Bachelor Thesis
A spectral approach to generating set separations

Michael Hermann

July 2022

Name	Michael Hermann
Matrikelnummer	6981007
Betreuer	Prof. Dr. Reinhard Diestel
Mitgutachter	M. Sc. Paul Knappe

Contents

1 Introduction

This work will investigate *spectral separations*, an important method to generate partitions of sets to analyse their internal structure.

Let V be a finite set and suppose every $v \in V$ can be described by a collection of binary features $q_1, \dots, q_k : V \to \{-1, 1\}$. Every feature q splits V into a subset $A \subseteq V$ of elements that have this particular feature and the complement $V \setminus A$ consisting of elements that do not have this feature, this can be encoded by letting $q(v) = 1$ or $q(v) = -1$ respectively. In this manner every feature induces a bipartition of V. Similarly, every bipartition can be interpreted as a binary feature. If $\{V \setminus B, B\}$ is an arbitrary bipartition of V, we can define a corresponding binary feature $q_B : V \to \{-1, 1\}$ by letting $q_B(v) = 1$ if and only if $v \in B$. This way, we can identify bipartitions and binary features.

If we equip a bipartition of V with a notion of *orientation*, by making the bipartition $\{V \setminus A, A\}$ either point in the direction of A or in the direction of $V \setminus A$, then such an oriented bipartition allows us to *localise* a feature within the set V.

The other way round, every $v \in V$ yields an orientation of all the bipartitions induced by its features' values. More generally, every possible assignment of values -1 or 1 to all of the features $q_1, \dots, q_k$ provides an orientation of all induced bipartitions. Such an assignment might correspond to a subset of elements that exactly match all the features but in general this need not be the case. Interestingly, if we impose restrictions on the set of allowed orientations, even feature combinations that do not appear anywhere in V can localise relevant substructures of V. The notion of a *tangle* provides a very general framework to formalise these restrictions and thereby offers a promising novel paradigm for data analysis. Informally, the idea is that relevant substructures cannot be cut into many irrelevant pieces by these bipartitions, because most of the elements of such a substructure share any given feature or its negation.

In the theory of tangles one works with the more general notion of *separations* instead of bipartitions. In our applied context a *set separation* is a segmentation of a set into two parts with possibly nonempty intersection. We still can interpret a set separation as a feature on V where we allow an unspecified feature value. For elements in the intersection we do not want to make a strict decision if they have a feature or not.

Usually a data set that we want to analyse does not come together with a set of meaningful binary features. Often only relations between data points are given, for example similarities or dissimilarities, or every data point is characterised by a number of real-valued measurements. In this cases we first have to generate a set of meaningful features

– or separations – that can describe relevant substructures of our data. Essentially, we need two things before we can analyse the tangles of a data set V. We first need a way to *measure* how descriptive a separation is of the structure. This is usually done with help of an *order function*, a function quantifying the 'usefulness' of a separation. Second, we need a way to *find* a set of separations that are reflective of the structure of V. These two tasks are of central importance and often impose a challenge of their own. The goal of this thesis is to first discuss an important class of order functions given by matrices containing pairwise information about the elements of V. We shall see that the problem to optimise these order functions then naturally leads to a spectral approach to generating set separations.

We start the discussion in section 2 where we formally introduce our notion of separation and informally describe how tangles can help to detect interesting substructures of data sets. The third section then describes the important class of *matrix order functions*. In the fourth section we finally delve into our main topic: the discussion of *spectral separations* as a generic way to find structure reflecting separations of data sets.

2 Prerequisites: Separations and Tangles

Our goal is to find separations of a set V that help us to unveil hidden structures induced by properties of – or relations between – the elements of a set V. In this chapter we want to briefly introduce the notion of *separation* we will use in the sequel and explain how separations can help us to analyse interesting substructures of data sets.

2.1 Systems of Set Separations

Let V be a nonempty, finite set of size $|V| = n \in \mathbb{N}$. We usually use V to denote a *ground set* that is a superset of all other sets encountered later. If $B \subseteq V$, we denote the complement of B in V by $B^c = V \setminus B$. The vector $\mathbb{1}_B \in \{0,1\}^V$ with $\mathbb{1}_B(v) = 1$ if and only if $v \in B$ is called *(ordinary) indicator vector* of the set B.
A *set separation* of V is a two-element set $\{A, B\}$ where $A, B \subseteq V$ and $A \cup B = V$. We call the sets A and B the *sides* of the separation $\{A, B\}$. If $A \cap B = \emptyset$, we call $\{A, B\}$ a *bipartition* of V. If $u, v \in V$, we say a separation $s = \{A, B\}$ *separates* u and v if $u \in A \setminus B$ and $v \in B \setminus A$ or $v \in A \setminus B$ and $u \in B \setminus A$.
The set of all set separations of V is denoted by $\mathcal{S}_V$, the set of all bipartitions of V by $\mathcal{B}_V$. Obviously we have $\mathcal{B}_V \subset \mathcal{S}_V$ and for a reasonably sized set V the sets $\mathcal{B}_V$ and $\mathcal{S}_V$ are both quite large. The total number of bipartitions is $|\mathcal{B}_V| = 2^{n-1}$ as every bipartition can be identified with a subset of V and every subset and its complement define the same bipartition. To compute the total number of set separations let $\mathrm{b}(k)$ be the number of bipartitions of a k-element set for $k \in \mathbb{N} \cup \{0\}$. Then $\mathrm{b}(0) = 1$ since the only set separation of the empty set is $\{\emptyset, \emptyset\}$. The total number of set separations can be computed by summing over the number of bipartitions of the sets $V \setminus A \cap B$ for all possible intersections $A \cap B$. We have

$$|\mathcal{S}_V| = \sum_{k=0}^{n} \binom{n}{k} \mathrm{b}(k) = 1 + \sum_{k=1}^{n} \binom{n}{k} 2^{k-1} = 1 + \frac{1}{2}\left(\sum_{k=0}^{n} \binom{n}{k} 2^k - 1\right) = 1 + \frac{3^n - 1}{2}$$

In the introduction we associated separations of V with *features* that the elements $v \in V$ either have or not have. It is intuitively clear that in real situations, i.e. if the elements of V correspond to real word objects or concepts, usually not all of this huge number of possible features are equally descriptive. Otherwise each element $v \in V$ can only be characterised by its plain identity and there is no subset that can be described by any distinctive properties except the identities of the contained elements. In other words, there is not really a *structure* on V in this case.

Since our goal is to detect meaningful substructures of V we can assume that not all features, or separations, are equally relevant. The relevance of a separation is usually given by a real valued function

$$|\cdot| : \mathcal{S}_V \to \mathbb{R},$$

called an *order function* on $\mathcal{S}_V$. Every order function induces a *quasi-order* on $\mathcal{S}_V$, a reflexive and transitive relation, that allows us to rank separations by their relevance or some kind of 'usefulness' to describe interesting substructures of V. To rank the separations we either sort a list of separations by ascending or by descending order value. Depending on the semantics of $|\cdot|$ sometimes one way is more intuitive than the other. If we are sorting by ascending order value, we can interpret the order value as the *cost* of splitting V in the two (possibly overlapping) parts of a separation.

We want to *localise* substructures of V, therefore we work with *oriented* separations. Every separation $s = \{A, B\} \in \mathcal{S}_V$ has two orientations $\vec{s}$ and $\overleftarrow{s}$, for example $\vec{s} = (A, B)$ and $\overleftarrow{s} = (B, A)$. [1]. If $T \subseteq \mathcal{S}_V$ is a set of (unoriented set separations), we denote the corresponding set of oriented separations by $\vec{T}$. And vice versa.
If for a set $C \subseteq V$ we have $C \subseteq B$, we say $\vec{s} = (A, B)$ is *pointing to* C, if $C \subseteq A$, we say $\vec{s}$ is *pointing away* from C. We call A the *small side* and B the *big side* of $\vec{s} = (A, B)$.
The set of all oriented separations of V is denoted by $\vec{\mathcal{S}}_V$, its subset of oriented bipartitions by $\vec{\mathcal{B}}_V$. The involution $^* : \vec{\mathcal{S}}_V \to \vec{\mathcal{S}}_V$ that maps $\vec{s}$ to $\overleftarrow{s}$ is called the *inversion* on $\vec{\mathcal{S}}_V$. If $\vec{s} = (A, B) \in \vec{\mathcal{S}}_V$, we call the vector $\mathbb{1}_{\vec{s}} \coloneqq \mathbb{1}_B - \mathbb{1}_A$ the *oriented indicator vector* of $\vec{s}$, the oriented indicator vector of $\overleftarrow{s}$ then is $\mathbb{1}_{\overleftarrow{s}} = -\mathbb{1}_{\vec{s}}$.
Every order function defined on $\mathcal{S}$ can be generalised to $\vec{\mathcal{S}}$ by simply setting $|\vec{s}| \coloneqq |s|$. To ensure that a candidate order functions $|\cdot| : \vec{\mathcal{S}}_V \to \mathbb{R}$ is well defined we therefore have to check independence of orientation, this means $|\vec{s}| = |\overleftarrow{s}|$ for all $s \in \mathcal{S}$.

We can define a partial order on $\vec{\mathcal{S}}_V$. Let $\vec{s} = (A, B)$ and $\vec{t} = (C, D)$ be two oriented separations, we set

$$\vec{s} \leq \vec{t} \iff A \subseteq C \text{ and } B \supseteq D$$

If $\vec{s} \leq \vec{t}$, the 'bigger' separation $\vec{t}$ is *more specific* as it points to less subsets of V than the 'smaller' separation $\vec{s}$. If two unoriented separations s and t have orientations $\vec{s}$ and $\vec{t}$ such that $\vec{s} \leq \vec{t}$, we say the separations are *nested*. Otherwise they *cross*.
If $\vec{\mathcal{T}}$ is a subset of $\vec{\mathcal{S}}_V$ that is closed under the inversion * on $\vec{\mathcal{S}}_V$, we call $\vec{\mathcal{T}}$ a *set*

[1] the notation $\vec{s}$ does not specify wich orientation is meant. It is *one* of the orientations of s and the only thing we know is that $\overleftarrow{s}$ is the *other* orientation of s.

separation system. If for every pair of separations $\vec{s}, \vec{t} \in \vec{\mathcal{T}}$ there exists an *infimum* $\vec{s} \wedge \vec{t}$ and a *supremum* $\vec{s} \vee \vec{t}$, we call the tuple $(\vec{\mathcal{T}}_V, \leq, \vee, \wedge, ^*, |\cdot|)$ a *universe of set separations* of V. The set $\vec{\mathcal{S}}_V$ obviously has this property, we call it therefore *the universe of all set separations of* V and denote it also by $\vec{\mathcal{S}}_V$.
If an order function $|\cdot|$ satisfies the inequality

$$|\vec{s} \vee \vec{t}| + |\vec{s} \wedge \vec{t}| \leq |\vec{s}| + |\vec{t}|$$

for all $\vec{s}, \vec{t} \in \vec{\mathcal{S}}$, then $|\cdot|$ is *submodular.* If the reverse inequality holds, then $|\cdot|$ is *supermodular.*
If two oriented separations $\vec{s}, \vec{t} \in \vec{\mathcal{S}}_V$ are represented by their oriented indicator vectors $\mathbb{1}_{\vec{s}}, \mathbb{1}_{\vec{t}} \in \{-1, 0, 1\}^V$, their supremum and infimum can be obtain by taking the *elementwise* minimum or maximum, respectively

$$\mathbb{1}_{\vec{s} \vee \vec{t}} = \min(\mathbb{1}_{\vec{s}}, \mathbb{1}_{\vec{t}}) \qquad \text{and} \qquad \mathbb{1}_{\vec{s} \wedge \vec{t}} = \max(\mathbb{1}_{\vec{s}}, \mathbb{1}_{\vec{t}})$$

If $\vec{\mathcal{T}}$ is a universe of separations, a subset $\vec{\mathcal{R}} \subseteq \vec{\mathcal{T}}$ closed under $\vee$ and $\wedge$ (as defined on $\vec{\mathcal{T}}$) is called a *subuniverse* of $\vec{\mathcal{T}}$. In this case we write $\vec{\mathcal{R}} \sqsubseteq \vec{\mathcal{T}}$.
If $R \subseteq \vec{\mathcal{S}}_V$ is any collection of oriented set separations, we call the intersection of all subuniverses of $\vec{\mathcal{S}}_V$ containing R, i.e.

