as \( g_k^\top S g_l = \mu_k \delta_{k-l}. \)

For the purpose of this work, we introduce the set of bandlimited graph signals

\[
\mathcal{B}_W = \{ x | \hat{x} = W \hat{x} \},
\]

such that we can then rewrite the Slepian criterion of Eq. (4.2) directly in the vertex domain as

\[
\mu = \frac{\hat{g}^\top S \hat{g}}{\hat{g}^\top \hat{g}} \quad \text{s.t.} \quad \hat{g} \in \mathcal{B}_W. \quad (4.3)
\]

An alternative Slepian design was also proposed in (Van De Ville et al., 2017)—see also (Huang et al., 2018), modifying the Laplacian embedded distance of Eq. (4.1) as follows:

\[
\zeta = \frac{\hat{g}^\top L^{1/2} S L^{1/2} \hat{g}}{\hat{g}^\top \hat{g}} \quad \text{s.t.} \quad \hat{g} \in \mathcal{B}_W. \quad (4.4)
\]

The Laplacian embedded distance \( x^\top L x \) is a measure of smoothness of the vector \( x \) over the graph, which is why eigenvectors of \( L \) with increasing eigenvalues are ordered according to
smoothness. Imposing the modification with the selection matrix $S$ focuses the smoothness on a certain subgraph, notwithstanding how the signal behaves outside it. Eq. (4.4) can also be seen as a generalization of Laplacian embedding, since $L^{1/2}SL^{1/2}$ reverts to $L$ for the special case of $S = I$. It is important to realize that the eigenvalues $\{\mu_k\}$ of the original design reflect the energy concentration in the subset $\mathcal{S}$, while the eigenvalues $\{\xi_k\}$ of the alternative design correspond to a modified embedded distance that can be interpreted as a “frequency value” localized in $\mathcal{S}$, in analogy to the global GFT case. Consequently, “interesting” eigenvectors correspond to those with high $\mu_k$, concentrated in the subset $\mathcal{S}$, or low $\xi_k$, showing the main localized low-frequency trends, respectively. However, the eigendecompositions, taken individually, do not necessarily lead to eigenvectors that combine both virtues.

4.1.2.3 Guiding spectral embedding using a new criterion

We hereby propose to further generalize the Slepian design in a number of ways. First, we relax the selection matrix $S$ to a cooperation matrix $M$ with diagonal elements that can take any non-negative real values $m_l \geq 0$, $l = 1, \ldots, N$. This allows to gradually change the impact of a node on the analysis, between an enhanced ($m_l > 1$), an unmodified ($m_l = 1$) and a reduced ($m_l < 1$) importance with respect to the selection matrix case. Second, we combine the criteria of both already existing Slepian designs by subtracting the modified embedded distance from the energy concentration:

$$
\zeta = \mu - \xi = \frac{g^\top Mg - g^\top L^{1/2}ML^{1/2}g}{g^\top g}, \quad \text{s.t. } g \in \mathcal{B}_W. \tag{4.5}
$$

Third, we remove the bandlimit constraint and allow $g$ to be any graph signal, which is an operational choice due to the joint optimization of both criteria, as will be illustrated and discussed later on.

Using the Taylor series approximation of the square root function, we derive $L^{1/2}$ in terms of the adjacency matrix $A$:

$$
L^{1/2} = (I - A)^{1/2} = I - \frac{1}{2} A - \frac{1}{8} A^2 - \frac{1}{16} A^3 - \ldots \tag{4.6}
$$

$$
= I - \sum_{k=1}^{\infty} c_k A^k, \tag{4.7}
$$

with $c_k = \frac{(2k)!}{2^k (k!)^2 (2k-1)}$. Details on the series expansion are discussed in Section 4.1.3.3. We can
Chapter 4. Processing global signals in local environment

then further rewrite the internal part of the criterion (4.5) as

\[
\begin{align*}
M - (I - A)^{1/2}M(I - A)^{1/2} &= \sum_{k=1}^{\infty} c_k \left( MA^k + A^k M \right) \\
- \sum_{k_1=1}^{\infty} \sum_{k_2=1}^{\infty} (c_{k_1} c_{k_2}) A^{k_1} M A^{k_2}.
\end{align*}
\]

(4.8)

By convention, the associated eigenvalues are sorted in decreasing order. Based on the fact that eigenvalues of the symmetric normalized Laplacian are greater or equal to 0 and lower or equal to 2, one can derive \( m_{\text{max}} \geq \zeta_1 \geq \zeta_2 \geq \ldots \geq -2m_{\text{max}} \), where \( m_{\text{max}} \) is the highest cooperation value appearing in \( M \), using bounds from Corollary 2.4 in (Lu and Pearce, 2000).

In what follows, we will be considering the linear and quadratic approximations of the new criterion’s eigenvalues:

\[
\begin{align*}
\zeta_{\text{lin}} &= \frac{g^T \left( \frac{MA + AM}{2} \right) g}{g^T g} \quad (4.9) \\
\zeta_{\text{quad}} &= \frac{g^T \left( \frac{MA + AM}{2} + \frac{MA^2 + A^2 M - 2AMA}{8} \right) g}{g^T g}.
\end{align*}
\]

(4.10)

Interestingly, the combination of both existing Slepian criteria leads to the emergence of the adjacency matrix \( A \) as the key player in our new formalism. In fact, when the cooperation matrix is the identity matrix, the criterion reverts to the eigendecomposition of \( A \) itself.

Let us now interpret the impact of the cooperation weights: obviously, an element \( a_{i,j} \) of the adjacency matrix contains the weight of a direct path from \( i \) to \( j \). The linear approximation \( \zeta_{\text{lin}} \) reweights such a direct path with the average \( (m_i + m_j)/2 \) of the cooperation weights that are attributed to nodes \( i \) and \( j \), as illustrated in Fig. 4.2A (left half). Notice that paths where only one node has a cooperation weight equal to 0 are still possible, as the other cooperation weight is then simply divided by two.

As for the quadratic approximation, it takes into account length-2 paths between nodes \( i \) and \( j \). For instance, the sum of all length-2 paths between \( i \) and \( j \) can be read out from the squared adjacency matrix:

\[
[A^2]_{i,j} = \sum_{l=1}^{N} a_{i,l} a_{l,j} = \langle a_{i}, a_{j} \rangle,
\]

where the inner product reveals the kernel interpretation of the length-2 walk matrix. Therefore, as illustrated in Fig. 4.2A (right half), the term

\[
[MA^2 + A^2 M]_{i,j} = (m_i + m_j) \sum_{l=1}^{N} a_{i,l} a_{l,j}
\]

66
4.1. Journal article: Guided graph spectral embedding: application to the \textit{c. elegans} connectome

Figure 4.2 – \textbf{A}. In the case of two nodes \(i\) and \(j\), the average of their cooperation weights yields the multiplying factor for \(a_{i,j}\) (blue term). When a third node \(l\) is added, the difference between average cooperation weight between nodes \(i\) and \(j\) (light blue term), and the cooperation weight of node \(l\) (salmon term), multiplies the length-2 path and then also contributes to the output entry. \textbf{B}. In an example three-node network, output entries for different examples where cooperation weights are either set to 0 (white nodes) or to 1 (black nodes). Edge thickness is proportional to the output entry weight. Red strokes denote positive edge values, while blue strokes highlight negative edge values. All non-zero entries of the normalized adjacency matrix of the example network equal \(1/2\).

Reweights all length-2 paths by the summed cooperation weight between the start and end nodes, while subtracting the term

\[
[AMA]_{i,j} = \frac{m_i + m_j}{2} a_{i,j} + \frac{1}{4} \left( \frac{m_i + m_j}{2} \sum_{l=1}^{N} a_{i,l} a_{l,j} - \sum_{l=1}^{N} m_l a_{i,j} a_{l,j} \right)
\]

penalizes the path according to the cooperation weight of node \(l\) through which it passes.

Analogously, the term \(A^k\) in the criterion introduces modifications of \(k\)-length paths in the graph. However, for \(k > N\), reweighting reduces to modifications of lower-length paths. The Cayley-Hamilton theorem implies that for every matrix \(A\) of size \(N \times N\), the matrix \(A^N\) can be written as a linear combination of matrices \(A^k\) for \(k = 0, 1, \ldots, N - 1\). By induction, it holds that \(A^k\) for every \(k > N\) can also be written as a linear combination of the same set of \(N\) matrices. Hence, modifications of paths longer than \(N - 1\) can be seen as a linear combination of additional modifications of paths of length 0 to \(N - 1\).

4.1.3 Mathematical considerations

This section provides mathematical foundations supporting the methods and the results presented in this work. We start by discussing the link between the selection matrix and the eigenspectrum associated to the energy concentration criterion, and the relationship with the modified embedded distance criterion, using full bandwidth. Then, we provide a formal
4.1.3.1 Eigenspectrum associated to the energy concentration criterion

For full bandwidth, the concentration matrix is defined as $C = U^T SU$, where $U$ is the matrix whose columns are eigenvectors of the graph Laplacian, and $S$ is a diagonal selection matrix. Hence, the eigendecomposition of $C$ is trivial: its eigenvectors are the rows of $U$, and the eigenvalues of $C$ correspond to the diagonal entries of $S$, as can be seen from Fig. 4.3A for $W = 279$.

4.1.3.2 Eigenspectrum associated to the modified embedded distance criterion

We show that for full bandwidth, the number of zero eigenvalues of the modified embedded distance matrix, denoted $z_\lambda$, is lower-bounded by the number of zeros on the diagonal of the selection matrix, denoted $z_S$. To see this, consider the following decomposition of the modified embedded distance matrix $C_{emb}$:

$$C_{emb} = L^{1/2} S L^{1/2} = \sum_{k=1}^{N - z_S} s_{i_k} l_{i_k} l_{i_k}^T,$$

where $i_k$ is the index of the $k^{th}$ non-zero entry of the selection matrix $S$, and $l_{i_k}$ denotes the $i_k^{th}$ column vector of the matrix $L^{1/2}$. From this expression, it can be seen that the rank of $C_{emb}$ is at most $N - z_S$ and hence, $z_\lambda \geq z_S$. Equality holds when the set of vectors $\{l_{i_k}\}$ corresponding to the non-zero entries of $S$ are linearly independent. This is the case for connected graphs, as any subset (with cardinality strictly less than $N$) of the columns of $L^{1/2}$ is linearly independent. This relationship is observed in Fig. 4.3B for $W = 279$.

4.1.3.3 Taylor series of matrix-valued functions

The Taylor expansion of $L^{1/2}$ proposed in Eq. (4.6) is derived using the scalar Taylor series of $f(x) = \sqrt{x}$ evaluated around the point $a = 1$:

$$\sqrt{x} = 1 + \sum_{k=1}^{\infty} t_k (x - 1)^k,$$
where $t_k = \frac{(-1)^{k-1}(2k)!}{2^{2k}(k!)^2}x^{2k}$ and $x \in \mathbb{R}, x > 0$. The square root matrix of $L$ then writes:

$$L^{1/2} = U_L \Lambda_L^{1/2} U_L^T$$

$$= U_L \left[ \begin{array}{cccc}
1 + \sum_{k=1}^{\infty} t_k (\lambda_1 - 1)^k \\
& \ddots \\
& & 1 + \sum_{k=1}^{\infty} t_k (\lambda_N - 1)^k
\end{array} \right] U_L^T$$

$$= U_L (I + \sum_{k=1}^{\infty} t_k (\Lambda_L - I)^k) U_L^T.$$  

Since the Laplacian and adjacency matrices are normalized, their eigenvalues verify $\Lambda_L = I - \Lambda_A$ and their eigenvectors are equal ($U_L = U_A$) when ordered following increasing and decreasing eigenvalues, respectively. The previous equation finally reduces to:

$$L^{1/2} = I + U_A (\sum_{k=1}^{\infty} t_k (\Lambda_A - I)^k) U_A^T$$

$$= I + \sum_{k=1}^{\infty} (-1)^k t_k U_A \Lambda_A^k U_A^T$$

$$= I - \sum_{k=1}^{\infty} c_k A^k,$$

where $c_k = \frac{(2k)!}{2^{2k}(k!)^2}x^{2k}$, which is the expression used in Eq. (4.6). 

Truncation of the Taylor series of a function $f(x)$ to a finite upper bound on $k \leq K$ leads to an approximation error which can be estimated by the Lagrange form of the remainder

$$R_K(x) = \frac{f^{(K+1)}(y)}{(K+1)!} (x-1)^{K+1},$$

where the $(K+1)^{th}$ derivative is evaluated at the point $y$ found between $x$ and 1. On the other hand, since the eigenvectors forming $U_L$ are unit-norm vectors, the distance $d_K$ between a finite sum approximation of $L^{1/2}$ and the true square root of the matrix is bounded as:

$$d_K = \|L^{1/2} - (I - \sum_{k=1}^{K} c_k A^k)\|_F \leq \sum_{i=1}^{N} |R_K(\lambda_i)|,$$

where $\| \cdot \|_F$ denotes the Frobenius norm. In the case of a first order Taylor approximation ($K = 1$), we get:

$$d_1 \leq \sum_{i=1}^{N} |f^{(2)}(y)| \frac{1}{2!} (\lambda_i - 1)^2.$$

The eigenvalues $\lambda_i$ range from 0 to 2, and all contribute to the total approximation error $d_1$. 