$$\bigcap_{R \subseteq \vec{\mathcal{T}},\, \vec{\mathcal{T}} \sqsubseteq \vec{\mathcal{S}}_V} \vec{\mathcal{T}} \tag{1}$$

the *subuniverse of* $\vec{\mathcal{S}}_V$ *generated by* R. The set of all oriented bipartitions of V, $\vec{\mathcal{B}}_V$, is obviously a subuniverse of $\vec{\mathcal{S}}_V$ and it can be generated by the set of all bipartitions that point to a single element of V

$$\big\{(V \setminus \{v\}, \{v\}) \;:\; c \in V\big\}$$

This follows directly from

Proposition 2.1. *Let V be a set of size $|V| = n$. A set of oriented bipartitions $\vec{S} \subseteq \mathcal{B}_V$ generates $\vec{\mathcal{B}}_V$ if and only if for every pair $u, v \in V$ there is an $s \in \vec{S}$ that separates u and v.*

Proof. Let us assume that we can separate every pair of elements $u, v \in V$ by a bipartition in $\vec{S}$. We have to show that for arbitrary subsets $B \subseteq V$ we can generate the oriented bipartition $\vec{s} = (B^c, B)$. Let $B \subseteq V$ be a an arbitrary subset of V. For every

$v \in B$ and $u \in B^c$ let $\vec{s}_{uv} \in S$ be the separation separating u and v oriented so that it points to v. Then for every $v \in B$ the supremum $\vec{s}_v = \bigvee_{u \in V \setminus \{v\}} \vec{s}_{uv}$ separates v from all other vertices in V and hence the infimum of these suprema is the wanted separation $\vec{s}_B = \bigwedge_{v \in B} \vec{s}_v = (B^c, B)$. As $B \subseteq V$ was arbitrary, we can generate the whole universe $\vec{\mathcal{B}}_V$.
For the other direction let $u, v \in V$ be inseparable by all bipartitions in $\vec{S}$. Then u, v are on the same side of every bipartition in $\vec{S}$. Let $\vec{\mathcal{U}} \subseteq \vec{\mathcal{B}}_V$ be the set of all bipartitions of V where u and v are on the same side. Then $\vec{\mathcal{U}}$ is a subuniverse of $\vec{\mathcal{B}}_V$. Indeed, if $\vec{s}, \vec{t} \in \vec{\mathcal{U}}$ do not separate u and v, then neither do $\vec{s} \vee \vec{t}$, $\vec{s} \wedge \vec{t}$ or one of the inverses $\overleftarrow{s}, \overleftarrow{t}$. Therefore $\vec{\mathcal{U}}$ is one of the intersected universes in (1) and the universe generated by $\vec{S}$ cannot contain any separation that separates u and v. Hence it is not equal to $\vec{\mathcal{B}}_V$. □

The generating set of $\vec{\mathcal{B}}_V$ we have seen above is canonical but it is not the most efficient way to generate $\vec{\mathcal{B}}_V$. We can generate $\vec{\mathcal{B}}_V$ using much fewer bipartitions:

Proposition 2.2. *Let V be a set of size $|V| = n$. The minimum number of bipartitions needed to generate $\vec{\mathcal{B}}_V$ is $\lceil \log_2 n \rceil$.*

Proof. Let $M \in \{0,1\}^{n \times \lceil \log_2 n \rceil}$ be the matrix that contains the binary representation of the integer $i - 1$ in the i-th row. We interpret the columns of M as indicator vectors on V. Let $u, v \in V$. Because every row in M is unique, we find a column $f \in \{0,1\}^V$ of M such that $f(u) \neq f(v)$. If $B \subset V$ is the subset indicated by f, then $(B^c, B) \in \vec{\mathcal{B}}_V$ separates u and v. As $u, v \in V$ are arbitrary, the previous proposition now assures that we can generate $\vec{\mathcal{B}}_V$.
Let now be $S \subset \vec{\mathcal{B}}_V$ a set of size $k < \lceil \log_2 n \rceil$. The matrix M containing the indicator vectors of the big sides has at most $2^k < n$ different columns hence there are different $u, v \in V$ corresponding to two identical rows of M. Therefore u and v are on the same side of all bipartitions in S. Hence, by the previous proposition S cannot generate all of $\vec{\mathcal{B}}_V$. □

Interestingly, the theory of separation systems can be worked out without any reference to sets that actually are separated leading to a theory of *abstract separation systems*. More details about the notions introduced in this section and an overview about the abstract theory of separations can be found in [4].

2.2 Tangles of Separation Systems

Now that we know the notions of systems and universes of separations, let us find out how separations can be used to detect interesting substructures in V. Our topic is mainly the *discovery* of separations and we will not need any technical details of how we use them to analyse data sets afterwards. Yet it is the main motivation of our considerations and we therefore want to describe some important parts of the theory and practice from a bird's eye perspective.

Let us assume we already have found a separation system $\vec{T} \subseteq \vec{\mathcal{S}}_V$, for example both orientations of the k most expressive separations according to an order function $|\cdot|$ defined on $\mathcal{S}_V$. Let $U \subset V$ be a set of data points sharing some interesting properties or revealing some high level information about the data. We explicitly do not want to specify in more detail what exactly makes a substructure of V 'interesting', the concrete interpretation of such a structure depends on the data and known or unknown relations between the data points. For example, if we have a measure of pairwise similarity between data points, an example of an interesting substructure is a *cluster*, i.e. a subset of very similar data points that can be clearly distinguished from its complement. But that is not all we want to be able to describe. Data points belonging to the same interesting substructure do not have to share any properties. Such a substructure can, for example, also be a set of *outliers*, consisting of objects that do not have anything in common, except that they neither have anything in common with any of the other (non-outlier) objects.
The central idea is to not describe *what* the interesting substructures exactly are but *where* they are [4] within the structure induced on V by the information we have about our data [2]. Our assumption is that our separation system $\vec{T}$ and, in particular, the order function we defined on $\mathcal{S}$ form a refined representation of this structure inducing information, and that the representation allows us to answer the question of the 'where' (but not necessarily the 'what'). The idea is that separations, marked as 'useful' by the order function, will not cut an interesting substructure easily into completely uninteresting parts, or vice versa, that subsets of V that can be *shattered* into a lot of small parts by the separations in T are not interesting at all.

In what follows, we want to give a very informal description of the *tangle* method for

[2] ... whatever form this information may take: It can be a combinatorial structure, like a graph or a matroid, it can be an explicit set of properties of each object, relations defined between the objects or any different concept that might make sense for the problem at hand.

data analysis. The interested reader is recommended to have a look at [4], [7] and [8] for more detailed theory, [9] and [6] for the use in practice and [11] for details on algorithmic aspects.

An *orientation* $\mathcal{O}$ of a separation system $\vec{T}$ is a subset of $\vec{T}$ that contains exactly one orientation of each separation in the corresponding set of unoriented separations T. We want to define a *tangle* of T as an orientation of T so that it describes describes an interesting substructure of V. Clearly not all orientations of $\vec{T}$ can be tangles, otherwise the notion would not contain much information. The minimum requirement for an orientation $\mathcal{O}$ to be a tangle is *consistency*, this means if $\vec{r}, \vec{s} \in \vec{T}$ are two different oriented separations and if $\vec{r} \leq \vec{s}$, then $\mathcal{O}$ does not contain the orientations $\overleftarrow{r}$ and $\vec{s}$ at the same time - in words, a consistent orientation does not contain orientations that *point away* from each other. Additionally we can define sets of *partial orientations* $\mathcal{F}$ of $\vec{T}$, these are orientations of subsets of T, that we want to forbid in a tangle. All together, *a tangle is a consistent orientation of* $\vec{T}$ *that does not contain any of the sets in* $\mathcal{F}$. An example of such a forbidden set $\mathcal{F}$ is the collection of three-element subsets $\{\vec{r}, \vec{s}, \vec{t}\} \subset \vec{T}$ such that the big side of the set separation $\vec{r} \vee \vec{s} \vee \vec{t}$ contains less than m elements of V, where $m \in \mathbb{N}$ is a pre-specified constant. A tangle of T then is an orientation $\mathcal{O}$ of $\vec{T}$ such that for all $\vec{r}, \vec{s}, \vec{t} \in \mathcal{O}$ with $\vec{r} = (A_1, B_1)$, $\vec{s} = (A_2, B_2)$ and $\vec{t} = (A_3, B_3)$ we have $|B_1 \cap B_2 \cap B_3| \geq m$.

Let us assume the separations in T are sorted, for example by the ranking induced by an order function. Let $T_1 \subset \cdots \subset T_k = T$ be the sequence of sets of separations where T_i contains the first i separations of T, this means the i most important separations of T. Then the tangles of all the oriented separations systems $\vec{T}_i$ can be represented by nodes of a binary tree whose edges between nodes at height $i-1$ and i correspond to the orientations of the i-th separations of T. Figure 1 shows an example. We call this tree the *tangle search tree (TST)* [3].
There is another tree that plays an important role in tangle theory. If the order function $|\cdot|$ is submodular, if the separation system $\vec{T}$ is in some sense 'rich' enough, and under some additional constraints on the set of forbidden orientations $\mathcal{F}$, we can find a *nested* subset of T that distinguishes all tangles of $\vec{T}$ [7]. This means, we can find a hierarchical representation of the data set and its structure by arranging nested separations in a tree.

[3]The name reflects its use in practice. When searching for tangles of a separation system T, such a tree is built iteratively by adding one separation after another, starting by the most relevant, until T is exhausted or no new tangle can be found.

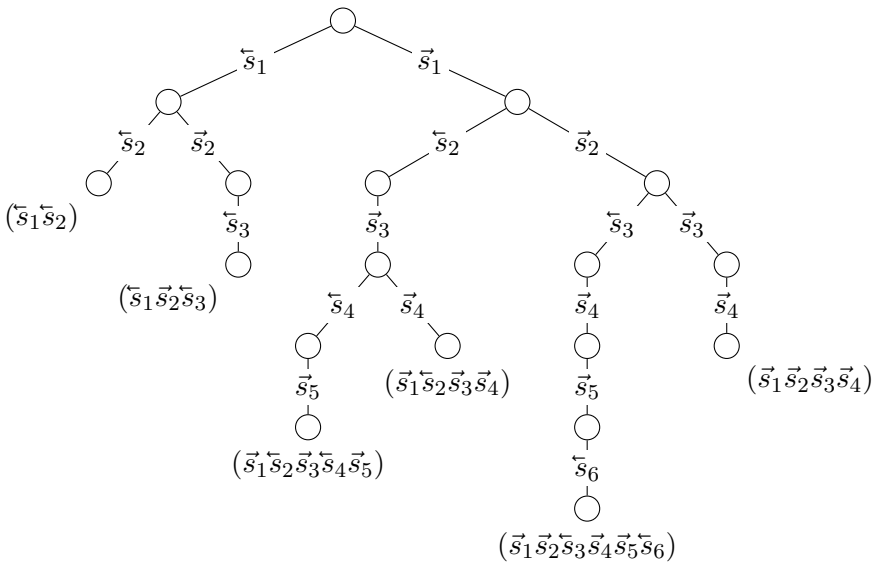

Figure 1: An example of a *Tangle Search Tree* of a separation system $\vec{T}$. Every edge corresponds to an orientation of a separation in $\vec{T}$. The path from root to a node defines the orientations corresponding to the node.

We call this tree the *Tree-of-Tangles (ToT)*. This tree is usually not a binary tree.

2.3 Further prerequisites

We assume the reader has basic knowledge of Linear Algebra and Graph Theory. A good overview of Linear Algebra, especially the 'spectral' part can be found in [18]. We will not need a lot of general graph theory, but we presume familiarity with the standard notations found in [5]. Details about spectral graph theory can be found in [14].

3 Matrix Order Functions

If we want to analyse the structure of a data set V with the help of tangles, the choice of an order function has a crucial influence on the characteristics of detectable substructures and therefore the meaning of a tangle. As we have seen, an order function should in general provide a way to assess a separation's usefulness or 'naturalness'. If our objects in V come with some natural notion of pairwise similarity, we intuitively would require a natural separation to rather separate dissimilar objects than similar objects. It turns out that in such situations matrices provide a general way to define an important class of order functions.

3.1 Motivation and Definition

Let us first look at a prominent example of a useful order function on (similarity) graphs. Let $G = (V, E)$ be a graph and let $A \in \mathbb{R}^{V \times V}$ be its adjacency matrix. Let $s = \{B^c, B\} \in \mathcal{B}_V$ be a bipartition of V and assume that both of B, B^c are nonempty. Then s induces a nontrivial cut in G. An important property of a cut is its size, i.e. the number of edges between the sides of s. Both orientations of s induce the same cut, so a cut's size is a good candidate for an order function on $\vec{\mathcal{B}}_V$.
We denote the size (or the weight) of the cut induced by a bipartition $\{B^c, B\}$ in G by

$$\mathrm{cut}_G(B^c, B) \coloneqq \sum_{u \in B, v \in B^c} A(u, v) = \mathbb{1}_{B^c}^t A \mathbb{1}_B$$

and the *volume* of a subset $B \subseteq V$ by

$$\mathrm{vol}_G(B) \coloneqq \sum_{u \in B, v \in V} A(u, v) = \mathbb{1}_B^t A \mathbb{1}$$

As the matrix A is symmetric, both of these notions are well-defined. The vector $A\mathbb{1} \in \mathbb{R}^V$ contains the degrees of G's vertices, hence the volume of $B \subseteq V$ is the sum of the degrees of the vertices in B and in particular $\mathrm{vol}_G(V) = 2|E|$.
We might drop the subscripts if the graph is clear from context. Let $D \in \mathbb{R}^{V \times V}$ be the degree matrix of G, this means the diagaonal matrix with diagonal $A\mathbb{1} = (\mathbb{1}^t A)^t$, and $L = D - A$ the laplacian of G. Then the size of the cut represented by an indicator vector $\mathbb{1}_B$ of one of its sides can be written as a quadratic form defined by the laplacian

matrix

$$\begin{aligned}\mathbb{1}_{B^c}^t A \mathbb{1}_B &= (\mathbb{1} - \mathbb{1}_B)^t A \mathbb{1}_B \\ &= \mathbb{1}^t A \mathbb{1}_B - \mathbb{1}_B^t A \mathbb{1}_B \\ &= \mathbb{1}_B^t D \mathbb{1}_B - \mathbb{1}_B^t A \mathbb{1}_B \\ &= \mathbb{1}_B^t L \mathbb{1}_B\end{aligned}$$

and because of $\mathbb{1}^t L = L\mathbb{1} = D\mathbb{1} - A\mathbb{1} = 0$ we have

$$\mathbb{1}_B^t L \mathbb{1}_B = (\mathbb{1} - \mathbb{1}_{B^c}^t) L (\mathbb{1} - \mathbb{1}_{B^c}) = \mathbb{1}_{B^c}^t L \mathbb{1}_{B^c}$$

We usually want to represent an oriented bipartition $\vec{s} = (B^c, B)$ by its *oriented* indicator vector $\mathbb{1}_{\vec{s}}$ instead of $\mathbb{1}_B$. Plugging $\mathbb{1}_{\vec{s}}$ into the quadratic form does not change anything qualitatively. Utilising the symmetry of L, the value of the quadratic form is

$$\begin{aligned}\mathbb{1}_{\vec{s}}^t L \mathbb{1}_{\vec{s}} &= (\mathbb{1}_B - \mathbb{1}_{B^c})^t L (\mathbb{1}_B - \mathbb{1}_{B^c})^t \\ &= \mathbb{1}_B^t L \mathbb{1}_B + \mathbb{1}_{B^c}^t L \mathbb{1}_{B^c} - 2 \cdot \mathbb{1}_B^t L \mathbb{1}_{B^c} \\ &= 4 \operatorname{cut}(B^c, B)\end{aligned}$$

because $\mathbb{1}_B^t L \mathbb{1}_B = \mathbb{1}_{B^c}^t L \mathbb{1}_{B^c}$ and $\mathbb{1}_B^t L \mathbb{1}_{B^c}$ is summing -1 once for every edge crossing the cut.

We can extend this example to weighted graphs: If G is an edge-weighted graph, then the adjacency matrix A contains the weights and we do not just count the cut's edges but sum their weights.

The question if it makes sense to replace the laplacian by a different matrix $M \in \mathbb{R}^{V \times V}$ seems very natural. The only property to be checked is independence of orientation. For any $\mathbb{1}_{\vec{s}} = (B_1, B_2) \in \vec{\mathcal{S}_V}$ we obviously have

$$(\mathbb{1}_{B_2} - \mathbb{1}_{B_1})^t M (\mathbb{1}_{B_2} - \mathbb{1}_{B^1}) = (\mathbb{1}_{B_1} - \mathbb{1}_{B_2})^t M (\mathbb{1}_{B_1} - \mathbb{1}_{B_2})^t (-1)^2$$

and we can go on and give such order functions their own name:

Definition 3.1. Let V be a finite set and $M \in \mathbb{R}^{V\times V}$ an arbitrary matrix. The function

$$\begin{aligned} |\cdot|_M : \vec{\mathcal{S}}_V &\to \mathbb{R} \\ \vec{s} &\mapsto \mathbb{1}_{\vec{s}}^{t} M \mathbb{1}_{\vec{s}} \end{aligned}$$

is called the *matrix order (function)* induced by M.

The definition allows non-symmetric matrices but we will usually focus on symmetric matrices. We do this for two reasons: The first is the word 'spectral' in the title. The symmetry of a matrix guarantees real valued eigenvalues - a property that we will need later. The second is practical: Most of the relevant matrices we will see encode pairwise information about elememts in V like similarities, correlations or distances, for example. Usually this information is symmetric.

The general interpretation of a matrix order function is very simple. Let $M \in \mathbb{R}^{V\times V}$ be a matrix and $\vec{s} = (B^c, B) \in \vec{\mathcal{B}}_V$ an oriented bipartition. If we write the oriented indicator vector of $\vec{s}$ in terms of ordinary indicator vectors, $\mathbb{1}_{\vec{s}} = \mathbb{1}_B - \mathbb{1}_{B^c}$, we can rearrange

$$\begin{aligned} \mathbb{1}_{\vec{s}}^{t} M \mathbb{1}_{\vec{s}} &= (\mathbb{1}_B - \mathbb{1}_{B^c})^t M (\mathbb{1}_B - \mathbb{1}_{B^c}) \\ &= \mathbb{1}_B^{t} M \mathbb{1}_B + \mathbb{1}_{B^c}^{t} M \mathbb{1}_{B^c} - \mathbb{1}_B^{t} M \mathbb{1}_{B^c} - \mathbb{1}_{B^c}^{t} M \mathbb{1}_B \\ \big(&= \mathbb{1}^t M \mathbb{1} - 2 \cdot \mathbb{1}_B^{t} M \mathbb{1}_{B^c} - 2 \cdot \mathbb{1}_{B^c}^{t} M \mathbb{1}_B \big) \end{aligned} \tag{2}$$

A matrix order thus always compares the sum of the matrix entries within each side with the sum of the matrix entries corresponding to edges between the sides. In the middle equation every matrix entry appears in the sum exactly once and with a factor of -1 if it corresponds to a pair of data points separated by $\vec{s}$. We see in the last equation that only separated pairs have influence on the order as $\mathbb{1}^t M \mathbb{1}$ is constant.

If $\vec{s} = (A, B) \in \vec{\mathcal{S}}_V$ is a set separation, the expression can be decomposed similarly

$$\begin{aligned} \mathbb{1}_{\vec{s}}^{t} M \mathbb{1}_{\vec{s}} &= (\mathbb{1}_{B\setminus A} - \mathbb{1}_{A\setminus B})^t M (\mathbb{1}_{B\setminus A} - \mathbb{1}_{A\setminus B}) \\ &= \mathbb{1}_{B\setminus A}^{t} M \mathbb{1}_{B\setminus A} + \mathbb{1}_{A\setminus B}^{t} M \mathbb{1}_{A\setminus B} - \mathbb{1}_{B\setminus A}^{t} M \mathbb{1}_{A\setminus B} - \mathbb{1}_{A\setminus B}^{t} M \mathbb{1}_{B\setminus A} \end{aligned} \tag{3}$$

The interpretation is very similar, except that the matrix entries corresponding to the intersection of the sides are ignored.
Please note that this doesn't necessarily mean that the entries in the intersection do not contribute to the order: Missing information also can be valuable information, and

moreover, the diagonal of M might contain summarised data about the relations between the objects including values ignored in the sum.

3.2 Examples

Now that we know that at least one matrix exists which induces a reasonable order function, namely the Laplacian of a graph, we should have a look at some more examples. Not all matrix orders assign the smallest order to the 'best' or in some sense most 'natural' separation. It depends on the the matrix entries' meaning:
For example, if the matrix entries represent *distances* on V, a 'good' separation should not separate a lot of points that are very close but should separate pairs of mainly very distant points. This means, a cut's weight, that is the sum of the distances between separated points, should be large. The cut's weight goes into the order negatively, hence we would like to minimise such an order function in order to find useful separations.
If the matrix entries describe *similarities*, a 'good' separation should not separate a lot of similar objects. Every separated pair should be as dissimilar as possible. Hence a cut's weight, i.e the sum of the similarities of separated pairs, should be small. Hence, we want to maximise the order function in this case [4].

3.2.1 Adjacency Matrix

Let $A \in \{0,1\}^{V \times V}$ be the adjacency matrix of a graph $G = (V, E)$. Let $\vec{s} = (B^c, B) \in \mathcal{B}_V$ be a bipartition of V. With equation (2) the order of $\vec{s}$ given by the adjacency matrix is

$$\begin{aligned} |\vec{s}|_A &= \mathbb{1}_B^t A \mathbb{1}_B + \mathbb{1}_{B^c}^t M \mathbb{1}_{B^c} - 2 \cdot \mathbb{1}_B^t M \mathbb{1}_{B^c} \\ &= \mathrm{vol}_{G[B]}(B) + \mathrm{vol}_{G[B^c]}(B^c) - 2\,\mathrm{cut}_G(B, B^c) \\ &= \mathrm{vol}_G(V) - 4\,\mathrm{cut}_G(B, B^c) \end{aligned}$$

The function takes a large value if the sides of $\vec{s}$ are highly connected and induce a small cut. This is an example of a matrix order taking higher values on more useful or more natural separations. As $\mathrm{vol}_G(V)$ is twice the number of edges, the order is negative if the cut contains more than a half of the edges.

[4] We can always negate a matrix to reverse the behaviour of the order function. This might result in an order function that takes only nonpositive values. Either we are comfortable with this fact or we have to subtract the resulting value from a big enough constant to shift all values to the positive side.

This order function also makes sense in $\mathcal{S}_V$. Please note that elements of $\mathcal{S}_V$ are in general not *separations* of G in the usual graph theoretic sense. We explicitly want to allow arbitrary splits of V into two possibly overlapping sets without any restrictions on edges that cross the separator. What type of 'separation' is meant will mostly be clear from the context and often we will use the term '(proper) vertex separation' if we talk about separations of graphs and the term 'set separations' if we mean separations of V as a set.
Comparing equation (2) and (3), we see that the function $|\cdot|_A$ takes the same value for a set separation $s = \{A, B\}$ as the bipartition version would for the bipartition $\{A \setminus B, B \setminus A\}$ on the reduced graph $G[V \setminus (A \cap B)]$. If all paths between the sides of the separation pass through the intersection, i.e. if the separation is a proper vertex separation, the subtracted terms in (3) are zero and the order depends only on the number of edges contained in the graph on the symmetric difference of A and B. In particular, a vertex separation of G, where the separator $T \subset V$ is an independent set with minimal number of neighbours on each side, has maximal order. The actual size of T only has an indirect effect on the order via the edges ignored in the sum [5]. The order of the degenerate set separation $\{V, V\} \in \mathcal{B}_V$ is zero. When we search for tangles or build a Tree-of-Tangles, natural separations have to be oriented first. For $|\cdot|_A$ this means we go through the list of separation in descending order and stop before we reach the degenerate set separation.

The adjacency order function can be generalised to weighted graphs like above: then A has real valued entries and we sum over the edges' weights instead of just counting the edges. The discussion of edge weighted graphs in graph clustering is often limited to graphs with nonnegative weights. This might have algorithmic benefits and most meaningful similarity measures are nonnegative, but there is an exception:
If our data set V is a set of vectors in $\mathbb{R}^k$, the *cosine similarity matrix* provides an interesting example of a graph with signed edge weights. The cosine similarity of two data points $u, v \in \mathbb{R}^k$ is the cosine of the angle between u, v and is defined by

$$\mathrm{sim}_{\cos}(u, v) \coloneqq \frac{u^t v}{\|u\|\|v\|}$$

The matrix $S \in \mathbb{R}^{V \times V}$ with entries $S(u, v) = \mathrm{sim}_{\cos}(u, v)$ is called *cosine similarity ma-*

[5] This allows, for example, a higher order for a separation given by a single highly connected vertex than for a separator consisting of two low-degree vertices. This unfortunate feature can be reduced by going over to the matrix $cI + A$ with some constant $c > 0$.

trix and can be interpreted as the adjacency matrix of a complete graph with signed edge weights in $[-1, 1]$.
Interpreting equation (2) for this matrix, the order of a bipartition $\vec{s} \in \mathcal{B}_X$ is high if the data points within the sides are very similar and the cut contains small, mostly negative edges. We can interpret the entries of S as correlations between the data points. A bipartition is rewarded by the order function if it combines positively correlated data points into the same group while at the same time elements of different groups tend to be negatively correlated.