Chapter 4. Processing global signals in local environment

with eigenvalues further from 1 contributing more. Since the second-order derivative of the square root function increases as its argument approaches 0, the most contributing factors of the error derive from Taylor approximation terms with near-zero eigenvalues. Hence, graphs whose Laplacian spectrum exhibits higher eigengaps in the lower band tend to have lower approximation error.

Finally, the Frobenius distance $d_{K,M}$ between the true proposed criterion $M - L^{1/2}ML^{1/2}$ and its approximation using a $K$th-order Taylor approximation of $L^{1/2}$ verifies:

$$d_{K,M} \leq d_K||M||_F d_K,$$

where $||M||_F$ corresponds to the Frobenius norm of the cooperation matrix. Hence, the upper bound on $d_{K,M}$ reduces as the nodes are given less importance; i.e., when the cooperation values get closer to 0.

4.1.4 Results

The C. elegans worm is an intensely studied model organism in biology. In particular, the wiring diagram of its 302 neurons has been carefully mapped during a long and effortful study (White et al., 1986). Here, we use the graph that summarizes data from 279 somatic neurons (unconnected and pharyngeal neurons were excluded from the full diagram of 302 neurons), and combined connectivity from chemical synapses and gap junctions (Chen et al., 2006). The binary adjacency matrix $A_{bin}$ with edge weights 0 or 1 has been symmetrically normalized with the degree matrix $D$ into $A = D^{-1/2}A_{bin}D^{-1/2}$, as described in Section 4.1.2.1. We retrieved the type of each neuron (sensory neuron, interneuron or motoneuron) from the WormAtlas database (http://www.wormatlas.org/).

In their modeling work, (Varshney et al., 2011) studied network properties of the worm connectome using different approaches, including Laplacian embedding. In particular, the topological view generated by mapping nodes on the first two eigenvectors with smallest non-zero eigenvalues already reveals interesting network organization (see Fig. 4.1). The horizontal dimension ($u_2$) mainly distinguishes the motoneurons from the head (right green circles) and from the ventral cord (left green circles). The vertical dimension ($u_3$) reflects information flow from sensory neurons and interneurons of the animal’s head (top) to the nerve ring and ventral cord circuitries (bottom).

4.1.4.1 Eigenvalues of different criteria

To illustrate the eigenvalues obtained with the existing Slepian designs, as well as the newly proposed criterion, we considered the 128 motoneurons and “unselected” them by setting their respective entries in $S$ to 0. We applied the original, concentration-based Slepian design for different bandwidths $W = 100, 150, 200, 279$, the latter corresponding to full bandwidth. The eigenvalues $\mu_k$, which reflect energy concentration in the 151 remaining neurons, are shown in
4.1. Journal article: Guided graph spectral embedding: application to the *c. elegans* connectome

Fig. 4.3A. The characteristic behavior of classical Slepians is preserved for the graph variant; *i.e.*, eigenvalues cluster around 1 and 0 for well and poorly concentrated eigenvectors, respectively, and the phase transition occurs more abruptly at higher bandwidth. For full bandwidth, perfect concentration becomes possible, and the problem degenerates in retrieving two linear subspaces of 151 and 128 dimensions spanned by eigenvectors with concentration 1 and 0, respectively (see Section 4.1.3.1 for a proof on the number of distinct eigenvalues). In practical terms, for high but not full bandwidth, the “interesting” eigenvectors with large concentration correspond to the part indicated by the green area on the plot, and become numerically indistinguishable. A few indicative examples of Slepian vectors across bandwidths are displayed in Supplementary Fig. S1C.

Next, we applied the modified Slepian design inspired by the Laplacian embedded distance. As shown in Fig. 4.3B, the eigenvalues $\xi_k$ reflect the modified embedded distance, which we now want to minimize. For increasing bandwidth (darker curves), its smallest values can be made lower; however, the subset of nodes with $S_{ij}$ entries set to 0 is also described by eigenvectors with small eigenvalues. This becomes even clearer at full bandwidth, a case for which a subspace of 128 dimensions spanned by eigenvectors with a modified embedded distance of 0 is retrieved, as indicated by the green area in Fig. 4.3B and explicitly demonstrated in Section 4.1.3.2. Some examples of Slepians across bandwidths can be seen in Supplementary Fig. S1D.

The degeneracies of the Slepian designs at full bandwidth are instructive about the opposing effects of maximizing energy concentration and minimizing modified embedded distance; *i.e.*, the subspaces indicated by the green areas in Figs. 4.3A and B, which are optimal for the corresponding criteria, are actually different ones, representing signals on sensory and interneurons (151 nodes) on the one hand, and on motoneurons (128 nodes) on the other hand (compare Supplementary Figs. S1C and S1D, first rows). This leads us to the eigenvalues $\zeta_k$ of the proposed criterion, as shown in Fig. 4.3C (black curve).

The maximum eigenvalue peaks close to 1, a case reflecting jointly high equivalent $\mu_k$ (blue curve) and low equivalent $\xi_k$ (purple curve); *i.e.*, a high energy concentration at the same time as a low modified embedded distance (low localized graph frequency) within $\mathcal{S}$. The low amount of such solutions shows that it is difficult to conceal high energy concentration and small modified embedded distance.

As values of $\zeta_k$ decrease, we first observe a rise in modified embedded distance (eigenvectors remain reasonably concentrated within $\mathcal{S}$, but rapidly exhibit a larger localized graph frequency), and then a decrease of both $\mu_k$ and $\xi_k$, which indicates that eigenvectors become less concentrated within the subset of interest. Afterwards, we observe a regime in which both quantities are null at the same time; that is, a subspace spanned by eigenvectors that are fully concentrated outside $\mathcal{S}$. Notice that this set of eigenvectors is now “pushed away” from the meaningful low $\xi_k$ ones, and lie in the middle of the spectrum. Finally, the sign of $\zeta_k$ switches, and the right hand side of Fig. 4.3C denotes eigenvectors of increasing concentration within.
Figure 4.3 – Plots of eigenvalues obtained using different Slepian criteria: (A) energy concentration $\mu$, (B) modified embedded distance $\xi$, and (C) our new proposed criterion $\zeta$. For the first two cases, in which the design depends on a bandwidth parameter, eigenvalue spectra are plotted for $W = 100$, 150, 200 and 279 with increasingly lighter blue or purple shades, respectively. In the full bandwidth case, the shaded green areas highlight eigenvalues linked to optimal solutions of the respective criteria (see Sections 4.1.3.1 and 4.1.3.2 for the associated mathematical derivations). In the third case, equivalent $\mu$ and $\xi$ eigenspectra are plotted in blue and purple on top of the $\zeta$ one. The full $\zeta$ eigenspectrum is also compared to approximations obtained through Taylor series of increasing order (D), from linear to order 20, as depicted by increasingly darker brown curves. The two smaller plots are insets sampled at the start and at the end of the main plot, respectively.

Interestingly, computing the eigenspectrum using a linear approximation of the criterion matrix (Fig. 4.3D, light brown curve) leads to very similar results, which only slightly vary for the largest eigenvalues. When the approximation order is increased up to 20 (increasingly dark brown curves), this low error further diminishes, although a mild difference remains with the ground truth. Inspection of the Slepian vectors related to several locations of the eigenspectrum (Supplementary Fig. S2) confirmed that the only salient differences actually involved the first Slepian vector (largest eigenvalue one).
4.1. Journal article: Guided graph spectral embedding: application to the *c. elegans* connectome

### 4.1.4.2 Topology revealed by guided spectral analysis

We now guide the spectral analysis to focus on the three different types of neurons. For instance, when focussing on the role of the sensory neurons, we gradually decrease the cooperation weights of interneurons and motoneurons from 1 to 0. For each setting, we then visualize the topology revealed by the guided analysis by projecting the nodes on the eigenvectors with the second and third largest eigenvalues. We build the trajectory of each node through this two-dimensional embedding, after applying the Procrustes transform (Schönemann, 1966) to compensate for any irrelevant global transformations. As a complementary visualization, note that we provide the start, intermediate and end points of each trajectory as separate figures in Supplementary Fig. S3. Finally, k-means clustering was performed on the nodes in focus at the end point embedding of trajectories, producing sets of clusters given in Supplementary Fig. S4 and Supplementary Tables 1-3 (see Supportive Information for details). Example visualizations when resorting to different Slepian vectors are provided in Supplementary Fig. S5.

In Figs. 4.4A and B, the trajectories are depicted when focussing on the sensory neurons by attributing cooperation weights to the other types of neurons ranging from 1 to 0.5, and from 0.5 to 0, respectively. During the first half (Fig. 4.4A), the network organization is only slightly altered with respect to the initial view of Fig. 4.1; i.e., the sensory neurons move slightly more to the periphery, while the interneurons and motoneurons move to the origin. In the second part of the trajectory (Fig. 4.4B), a major split occurs in the bottom right branch of Fig. 4.4A between the left and right versions of a whole series of neurons, while the bottom left branch neurons move back to the center of the coordinate frame. The cell types found in the top branch are amphid neurons, whereas the rest of the sensory neurons split into their left and right counterparts located in the left and right bottom branches. The clusters found by the k-means approach (see Supplementary Table 1) include a group of 5 bilateral amphid neurons (AWA, AWC, ASE, ASI and AFD; cluster C\(_3\)) and 6 other clusters, 2 of which span the bottom left and right sub-branches (clusters C\(_5\) and C\(_2\)).

As described in Section 4.1.2.3, since paths through nodes with cooperation weights set to 0 are still considered by the proposed criterion, the embedding focusing on a particular subtype of neurons can still include functionally distinct cells as clearly standing out in the visualization. For instance, in addition to the above clustering of sensory neurons in Fig. 4.4B, we notice the segregation of the bilateral RIP interneurons towards the left and right branches. This shows that the embedding does not neglect nodes outside the focus, even when their cooperation weight is set to 0.

In Figs. 4.5A and B, we then focus on the interneurons by reducing the cooperation weights of sensory neurons and motoneurons in two steps. As expected, the interneurons move towards the periphery. Their organization does not seem to be dominated by left versus right variants, as we found for sensory neurons, but rather by a set of well-defined clusters related to their functional involvement in the *C. elegans* neuronal circuitry (see Supplementary Table 2): in
Chapter 4. Processing global signals in local environment

Figure 4.4 – Focussing on the sensory neurons by reducing the cooperation weights of the interneurons and motoneurons (A) from 1 to 0.5, and (B) from 0.5 to 0. The trajectory of a neuron is represented by a colour change from light to dark tones, and dots represent final positions. Note that the starting configuration in (A) is identical to the representation in Fig. 4.1. Cells are labeled according to (Varshney et al., 2011).

In the first quadrant, we find the isolated AIA bilateral pair (cluster C4). Moving clockwise, a larger cluster of neurons includes the bilateral AIY, AIZ, AIN, AIB, RIA, RIB, AUA and the single neurons RIR and RIH (cluster C3). Next we find a cluster including AVE, AVK, RIG, PVT, DVA and other neurons located closer to the origin of Fig. 4.5 (cluster C5), before reaching another large ensemble of neurons including the bilateral AVA, AVD, LUA, PVC, PVW, and the single neuron PVR (cluster C6). Moving back upwards, cluster C1 contains the bilateral AVB, AVJ, BDU, the single neuron AVG, and PVPR, whose left counterpart PVPL belongs to cluster C5, thus standing as the only bilateral pair of neurons split into different clusters. Finally, we reach the last group of cells containing the bilateral RIF, AVH, AIM, PVQ and AVF (cluster C2).

Finally, in Figs. 4.6A and B, the organization of motoneurons is examined. Already in the first step (Fig. 4.6A), when reducing the cooperation weights of the sensory and interneurons from 1.0 to 0.5, we observe much stronger changes than in the previous cases. In particular, the initial organization completely collapses and the left branch of the motoneurons spreads out. This branch then develops into a peripheral organization when further decreasing the cooperation weights (Fig. 4.6B), with three main subsets of neurons and ambiguous positioning of the cell DVB between the left and the right bottom branches. K-means clustering into optimal cell groups captured this architecture into 7 smaller clusters (Supplementary Table 3): clusters C4 and C7 spanned top neurons, clusters C2 and C4 included the bottom left branch neurons, and clusters C5 and C6 contained the bottom right branch cells.
4.1. Journal article: Guided graph spectral embedding: application to the *c. elegans* connectome

Figure 4.5 – Focussing on the interneurons by reducing the cooperation weights of the sensory neurons and motoneurons (A) from 1 to 0.5, and (B) from 0.5 to 0. The trajectory of a neuron is represented by a colour change from light to dark tones, and dots represent final positions. Note that the starting configuration in (A) is identical to the representation in Fig. 4.1. Cells are labeled according to (Varshney et al., 2011).
Figure 4.6 – Focussing on the motoneurons by reducing the cooperation weights of the sensory neurons and interneurons (A) from 1 to 0.5, and (B) from 0.5 to 0. The trajectory of a neuron is represented by a colour change from light to dark tones, and dots represent final positions. Note that the starting configuration in (A) is identical to the representation in Fig. 4.1. Cells are labeled according to (Varshney et al., 2011).

4.1.5 Discussion

4.1.5.1 Beyond original Slepian designs

The originality of our approach lies in providing a new and simple way to guide graph spectral analysis. Inspired by graph Slepians, we propose a novel criterion that combines energy concentration and modified embedded distance, taking into account cooperation weights that can gradually increase or decrease the importance of selected nodes. The new criterion lets the adjacency matrix emerge as the central graph operator, instead of the Laplacian, and is operational at full bandwidth.

This is surprising at first sight, because neither of the conventional Slepian criteria is practical without the bandlimit constraint. For the energy concentration with binary cooperation weights, as shown in Fig. 4.3A for an illustrative example on the *C. elegans* connectome, full bandwidth leads to two eigenvalues (1 and 0), the dimensionality of the corresponding subspaces being the number of nodes with cooperation weight 1 and 0, respectively. For the modified embedded distance, as shown in Fig. 4.3B, full bandwidth creates a subspace with eigenvalue 0 of dimensionality equal to the number of nodes with cooperation weight 0. Therefore, subtracting both criteria leads to opposing objectives; *i.e.*, at full bandwidth, an energy concentration of 1 encodes the subspace for nodes with weight 1, while a modified embedded distance of 0 encodes the subspace for nodes with weight 0.
4.1. Journal article: Guided graph spectral embedding: application to the \textit{c. elegans} connectome

The obtained eigenspectrum for the new criterion, shown in Fig. 4.3C, illustrates that only a few eigenvectors are able to combine high energy concentration with low modified embedded distance, a counterbalance that can be further revealed by measuring $\mu$ and $\xi$ separately for these new eigenvectors. Such a large eigengap is also good news for numerical computation of the leading eigenvectors for large graphs when relying upon efficient large-scale solvers (Lehoucq and Sorensen, 1996) implemented in widely available software libraries such as ARPACK.

Intriguingly, the approximation error was already low using a linear approximation, and did not noticeably decrease further, except for the first Slepian vector, when resorting to higher-order terms (see Fig. 4.3 and Supplementary Fig. S2). Modifying the importance of a node via the corresponding cooperation value affects all-length paths through that node according to the series expansion from Eq. (4.8), where the power of $A$ in each term corresponds to the affected path length. Once we restrict the criterion to a linear approximation, the only paths whose importance is changed are those of length 1. This does not mean that other paths are not included in the graph analysis, but rather that they are included with their original (unmodified) effect on the topology. Low error of linear approximation suggests that the highest percentage of topological importance of a node falls into the importance of its length-1 paths. Further, a slightly higher error at eigenvectors with the highest $\xi$ may be explained similarly: not modifying higher order paths produces greater error at these eigenvectors because of their increased relative importance due to the fact that high $\xi$ eigenvectors tend to be very smooth (even approaching a constant signal); thus, in order to even out the values at all nodes in the process, one needs to "reach" far enough.

The proposed criterion should not be confused with the Sobolev norm that is sometimes used to regularize graph signals (Mahadevan and Maggioni, 2006). Specifically, in the case of $M = I$, our criterion of Eq. (4.5) applied to $g$ reverts to $g^\top g - g^\top L g$, whereas the Sobolev norm of $g$ reads $g^\top g + g^\top L g$. The difference in the sign of the second term introduces significantly distinct optimization goals regardless of the apparent similarity of the two expressions.

As for future extensions of our approach, one could envisage to dig into the relationship with graph uncertainty principles (Agaskar and Lu, 2013; Tsitsvero et al., 2016; Teke and Vaidyanathan, ress), to consider statistical resampling for graphs (Pirondini et al., 2016), or to focus on the discovery of hierarchical graph structure (Arenas et al., 2008; Irion and Saito, 2014) by gradual refinement of the subgraph. The design could also be extended to directed graphs using recent extensions of spectral decompositions in this context (Sandryhaila and Moura, 2013; Mhaskar, 2018).

4.1.5.2 Gaining insights on \textit{c. elegans}

The application of our newly developed approach to the \textit{C. elegans} connectome enabled to confirm past findings from the literature, and to shed light on additional cellular targets and groupings that may deserve further experimental analyses. At the level of sensory neurons
(Fig. 4.4, Supplementary Fig. S3A and Supplementary Fig. S4A), seven clusters were extracted, collectively accounting for the three branches evident in Fig. 4.4: the top branch made of twelve (including the thermosensor AFD) pairs of amphid neurons (at \( y \)-coordinate greater than 0.04), and other cells split into the left and right bottom branches. Interestingly, one of the clusters found by k-means included five pairs of bilateral amphid neurons: AWA and AWC involved in odor taxis (Bargmann et al., 1993; Li et al., 2012), the thermosensor AFD (Mori and Ohshima, 1995), and ASE and ASI implicated in chemotaxis (Bargmann and Horvitz, 1991; Luo et al., 2014). These neurons act as low-order sensors, whose extraction as a separate cluster inside the amphid group may suggest new information worth further exploration.

The lower branches in Fig. 4.4 split the neurons into their right and left counterparts, thus extracting relevant somatic information. These neurons act as higher-order sensing apparatus as compared with amphid neurons: IL1 and OLQ have jointly been implicated in the worm foraging response (Hart et al., 1995); CEP and ADE are involved in the response upon food sensing (Sawin et al., 2000); URX, URY and OLL are linked to the reproductive drive (Barrios et al., 2012), and so on. The split between low and high order sensing is summarized in Fig. 4.7A.

Further inspection of the branches (Supplementary Fig. S6A) showed that the left-right segregation involved chemical synapses, but not gap junctions. Also, Supplementary Fig. S5 (second row) shows that for higher-order Slepian vectors (fourth and fifth), additional contributors emerge, such as the bilateral PHA/PHB. This suggests that the approach finds different subgroups of higher-order sensory neurons depending on the choice of the embedding eigenvectors. The biological/functional interpretation of the exact clusters asks for a more detailed analysis of the subgroups of neurons. Finally, the emergence of RIP interneurons in the embedding (Fig. 4.4) points towards an important role of the sensory neurons yet to be explained, possibly in connection with their presynaptic inputs from IL1 (White et al., 1986).

Turning to interneurons (Fig. 4.5, Supplementary Fig. S3B and Supplementary Fig. S4B), we notice a trend of grouping neurons at the same command-chain level. Starting from the top of Fig. 4.5, we find AIA, AIB, AIY and AIZ jointly known for their role on locomotory behaviour and acting as a first-relay drives (Wakabayashi et al., 2004; Gray et al., 2005). Moving clockwise, we find RIA and RIB acting as second-layer intermediates, and further on, neurons such as AVE, and in the next cluster AVA and AVD, all being command interneurons (Hobert, 2003; Haspel et al., 2010; Kawano et al., 2011). The trend of following the locomotory pathway clockwise in the embedding space suggests that the approach targets relevant information about the neural system. However, the exact compact clusters in Supplementary Fig. S4B need further elaboration. Some of the interesting findings worth exploring would be the unexplained grouping of the scarcely studied RIR neuron (Hobert et al., 2002) with the cluster of cells including AIB and AIY, or the grouping of PVR and LUA (Chalfie et al., 1985; Wicks and Rankin, 1995) with locomotion-regulating neurons such as AVD and AVA.

Considering motoneurons (Fig. 4.6, Supplementary Fig. S3C and Supplementary Fig. S4C),
Figure 4.7 – Summary of the main functions operated by the sensory neurons (A), interneurons (B) and motoneurons (C) unraveled by guided spectral analysis. Clusters of neurons discussed in Section 4.1.5.2 are delineated and color coded according to their main roles: this may be in sensing (thermosensation in red, olfactory sensation in yellow, chemosensation in green and mechanosensation in blue), higher-order functions (reproduction in pink, food responses in brown), or locomotion (from first cellular relays to effector motoneurons in increasingly darker shades of gray). A gradient in the color coding indicates that more than one function is performed by neurons from a given cluster. Neurons that could not be clearly related to the rest of the unraveled circuitry are encircled in white.
the embedding positions fit somatic location (see Supplementary Fig. S7): a spiral beginning at the origin, turning right, then moving clockwise and ending in the top branch follows the postero-anterior direction. This confirms that the approach has extracted meaningful information. However, the exact split between the three branches as well as the k-means clustering into the seven ensembles remains unclear, since, from preliminary explorations, we find both A-type and B-type cholinergic motoneurons and the inhibitory D-type motoneurons in all clusters. Finally, DVB deserves further attention (Schuske et al., 2004) due to its isolated location between the two bottom branches.

In Fig. 4.6B, two sensory nodes stick out the furthest away from the center; i.e., towards the lower left and right branches of motoneurons. These are PVD and PHC neurons, responsible for nociceptive mechano- and thermosensation, respectively. The locations of these nodes in the embedding may be linked to the fact that harmful nociceptive stimuli induce a locomotory response. As in the case of RIP interneurons emerging in the focused embedding of sensory neurons, we once again confirm the ability of the proposed approach to extract important nodes even when their cooperation weight was initially set to 0.

In summary, as illustrated in Fig. 4.7, all three types of neurons found in the C. elegans nematode could be arranged in a meaningful hierarchy thanks to the introduced guided graph spectral embedding. Sensory neurons were separated between first-order and higher-order sensors. Different levels of processing of motor functions were distinguished (see the gradient from white to dark gray tones going clockwise in Fig. 4.7B), with the eventual recruitment of motoneurons, which have been separated on the basis of somatic location. Future analyses will allow the study of different types of neurons through more elaborate combinations of focused nodes. In addition, it will be interesting to see whether future experimental work can shed light on some of the neurons that were extracted here without being yet extensively documented in the literature, such as AVKL or RIR.

### 4.1.5.3 Perspectives for future uses

The proposed graph embedding provides a simple, yet powerful approach to visualization and, if combined with clustering techniques, to the extraction of meaningful subgraphs from any graph-modeled dataset. In neuroimaging, focusing on a specific subgraph of interest (by setting the appropriate cooperation values) can direct research onto clinically relevant concepts, such as the medial temporal lobe and limbic structures for human brain imaging studies comparing healthy controls and Alzheimer patients (Krusinski et al., 1998). Be it using the structural or the functional connectome for analyses (Contreras et al., 2015), features such as cluster size and/or the inclusion of specific nodes (brain regions) in a cluster may become biomarkers for an early diagnosis or prediction of the disease.

Furthermore, graph modeling of the human brain is frequently employed to extract important nodes/brain regions and to identify their topological roles, such as a provincial/connector hubs suggesting clinically significant functional roles (van den Heuvel and Sporns, 2013).
4.1. Journal article: Guided graph spectral embedding: application to the *C. elegans* connectome

Doing so requires the use of diverse node centrality measures, such as degree or betweenness centrality. On the other hand, entries of the proposed Slepian eigenvectors may be interpreted as higher-order spectral centrality measures relative to the focused subgraph, and for the special case $M = I$, the eigenvector corresponding to the highest positive eigenvalue reverts to the eigenvector centrality (Newman, 2010). Hence, if clustering of a dataset based on the proposed embedding coordinates reveals nodes distant from the rest of the graph, it is suggested that those nodes exhibit a hub-like role when the focused subgraph is considered more important than the rest of the graph. For example, the AIA pair in the discussed *C. elegans* example emerges as a separate cluster in Fig. 4.5 and Supplementary Fig. S4B, where the focus is set on interneurons. Its role as a hub can be confirmed by the high number of connections to the set of amphid neurons, and a small number of connections to the other cells, as compared to the rest of the interneurons. Identification of hubs and/or peripheral nodes with respect to other similar type nodes may lead to a better understanding of the functional role of both neurons and brain regions, depending on the inspected dataset.

4.1.5.4 Acknowledgments

This work was supported in part by each of the followings: the CHIST-ERA IVAN project (20CH21_174081), the Swiss National Science Foundation (project 200021_175506), the Center for Biomedical Imaging (CIBM) of the Geneva - Lausanne Universities and the EPFL, and the Leenaards and Louis-Jeantet Foundations.
4.2 Conference proceedings article: Slepian guided filtering of graph signals

(This article has been published in Proceedings of SPIE, DOI: 10.1117/12.2528827)

Miljan Petrović\textsuperscript{a,b}, Dimitri Van De Ville\textsuperscript{a,b}

\textsuperscript{a} Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, Campus Biotech, Geneva, Switzerland
\textsuperscript{b} Department of Radiology and Medical Informatics, University of Geneva, Geneva, Switzerland

Abstract

Joint localization of graph signals in vertex and spectral domain is achieved in Slepian vectors calculated by either maximizing energy concentration ($\mu$) or minimizing modified embedded distance ($\xi$) in the subgraph of interest. On the other hand, graph Laplacian is extensively used in graph signal processing as it defines graph Fourier transform (GFT) and operators such as filtering, wavelets, etc. In the context of modeling human brain as a graph, low pass (smooth over neighboring nodes) filtered graph signals represent a valuable source of information known as aligned signals. Here, we propose to define GFT and graph filtering using Slepian orthogonal basis. We explored power spectrum density estimates of random signals on Erdős–Rényi graphs and determined local discrepancies in signal behavior which cannot be accessed by the graph Laplacian, but are detected by the Slepian basis. This motivated the application of Slepian guided graph signal filtering in neuroimaging. We built a graph from diffusion-weighed brain imaging data and used blood-oxygenation-level-dependent (BOLD) time series as graph signals residing on its nodes. The dataset included recordings of 21 subjects performing a working memory task. In certain brain regions known to exhibit activity negatively correlated to performing the task, the only method capable of identifying this type of behavior in the bandlimited framework was $\xi$-Slepian guided filtering. The localization property of the proposed approach provides significant contribution to the strength of the graph spectral analysis, as it allows inclusion of a priori knowledge of the explored graph’s mesoscale structure.