3.2.2 Modularity Matrix

A slightly different view on the characterisation of interesting substructures of networks (or graphs) [6] is proposed in [19]. These substructures are called *communities* and are subgraphs whose connectivity differs significantly from what we would expect in a random graph with the same degree distribution. The random graph model underlying the idea is the *configuration model* [15]. The model provides a distribution over multigraphs with predefined degree sequence. Therein the expected number of edges between two nodes $u, v \in V$ with degrees $d_u, d_v \in \mathbb{N}$ is given by $\frac{d_u d_v}{2|E|}$.
Let $G = (V, E)$ be a graph and $d \in \mathbb{N}^V$ its degree sequence. If the edges of G are distributed according to the configuration model, the expected adjacency matrix is

$$\mathbb{E}(A) = \frac{dd^t}{\mathrm{vol}(G)}$$

The *modularity matrix* M of G is defined as the deviation of the adjacency matrix A of G from $\mathbb{E}(A)$

$$M \coloneqq A - \mathbb{E}(A) = A - \frac{dd^t}{\mathrm{vol}(G)}$$

Let $B \subset V$, then $\mathbb{1}_B^t M \mathbb{1}_B$ is the sum over the matrix entries corresponding to edges between vertices in B and measures the deviation of B's connectivity from the one we would expect from the configuration model. The value should be high, if B really is a community of G, i.e if we have significantly more edges within the subgraph induced by B than expected. The idea is the following:
We already know the degrees of all vertices in B, so we know how much edges have *one*

[6] A network is the same as a graph. The notion is sometimes used if a graph directly represents a concrete real world system of interconnected entities, for example, people usually refer to social *networks* instead of social *graphs*. In our context the difference is futile.

end vertex inside B. If the edges were random and if B was not too big, we would expect that a relatively large number of edges leaves B. If they do not, we have good reason to think that B is a community in G.

The *modularity* of a bipartition $\vec{s} \in \vec{\mathcal{B}}$ is given by

$$\mathrm{mod}(\vec{s}) \coloneqq \frac{1}{2\,\mathrm{vol}(G)} \mathbb{1}_{\vec{s}}^{t} M \mathbb{1}_{\vec{s}}$$

and measures, informally, to what degree the sides form communities and to what degree other communities of G are torn apart by the split.

We ignore the constant factor [7] and use the remaining part to define a new matrix order function $|\cdot|_M$, the *modularity order function*.

Given that a bipartition of high modularity should split the graph into parts consisting of tight communities, one could assume on first sight that this order function evaluates a very different criterion than the order functions we have seen before. The definition of the matrix might already give a hint that this is not really the case. This becomes more clear if we explicitly write the modularity order in terms of the adjacency order

$$\begin{aligned} |\vec{s}|_M &= |\vec{s}|_A - |\vec{s}|_{\mathbb{E}(A)} \\ &= |\vec{s}|_A - \frac{(d^t \mathbb{1}_{\vec{s}})^t (d^t \mathbb{1}_{\vec{s}})}{\mathrm{vol}(G)} \\ &= |\vec{s}|_A - \frac{(\mathrm{vol}(B) - \mathrm{vol}(B^c))^2}{\mathrm{vol}(G)} \end{aligned}$$

We see that the order function is a slight variation of the adjacency order including a term favouring splits into parts that are more balanced in volume.

3.2.3 Laplacian Matrices

We already have seen the order induced by the (combinatorial) Laplacian

$$L = D - A$$

[7] ... that was introduced in [19] merely for compatibility with previous definitions

of an edge-weighted graph $G = (V, E)$ on $\vec{\mathcal{B}_V}$ with adjacency matrix A and degree matrix D. Now we can derive it in a different way using what we already know about $|\cdot|_A$: The order of a bipartition $\vec{s} = (B^c, B) \in \vec{\mathcal{B}_V}$ is

$$|\vec{s}|_L = \mathbb{1}_{\vec{s}}^t D \mathbb{1}_{\vec{s}} - \mathbb{1}_{\vec{s}}^t A \mathbb{1}_{\vec{s}} = \mathrm{vol}(G) - |\vec{s}|_A = 4\,\mathrm{cut}_G(B, B^c)$$

We call this order function *Laplacian order* or *cut weight order*.

Like the adjacency order, this order function is also meaningful for set separations. Let $\vec{s} = (A, B) \in \vec{\mathcal{S}_V}$ and $\mathring{G} = G[V \setminus A \cap B]$. Then

$$\mathbb{1}_{\vec{s}}^t L \mathbb{1}_{\vec{s}} = 4\,\mathrm{cut}_{\mathring{G}}(A \setminus B, B \setminus A) + \mathrm{cut}_G(A \cap B, V \setminus A \cap B)$$

this means a set separation might be strongly preferred to a bipartition if most of the paths between the sides meet a vertex in the intersection.

Another important type of Laplacian matrix is the normalised Laplacian

$$L_{norm} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$$

The induced matrix order function has a slightly different interpretation. It can be explained easiest as *importance weighting* of the edges in G. We interpret $D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ as an adjacency matrix of a weighted graph $G^* = (V, E, W)$ with the same vertices and edges but each edge uv of G^* is weighted by $W(u, v) = d(u)^{-\frac{1}{2}} d(v)^{-\frac{1}{2}}$, this means the adjacency matrix of G^* is $A^* = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$. Let D^* be the degree matrix of G^*, then

$$\begin{aligned} L_{norm} &= D^{-\frac{1}{2}}(D - A)D^{-\frac{1}{2}} \\ &= I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \\ &= (D^* - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}) + (I - D^*) \\ &= L^* + C \end{aligned}$$

with a diagonal matrix $C = I - D^*$ and the combinatorial Laplacian L^* of G^*. The edge weights quantify the importance to the connectedness of their end vertices. The order function favours cutting edges between high degree vertices. If a graph contains two sparsely inter-connected but highly inner-connected subgraphs, the order function $|\cdot|_{L_{norm}}$ will reward separating these two regions more than $|\cdot|_L$ does by counting the edges alone. Hence the order function takes edges inside the sides of a separation into

account.

3.3 Sub- and Supermodularity

An important property of an order function for the theory of tangles is sub- or supermodularity. In general, matrix orders are neither sub- nor supermodular, but whether they are, can easily be read off the signs of the matrix entries, at least if the matrix is symmetric.

Lemma 3.2. *Let $M \in \mathbb{R}^{V\times V}$ and $f, g \in \mathbb{R}^V$ arbitrary. Then*

$$\begin{aligned} f^tMf + g^tMg = &\max(f,g)^tM\max(f,g) + \min(f,g)^tM\min(f,g) \\ &- (g - \min(f,g))^tM(\max(f,g) - g) \\ &- (\max(f,g) - g)^tM(g - \min(f,g)) \end{aligned}$$

where minima and maxima are taken element-wise.

Proof. We have

$$f = \min(f,g) + \max(f,g) - g$$

and therefore

$$\begin{aligned} f^tMf + g^tMg &= (\min(f,g) + \max(f,g) - g)^tM(\min(f,g) + \max(f,g) - g) \\ &\quad + g^tMg \\ &= \min(f,g)^tM\min(f,g) + \min(f,g)^tM(\max(f,g) - g) \\ &\quad + (\max(f,g) - g)^tM\min(f,g) + (\max(f,g) - g)^tM(\max(f,g) - g) \\ &\quad + g^tMg \\ &= \min(f,g)^tM\min(f,g) + \max(f,g)^tM\max(f,g) \\ &\quad + (\min(f,g) - g)^tM(\max(f,g) - g) \\ &\quad + (\max(f,g) - g)^tM(\min(f,g) - g) \\ &= \max(f,g)^tM\max(f,g) + \min(f,g)^tM\min(f,g) \\ &\quad - (g - \min(f,g))^tM(\max(f,g) - g) \\ &\quad - (\max(f,g) - g)^tM(g - \min(f,g)) \end{aligned}$$

□

Please note that for every $v \in V$ and every $f, g \in \mathbb{R}^V$ at most one of the values $g(v) - \min(f,g)(v)$ and $\max(f,g)(v) - g(v)$ is nonzero. Therefore the diagonal entries of M do not contribute to the subtracted terms in the lemma.
If the matrix M is symmetric, the equation simplifies to

$$\begin{aligned} f^t M f + g^t M g = &\max(f,g)^t M \max(f,g) + \min(f,g)^t M \min(f,g) \\ &- 2(g - \min(f,g))^t M (\max(f,g) - g) \end{aligned} \tag{4}$$

Now the claimed connection between signs of the matrix elements and sub- or supermodularity is apparent:

Corollary 3.3. *Let $M \in \mathbb{R}^{V \times V}$ be a symmetric matrix and $|\cdot|_M : \vec{\mathcal{S}} \to \mathbb{R}$ the induced matrix order function. Then $|\cdot|$ is submodular if and only if all off-diagonal entries of M are non-positive and $|\cdot|_M$ is supermodular, if and only if all off-diagonal entries of M are non-negative.*

Proof. Let all off-diagonal entries of M be non-negative. Then the subtracted term in equation (4) also is non-negative, hence the inequality

$$\begin{aligned} |\vec{s}|_M + |\vec{t}|_M &= \mathbb{1}_{\vec{s}}^t M \mathbb{1}_{\vec{s}} + \mathbb{1}_{\vec{t}}^t M \mathbb{1}_{\vec{t}} \\ &\leq \max(\mathbb{1}_{\vec{s}}, \mathbb{1}_{\vec{t}})^t M \max(\mathbb{1}_{\vec{s}}, \mathbb{1}_{\vec{t}}) + \min(\mathbb{1}_{\vec{s}}, \mathbb{1}_{\vec{t}})^t M \min(\mathbb{1}_{\vec{s}}, \mathbb{1}_{\vec{t}}) \\ &= \mathbb{1}_{\vec{s}\wedge\vec{t}}^t M \mathbb{1}_{\vec{s}\wedge\vec{t}} + \mathbb{1}_{\vec{s}\vee\vec{t}}^t M \mathbb{1}_{\vec{s}\vee\vec{t}} \\ &= |\vec{s} \wedge \vec{t}|_M + |\vec{s} \vee \vec{t}|_M \end{aligned}$$

holds for arbitrary separations $\vec{s}, \vec{t} \in \vec{\mathcal{S}_V}$. Therefore the matrix order $|\cdot|_M$ is supermodular.
The same reasoning applies if all off-diagonal entries of M are non-positive. The submodulary of $|\cdot|_M$ follows since the subtracted term are non-positive and hence the inverse inequality holds.
We noticed already that elements on M's diagonal do not contribute to the difference between the sides of the inequality, but all other entries do: Let $M(u_1, v_1) > 0$ and $M(u_2, v_2) < 0$ for $u_1, u_2, v_1, v_2 \in V$. Then the bipartitions (A^c, A) and (B^c, B) with corners $A \cap B^c = \{u_1\}$, $A^c \cap B = \{u_2\}$, the remaining corners chosen nonempty, but arbitrarily, and the bipartitions $(C^c, C), (D^c, D)$ with corners $C \cap D^c = \{v_1\}$, $C^c \cap D = \{v_2\}$, the remaining corners again chosen arbitrarily nonempty, are examples, where for the first pair of separations the submodularty inequality and for the second pair the super-

modularity inequality holds. The inequalities are strict in both cases. □

The if-statements of the proposition hold for non-symmetric matrices, too. The other direction doesn't hold in general as the subtracted terms in the lemma might counteract each other. For example, if the matrix is skew-symmetric, the subtracted terms cancel out and the induced matrix order is both super- and submodular [8].

For graphs with nonnegative edge weights, for example similarity or distance graphs, we obtain

Corollary 3.4. *Let $G = (V, E)$ be a graph with nonnegative edge weights. Let $A \in \mathbb{R}^{V \times V}$ be the adjacency matrix and $L \in \mathbb{R}^{V \times V}$ the Laplacian of G. Then the matrix order $|\cdot|_A$ is supermodular and the matrix order $|\cdot|_L$ is submodular.*

Let us conclude this section by rephrasing its observations in a more general way:

Corollary 3.5. *Let V be a finite set and $\sigma : V \times V \to [0, \infty)$ a nonnegative function on $V \times V$. Then the order function*

$$|\cdot|_\sigma : \vec{\mathcal{B}}_V \to \mathbb{R},\ (B^c, B) \mapsto \sum_{u \in B, v \in B^c} \sigma(u, v)$$

is submodular.