Keywords: Slepian, graph, filtering, neuroscience, brain, spectrum.

4.2.1 Introduction

Complex systems and dynamic processes existing on top of them are successfully modeled as graphs with nodes connected by edges reflecting the intrinsic relationships between them, and signals whose samples are associated with the graph nodes. This framework referred to as
4.2. Conference proceedings article: Slepian guided filtering of graph signals

Graph signal processing (GSP) has brought insight in many different scientific disciplines (Ortega et al., 2018) including image processing, neuroscience, geoscience, etc. The key concepts of the framework are generalizations of the Fourier transform, filtering, convolution, wavelets, and other operators, from 1-dimensional to the domain of graph vertices (Sandryhaila and Moura, 2013, 2014b; Shuman et al., 2013). Strength of this methodology is the ability to process signals with respect to the underlying pattern of connectivity between signal origins.

On the other hand, the question of what would be the graph signal mostly concentrated in a certain subset of nodes has been addressed by graph Slepian vectors (Tsitsvero et al., 2016; Van De Ville et al., 2017) deriving from a more general continuous framework (Slepian and Pollak, 1961). Slepian functions allowed for an appropriate signal sampling and reconstruction (Tsitsvero and Barbarossa, 2015) as well as sophisticated graph embedding solutions (Petrovic et al., 2019).

Whereas GSP focuses on processing signals living on the graph, Slepian analysis explores the properties of the graph itself. The idea of merging the two methodologies seems beneficial as it would provide a framework for processing signals on the nodes while taking into account the mesoscale local features of the underlying graph besides the immediate neighborhood node connectivity. Here, we investigate one possible ingredient of such a framework and we will refer to it as Slepian guided filtering of graph signals. However, this kind of framework comes with certain precautions regarding theoretical conditions for a valid spectral graph analysis. Furthermore, depending on the explored graph, one may indeed benefit from including Slepian vectors into the filtering procedure, but this is for sure not guaranteed as the graph itself may not exhibit a modular structure (Fortunato and Hric, 2016). Hence, we devoted the work presented in this paper to explore the benefits of Slepian guided filtering in the context of random graphs and human brain models, while considering the theoretical circumstances of such a procedure.

In the next sections of the paper we briefly recall the mathematical basis of GSP (Sec. 4.2.2.1) and graph Slepians (Sec. 4.2.2.2), and build upon this to present Slepian guided filtering (Sec. 4.2.3). Then, we demonstrate the results of a power spectrum analysis of random signals on graphs with random connectivity (Sec. 4.2.4), in order to provide motivation for an application to real-world data. The main significance and advantages of the proposed approach are emphasized through our example including a graph which models the human brain, and signals which reflect neural activity (Sec. 4.2.5). We conclude with open questions and further possibilities to refine the described framework (Sec. 4.2.6).

4.2.2 Background

4.2.2.1 Graph signal processing framework

Graphs provide an invaluable mathematical model of real-world networks for analyzing data on an irregular domain. They constitute a pair of the set $\mathcal{V}$ of $N$ vertices usually labeled as
Chapter 4. Processing global signals in local environment

\[ \mathcal{V} = \{1, 2, 3, \ldots, N\}, \]  
and the set \( \mathcal{E} \) of edges denoted with tuples \((i, j)\) when there is an edge between the \( i \)th and \( j \)th vertex and weights \( a_{ij} \). Edge weights together build the adjacency matrix \( A \) of size \( N \times N \) with entries \( a_{ij} \). Graph signal is defined as a set of scalar values associated with each node of the graph, are then processed taking into account the underlying graph edges (Shuman et al., 2013). This framework is built upon the graph Laplacian matrix \( L = D - A \), where \( D \) is the diagonal degree matrix with entries \( d_{ii} = \sum_{j=1}^{N} a_{ij} \). A vector \( x \) of size \( 1 \times N \) corresponds to the graph signal (for a fixed chosen node labeling/ordering) and its graph Fourier transform (GFT) pair is further defined as

\[
\hat{x} = U^\top x \quad \text{and} \quad x = U\hat{x}, \quad (4.11)
\]

where \( \hat{x} \) is the vector of spectral coefficients, and \( U \) is the eigenbasis matrix containing orthogonal eigenvectors of \( L \) as its columns, with respect to the eigendecomposition \( L = U\Lambda U^\top \). Assuming the eigenvectors are ordered by increasing eigenvalues, a low pass filtered (with cutoff “frequency” at the \( w \)th eigenvector) signal equals \( y = UH^\top x \), where \( H \) is a diagonal matrix with first \( w \) entries equal to 1 and others to 0.

In this context, the graph Laplacian matrix \( L \) is the chosen matrix for a more general operator called the shift \( \Sigma \) (Ortega et al., 2018). This operator allows to define a space of polynomial filters whose implementation in the vertex domain is highly beneficial from the computational point of view, as an eigendecomposition of a large matrix is time- and memory-consuming. With an assumption of a distinct set of matrix eigenvalues, every shift-invariant (commuting with the shift matrix) filter can be found as a polynomial of the shift operator (Chen et al., 2018) \( p(L) \) where the polynomial \( p \) is designed (often using Chebyshev recursion) to approximate the cutoff characteristic of the filter’s response in the range of the relevant eigenvalues.

4.2.2.2 Graph Slepians

Approaching a different problem, one may wish to find signals which, for a given graph and its subgraph (denoted with diagonal selection matrix \( S \) with entries 1 for nodes within the subgraph, and 0 otherwise), exhibit maximally concentrated energy in the subgraph while having a localized (bandlimited) spectrum (at the first \( w \) low-pass spectral components). These signals are known as the graph Slepians (Tsitsvero et al., 2016), and derive from a more general class of prolate spheroidal functions in the time domain (Slepian and Pollak, 1961; Slepian, 1978). Graph Slepians are found by maximizing (over coefficients vector \( \hat{x} \)) the energy concentration criterion \( \mu \) or minimizing the modified embedded distance criterion \( \xi \) (Van De Ville et al., 2017):

\[
\mu = \frac{\hat{x}^\top W^\top U^\top S W x}{\hat{x}^\top \hat{x}} \quad \text{and} \quad \xi = \frac{\hat{x}^\top \Lambda^{1/2}_W W^\top U^\top S W \Lambda^{1/2}_W W x}{\hat{x}^\top \hat{x}}, \quad (4.12)
\]  
\[ Certain authors prefer to use the adjacency matrix directly to define graph signal operators.

where $W$ is a diagonal degree matrix with first $w$ entries equal to 1, and $\Lambda_W = W^T \Lambda W$. Hence, this reverts to finding the eigenvectors of the concentration matrix $C = W^T U^T S U W$ or the modified concentration matrix $C_{emb} = \Lambda_W^{1/2} C_W \Lambda_W^{1/2}$, and subsequently calculating Slepian $x_k = U \hat{x}_k$ for $k = 1, \ldots, w$. Even though both Slepian designs ($\mu$ and $\xi$) share the properties of vertex-domain localization and spectral bandlimitedness, the key difference between them lies in the notion of smoothness. Whereas $\xi$-Slepian signals are found as smooth and localized, $\mu$-Slepian signals are lacking this constraint as their energy is locally concentrated regardless of what is the oscillatory origin of that energy.

### 4.2.3 Slepian guided graph signal filtering

The main question being explored in this paper is whether it is beneficial to define graph signal filtering and signal’s spectral representation through the Slepian vectors as an orthogonal basis. Mathematically, if we have graph Slepian signals derived using the two criteria in Eq. (4.12) and stored in matrices $U_\mu$ and $U_\xi$, respectively, of sizes $N \times w$, then for a signal $x$ we can define the $\mu$-Slepian GFT pair as:

$$\hat{x} = U_\mu^T x \quad \text{and} \quad x = U_\mu \hat{x}, \quad (4.13)$$

and the $\xi$-Slepian GFT pair as

$$\hat{x} = U_\xi^T x \quad \text{and} \quad x = U_\xi \hat{x}. \quad (4.14)$$

Note that these transforms are “almost unitary”. The columns of $U_\mu$ and $U_\xi$ are orthonormal, but mapping from $x$ to $\hat{x}$ reduces data dimensionality, i.e. the size of $\hat{x}$ is $w \times 1$ and $w < N$. Hence the vector norm is not preserved entirely and this depends on the choice of $w$. Indeed, the total power of the spectral representation $\hat{x}$ accounts for $(100 \cdot w/N)$ percent of the power of $x$. Even though a certain amount of information is lost with this transform, it was shown that Slepian is useful for exploring localized behavior. Indeed, for broadband signals, a local excerpt/subgraph (in time/graph domain) of the signal may not include all the spectral components and thus actually be bandlimited.

Once GFT has been defined, we can proceed to construct operators for graph signal processing. Here we explore only one – filtering. In the analysis performed in Sec. 4.2.5 only spectral domain filtering was performed, though we will give certain notes on the vertex domain filtering. A filtered signal $\tilde{x}$ is derived from $x$ through spectral domain as:

$$\tilde{x} = U_\Box H U_\Box^T x, \quad (4.15)$$

where $\Box$ can stand for either $\mu$ or $\xi$ for a given value of $w$ and a fixed subgraph of interest.
(represented in the selection matrix $S$ when deriving Slepian vectors). Depending on the diagonal filter response matrix $H$ we can have a low pass filter (with nonnegative entries in the top left corner and zeros elsewhere), or a high pass filter (with nonnegative entries in the bottom right corner and zeros elsewhere). The terminology assumes ordering of eigenvectors according to decreasing $\mu$ and increasing $\xi$. The size of $H$ is $w \times w$, so in this case the "high"-$\xi$ frequencies may still appear smooth if $w \ll N$. They are only high w.r.t. to the bandwidth extracted by the Slepian basis. Eq. 4.15 shows the analogy with the multiplication property of the Fourier transform. Indeed, the filtered output graph signal is derived after calculating the GFT of the input $(U^T x)$, multiplying with the frequency response of the filter $(H)$ and finally taking the inverse GFT of the result (multiplying with $U$).

As mentioned in Sec. 4.2.2.1, linear shift-invariant filters operating in the vertex domain as shift matrix polynomial multiplication $\tilde{x} = p(\Sigma)x$ correspond to spectral domain filtering defined through eigenvectors of the shift matrix. In the case of Slepian bases $U_\mu$ and $U_\xi$, the shift operator corresponds to expressions of concentration matrices $UCU^T$ and $UC_{emb}U^T$, respectively, which for full bandwidth ($w = N$) revert to the selection matrix $S$, and selection modified Laplacian $L^{1/2}SL^{1/2}$. Note that the calculation of Slepian vectors usually considers a column-truncated $U$ instead of indicator matrix $W$ applied to $U$, in order to get concentration matrices of sizes $w \times w$, thus speeding up the subsequent eigendecomposition. However, if we use $C$ and $C_{emb}$ to find the shift operators, we need to comply with dimensionality condition for matrix-vector multiplication and use indicator matrices to build $C$ and $C_{emb}$ of size $N \times N$ as in Eq. 4.12.

The condition for a distinct set of eigenvalues of $\Sigma$ cannot be ignored (Chen et al., 2018), which calls for a cautious definition of Slepian guided filtering. In general, there are around as many $\mu$ eigenvalues equal to 1 as there are nodes in the selected subgraph for close to full bandwidth. Indeed, for bandlimited decompositon this number reduces but there may still exist a degenerate subspace (corresponding to $\mu = 1$ with multiplicity strictly greater than 1) of the concentration matrix. Similarly, there is a degenerate subspace $\xi = 0$ whose dimensionality depends on the size of the subgraph and the bandwidth. Hence, if we consider concentration matrices as shift operators, there may be spectral domain operators that do not have their analogues in the vertex domain, at least not in the class of polynomial-shift operators. In order to make sure that the spectral filter can be represented as $p(UCU^T)$ (or $p(U_{emb}U^T)$), the same value of entries of $H$ must be used for all spectral components spanning the same eigenvalue subspace. In practice, this means that one should be particularly careful for $\mu$-Slepian filtering as there may even exist two degenerate subspaces at far ends ($\mu = 1$ and $\mu = 0$) so that the filtering cutoff frequency (eigenvector index) is only "allowed" in the mid-band. The situation is somewhat simpler in the case of $\xi$-Slepian guided filtering since the only degenerate subspace is at the low end ($\xi = 0$); in this case, filtering can be performed with the cutoff eigenvector inside the strictly positive range $\xi > 0$. 
4.2. Conference proceedings article: Slepian guided filtering of graph signals

4.2.4 Power spectrum density estimation on random graphs

Spectral domain representation of a graph signal provides insight into oscillatory patterns of the signal along the neighboring nodes of the graph. Connectivity of the graph nodes is employed by using the Laplacian matrix to define GFT, hence incorporating the geometrical/topological properties of the graph. However, as we show in the following example, certain signal behavior may remain hidden in the noise and appear to have a spectrum corresponding to a random signal. The approach of defining the spectral domain with Slepian eigenbasis allows to properly differentiate between these types of behavior. We decided to use random graphs and estimate power spectrum density of signals defined on their vertices in order to show how the Slepian eigenbasis overcomes the problem existing in the case of the Laplacian spectrum.

An ensemble of 10,000 Erdős–Rényi random graphs (Erdős and Rényi, 1959) with $N = 100$ nodes and edge probability $p = 0.2$ was generated. Furthermore, for each of them, an instance of three types ($s_1$, $s_2$, and $s_3$) of random signals was generated on the vertices. Signal $s_1$ is a zero-mean unit-norm random signal with uniformly distributed values. Signal $s_2$ is uniformly distributed, has zero mean and unit norm, and its values within the subset of 50 nodes (right semicircle in Fig. 4.8) exhibit higher variance than values on the rest of the vertices. Signal $s_3$ is similar to $s_2$ but with only the mean being different between the two subsets of nodes. One instance of each of the three signals is illustrated in Fig. 4.8 with the underlying graph.

In analogy with periodogram of time series, the power spectrum density of graph signals are estimated as the average (over the ensemble of pairs graph/signal) of the spectral powers, i.e., element-wise square magnitude of the GFT of signals. In order to compare the proposed approach with the existing one, we repeated this for three definitions of GFT: using full Laplacian eigenbasis (100 eigenvectors), $\mu$-Slepian eigenbasis (75 eigenvectors due to chosen finite bandwidth $W = 75$), and $\xi$-Slepian eigenbasis (75 eigenvectors). The PSDs are plotted in blue, red, and cyan lines, respectively, assuming the ordering of spectral components w.r.t. increasing eigenvalues for Laplacian eigenvectors and $\xi$-Slepian, and decreasing eigenvalues for $\mu$-Slepians (Fig. 4.8, bottom row). Note that due to $W = 75$ and the number of nodes in the subgraph being 50, the spectral bands encompass 25 eigenvalues $\xi = 0$, 25 eigenvalues $\mu = 0$ (both gray shaded areas), and 25 eigenvalues $\mu = 1$ (yellow shaded area). We also explored how the PSD estimates degrade if the selection matrix in the definition of Slepian basis does not accurately target the nodes of the subgraph which indeed exhibits localized behavior. We shifted the selection for 20% (10 nodes) around the topological circle of the graph representation (Fig. 4.8), therefore excluding 10 nodes with relevant subgraph behavior and including 10 irrelevant ones. These PSD estimates are presented in Fig. 4.8 with dashed magenta line ($\mu$-Slepian basis) and dashed green line ($\xi$-Slepian basis).

Evidently, Laplacian PSD does not differentiate between $s_1$, $s_2$, and $s_3$. All three have broadband and almost flat PSDs (note the corresponding ranges of values in Fig. 4.8) which is expected knowing that the signal is (globally) random. Even though the special localized
Figure 4.8 – On the left: an instance of random graph \( (N = 100) \) and three different signals with values encoded by color and length of vertical lines: \( s_1 \) - a random signal, \( s_2 \) - random signal with higher variance on the nodes forming the right semicircle of the topology, \( s_3 \) - random signal with higher mean on the nodes forming the right semicircle. On the right: PSD estimates of these types of signals defined as the average power spectrum over an ensemble of 10000 graph/signal pairs defined with Laplacian (blue), \( \mu \)-Slepian (red and magenta), and \( \xi \)-Slepian (cyan and green). In the case of Slepian bases, the solid lines (red, cyan) correspond to cases when the selection matrix accurately extracts the right semicircle nodes, whereas dashed lines (magenta, green) correspond to cases where the selection is shifted by 20% of the number of nodes in the relevant subgraph. The shaded areas represent the spectral bands defined by the high-multiplicity eigenvalue equal to 0 (gray) or 1 (yellow). For visual clarity, plots of Laplacian spectra are zoomed around 0.01 at \( y \)-axis so that the first component (with eigenvalue equal to 0) cannot be seen, but as it reflects the global signal mean, it is equal to 0 in all instances. The plots in the last row represent the mean graph eigenvalue spectrum for each of the three used bases.
4.2. Conference proceedings article: Slepian guided filtering of graph signals

properties in the context of mean \( (s_3) \) and variance \( (s_2) \) are introduced in the half of the vertex-domain of the signal, the Laplacian approach is not powerful enough to grasp this kind of signal behavior. Furthermore, the random signal \( s_1 \) shows Slepian-based PSDs with the same features (broadband, almost flat) confirming that the signal \( s_1 \) is indeed random (top row in Fig. 4.8).

On the other hand, \( \mu \)- and \( \xi \)-Slepian PSDs of \( s_2 \) and \( s_3 \) are different from the one of the random signal \( s_1 \), as well as from each other. As expected, Slepian eigenvectors are able to extract relevant information on localized signal behavior. Energy concentration Slepian PSD of \( s_2 \) has higher values in range \( \mu = 1 \), lower in \( \mu = 0 \), and the density is decreasing in the middle, in overall resembling the (scaled) eigenvalue spectrum of the corresponding concentration matrix. The \( \mu \)-Slepian PSD of \( s_3 \) exhibits a different shape, having high values in both extremal eigenvalue ranges, with a significant trough in the middle band. As compared to the PSD estimate of \( s_1 \), this suggests that the difference of local signal variance is encoded in reducing the relative contribution of nonconcentrated spectral components (\( \mu \approx 0 \)). Since the variance of the signal is indeed locally diminished in the unselected subgraph, that is where the “nonconcentrated” components exist. Furthermore, reducing the local mean in the unselected subgraph reduces the contribution of the spectral band \( 0 \ll \mu \ll 1 \), which in general consists of components concentrated in both selected and unselected subgraphs. If we represent the difference of local means as a piecewise-constant signal over the two subgraphs, it becomes more intuitive that this feature would be encoded in the class of equally concentrated and nonconcentrated signals, \( i.e. \) those existing in the midband of \( \mu \).

Significant differences of PSDs of \( s_2 \) and \( s_3 \) are also present in the case of \( \xi \)-Slepian eigenbasis. Signal \( s_2 \) exhibits low contribution of smooth low-frequency (low modified embedded distance) spectral components. Indeed, variance in the selected subgraph is higher relative to the rest of the graph, so one should not expect smooth components in the Slepian spectrum. On the other hand, PSD of \( s_3 \) as compared to that of \( s_1 \) exhibits an increase of the middle frequency components, seen as a prominent peak in the mid-spectrum. This can be explained similarly as in the case of \( \mu \)-spectrum. A piece-wise constant signal can hardly be classified as neither smooth or nonsmooth, but rather as piecewise-smooth, hence it is explained by components corresponding to moderate values of \( \xi \).

The ability to extract otherwise hidden signal behavior makes the Slepian GFT basis a powerful tool for detection of localized processes on graphs. Even in cases when the selection matrix does not accurately (with 20% of subgraph nodes missed) target the subgraph of interest (dashed magenta and green PSDs in Fig. 4.8), the spectral contents of the signals are distinct enough to suggest an existing discrimination between the explored types of graph signals. This result points to a potential benefit of processing signals with Slepian-defined GFT – manipulation of signals with complex behavior which exist on segregation-modeled graphs (with partial a priori knowledge of meaningful subgraphs). Hence, in the next section, we explore graph signal filtering in Slepian spectral domain.
4.2.5 Guided filtering of brain signals

Neuroscience has particularly advanced since the development of graph-based methods and GSP (Huang et al., 2016, 2018). Exploring the human brain using graph models allowed for high-level abstraction approaches thus simplifying the complexity of the problem at hand. By parceling the brain into relevant regions of interest one can analyze segregation/integration levels of both anatomical and functional nature (Sporns and Betzel, 2016). The two most common approaches are to build a graph by connecting anatomically linked regions (as nodes) as derived from diffusion weighted imaging (Bihan and Lima, 2015), or by connecting regions with correlated time courses of BOLD (Blood-Oxygenation-Level-Dependent (Leonardi et al., 2013)) signals recorded in these regions with functional magnetic resonance imaging (fMRI) (Logothetis, 2008). Efforts to jointly explore both structure and function of the brain (Ajilore et al., 2013; Becker et al., 2018; Abdelnour et al., 2018) have been recently boosted by GSP techniques through building a graph out of anatomical connections, but processing BOLD signals residing on that graph in order to examine how the dynamics of neural activity are aligning with the underlying structure (Huang et al., 2018; Medaglia et al., 2017). Here, we follow this joint approach.

The dataset we use in the following examples consists of publicly available Human Connectome Project (Essen et al., 2012; Glasser et al., 2013) structural and functional MRI recordings of the human brain. We constructed a weighted graph representing neural pathway strengths between 360 brain regions (Glasser et al., 2016) for 21 subjects and averaged the results into a unique graph for all available subjects. The BOLD signals (sampling rate 0.72 Hz, duration approx. 5 min) acquired from fMRI were averaged over voxels belonging to the same region/node and z-scored, resulting in a signal matrix of size $360 \times 395$ per subject; i.e. we have as many graph signals as there are time points in the functional data. The signals were recorded while subjects performed a working memory task involving visual stimuli (Essen et al., 2012). Furthermore, we acquired resting-state BOLD signals in matrices of size $360 \times 1990$ per subject.

In order to justify the Slepian guided filtering in the context of the available dataset, we have performed a PSD estimation analysis analogous to the random graph analysis in Sec. 4.2.4. The subgraphs for focusing graph Slepians were chosen according to the general knowledge on structural and functional neuroanatomy (Fig. 4.9). These are occipital (58 nodes), sensorimotor (42 nodes), temporal (82 nodes), parietal (38 nodes), cingulate (54 nodes), and frontal (86 nodes) regions. For each of the subgraphs and the two Slepian designs ($\mu$ and $\xi$), we have averaged the power spectra based on Slepian eigenbasis over all time points per subject, considering separately resting-state and task-based signals. The resulting eigenvalue spectra of the Slepian ($\mu$ and $\xi$) and Laplacian ($\lambda$) bases are shown in Fig. 4.10, whereas the final PSD estimates are plotted in Fig. 4.11 (full and dashed lines denote median across subjects, and the shaded dark gray areas reflect the 25th and 75th percentiles). All Slepian signals were designed with bandwidth $W = 100$ so that $W$ is higher than the maximal number of nodes in a focused subgraph, thus making sure the spectra would encompass both degenerate and non-degenerate subspaces.
Comparing the PSD estimates to those of random graph signals in Fig. 4.8 reveals certain similar shapes. Higher contribution of components in range $\mu \gg 0$ than in $\mu = 0$ are evident in all $\mu$-PSDs with the clear split coinciding with nonzero and zero eigenvalues for the cases of occipital and cingulate areas. This reflects variance of a graph signal higher in the focused subgraph than in the rest of the graph. On the other hand, difference of local mean is encoded by higher contribution of components with $\xi$ close but not equal to 0 than by the rest of the spectrum (Fig. 4.8). To a certain extent, this behavior is reflected by a concave arc-shaped trend of PSD estimates in the case of occipital and temporal subgraphs. Apparent feature (spatial localization) of graph signals living on the brain network model introduce a caveat for Laplacian graph signal filtering. As in the example including random graphs, one could expect certain signal behaviors to remain hidden from the Laplacian spectra and thus not be suitable to filtering in a conventional GSP manner (Shuman et al., 2013). This motivated us to perform Slepian guided graph signal filtering on the available dataset and check whether there is any information accessed by the Slepians and missed by the graph Laplacian approach.

We performed graph signal filtering of task-based data using Laplacian spectra as described in Sec. 4.2.3 and using $\mu$ and $\xi$ Slepian spectra as in Eq. 4.15. Filters were designed as ideal spectral domain operators (values of $H$ are binary). In the case of Laplacian filtering we defined a low pass filter preserving components for which $\lambda \leq 4$ resulting in $W = 34$. In the case of Slepian filtering, the bandwidth of Slepian eigenvectors was $W = 100$, whereas the filter passband widths were set to include components at least as concentrated as $\mu \geq 0.5$ and at most as smooth as $\xi \geq 1$ (see dashed lines in Fig. 4.10). This way, the cutoff frequency of all filters was chosen so that it is not splitting any of the degenerate subspaces ($\mu = 1$, $\mu = 0$, or $\xi = 0$), thus complying to the shift-enabled condition (Chen et al., 2018), i.e. the passband always includes the whole subspace $\mu = 1$, and always excludes all the spectral components in
Figure 4.10 – Eigenvector spectra for $\mu$- (top left), $\xi$-Slepian (top right) and Laplacian eigenbasis (bottom). Different shades of red ($\mu$) and blue ($\xi$) denote cases when Slepians are concentrated in different subgraphs (labeled in the legend). For clarity $x$-axes are in log-scale for Slepian designs. Horizontal dashed lines cross the spectra at points used as cutoff frequencies of filters used in the example in Fig. 4.12. These result in cutoff eigenvector indices (ordered as subgraphs in the legend from top to bottom): 11, 7, 34, 4, 21, and 26 for $\mu$-Slepian, and 12, 8, 36, 6, 21, and 30 for $\xi$-Slepian, and 34 for Laplacian filtering.