Proof. Let L be the Laplacian of a graph $G = (V, V \times V)$ where the edges are weighted by σ. Then $|\cdot|_\sigma = \frac{1}{4}|\cdot|_L$. □

3.4 Shifting and Balance

We have seen that the signs of a matrix' entries determine wether its matrix order is sub- or supermodular or neither. If we use the order function just to evaluate the quality of a separation, it does not play a big role. For example, if we are satisfied with creating a tangle search tree, we can use any matrix order function. But as soon as we want to build a Tree-of-Tangles, the property of an order function to be sub- or supermodular is very important. An obvious idea to enforce this property is to simply change the matrix entries so that the off-diagonal entries have the same sign and, hopefully, as much as

[8] These matrix orders might sound interesting at first, but a closer look unveils that they are completely useless in our scenario: the induced quadratic form takes the same value on every oriented bipartition vector, namely the sum of the diagonal, the only matrix elements that do not cancel out

possible of the order's effect on $\mathcal{B}_V$ is retained.
The simplest way to achieve non-positivity or non-negativity, and therefore submodularity or supermodularity, is to add a fixed constant to all matrix entries. We call this process *shifting of a matrix order (function)*. Shifting a matrix order has an interesting effect, *it changes the order's preference for balanced separations*:
Let $A \in \mathbb{R}^{V \times V}$ and $|\cdot|_A : \vec{\mathcal{S}}_V \to \mathbb{R}$ be the induced matrix order function. Let $c \in \mathbb{R}$ be a constant and $J = (1)_{u,v \in V} \in \mathbb{R}^{V \times V}$ be the matrix that has all entries equal to one. Then the shifted order of $\vec{s} = (A, B) \in \vec{\mathcal{S}}_V$ is

$$\begin{aligned} |\vec{s}|_{A+cJ} &= \mathbb{1}_{\vec{s}}^{t} A \mathbb{1}_{\vec{s}} + c \mathbb{1}_{\vec{s}}^{t} J \mathbb{1}_{\vec{s}} \\ &= |s|_A + c(|B \setminus A| - |A \setminus B|)^2 \end{aligned}$$

Hence, the order of an unbalanced separation will increase if we shift the matrix in positive direction using $c > 0$ and decrease if $c < 0$. The less balanced $\vec{s}$ is, the greater is the effect.
If we interpret A as an adjacency matrix of a similarity graph, more useful splits of V typically have higher order, so a positive shift of A results in an order function that tends to prefer unbalanced separations. This means that if we want our order function to be supermodular, we have to, at least partially, desist from desires about balanced cuts. Clearly, neither functions that ignore the balance nor functions that ignore everything except the balance are usually very helpful. If we are not forced to use sub- or supermodular order functions, shifting allows to gradually change the balancing preference.

The effect is quite interesting in situations where the relationships between objects of a set V are given by a distance function $\text{dist} : V \times V \to \mathbb{R}_{\geq 0}$. Let us assume we are searching for an order function on bipartitions of V. A good bipartition should split V into subsets with distances between elements on the same side rather small and distances between elements on different sides fairly large. Matrix order functions provide an easy way to evaluate such an objective. We could either directly use an order function defined by the distance matrix or transform the distance information into a similarity measure. Intuitively, these two approaches should be quite similar, but the effect of matrix shifting can make a significant difference.
There are multiple ways to create similarity functions based on distances, the simplest is to just negate the distance function and thereby turn minimal distances into maximal similarities. Similarities are usually nonnegative, so we might be tempted to additionally

add a constant $c \in \mathbb{R}$ and obtain the similarity function

$$\begin{aligned} \sigma : V \times V &\to \mathbb{R} \\ (u, v) &\mapsto c - \operatorname{dist}(u, v) \end{aligned}$$

We combine all values of σ into a similarity matrix $A = (\sigma(u, v))_{u \in V, v \in V}$ to assess bipartitions of V by the induced matrix order function $|\cdot|_A$ on $\vec{\mathcal{B}_V}$.
Now a change of the constant c corresponds to a shift of the matrix order function and therefore changes the balance preferences. Hence, working with distances seems to have a significantly different meaning than working with similarities, at least in the context of matrix order functions and if we used a naive way to make the similarities nonnegative. The two approaches are equivalent only if $c = 0$ and all similarities are nonpositive. If we shift the values of the similarity function into a positive range, we change our matrix order to prefer more unbalanced separations. This effect does not have to be detrimental, in this example using the distance matrix without any changes leads to a matrix order that strongly prefers balanced cuts. This is due to combinatorial reasons: There are much more pairs of separated points if a bipartition is balanced.
As we have noted above, an important motivation for such shifts can also be the sub- or supermodularity of the implied matrix order. From an optimisation perspective there seems to be *no free lunch*: in our example above, we either maximise a submodular set function that takes the balance into account ($c = 0$) or a supermodular function that prefers unbalanced separations ($c = \max_{u,v \in V} \operatorname{dist}(u, v)$). The maximisation of supermodular set functions without constraints is considered easy but as soon as constraints come into play, for example the balance of a cut, the maximisation might become difficult [3].

We have seen a similar balancing effect in example 3.2.2. There, the order function preferred slightly more balanced bipartitions by introducing a term comparing the volumes of the sides. This was achieved by weighting the vertices non-uniformly by their degree via using the matrix dd^t instead of $J = \mathbb{1}\mathbb{1}^t$. Both approaches follow the same principle and we can summarise the effect of matrix shifts on a matrix order's balancing preferences.

Proposition 3.6. *Let V be a set, $M \in \mathbb{R}^{V\times V}$ a matrix, $c \in \{-1, 1\}$ and $f \in \mathbb{R}^V$. Let $\vec{s} = (B^c, B) \in \vec{\mathcal{S}}_V$ be a separation, the order induced by the matrix*

$$M + cff^t$$

computes to

$$|\vec{s}|_{M+cff^t} = |\vec{s}|_M + c\left(\sum_{u\in B^c} f(u) - \sum_{v\in B} f(v)\right)^2$$

resulting in a change of the preference for balanced partitions.

3.5 Clamped matrices

Another way to influence the balancing preference of a matrix order function $|\cdot|_M$ is to limit the range of the matrix' entries. We want to call this operation *clamping* a matrix order function. For matrices $M, N \in \mathbb{R}^{V\times V}$ we define the element-wise maximum and minimum by

$$\min(M, N) \coloneqq \Big(\min\big(M(u, v), N(u, v)\big)\Big)_{(u,v)\in V\times V}$$
$$\max(M, N) \coloneqq \Big(\max\big(M(u, v), N(u, v)\big)\Big)_{(u,v)\in V\times V}$$

Let $c_* < c^*$ be two real numbers. Then the result of clamping the matrix M to the range $[c_*, c^*]$ is the matrix

$$\min\Big(\max\big(M, c_*J\big), c^*J\Big)$$

where $J \in \mathbb{R}^{V\times V}$ is the all-one-matrix.

We illustrate the effect by revisiting the previous example about data points in euclidean space and similarities arising from distances.
Let V be a finite set of points in $\mathbb{R}^k$ and $\Delta \in \mathbb{R}^{V\times V}$ matrix containing the distance between the points, i.e. $\Delta(u, v) = \mathrm{dist}(u, v)$ for all $u, v \in V$. The matrix order $|\cdot|_\Delta$ directly induced by Δ takes low values on 'good' cuts and strongly favours separating distant points. This is generally what we want, but the effect can be quite strong, making the order function sensitive to outliers and exaggerating the preference for balanced cuts. As we have seen above, we can try to counterbalance this property by simply shifting

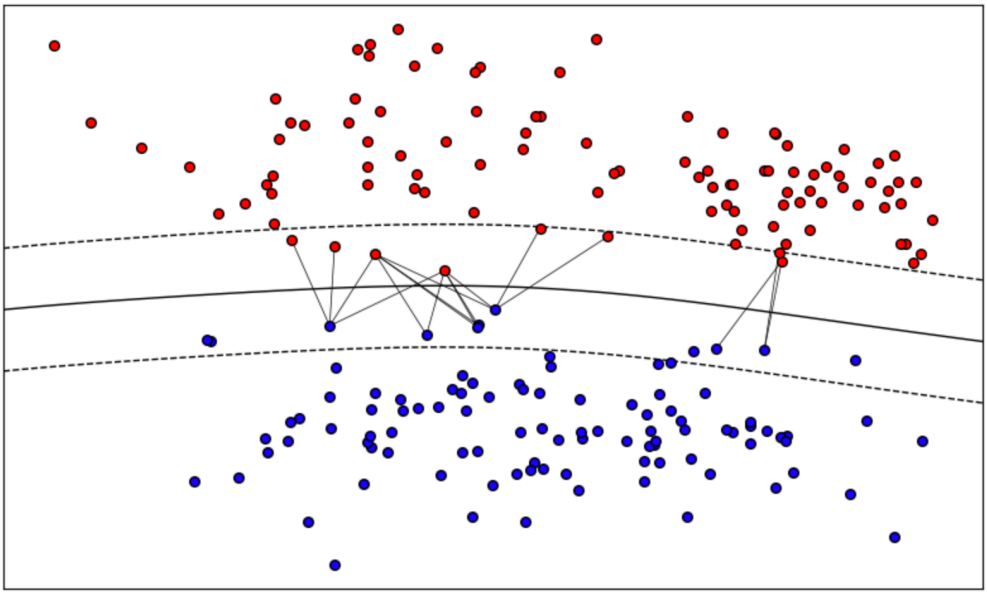

Figure 2: A randomly generated set of points in $\mathbb{R}^2$, partitioned into two groups. The solid line is soft max-margin decision boundary found by a Support Vector Maschine. The dashed lines illustrate the margin around the decision boundary. The bipartite graph inside the margin is induced by points on opposite sides of the bipartition that are close enough.

the matrix. A different possibility is trying to reduce the effect of far distant pairs by limiting the maximal entry of Δ, i.e by using the matrix

$$\Delta' := \min(\Delta, \delta_{max} J)$$

for some $\delta_{max} \in \mathbb{R}$. The modified matrix induces an order function where the linearly growing influence of sufficiently distant points is reduced to a constant one. If we combine both operations, shifting and clamping, we obtain, after scaling and negating the matrix, a very simple and natural similarity graph G^* with nonnegative edge weights. The adjacency matrix of G^* is given by

$$\Delta^* := -\frac{\min(\Delta, \delta_{max} J) - \delta_{max} J}{\delta_{max}} = \max(J - \frac{\Delta}{\delta_{max}}, 0)$$

As usual we treat edges of weight zero as nonexistent, so in G^* two vertices are connected if their corresponding points in $\mathbb{R}^k$ are close together and the edge's weight is higher if their distance is smaller, i.e. if they are more similar.

The induced matrix order $|\cdot|_{\Delta^*}$ ignores big distances (or the corresponding small similarities). Loosely speaking, the order function behaves more *locally* in $\mathbb{R}^k$ around a hypothetical boundary curve of the separations' sides than the original one induced by Δ. According to this order function better bipartitions have higher order and with equation (2) we see that a bipartition $s = \{B, B^c\} \in \mathcal{B}_V$ has optimal order, if

$$\sum_{u\in B, v\in B^c} \max(1 - \frac{\mathrm{dist}(u,v)}{\delta_{max}}, 0) = \sum_{\substack{u\in B, v\in B^c \\ \mathrm{dist}(u,v)<\delta_{max}}} 1 \; - \sum_{\substack{u\in B, v\in B^c \\ \mathrm{dist}(u,v)<\delta_{max}}} \frac{\mathrm{dist}(u,v)}{\delta_{max}}$$

is minimal. The expression balances the number of separated close pairs and the distances of these pairs. There might be a bipartition that separates more close pairs than another bipartition but nevertheless can have higher order if the sum of the distances between close points is sufficiently larger. The order is minimal if only a few close points are separated and if these points are not too close to each other.
Intuitively, we can imagine a soft *margin* given by δ_{max} around an hypothetical optimal decision boundary between the sides of s in $\mathbb{R}^k$. The deeper a pair of separated points is inside this margin the lower is the order. Points outside the margin are ignored.
Please note that unlike to common max-margin methods, not only the distance of a pair from the boundary is important but also the connectivity of the bipartite graph induced on the data points within the margin. Figure 2 shows an example.