The subspaces $\mu = 0$ and $\xi = 0$. Values of the cutoff eigenvector indices $\Omega$ are: 11, 7, 34, 4, 21, and 26 for $\mu$-Slepian, and 12, 8, 36, 6, 21, and 30 for $\xi$-Slepian, and 34 for Laplacian filtering (ordered as subgraphs in Fig. 4.11 from top to bottom). The passbands are marked with red, blue, and dark gray flags above PSD estimates in Fig. 4.11. We note that although the idea of Slepian guided graph signal filtering was briefly introduced in Ref. 79 for the purpose of surrogate-based signal excursions analysis, this condition for valid shift-invariant filters was not previously considered.

Inspection of the filtered signals was done in terms of the known experimental task paradigm of the fMRI recordings. After filtering graph signals at all time points for all subjects, we reconstructed the BOLD time series and explored how well they reflect subject’s response to the task stimuli. Results are given in Fig. 4.12. We present the outcomes of 3 cases guiding the Slepian vectors to the occipital, temporal, and cingulate subgraph (from top to bottom in Fig. 4.12). For each subject’s filtered (and non-filtered) BOLD time courses we calculated the mean signal over all nodes in the corresponding subgraph. The plots in Fig. 4.12 then show the
4.2. Conference proceedings article: Slepian guided filtering of graph signals

Figure 4.11 – Power spectrum density estimates (averaged over available time points) of brain graph signals calculated using Slepian GFT focused on 6 subgraphs: (from top to bottom) occipital, sensori-motor, temporal, parietal, cingulate, and frontal region (see Fig. 4.9), and using Laplacian eigenbasis (bottom middle). The estimation was performed separately for Laplacian, $\mu$- (left) and $\xi$-Slepian's (right), and for resting-state and task-based BOLD time courses (in the appropriately labeled columns). Zero-eigenvalue subspaces are denoted with light gray shaded areas and degenerate $\mu = 1$ subspaces with yellow shaded areas. Dark gray shaded areas reflect 25th and 75th percentiles of PSD estimates over subjects. Bandwidth of Slepian signals was set to $W = 100$. Red, blue, and dark gray flags mark the passbands of filters used in the following example in Fig. 4.12.
Figure 4.12 – Mean (over nodes in a subgraph) task-based BOLD signals before (black lines) and after graph filtering (blue lines: Laplacian, cyan lines: ξ-Slepian, and red lines: µ-Slepian). The lines denote median signal over 21 subjects with the shaded areas around representing the range between 25th and 75th percentiles. The shades of yellow follow the task paradigm from bright to dark color denoting fixation time, 0-back task and 2-back task. Values of correlation coefficient $R$ (and corresponding $p$-value) between the median signals and 2-back task paradigm (convolved with HRF) are reported in dark yellow font. The results are shown for 3 subgraphs: (from top to bottom) occipital, temporal, and cingulate regions.
median (solid lines) and the range between 25th and 75th percentile (gray shaded areas) over all 21 subjects. Here we chose the median instead of mean as a more robust metric in order to account for the heterogeneous nature of the neuroimaging data across different people. We compared these median signals for all filtering cases: Laplacian, \( \xi \)-Slepian, \( \mu \)-Slepian, and no (allpass) filtering. In each case we report the correlation coefficient \( R \) (and the corresponding \( p \)-value) with the binary indicator function of the 2-back working memory task\(^2\). In order to emphasize (dis)similarity between the two, in Fig. 4.12 we reflect the full task paradigm in time as shaded areas with light to dark yellow background colors denoting fixation (rest), 0-back, and 2-back task, in that order.

In the occipital region, values of \( R \) indicate a moderate positive and statistically significant correlation (above 0.4) regardless of the chosen filtering. Since the task in question includes visual stimuli (images were shown to subjects as a part of the working memory task (Essen et al., 2012)), one would indeed expect to capture neural activity following the paradigm in the occipital region, as it is known that visual processing in the human brain is highly localized in this area. However, focusing on the occipital subgraph does not show the advantages of using Slepian guided filtering over other approaches. Here it was included to emphasize the fact that the choice of the appropriate type of filtering should be guided by the nature of the graph and of the subgraph of interest. On the other hand, in the case of temporal and cingulate subgraphs, all types of filtering except \( \xi \)-Slepian exhibit very low (up to 0.1 in absolute value) and/or statistically insignificant value of \( R \). In fact, Laplacian filtering gives somewhat higher and significant \( R = 0.119 \), but this may be attributed to a response to acoustic noise, since fMRI recordings are very loud and the auditory processing is localized in the temporal lobe of the human brain (Tomasi et al., 2005). Still, \( \xi \)-Slepian filtering manages to extract significant negative correlation (\( R = -0.282 \)) of the mean BOLD signal in the temporal region, as well as in the cingulate subgraph (\( R = -0.269 \)). This result has not been put forward by the other approaches even though it might have been anticipated from the neuroscientific point of view. In fact, a similar filtering approach repeated on a finer scale with 22 nonoverlapping brain regions (Glasser et al., 2016) has revealed that the biggest contributors to this “correlation boosting” are the lateral temporal and posterior cingulate regions. These are parts of the so called default mode network (DMN) (Raichle, 2015) which represents a pattern of brain regions known to often exhibit neural activity negatively correlated with whatever task is being performed by the subject. Hence, the only way one can confirm the “presence” of the activity in these parts of the DMN at smooth frequencies (low-pass bandlimited behavior) is by employing \( \xi \)-Slepian guided filtering. Energy concentration based Slepians do not show clear advantages in this example, except just slightly higher \( R \) and lower significant \( p \) value (for alpha level 0.05) in the case of the cingulate region, as compared to no filtering. However, the structured shape of \( \mu \)-based PSD estimates in Fig. 4.11 do indicate that there may still be an application where this type of Slepians would be invaluable.

\(^2\)More precisely we used the binary indicator convolved with haemodynamic response function (HRF) mathematically modeled with gamma functions. HRF is used to account for the time delay between the actual neural response to stimulus and the cardiovascular response proxy that is in fact being captured by fMRI (Lindquist et al., 2009).
Chapter 4. Processing global signals in local environment

The ability of $\xi$-Slepian basis to extract information otherwise inaccessible by classical graph signal filtering could be partially understood as a consequence of the specific graph connectivity pattern. The main difference between Laplacian eigenvectors and $\xi$-Slepian signals lies in the localization property. Laplacian eigenvectors reflect signals changing smoothly across nodes with highly weighted edges connecting them, and nonsmoothly across loosely connected or disconnected nodes. In this context, for a given bandwidth $W$, a lowpass Laplacian filter can extract globally smooth signals across the whole graph. However, if a signal is locally smooth across a certain subgraph, and outside it is either nonsmooth or smooth but in an anticorrelated fashion, this type of behavior would be considered high frequency by the graph Laplacian. Note that this is the case in the example of the cingulate region (Fig. 4.13). Cingulate subgraph is loosely connected to the whole graph, and just slightly more connected to the frontal region. We would expect both cingulate and frontal region activity to be correlated with the task indicator function, with negative and positive coefficient values, respectively (cingulate as it includes part of the DMN, and frontal as it is generally linked to human cognitive function). Despite this, a certain degree of anticorrelated behavior across the border between these two subgraphs could lead to such oscillations to be identified as high frequency by the Laplacian, thus being cut off during filtering and yielding a non-significant correlation value when focusing on the cingulate region (Fig. 4.12). On the other hand, the nature of Slepian designs includes the notion of localized behavior, whereas the $\xi$ criterion exactly optimizes for locally smooth patterns inside the selected subgraph. Hence, for the same bandwidth $W$, Slepian basis preserves the task anticorrelated behavior in the cingulate region as it is considered low frequency. The same holds for the case of focusing on the temporal subgraph. Indeed, temporal subgraph is highly connected to the occipital and frontal regions where we identify positive correlation with the task. As a consequence, even a small amount of task-negatively correlated activity in the temporal region renders the pattern globally high frequency as seen by the graph Laplacian. Nonetheless, $\xi$-Slepian filtering successfully extracts this activity in the temporal lobe due to subgraph selection. Finally, in the case of occipital region, even Laplacian filtering does extract task-positively correlated activity. This may be explained by the fact that the same type of activity can be seen in the parietal region (known to be associated with the working memory) so that Laplacian in fact sees the occipital activity as a part of spatially larger pattern (including both occipital and parietal subgraphs moderately connected – see Fig. 4.13) which is supposedly big enough to be considered as a globally smooth pattern and thus be included in the filter’s passband.

4.2.6 Conclusions and future work

Slepian guided graph signal filtering benefits from the localization step by using a priori knowledge of meaningful subgraphs. This leads to the ability to extract localized bandlimited signal behavior that cannot be identified by the graph Laplacian due to its affinity for global patterns. Graph signal PSD estimation on both random and brain graphs have suggested a big amount of signal’s information is stored in discrepancies between local activities. Furthermore, brain signal filtering example pointed to the importance of the underlying graph connectivity
as it constraints what signal patterns can and cannot be detected by different techniques.

Still, there remain several open questions and possibilities to further extend our understanding and capabilities of graph Slepians in GSP. Energy concentration Slepians do show intriguing PSD estimates and yet no apparent uniqueness of the results in the case of filtering. This may be explained by the fact that our example explores the slow task paradigm, whereas \( \mu \)-Slepians do not take into account the maximal possible smoothness of signals, but rather energy concentration regardless of what is its spectral nature, even when the energy is concentrated around smooth spectral components. Hence, one of possible directions to search for an on-hands application of \( \mu \)-guided filtering could be in resting-state BOLD signals, since the nature of these is still widely unclear. In this context, \( \mu \)-Slepians may provide a generalized approach to finding localized intrinsic brain activity.

Joint extraction of both locally smooth and energy concentrated Slepian signals was recently proposed (Petrovic et al., 2019) as optimization of the criterion \( \mu - \zeta \). Extending this type of signals, called \( \zeta \)-Slepians, to the framework of guided graph signal filtering is certainly worth exploring. First, the joint criterion would search for more relevant signal behavior. Secondly, \( \zeta \)-Slepians are found by eigendecomposition of a polynomial of normalized adjacency matrix excluding any bandwidth parameter. This simplifies the filtering procedure since we would no longer have two parameters (\( W \) and \( \Omega \)). Slepian bandwidth is implicitly defined, therefore the only free parameter is the filtering cutoff index \( \Omega \) (s.t. \( \Omega < N \)).

Further in the context of neuroscience, the described Slepian filtering may find use on graphs known as functional connectomes. There, BOLD signals are considered both as graph signals and as basis to calculate the edge strengths between corresponding nodes w.r.t. to temporal coupling of the time series (Huang et al., 2016). However, as a general framework, Slepian guided graph signal filtering is not constrained by a specific application domain. In all disciplines inclined towards GSP analysis, one could further benefit from guided filtering, i.e. geoscience and radar imaging (Gishkori and Mulgrew, 2019), sensor networks and smart grids (Jablonski, 2017), image processing (Cheung et al., 2018), etc. Finally, before applying more sophisticated techniques, one should further elaborate on the theoretical aspect of the Slepian guided filtering, possibly by investigating the localization feature from a graph wavelet perspective (Leonardi and Van De Ville, 2013; Hammond et al., 2011).

Acknowledgments

This work was supported in part by the Swiss National Science Foundation (project 200021_175506).

---

\( ^3 \)Here the term slow refers to temporal smoothness. However, in the irregular setting such as graphs, temporal and spatial frequencies may have interdependencies strongly affecting the performed analysis.
Chapter 4. Processing global signals in local environment

Figure 4.13 – Illustration of the brain graph based on anatomical connections. The 6 subgraphs used in Slepian designs are shown in red. The total strength of all edges between nodes in two subgraphs is encoded in the width and color of an arrow line between the subgraphs. Edges (arrows) between the most weakly connected pairs of subgraphs are excluded. The values $R$ at each subgraph indicate the correlation coefficient between the mean BOLD signal in the subgraph (nonfiltered) and the 2-back task paradigm. Significant values w.r.t. alpha level of 0.05 are colored in red.
The present dissertation explored the effects of community structure on the properties of graph signals and operators on them, as well as the relation between graph signals and certain subgraphs of interest. In this chapter, I summarized the main findings, along with potential avenues for future work that would build upon the foundations set in this thesis.

5.1 Summary of findings

Informing signal processing techniques with community structure: The framework of GSP with the modularity matrix as the shift operator provided a perspective on signals complementary to the conventional one based on the graph Laplacian. It allowed assessing to what extent the signal values distribute over vertices in alignment with the community structure, both assortative and disassortative (Petrovic et al., 2020). Indeed, as shown through the analysis of a transportation network, the assumption of nodes having similar signal values if they belong to the same community should be taken following the specific interpretation of edges in the given graph. Furthermore, the underlying community structure drove localization properties of graph signals. In the described work, the concentration of signal energy on a community was facilitated by assuming bandlimited signals in the modularity Fourier domain, whereas Laplacian-based bandlimitedness promoted energy concentration on a random subset of nodes.

Processing global signals in local environment: A novel design of Slepian signals was introduced – one that combines the criteria of high energy concentration and low modified embedded distance. The main advantages of these Slepians turned out to be computational stability and the loss of bandwidth as a design parameter, which was implicit in the criterion. Due to the very narrow range of solutions for the new criterion, the truncated set of $\zeta$-Slepians emerged as a great opportunity for efficacious dimensionality reduction. Furthermore, extracted local oscillatory patterns from a signal – by means of graph filtering based on Slepians – were deemed otherwise inaccessible, as demonstrated on human brain networks.
5.2 Perspectives of future work

Extension to directed and signed graphs

The work presented in this thesis is restricted to undirected graphs with their adjacency matrix symmetric and containing positive weights. An extension to graphs with directed and/or signed edges would grant the processing of signals on the many networks that are inherently such.

One class of approaches for this is to initially transform the graph adjacency matrix into one of the symmetric equivalents (Satuluri and Parthasarathy, 2011), though in this way information on edge directionality is at least partly lost. A somewhat more forthright manner is to explicitly define crucial graph description matrices in the case of directed graphs. Indeed, such Laplacian matrix can be formalized to include only in-degrees in the constituting degree matrix and then incorporated into GSP framework (Singh et al., 2016; Sardellitti et al., 2017). Similar extensions exist for the modularity matrix, including one that combines in-degrees \( k_{\text{in}} \) and out-degrees \( k_{\text{out}} \) to build the null model matrix (Malliaros and Vazirgiannis, 2013):

\[
Q_{\text{directed}} = A - \frac{1}{m} k_{\text{out}} k_{\text{in}}^T
\]

where \( m \) is the sum of all weights in \( A \). Community detection can be performed in the spectral domain using eigenvectors of symmetrized form \( Q_{\text{directed}} + Q_{\text{directed}}^T \). Within the GSP framework, the Fourier transform can be defined as the projection onto the eigenvectors from Jordan decomposition of \( Q_{\text{directed}} \). Such definition would preserve edge directionality.

Signed graphs have both positive and negative edge weights as entries of their adjacency matrices, and they are especially prominent in brain functional connectivity networks, where an edge reflects the correlation between two signals (nodes) (Sporns and Betzel, 2016). The sign property perplexes even the basic network metrics such as the node importance via eigenvector centrality since Perron–Frobenius theorem does not hold. Nevertheless, different generalizations of the Laplacian to signed graphs (Gallier, 2016; Roy et al., 2020) do have potential use in GSP (Dittrich and Matz, 2020).

The simplest form of modularity matrix for signed graphs derives from splitting the adjacency matrix \( A \) into \( A^+ \) and \( A^- \) with degree vectors \( k^+ \) and \( k^- \) and total degree volumes \( m^+ \) and \( m^- \), respectively. These subgraphs contain only positive or only negative (with the sign changed) weights. Then, the modularity matrix of a signed graph writes (Gómez et al., 2009).

\[
Q = \frac{1}{2m^+ + 2m^-}(A - \frac{1}{2m^+} k^+ k^+^T - \frac{1}{2m^-} k^- k^-^T)
\]

Alternatively, \( Q \) builds as a linear combination of the two corresponding matrices \( Q^+ \) and \( Q^- \) of the unsigned parts:
5.2. Perspectives of future work

\[ Q = \frac{2m^+}{2m^+ + 2m^-}Q^+ - \frac{2m^-}{2m^+ + 2m^-}Q^- \]  
\[ = \frac{2m^+}{2m^+ + 2m^-}(A^+ - \frac{1}{2m^+}k^+k^{+T}) - \frac{2m^-}{2m^+ + 2m^-}(A^- - \frac{1}{2m^-}k^-k^{-T}) \]  

The weighting of \( Q^+ \) and \( Q^- \) in terms of expressions of \( m^+ \) and \( m^- \) is sometimes modified w.r.t. relative importance of positive and negative weights for the particular network (Rubinov and Sporns, 2011). When edges are indeed correlations across time series, a different definition of modularity might be more appropriate (MacMahon and Garlaschelli, 2015).

**Considering communities beyond modularity matrix**

Directions for future research aforementioned in the concluding sections of manuscripts in preceding chapters include alternative null models for the modularity matrix (Massen and Doye, 2005; Newman, 2002) and the extension to overlapping communities (da Fonseca Vieira et al., 2020; Shen, 2013). Here, we discuss the community-aware framework beyond modularity-based approach.

The main weakness of modularity maximization as the community detection algorithm is the resolution limit (Fortunato and Barthélemy, 2007). Some ways to resolve it are to scale the null model matrix (Reichardt and Bornholdt, 2006) or add weighted self-loops in the adjacency matrix (Arenas et al., 2008). These approaches introduce a new parameter in the framework, used to control for the size of detected communities. Other methods that do not impose such a parameter should be preferable, such as maximization of the metric of asymptotical surprise (Nicolini et al., 2017), which is indeed resolution-free. However, one needs to first derive the expressions of edge probabilities across all pairs of nodes in the null model. Only with such matrix form of the model, it is consequently possible to define the shift operator. The same holds for other objective functions that might be used for community detection, e.g. significance (Traag et al., 2013).

In this thesis, the work is centered around informing the GSP framework with the communities **without their detection**. However, an alternative could be to first detect communities by an algorithm of choice and then modify the graph’s adjacency matrix by weighting each edge according to community memberships of the nodes it connects. Alternatively, the weight might also relate to the certainty that the detected community is close to the ground truth. Consequently, one can proceed with the conventional graph signal processing with adjacency or graph Laplacian as the shift operator. In this way, the community structure is emphasized within the adjacency, but careful work would be needed for specifying the appropriate weighting procedure.
Chapter 5. Summary and future perspectives

Graph signal processing guided by $\zeta$-Slepians

The purpose of $\zeta$-Slepians goes beyond dimensionality reduction, e.g. they can be regarded as a generalized eigenvector centrality (Petrovic et al., 2019). Here, I mention some research perspectives that aim to merge the GSP framework with Slepian signals.

Besides graph filtering in the context of $\zeta$-Slepiant based Fourier domain, other signal operators are worth exploring from a similar perspective. Graph wavelets are a convenient construct to tackle manifold denoising (Deutsch et al., 2016) or graph subsampling (Narang and Ortega, 2012). By detecting communities of a graph and then deriving Slepian vectors focused on each of the communities, one can yield a combined set of Slepians that might serve a similar purpose as wavelets, or more accurately – decomposition into frames (Leonardi and Van De Ville, 2013). However, only subsets of Slepians would be mutually orthogonal, so research is needed to pinpoint which vectors can be excluded to reach tight frames, without losing relevant information about a signal. Another option is to impose orthogonality by combining matrix Slepian criteria for all given communities in a certain way. This would produce frames for accessing signal localization properties across the underlying community structure.

On the other hand, communities could be sought by graph Slepians. The matrix criterion might be modified to include multiple subgraphs of interest, whose union is the whole graph, possibly through the use of the unit-circle complex-valued indicator function (Liergeois et al., 2019). Then, the exact membership of a node to a subgraph of interest is found by maximizing the Slepian criterion across all indicator functions. The described method would serve to identify community structure, with communities regarded as subgraphs whose local connectivity pattern allows high energy concentration of bandlimited signals.
A1 Supplementary material for Chapter 3

A.1 Supplementary material for Section 3.2

Proof of Lemma 1

For an expected graph under an SBM model with $N$ nodes and $K$ communities (cf. Eq. (3.21) and Prop. 2), a spectral-limiting operator $B$ that preserves all smooth eigenvectors $L$ or modular eigenvectors of $Q$ has block-diagonal form. All entries within the $K$ blocks of size $N/K$ equal $(K - 1)/N$ and the remaining matrix entries are $-1/N$.

The described block-diagonal operator $B$ relates to the expected Laplacian matrix:

$$B = (L - (a + c)I_N + bN1_N1_N^\top) \frac{K}{N(b - a)} - \frac{1}{N}1_N1_N^\top.$$  (A.1)

where $c$ is given by (3.22).

Therefore, $B$ has the same eigenvectors as $L$ with different eigenvalues. The eigenvalue of $B$ corresponding to the constant-entries vector is then given by:

$$\lambda_{constant}^{(B)} = (0 - (a + c) + bN) \frac{K}{N(b - a)} - \frac{N}{N} = 0.$$  (A.2)

Plugging eigenvalues of $L$ corresponding to smooth $(Nb)$ and non-smooth eigenvectors $(\frac{N}{K}(a + (K - 1)b))$ into a slightly different expression (disregarding outer products of the constant in (A.1)) gives:

$$\lambda_{smooth}^{(B)} = (Nb - (a + c)) \frac{K}{N(b - a)} = 1$$  (A.3)

$$\lambda_{non-smooth}^{(B)} = \left(\frac{N}{K}(a + (K - 1)b) - (a + c)\right) \frac{K}{N(b - a)} = 0$$  (A.4)

Since the only non-zero eigenvalues of $B$ correspond to all smooth eigenvectors of $L$ and are
equal to 1, it follows that $B$ is derived as the sum of outer products of these eigenvectors with themselves. In other words, $B$ preserves only the smooth eigenvectors. The same conclusion is reached for the modular spectral components of $Q$, since $L$ and $Q$ have the same eigenvectors (Prop. 2).
B Supplementary material for Chapter 4

B.1 Supplementary material for Section 4.1

Results of k-means clustering

In Supplementary Fig. S4, we present the proposed embedding from Figs. 4.4-4.6 and clusters of nodes with cooperation weight 1 derived by the k-means approach with 20 repetitions. Dimensionality of the considered data points was set to 2, i.e. entries of the two Slepian eigenvectors were used for clustering – the second and the third. The Silhouette method (Rousseeuw, 1987) was used to estimate the optimal number of clusters as the one which produces the minimal number of negative silhouette values. Convex hulls of each found cluster are represented by dashed black lines. The exact lists of neurons assigned to each cluster, for the three investigated cell types, are provided in Supplementary Tables 1, 2 and 3.

Evaluation of the clustering

In order to evaluate the inspected clusters of sensory neurons (Fig. 4.4B), interneurons (Fig. 4.5B) and motoneurons (Fig. 4.6B), we used statistical testing of communities (clusters). In all three cases, the nodes with importance \( m_i = 0 \) are considered as one additional cluster. We use the Newman-Girvan modularity as statistic (Newman, 2006). A vector of nodal assignments to clusters expresses its goodness of fit to the underlying adjacency matrix through the value of modularity \( Q \). It is calculated as:

\[
Q = \frac{1}{2w} \sum_{i,j}^N (|A_{bin}|_{i,j} - \frac{d_i d_j}{2w}) \delta_{C_i, C_j},
\]

where \( N \) is the number of nodes, \( w \) is the total strength of edges in the graph, \( A_{bin} \) is the graph binary adjacency matrix, \( d_i \) denotes the degree of the \( i \)th node, \( \delta \) is the Kronecker delta function, and \( C_i \) denotes the cluster to which the \( i \)th node belongs.
Appendix B. Supplementary material for Chapter 4

Table B.1 – Sensory neurons of the *C. elegans*. Columns correspond to clusters derived by optimized k-means.

<table>
<thead>
<tr>
<th>C_1</th>
<th>C_2</th>
<th>C_3</th>
<th>C_4</th>
<th>C_5</th>
<th>C_6</th>
<th>C_7</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVM</td>
<td>CEPDR</td>
<td>AFDL</td>
<td>PHAL</td>
<td>ADEL</td>
<td>ADFR</td>
<td>ADFL</td>
</tr>
<tr>
<td>PDEL</td>
<td>CEPVR</td>
<td>AFDR</td>
<td>PHAR</td>
<td>ADER</td>
<td>ADLR</td>
<td>ADLL</td>
</tr>
<tr>
<td>PDER</td>
<td>IL1DR</td>
<td>ASEL</td>
<td>PHBL</td>
<td>ALMR</td>
<td>ASGL</td>
<td>ALA</td>
</tr>
<tr>
<td>PHCL</td>
<td>IL1R</td>
<td>ASER</td>
<td>PHBR</td>
<td>CEPDL</td>
<td>ASGR</td>
<td>ALML</td>
</tr>
<tr>
<td>PHCR</td>
<td>IL1VR</td>
<td>ASIL</td>
<td>CEPVL</td>
<td>ASHR</td>
<td>ALNL</td>
<td></td>
</tr>
<tr>
<td>PLML</td>
<td>IL2DR</td>
<td>ASIR</td>
<td>IL1DL</td>
<td>ASKL</td>
<td>ALNR</td>
<td></td>
</tr>
<tr>
<td>PLMR</td>
<td>IL2R</td>
<td>AWAL</td>
<td>IL1L</td>
<td>ASKR</td>
<td>AQR</td>
<td></td>
</tr>
<tr>
<td>PVM</td>
<td>IL2VR</td>
<td>AWAR</td>
<td>IL1VL</td>
<td>AWBL</td>
<td>ASHL</td>
<td></td>
</tr>
<tr>
<td>OLLR</td>
<td>AWCL</td>
<td>AWCR</td>
<td>IL2DL</td>
<td>AWBR</td>
<td>ASIL</td>
<td></td>
</tr>
<tr>
<td>OLRQVR</td>
<td>URIXR</td>
<td>IL2L</td>
<td>ASIR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLL</td>
<td>BAGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLRQDL</td>
<td>FLPL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLRQDR</td>
<td>FLPR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLRQVL</td>
<td>PLNL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URYDL</td>
<td>PLNR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URYDR</td>
<td>PQR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URYVL</td>
<td>PVDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URYYR</td>
<td>PVDR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URXL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Supplementary Fig. S8, we present the results of the statistical approach for the case of sensory (red plots), inter- (grey plots) and motoneurons (green plots). The modularity values for the assignments to clusters as found by k-means clustering are marked with the dashed lines and labeled with \( Q_{\text{sensory}} \), \( Q_{\text{inter}} \), and \( Q_{\text{moto}} \) (Supplementary Fig. S8B). For the number of clusters estimated by the Silhouette method, we generated 999 random assignment vectors and calculated \( Q \) each time, in order to build a null distribution (Supplementary Fig. S8A).