3.6 Monotonic Transformations

Let $\vec{\mathcal{S}}$ be a universe of separations. Every order function $|\cdot| : \vec{\mathcal{S}} \to \mathbb{R}$ induces a quasi-order on $\vec{\mathcal{S}}$. We have already seen examples of order functions inducing the same or a reversed quasi-order. It will simplify our following considerations if we give this relationship a name:

Definition 3.7. Let $|\cdot|_1 : \vec{\mathcal{S}} \to \mathbb{R}$ and $|\cdot|_2 : \vec{\mathcal{S}} \to \mathbb{R}$ be two order functions. We say that $|\cdot|_2$ is a *monotonic transformation* of $|\cdot|_1$, if there is a strictly monotonic function $h : \mathbb{R} \to \mathbb{R}$ such that $|\cdot|_2 = h \circ |\cdot|_1$. If h is monotonically increasing, we say $|\cdot|_2$ is an *order preserving* (monotonic) transformation of $|\cdot|_1$, if h is monotonically decreasing, we say $|\cdot|_2$ is an *order reversing* (monotonic) transformation of $|\cdot|_1$.

It is intuitively obvious that the property of two order functions being monotonic transformations of each other is an equivalence relation:

Proposition 3.8. *Let $\mathcal{O}$ be the set of all order functions on $\vec{\mathcal{S}}$. The relation $\sim \in \mathcal{O} \times \mathcal{O}$ defined by*

$$|\cdot|_1 \sim |\cdot|_2 \iff |\cdot|_2 \text{ is an order preserving transformation of } |\cdot|_1$$

is an equivalence relation.

Proof. The reflexivity is clear, for every order function we have $|\cdot| = I \circ |\cdot|$ where $I : \mathbb{R} \to \mathbb{R}$ is the identity function.
If $|\cdot|_2$ is an order preserving transformation of $|\cdot|_1$, then there is a strictly increasing function $h : \mathbb{R} \to \mathbb{R}$ and $|\cdot|_2 = h \circ |\cdot|_1$. As h is strictly increasing it is invertible, so $|\cdot|_1 = h^{-1} \circ |\cdot|_2$. The inverse of a strictly increasing function is also strictly increasing, hence $|\cdot|_1$ is an order preserving transformation of $|\cdot|_2$.
Let $|\cdot|_1, |\cdot|_2, |\cdot|_3$ be three order functions. If $|\cdot|_1 \sim |\cdot|_2$ and $|\cdot|_2 \sim |\cdot|_3$, then there are monotonic functions $h_1, h_2 : \mathbb{R} \to \mathbb{R}$ with $|\cdot|_1 = h_1 \circ |\cdot|_2$ and $|\cdot|_2 = h_2 \circ |\cdot|_3$ and hence $|\cdot|_1 = (h_1 \circ h_2) \circ |\cdot|_3$. The composition of monotonically increasing functions is monotonically increasing, so $|\cdot|_1 \sim |\cdot|_3$. □

We have already seen that the diagonal of a matrix $M \in \mathbb{R}^{V \times V}$ does not have a qualitative effect on the induced order function $|\cdot|_M$ on $\vec{\mathcal{S}}_V$. In the world of graphs we can state this fact quite nicely using the notion of monotonic transformations:

Definition 3.9. Let $G = (V, E, A)$ be a weighted graph with adjacency matrix A and Laplacian matrix L. Let $D \in \mathbb{R}^{V \times V}$ be an arbitrary diagonal matrix. Then we call $A' \coloneqq D + A$ a *generalised adjacency matrix* and $L' \coloneqq D + L$ a *generalised Laplacian matrix* of G.

Proposition 3.10. *$G = (V, E)$ be an arbitrarily weighted graph. Then every pair of generalised adjacency matrices of G is order preserving, every pair of generalised Laplacian matrices of G is order preserving and every pair consisting of a generalised adjacency matrix of G and a generalised Laplacian matrix of G is order reversing.*

Proof. Let $A_1 = A + D_1, A_2 = A + D_2 \in \mathbb{R}^{V \times V}$ be two generalised adjacency matrices of G. Then $|\cdot|_{A_1} = |\cdot|_A + |\cdot|_{D_1} = |\cdot|_A + |\cdot|_{D_2} + |\cdot|_{D_1 - D_2} = |\cdot|_{A_2} + C$ where $C \in \mathbb{R}$ is constant as $D_1 - D_2$ is diagonal. Hence $|\cdot|_{A_1} = h \circ |\cdot|_{A_2}$ with $h : x \mapsto x + C$. The same argument applies if $L_1 = D_1 + L$ and $L_2 = D_2 + L$ are generalised Laplacian matrices of G.
Finally, we compute $|\cdot|_{L_1} = |\cdot|_L + |\cdot|_{D_1} = |\cdot|_D - |\cdot|_A + |\cdot|_{D_1} = -|\cdot|_A + C'$ for a constant $C' \in \mathbb{R}$ and hence $|\cdot|_{L_1} = h' \circ |\cdot|_{L_1}$ with $h' : x \mapsto -x + C'$. □

4 Spectral Separations

An order function should provide a way to rank separations according to their usefulness for detecting meaningful structures in a dataset. Given any order function it is interesting to ask how the most useful separations might look like, how we can find them and, if this turns out to be difficult, if there is an easy way to approximate them. In this section we want to discuss these questions for order functions given by real symmetric matrices and want to motivate the use of spectral methods.
Spectral approaches in data analysis can be motivated in multiple ways. A nice discussion, in particular about *Spectral Clustering*, can be found in [22]. We will look at the topic mainly from an *optimisation perspective*.

4.1 Definitions and basic facts

In what follows we need some basic properties of eigenvalues and eigenvectors of symmetric matrices. We assume the reader already has basic knowledge but want to briefly repeat the most important definitions and some basic facts. We will state all properties without proofs, they can be found for example in [14] or [18].
Let $M \in \mathbb{R}^{n\times n}$ be a matrix. A number $\lambda \in \mathbb{R}$ is an eigenvalue of M if there exists a vector $0 \neq f \in \mathbb{R}^n$ solving the equation

$$(M - \lambda I)f = 0$$

where I is the $n \times n$-identity matrix. If such a vector exists, it is called an *eigenvector* corresponding to the eigenvalue λ. An eigenvector f is not zero, hence the equation has a solution if and only if the determinant of the matrix $M - \lambda I$ is zero. The polynomial

$$P_M(\lambda) = \det(M - \lambda I)$$

of degree n is called *characteristic polynomial* of M. Every eigenvalue of M is a root of the characteristic polynomial of M and therefore M can have at most n different eigenvalues. The set of all different eigenvalues,

$$\sigma(M) \coloneqq \{\lambda \in \mathbb{R} : P_M(\lambda) = 0\},$$

is called the *spectrum* of M and the real number

$$\rho(M) \coloneqq \max_{\lambda \in \sigma(M)} |\lambda|$$

is called the *spectral radius* of M.
If λ is an eigenvalue of M, we call the multiplicity of λ as a root of P_M the *algebraic multiplicity* of λ. The subspace

$$E_\lambda = \{f \in \mathbb{R}^n \,|\, (M - \lambda I)f = 0\}$$

of $\mathbb{R}^n$ is called *eigenspace* of λ and its dimensionality is called the *geometric multiplicity* of λ.
In general, these two types of multiplicity need not necessarily be equal. But for symmetric matrices they are, we therefore will talk simply about the *multiplicity* of an eigenvalue of a symmetric matrix.
Symmetric matrices are very friendly matrices, we summarise some important properties in a theorem we want to call the *Spectral Theorem for symmetric matrices*:

Theorem 4.1. *Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix. If we count the eigenvalues of M with multiplicity, then M has exactly n real eigenvalues $\lambda_1 \leq \cdots \leq \lambda_n$ and there is an an orthonormal basis of $\mathbb{R}^n$ consisting of the corresponding eigenvectors $f_1, \ldots, f_n$. The matrix M therefore is similar to a diagonal matrix Λ containing the eigenvalues of M on the diagonal. This means*

$$M = U\Lambda U^t$$

where $U \in \mathbb{R}^{n\times n}$ is an orthogonal matrix that contains the eigenvectors $f_1, \ldots, f_n$ as columns. □

Every vector $f \in \mathbb{R}^n$ can be written as a linear combination of the orthonormal basis vectors $f_1, \ldots, f_n$ with a vector $a = (\alpha_1, \ldots, \alpha_n)^t \in \mathbb{R}$ of coefficients given by

$$a = U^t f$$

Hence, the matrix vector product Mf can be computed very intuitively by simply multiplying the coefficients by the eigenvalues

$$Mf = U\Lambda U^t f = U\Lambda a = \sum_{i=1}^{n} \alpha_i \lambda_i f_i$$

If we write out the equation in theorem 4.1 a bit differently,

$$M = U\Lambda U^t = \sum_{i=1}^{n} \lambda_i f_i f_i^t$$

we can see that every symmetric matrix can be written as a linear combination of the rank-one-matrices $f_i f_i^t \in \mathbb{R}^{n\times n}$ weighted by the eigenvalues of M. Hence, we could truncate the sum after a couple of terms to work with a low rank approximation of M. Similarly, we can remove or alter some of the terms to create new matrices that still contain parts of the original matrix' information but differ in other aspects.

There is one small observation that might appear trivial but is of central importance for our topic. Let $\lambda \neq 0$ be an eigenvalue of $M \in \mathbb{R}^{V\times V}$. Then the entries of a corresponding eigenvector $f \in \mathbb{R}^V$ are *coupled* by the equation $Mf = \lambda f$. The equation for a single component is

$$\big(\lambda - M(v,v)\big) f(v) = \sum_{u\neq v} M(v,u) f(u) \tag{5}$$

In particular the coupling has consequences for the possible signs of the entries corresponding to significant values $M(u,v)$, this means values of large absolute value.

This is a bare minimum of eigenvalue theory we need repeatedly in the sequel. We will see some more specialised results in places where they are used.

4.2 Matrix Order Optimisation

Let V be a finite set and $M \in \mathbb{R}^{V\times V}$ be a symmetric matrix. We assume that the matrix order given by M takes large values on good partitions, so we are interested in a separation $s \in \mathcal{B}_V$ with maximal order $|s|_M$ [9]. Our general goal is to solve the optimisation problem

$$\underset{f \in \{-1,1\}^V}{\text{maximize}} \quad f^t M f \tag{6}$$

[9] There are no substantial differences for matrix order functions that are to be minimised. We want to focus on the maximisation case because a lot of the needed classic matrix theory is about nonnegative matrices and we want to minimise the situations where we artificially talk about the negation of our matrix of interest.

The optimisation problem (6) is a form of a *quadratic unconstraint binary optimisation* problem (*QUBO*-problem [10]) and very difficult to solve for general matrices [16]. But for some special matrices the problem is very easy to solve. For example also in our case: if we require non-negativity of all entries of M, then the vectors $\mathbb{1}$ and $-\mathbb{1}$ both produce maximal solutions. We have seen above (section 3.6) that the diagonal of a matrix does not play a role if we compare the matrix's induced order of two separations. Hence, if a matrix order is supermodular, the trivial bipartition $\{\emptyset, V\}$ has maximal order. Similarly, if we were to minimise a submodular order function given by a symmetric matrix with nonpositive off diagonal entries, the trivial bipartition has minimal order.
The trivial bipartition is usually not what we are interested in. We can avoid it as the solution of (6) by adding a balancing term like in section 3.4, but, as we have seen, the order function will then loose its supermodularity and the optimisation problem will become difficult. The latter applies also if we use different constraints to enforce more balanced solutions, for example if we optimise the *RatioCut* objective function for graph cuts [22].
However, the potential difficulty should not discourage us from taking a further look into the problem: often satisfying approximations to the optimal solution can be found by applying suitable heuristics.

A particularly appealing heuristic is to relax the constraint forcing the solution to be a bipartition indicator vector with entries in $\{-1, 1\}$. Dropping this constraint, the relaxed optimisation problem reads

$$\begin{aligned} &\underset{f \in \mathbb{R}^V}{\text{maximize}} && f^t M f \\ &\text{subject to} && \|f\| = 1 \end{aligned} \tag{7}$$

where the length constraint is added to enforce the existence of a maximum.
This problem can be solved efficiently by computing the eigenvector corresponding to the smallest eigenvalue of M. This follows from a slightly more general fact that we will need often in what follows. These inequalities, or some variations of them, are sometimes called *Rayleigh's Inequalities* [14]:

Theorem 4.2. *Let $M \in \mathbb{R}^{n \times n}$ be a symmetric matrix and $f_1, \dots, f_n$ an orthonormal basis of eigenvectors of M corresponding to eigenvalues $\lambda_1 \leq \dots \leq \lambda_n$. Let $I \subseteq \{1, \dots, n\}$*

[10] sometimes these problems are called *unconstrained binary quadratic programming*-problems, but the acronym *UBQP* is much less appealing. Some authors switched to *QUBO* in a later paper [17]

be a nonempty subset of indices and $f \in \operatorname{span}\{f_i \mid i \in I\}$. *Then*

$$\lambda_{\min I} \leq \frac{f^t M f}{\|f\|^2} \leq \lambda_{\max I}$$

with equalities if and only if

$$f \in E_{\lambda_{\min I}} \cap \operatorname{span}\{f_i \mid i \in I\}$$

or

$$f \in E_{\lambda_{\max I}} \cap \operatorname{span}\{f_i \mid i \in I\}$$

respectively.