As \( Q_{\text{sensory}} \), \( Q_{\text{inter}} \), and \( Q_{\text{moto}} \) are above the corresponding distributions of modularity for random assignments, we conclude that the found clustering is significant. Since these modularity values are strictly greater than all other \( Q \) values for random assignments, and, consequently, from any chosen percentile of the calculated distributions, the test rejects the null hypothesis that the chosen clustering is random at even very small significance levels. Finally, we note that the distribution of \( Q \) in the case of interneurons is slightly closer to the corresponding value of \( Q_{\text{inter}} \) than in the case of sensory or motoneurons. This can be expected, since interneurons are more strongly connected to other cell types, and thus, do not impose as strong communities as for the clusters formed from sensory or motoneurons.

Supplementary figures
B.1. Supplementary material for Section 4.1

Figure B.1 – For energy concentration (A) and modified embedded distance (B) criteria, eigenspectra at bandwidth $W = 100, 150, 200, 279$, as depicted by increasingly darker blue or purple shades, respectively. Yellow and orange vertical bars map locations of the eigenspectra at which Slepian vectors are shown (C and D). They are displayed for increasing bandwidth going from left ($W = 100$) to right ($W = 279$, full bandwidth). For energy concentration (C), the first row illustrates two Slepian vectors mapping the start of the spectrum (normalized indices of 0.1 — strongly concentrated in $S$ — and 0.4 — still concentrated, but less for lower bandwidth). The second row denotes two Slepian vectors from the second half of the spectrum (normalized indices of 0.6 — mildly concentrated in $S$ using a smaller bandwidth — and 0.9 — not concentrated at all). Visualizations are similar for modified embedded distance (D), but in this case, low eigenvalues imply either non-concentrated (e.g., X axis, first row of plots) or mildly concentrated but low localized spatial frequency Slepian vectors (for instance, Y axis, first row of plots, $W = 200$), while high eigenvalues relate to high localized spatial frequency Slepian vectors (see Y axis, second row of plots). See (Van De Ville et al., 2017) for another preliminary analysis of the dataset from the modified embedded distance viewpoint. $\mu$ and $\xi$ values of the shown Slepian vectors are provided in parentheses on each axis.
Appendix B. Supplementary material for Chapter 4

Figure B.2 — Eigenspectrum of the newly developed $\zeta$ criterion (A), with vertical bars highlighting the locations of the spectrum at which four pairs of Slepian vectors were sampled for display (B, from first to fourth row as respectively depicted by yellow, orange, brown and red color codes). Results obtained with linear, quadratic and order 10 approximations, as well as from a full computation of $M - L^{1/2}M^{1/2}$, are respectively shown from left to right. $\zeta$ values of the shown Slepian vectors are provided in parentheses on each axis.
B.1. Supplementary material for Section 4.1

Figure B.3 – Start (left column), intermediate (middle column) and end (right column) representations of sensory neuron (A), interneuron (B) or motoneuron (C) trajectories, respectively, setting cooperation weights for other neuron types to 1, 0.5 or 0. Cells are labeled according to (Varshney et al., 2011). The start representation is the same across cases, since then $M = I$ and the problem boils down to the eigendecomposition of the adjacency matrix $A$, or equivalently of the Laplacian $L = I - A$ highlighted in Fig. 4.1.
Appendix B. Supplementary material for Chapter 4

Table B.2 – Interneurons of the *C. elegans*. Columns correspond to clusters derived by optimized k-means.

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVBL</td>
<td>AIML</td>
<td>ABL</td>
<td>AIAL</td>
<td>ADAL</td>
<td>AVAL</td>
</tr>
<tr>
<td>AVBR</td>
<td>AIMR</td>
<td>ABR</td>
<td>AIAR</td>
<td>ADAR</td>
<td>AVAR</td>
</tr>
<tr>
<td>AVG</td>
<td>AVFL</td>
<td>AINL</td>
<td>AVEL</td>
<td>AVER</td>
<td>AVDR</td>
</tr>
<tr>
<td>AVIL</td>
<td>AVFR</td>
<td>AINR</td>
<td>AVEL</td>
<td>AVKL</td>
<td>LUAL</td>
</tr>
<tr>
<td>AVJR</td>
<td>AVHL</td>
<td>AIYL</td>
<td>AVKR</td>
<td>LUAL</td>
<td>LIAR</td>
</tr>
<tr>
<td>BDUL</td>
<td>AVHR</td>
<td>AIYR</td>
<td>DVA</td>
<td>PVCL</td>
<td></td>
</tr>
<tr>
<td>BDUR</td>
<td>PVQL</td>
<td>AIZL</td>
<td>PVCR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PVQR</td>
<td>AZFR</td>
<td>DVC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PVPR</td>
<td>RIFL</td>
<td>AJAL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIFR</td>
<td>AJAR</td>
<td>RIAL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIAL</td>
<td>RIAR</td>
<td>RIBL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIBL</td>
<td>RIBR</td>
<td>RIGR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIBR</td>
<td>RIH</td>
<td>RIPL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIH</td>
<td>RIR</td>
<td>RIPS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIR</td>
<td>RIS</td>
<td>RMGL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIS</td>
<td>RMGR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RIS</td>
<td>URBL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URBL</td>
<td>URBR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure B.4 – Clusters derived by repeated k-means clustering of the focused nodes in the case of sensory neurons (A), interneurons (B) or motoneurons (C). The optimal number of clusters was estimated with the Silhouette approach. Nodes constituting the border of each cluster’s convex hull are connected by a dashed black line to visualize the clusters.
Table B.3 – Motoneurons of the *C. elegans*. Columns correspond to clusters derived by optimized k-means.

<table>
<thead>
<tr>
<th>C1</th>
<th>C2</th>
<th>C3</th>
<th>C4</th>
<th>C5</th>
<th>C6</th>
<th>C7</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS10</td>
<td>DD04</td>
<td>DD05</td>
<td>DA04</td>
<td>AVL</td>
<td>AS11</td>
<td>AS01</td>
</tr>
<tr>
<td>AS06</td>
<td>VA07</td>
<td>VA08</td>
<td>DB03</td>
<td>DA08</td>
<td>DA09</td>
<td>AS02</td>
</tr>
<tr>
<td>AS07</td>
<td>VB06</td>
<td>VB09</td>
<td>DB04</td>
<td>DB07</td>
<td>DD06</td>
<td>AS03</td>
</tr>
<tr>
<td>AS08</td>
<td>VB07</td>
<td>VB08</td>
<td>DD02</td>
<td>PVNR</td>
<td>DVB</td>
<td>AS04</td>
</tr>
<tr>
<td>AS09</td>
<td>VS07</td>
<td>VS09</td>
<td>DD03</td>
<td>RID</td>
<td>PDA</td>
<td>AS05</td>
</tr>
<tr>
<td>DA07</td>
<td>VS08</td>
<td>VS10</td>
<td>VA06</td>
<td>RMVL</td>
<td>PD8</td>
<td>DA01</td>
</tr>
<tr>
<td>DB05</td>
<td>VS09</td>
<td>VS10</td>
<td>VB02</td>
<td>RMVR</td>
<td>VA11</td>
<td>DA02</td>
</tr>
<tr>
<td>DB06</td>
<td>VB03</td>
<td>RIVL</td>
<td>VA12</td>
<td>DA03</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HSNL</td>
<td>VB04</td>
<td>RIVR</td>
<td>VB10</td>
<td>DA05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PVNL</td>
<td>VB05</td>
<td>RMDDL</td>
<td>VB11</td>
<td>DA06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMHL</td>
<td>VC01</td>
<td>RMDDR</td>
<td>VS12</td>
<td>DB01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RMHR</td>
<td>VC02</td>
<td>RMDL</td>
<td>VS13</td>
<td>DB02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SARD</td>
<td>VC03</td>
<td>RMDR</td>
<td>DD01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SABVL</td>
<td>VE02</td>
<td>RMDVL</td>
<td>HSNR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAVR</td>
<td>VE03</td>
<td>RMDVR</td>
<td>VA01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAVR</td>
<td>VE04</td>
<td>RMDD</td>
<td>VA02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAVR</td>
<td>VE05</td>
<td>RMEL</td>
<td>VA03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAVR</td>
<td>VE06</td>
<td>RMR</td>
<td>VA04</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAVR</td>
<td>RMEL</td>
<td>VB05</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SADRL</td>
<td>SAADL</td>
<td>SADDL</td>
<td>VA01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAVR</td>
<td>SADDL</td>
<td>SADDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMBDR</td>
<td>SAAV</td>
<td>SAAV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URA</td>
<td>SAAV</td>
<td>SAAV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VA10</td>
<td>SMBL</td>
<td>SMBL</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VC05</td>
<td>SMB</td>
<td>SMB</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMDD</td>
<td>SMDD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMDD</td>
<td>SMDD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMD</td>
<td>SMD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SMD</td>
<td>SMD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URA</td>
<td>URA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URA</td>
<td>URA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>URA</td>
<td>URA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

111
Figure B.5 – Focussing on (A) sensory neurons (red), (B) interneurons (black) or (C) motoneurons (green), two-dimensional visualization using alternative sets of Slepian vectors: first and second (first row), fourth and fifth (second row), or last two (third row). Cells are labeled according to (Varshney et al., 2011). $\zeta$ values of the shown Slepian vectors are provided in parentheses on each axis.
Figure B.6 – Separate two-dimensional visualizations when only considering chemical synapses (left column) or gap junctions (right column) for sensory neurons (A), interneurons (B) or motoneurons (C). Cells are labeled according to (Varshney et al., 2011). ζ values of the shown Slepian vectors are provided in parentheses on each axis.
Figure B.7 – Two-dimensional visualizations for sensory neurons (A), interneurons (B) or motoneurons (C) when representing each neuron as a function of its position along the X direction. A value of 0 indicates the location of the nerve ring, and positional data was retrieved from (Varier and Kaiser, 2011). \( \zeta \) values of the shown Slepian vectors are provided in parentheses on each axis.

Figure B.8 – Testing of clustering assignments when the focus is on sensory neurons (red), interneurons (grey) or motoneurons (green). For each case, all nodes of other types are considered as one additional cluster. The null distributions of the modularity of random assignments to clusters are given on the left side of the plot (A). The dashed straight lines on the right represent values of modularity \( Q \) for the clusters in Fig. S4 derived by the k-means approach (B). The x-axis is broken at 0.06 for better visualization.
Bibliography


Bibliography


127


Miljan Petrović

Nationality: Serbian
E-mail: miljan.petrovic@epfl.ch

EDUCATION

- **2018/04–present** PhD Student in Electrical Engineering
  École Polytechnique Fédérale de Lausanne
  Lausanne, Switzerland
- **2016/10–2017/07** Master in Electrical Engineering and Computing
  University of Nis
  Nis, Serbia
- **2016/09–2016/09** Summer School Robust Signal Processing
  IEEE SPS & EURASIP
  Rudesheim am Rhein, Germany
- **2012/09–2016/06** Bachelor with Honours in Electrical Engineering and Computing
  University of Nis
  Nis, Serbia

RESEARCH EXPERIENCE

- **2018/04–present** Doctoral Candidate
  École Polytechnique Fédérale de Lausanne
  Lausanne, Switzerland
  - *Research project title: Graph Slepian Designs for Directed and Tunable Network Analysis*
- **2016/10–2017/07** Master thesis
  University of Nis
  Nis, Serbia
  - *Thesis title: On The Design of Short Wordlength FIR Filters*
- **2015/07–2015/08** Research intern
  École Polytechnique Fédérale de Lausanne
  Lausanne, Switzerland
  - *Responsibilities: Investigation of quantitative metrics of baroreflex sensitivity based on biomedical recordings*
- **2012/09–2016/06** Bachelor Thesis
  University of Nis
  Nis, Serbia
  - *Thesis title: Selective digital filters with quadratic phase*

AWARDS AND DISTINCTIONS

- Best BSc Student of the Generation enrolled in 2012, issued by Faculty of Electronic Engineering in Nis, Serbia
Best MSc Student of the Generation enrolled in 2016 for module Electronics and Microsystems, issued by Faculty of Electronic Engineering in Nis, Serbia


**TEACHING ACTIVITIES**

2018–2022 (Master courses at EPFL, Lausanne)

» Image Processing I (MICRO-511)

» Image Processing II (MICRO-512)

» Signal processing for functional brain imaging (MICRO-513)

**PROFESSIONAL EXPERIENCE**

📅 2018–present  
Ambassador of the EPFL Excellence in Engineering internship program  
École Polytechnique Fédérale de Lausanne

📍 Lausanne, Switzerland

**LANGUAGES**

» Serbian (native)

» English (fluent)

» French (basic)

**PUBLICATIONS**


» **M. Petrovic**. A Graph Perspective on Logic Circuits: Faults, Robustness & Eigenspectrum. 8th Small Systems Simulation Symposium SSSS, Nis, Serbia, Feb 2020.