Proof. By the Spectral Theorem (4.1) we have $M = U\Lambda U^t$ where U is an orthogonal matrix containing the eigenvectors $f_1, \ldots, f_n$ as columns and Λ is a diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$. Then $f^t M f = (U^t f)^t \Lambda (U^t f)$ where $(a_1, \ldots, a_n)^t \coloneqq U^t f$ is a vector containing the coefficients of f written as a linear combination of the orthonormal vectors $f_1, \ldots, f_n$. Because $f \in \operatorname{span}\{f_i : i \in I\}$ the coefficients a_i for $i \notin I$ are zero. If we set

$$b_i \coloneqq \frac{a_i^2}{\sum_{j \in I} a_j^2}$$

for $i \in I$ we have $b_i \geq 0$ for all $i \in I$ and $\sum_{i \in I} b_i = 1$. Then the quotient

$$\frac{f^t M f}{\|f\|^2} = \frac{\sum_{i \in I} a_i^2 \lambda_i}{\sum_{i \in I} a_i^2} = \sum_{i \in I} b_i \lambda_i$$

is a convex combination of the eigenvalues corresponding to indices in I and

$$\lambda_{\min I} = \sum_{i \in I} b_i \lambda_{\min I} \leq \sum_{i \in I} b_i \lambda_i \leq \sum_{i \in I} b_i \lambda_{\max I} = \lambda_{\max I}$$

is obvious. We have equality on the left if and only if all weights b_i for indices $i \in I$ where $\lambda_1 > \lambda_{\min}$ are zero. This holds if and only if $f \in E_{\lambda_{\min I}} \cap \operatorname{span}\{f_i \mid i \in I\}$. The statement about equality on the right side follows similarly.

□

Well, the eigenvector corresponding to the largest eigenvalue solves our relaxed optimisation problem but a critical step is missing: We have to turn a solution into a bipartition of V.

Every component of an eigenvector $f \in \mathbb{R}^V$ of M corresponds to a data point $v \in V$, so

the most straightforward way to transform the solution of (7) into an approximate solution of (6) is to interpret the signs of the eigenvector's entries: We create a bipartition $(B^c, B) \in \vec{\mathcal{B}_V}$ by combining all $v \in V$ with $f(v) > 0$ to the set B. This idea is simple but is not at all clear that it makes any sense. We will try to justify this simple method in the following sections. Before we start, we optimistically anticipate the definition of our central notion:

Definition 4.3. Let $M \in \mathbb{R}^{V \times V}$ be a symmetric matrix and $f \in \mathbb{R}^V$ an arbitrary eigenvector of M. Any separation $\vec{s} \in \vec{\mathcal{S}_V}$ that is derived from f is called a *spectral separation*. If $\vec{s} = (B^c, B)$ with $B = \{v \in V \mid f(v) > 0\}$ we say that $\vec{s}$ is the *canonical spectral separation* of f. If the eigenvalues of M are sorted and the eigenvalue belonging to f appears in the k-th position, we say that $\vec{s}$ is of rank k.

Please note that this is a definition of a class of separations. There are multiple reasonable ways to derive a bipartition from an eigenvector. We will see some of them later but want to have a strong focus on canonical spectral separations. If the term 'spectral separation' is used without further specification, we mostly mean 'canonical spectral separations'.
Please note also, that the rank of a spectral separation depends on a sorting that is not specified in the definition. There are situations where we want to sort the eigenvalues ascending and in other situations descending. Moreover, we have to be careful with the rank if we have eigenvalues of multiplicity greater than one. In this case we have multiple canonical spectral separations with the same rank that are not inverses of each other.

We start with a few fundamental observations that follow immediately from the orthogonality of eigenvectors. We are interested in the signs of the eigenvectors' entries; if $f \in \mathbb{R}^V$ is an eigenvector of M then so is $-f$, therefore we can simplify the statements by following the convention that an eigenvector where all nonzero entries have the same sign is scaled such that all nonzero entries are positive.

Proposition 4.4. *Let V be a finite set, $|V| = n$, and $M \in \mathbb{R}^{V \times V}$ a symmetric matrix. Let $f_1, \ldots, f_n$ be an orthonormal set of eigenvectors of M. Then*

(i) there is at most one $1 \leq i \leq n$ such that $f_i(v) > 0$ for all $v \in V$

(ii) if there are $1 \leq i < j \leq n$ such that $f_i(v) \geq 0$ and $f_j(v) \geq 0$ for all $v \in V$ then the sets $\{v \in V | f_i(v) > 0\}$ and $\{v \in V | f_j(v) > 0\}$ are disjoint.

□

The proposition does not yet tell us much about the usefulness of spectral separations, but it serves as a first sanity check:
If an orthonormal eigenvector basis of $\mathbb{R}^V$ contains an eigenvector giving rise to the trivial canonical spectral separation, it is the only one with this property.
If there are multiple eigenvectors that do not contain entries of both signs, then each one of them contains information about a set of data points disjoint from the support of the others. We will see that this happens in particular easy situations where each of these vectors induces an optimal bipartition of V in a quite trivial way.

4.3 Supermodular Order Functions, Perron-Frobenius and Essential Positivity

Let $M \in \mathbb{R}^{V \times V}$ be a matrix that induces a supermodular order function $|\cdot|_M$. Then the trivial bipartition $(\emptyset, V)$ has optimal order, so a solution of the optimisation problem (6) will not help us to find good bipartitions. We have to add an additional constraint. Proposition 4.4, together with Rayleigh's Inequalities 4.2, provides an elegant way to find approximate optimal bipartitions regarding supermodular (or submodular) matrix order functions. To see how and why, we need some more notions and a powerful theorem from Linear Algebra, the *Perron-Frobenius-Theorem* about nonnegative matrices.

A matrix $M \in \mathbb{R}^{V \times V}$ is called *reducible* if there exists a permutation matrix P such that

$$P^t M P = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix} \tag{8}$$

where $X \neq 0$ and $Z \neq 0$ are square matrices [18]. If this is not the case, M is called *irreducible*.
Multiplying M by P^t and P results in a simultaneous reordering of rows and columns, hence the symmetry is preserved. Therefore a symmetric matrix M is reducible if we find a permutation matrix such that $P^t M P$ is block-diagonal (with more than one block), this means $Y = 0$ in (8). Intuitively, a symmetric nonnegative matrix M is irreducible if and only if the graph with adjacency matrix M is connected.

Theorem 4.5 (Perron-Frobenius-Theorem)**.** *Let $M \in \mathbb{R}^{V\times V}$ be a nonnegative, irreducible matrix. Then $\rho(M)$ is an eigenvalue of M with algebraic multiplicity one and there is a corresponding eigenvector $f \in \mathbb{R}^V$ with $f(v) > 0$ for all $v \in V$.* □

A proof of the theorem can be found for example in [18]. The unique vector $p \in \mathbb{R}^V$ with $Mp = \rho(M)p$ and $\|p\|_1 = 1$ is called the *Perron-vector* of M.

If an irreducible matrix $M \in \mathbb{R}^{V\times V}$ induces a supermodular order function, all its off-diagonal entries are nonnegative. Then there is an everywhere nonnegative matrix $M' = M + D$ with $D \in \mathbb{R}^{V\times V}$ diagonal that induces an order preserving transformation $|\cdot|_{M'}$ of $|\cdot|_M$. The matrix M' clearly still is irreducible and symmetric. The Perron-vector $p_{M'}$ of M' induces the trivial bipartition $(\emptyset, V)$ and every eigenvector corresponding to a different eigenvalue than $\rho(M)$ is orthogonal to $p_{M'}$ and hence must have both positive and negative entries. It therefore induces a nontrivial bipartition of V. In particular, the unit length eigenvector corresponding to the second largest eigenvalue is an optimal solution of (7) for the matrix M' with the additional constraint that the solution must be orthogonal to $p_{M'}$.

We can transfer this optimsation directly to matrices inducing supermodular order functions without explicitly creating a new order function. The important parts of the Perron-Frobenius Theorem can be generalised to matrices that have negative entries on the diagonal.
An irreducible matrix with nonnegative off-diagonal elements is called *essentially positive* [18]:

Definition 4.6. Let $M \in \mathbb{R}^{V\times V}$ be a symmetric matrix. If $M(u, v) > 0$ for all $u \neq v \in V$, we call M *essentially positive*.

Any essentially positive matrix has an eigenvector where all entries are positive:

Proposition 4.7. *Let $M \in \mathbb{R}^{V\times V}$ be essentially positive and symmetric. Then the multiplicity of the largest eigenvalue $\lambda_{\max}$ of M is one and there exists a corresponding eigenvector $f \in \mathbb{R}^V$ where all entries are positive.*

Proof. Because M is essentially positive, there is $\alpha \in \mathbb{R}$ such that $M' \coloneqq M + \alpha I$ is nonnegative, irreducible and symmetric. By the Perron-Frobenius Theorem there is a positive eigenvector p of M' corresponding to the maximal eigenvalue $\lambda_{\max} = \rho(M')$. Then $Mp = (M' - \alpha I)p = \lambda_{\max}p - \alpha p = (\lambda_{\max} - \alpha)p$, so p is an eigenvector of M with eigenvalue $\lambda_{\max} - \alpha$. The multiplicity of $\lambda_{\max}$ as an eigenvalue of M' is one and hence

the multiplicity of $\lambda_{\max} - \alpha$ as an eigenvalue of M.
If λ is any other eigenvalue of M, then it is a real number because of the symmetry and $\lambda+\alpha$ is an eigenvalue of M' since $M'f = (M+\alpha I)f = (\lambda+\alpha)f$. Therefore $\lambda+\alpha < \lambda_{\max}$ and hence $\lambda_{\max} - \alpha$ is a maximal eigenvalue of M. □

If the vector f in the proposition sums to one, i.e. $|f|_1 = 1$, we call f the *Perron-vector* of M, similar to the nonnegative case.

If a symmetric and off-diagonal nonnegative matrix M is reducible, it makes sense to analyse the blocks separately:

Proposition 4.8. *Let $M \in \mathbb{R}^{V\times V}$ be a symmetric and reducible matrix with nonnegative off-diagonal entries. Then there is a partition $V = \bigcup_{i=1}^{k} V_i$ in sets of sizes $|V_i| = k_i > 0$, $\sum_{i=1}^{k} k_i = |V|$, and an orthonormal set of eigenvectors of M*

$$\mathcal{B} \coloneqq \{f_{ij_i} \in \mathbb{R}^V \mid 1 \le i \le k, 1 \le j_i \le k_i\},$$

such that

(i) the matrices $M|_{V_i\times V_i}$ are irreducible for $1 \le i \le k$

(ii) $f_{ij_i}|_{V\setminus V_i} = 0$ for $1 \le i \le k$, $1 \le j_i \le k_i$

(iii) for each $1 \le i \le k$ exactly one of the eigenvectors in $\mathcal{B}$ is positive on V_i and the corresponding eigenvalue is the maximal eigenvalue of $M|_{V_i\times V_i}$

Proof. As M is reducible there is a permutation matrix P such that P^tMP is blockdiagonal with $k > 1$ blocks. We can assume that P is chosen so that k is maximal. Let $V_1 \ldots, V_k \subset V$ be subsets of V corresponding to the blocks. Then clearly $V = \bigcup_{i=1}^{k} V_i$ is a partition of V.
Let the i-th block be $M_i \coloneqq M|_{V_i\times V_i} \in \mathbb{R}^{V_i\times V_i}$ and $k_i = |V_i|$ for $1 \le i \le k$. Since k is maximal, each block M_i is irreducible, otherwise we could find a second permutation matrix P_2 such that $P_2^tP^tMPP_2$ has more than k blocks. So the first statement holds. The other statements follow directly from the irreducibility of the matrices M_i for all $1 \le i \le k$: If $g_{ij_i} \in \mathbb{R}^{V_i}$ is an eigenvector of M_i with eigenvalue λ_{ij_I}, then $f_{ij_i} \in \mathbb{R}^V$ with $f_{ij_i}|_{V_i} = g_{ij_i}$ and $f_{ij_i}|_{V\setminus V_i} = 0$ is an eigenvector of M with the same eigenvalue. this is the second statement, the third then follows from proposition 4.7 for the matrix M_i. □

As seen in the proof, the spectrum of a blockdiagonal matrix is the union of the spectra of the blocks, so if the spectra of the blocks are disjoint, all orthogonal sets of eigenvectors have the form given in the proposition. If the spectra intersect, there might be a different set of orthogonal eigenvectors that are nonzero on multiple V_i's at the same time - in this case an eigenspace might arise from multiple blocks.

If the matrix M is reducible, there are pairwise disjoint subsets $V_1, \dots, V_k \subset V$ that are kind of *independent* from each other. By equation (2), the order $|\cdot|_M$ does not change if one or multiple of the sets V_i change sides as a whole. Therefore every bipartition that does not cut through one of the V_i's has maximal order or, in other words, the order of a bipartition is only determined by the sets V_i that are split.
We summarise the results in the words of matrix orders and bipartitions:

Corollary 4.9. *Let $M \in \mathbb{R}^{V \times V}$ be a symmetric matrix inducing a supermoduler matrix order function $|\cdot|_M$.*

(i) If M is irreducible, then the trivial bipartition $\{\emptyset, V\}$ is a canonical spectral separation of the Perron-vector of M and has maximal order $|\cdot|_M$. The order of all other bipartitions is strictly smaller. The canonical spectral separation of an eigenvector of the second largest eigenvalue of M is an approximation to the optimal nontrivial bipartition according to $|\cdot|_M$.

(ii) If M is reducible, then there are multiple nontrivial bipartitions of maximal order $|\cdot|_M$. Each one is a canonical spectral separation of a suitable eigenvector of M.

□

4.4 Importance and Orthogonality

In this section we want to give an heuristic interpretation of the information contained in eigenvectors and discuss some consequences of their orthogonality.

The absolute values of an eigenvector's entries can be interpreted as weights of the data points in V. Let $A \in \mathbb{R}^{V \times V}$ be an essentially positive matrix and let $p \in \mathbb{R}^V$ be its Perron-vector. For every $v \in V$ the (positive) value $p(v)$ is called *eigenvector centrality* of v .

Eigenvector centrality is a widely used measure of the *importance* of nodes in networks or the *influence* a node has on other nodes in a network [20]. It measures the importance of each data point $v \in V$ by evaluating the relations to other data points given by A. What is interesting about eigenvector centrality is that it takes the importances of closely related data points into account. A data point has a high score if it is closely related to other points that themselves have a high score. The score of each v therefore depends not only on local relations of v but takes the global structure given by the matrix A on V into account [11].
Let now $f \in \mathbb{R}^V$ be another eigenvector of A. Then f is orthogonal to p and we can try to extend the interpretation of the values $|f(v)|$ for $v \in V$ as importances of the data points. If we create the canonical spectral separation of f, the signs of f's entries determine the side a data point belongs to and the absolute value of an entry can similarly be interpreted as some kind of centrality *inside the partition's side*. The coupling of the eigenvector entries given by equation (5) in section 4.1 has the effect that also inside each side heavily weighted data points have strong relations to other heavily weighted data points with the same sign. If A is an adjacency matrix, this means the weights of a positively weighted vertex' neighbours are mostly positive. This suggests that canonical spectral separations corresponding to positive eigenvalues tend to cut through bottlenecks of the graph, at least if the weighting is taken into account.
The orthogonality has the effect that the 'total importance' on each side sums to the same value, where each data point v's contribution $|f(v)|$ to a side is additionally weighted by the a-priori centrality given by the Perron-vector. Hence every spectral separation of such a matrix is *balanced*, in the sense of a centrality related importance [12].

The interpretation of the absolute values of an eigenvector's entries as some kind importance weighting of V is quite interesting in connection with the equivalence defined by order preserving transformations. We have seen that two matrix order functions given by matrices only differing on the diagonal induce the same quasi-order on $\vec{\mathcal{B}}_V$. If we are only interested in this induced quasi-order, we can change the diagonal of a matrix arbitrarily. Then the eigenvectors (and eigenvalues) certainly change and, in particular, the eigenvector centralities change. We therefore can influence the importance of the vertices and hence the spectral separations without changing the effect of the associated

[11] A variant of eigenvector centrality is one of the central building blocks of the *page rank algorithm* [21], a search result ranking algorithm that revolutionised web search.

[12] Please note that we cannot conclude that a spectral separation necessarily is balanced in any other sense: There are extremely unbalanced examples of spectral separations. Usually this is a result of a very special structural feature of the data, for example extreme outliers.

order function on $\vec{\mathcal{B}}_V$. On the one hand this emphasises the heuristic and possibly sub-optimal nature of spectral separations in general, on the other hand it might provide a tool to influence the results of the spectral approach in an advantageous way.

Any symmetric matrix $A \in \mathbb{R}^{V \times V}$ associated with a data set V induces a structure on V by interpreting its entries as similarities or dissimilarities, depending on the data model one primarily started with. The orthogonality has the effect that eigenvectors tend to *oscillate* over this induced structure. Like the basis vectors of the *discrete fourier transform (DFT)* [13] , they are orthogonal to each other and hence we can assume that each eigenvector captures a different aspect of the data that is uncorrelated to all other aspects. This makes them especially suitable for *divisive* data analysis techniques utilising combined partitioning information from multiple levels, like for example tangle-analysis. Moreover, by Rayleigh's Inequalities (theorem 4.2) the k-th eigenvector [14] is an optimal solution to the relaxed optimisation problem (7) under the additional constraint that the solution should be orthogonal to the space spanned by all previous eigenvectors. This means, heuristically, if the first $k-1$ eigenvectors capture the $k-1$ most important aspects of the data, then the k-th eigenvector describes the most important aspect that is still missing. In different words, a rank-k-spectral separation can be seen as a simultaneous refinement of all rank-$(< k)$-spectral separations at once. By orthogonality, at least one of the sides of each previous separation is split further. On real data, pairs of nested spectral separations, this means where only one side is split further, are very rare. It is not at all easy to randomly create a nonnegative matrix of medium size where such a pair can be found within a reasonable number of the first canonical spectral separations. The fact that spectral separations tend to cross makes them powerful generators of separation universes.

Theorem 4.10. *Let $M \in \mathbb{R}^{V \times V}$ be a symmetric matrix. Then there is a set of canonical spectral separations that generates the whole universe of bipartitions $\mathcal{B}_V$ over V.*

Proof. Let U be a matrix that has the eigenvectors of M as columns and $S \subseteq \mathcal{B}_V$ the set of canonical spectral separations derived from the columns of U. Since the columns of U are orthonormal, U is an orthogonal matrix and hence the rows also are orthonormal. We can think of the rows of U being indexed by elements of V and the columns by elements of S, so $U \in \mathbb{R}^{V \times S}$. Then column s of U is the eigenvector that induces one of

[13] ... which are closely related to the Laplacian eigenvectors of cycle graphs [10]

[14] The sorting of the eigenvalues and eigenvectors can be either ascending or descending - we reference it by the rank of the induced spectral separation for simplicity.

the orientations $\vec{s}$ or $\overleftarrow{s}$ of s. We denote the rows by r_v for $v \in V$ and their components by $r_v(s)$ for $s \in S$.

We show, that every two data points $u, v \in V, u \neq v$ can be separated by a canonical spectral separation of one of U's columns. The theorem then follows with proposition 2.1.

A small difficulty arises because exactly one of the sides of every separation includes entries equal to 0. We start with an easy case and afterwards show that in the more difficult situation we can negate some of the eigenvectors and are in the easy case.

Let us assume, that U does not contain two nonpositive rows, that's the easy case. Let $u, v \in V$ be two arbitrary data points. Then u and v are on different sides of a separation $s \in S$ if and only if either $r_u(s)r_v(s) < 0$ or $r_u(s)r_v(s) = 0$ and one of the values $r_u(s)$, $r_v(s)$ is strictly positive.

Let $s \in S$ be a spectral separation such that $r_u(s) > 0$ or $r_v(s) > 0$. Such a row must exist because U has at most one nonpositive row. We can assume $r_u(s)r_v(s) > 0$ otherwise u and v are already separated by s. The rows r_u and r_v are orthogonal and hence there must be $t \in S$, such that $r_u(t)r_v(t) < 0$. This t separates u and v.

We now assume, that there are multiple nonpositive rows in U. The matrix U has full rank and hence cannot contain a row that is completely zero. Therefore we find a column of U where one of the rows is negative. We replace this column by its negation. The negation of the column is still an eigenvector and at least one of the nonpositive rows has a positive entry in this position, i.e. its data point can be separated from the other nonpositive rows. If the new matrix still contains multiple nonpositive rows, we repeat the operation but take care, that we never negate the same column twice. Then no formerly nonpositive row will turn nonpositive a second time and therefore we stop after maximally $n - 1$ steps.

It remains to show that we can always find a column that was not negated before. Let us assume this is not the case. Let r_u be a nonpositive row of U where every negative entry belongs to a column that was already negated in an earlier step. Let s be the separation of such a column. Every already negated column contains a positive entry, so there is another row r_v such that $r_u(s)r_v(s) < 0$ and by orthogonality of r_u and r_v both rows must contain a positive entry somewhere. This is a contradiction.

□

This result might not be that impressive, even if we had to clear a hurdle mounted by our simple thresholding process. Please note that by proposition 4.8 the more difficult case in the proof can only happen for matrices that induce order functions that are neither

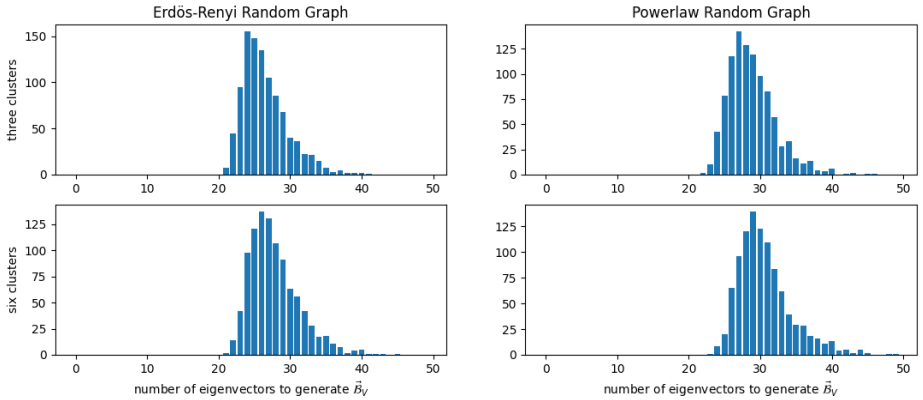

Figure 3: Number of (adjacency) spectral separations needed to generate the whole universe of bipartitions of different random graphs on 1000 vertices.

super- nor submodular.

For real data usually much less spectral separations are needed to generate the universe. Figure 3 shows empirically results obtained from experiments on different types of random graphs. Every figure was generated by sampling 1000 random graphs on 1000 vertices and counting how much canonical separations of the graphs adjecency matrix are needed to generate the complete universe $\vec{\mathcal{B}}_V$. The graphs were created as randomly interconnected unions of random graphs to avoid results biased towards graphs without clear structure, on the left each cluster is an Erdös-Renyi random graph on the right random graphs with a powerlaw degree distribution. The graphs in the top row were composed of three, in the bottom row of six clusters. The numbers look quite stable, the mean number of spectral bipartitions needed to generate $\vec{\mathcal{B}}_V$ seems to be larger if the graph has more structure.

The generating power of spectral separation sounds like a nice property. On real data we typically need only very few canonical spectral separations to separate arbitrary subsets from their complement. If our order function is reasonable and if we have a tractable way to generate 'good' suprema and infima, spectral separations theoretically allow to solve every thinkable partitioning problem [15]. However, unfortunately this property is not without drawbacks. To be able to generate the whole universe of bipartitions out of only a small number of bipartitions, these bipartitions must contain *global information*. This

[15] Clearly, the method is not a 'silver bullet': Finding a systematic way to generate exactly the right infima and suprema for the problem at hand can be expected to be difficult...

means, an orientation of such a bipartition might point to multiple substructures that are internally strongly connected but completely unrelated among each other. Again the analogy to the DFT-basis vectors might help our intuition: The DFT-basis-vectors are not localised in time, the eigenvectors of a matrix are not localised in the structure on V that is induced by the matrix. As a consequence, one usually has to extract local information from the canonical spectral separations in practice. We will see an important example in the next section.

4.5 Spectral Separations of Graph Laplacians

Everything we have seen about nonnegative matrices clearly applies to order functions based on the adjacency matrix of a graph [16]. We already mentioned that we can arbitrarily change the diagonal of a matrix without changing the quasi-order induced by the corresponding matrix order function on V. Let us use this fact and remove the a-priori weighting of the vertices given by the Perron-vector of a graph's adjacency matrix:
Let $G = (V, E)$ be a connected graph and $A \in \mathbb{R}^{V \times V}$ its adjacency matrix. Then A is irreducible and the entries of the Perron-vector p of A define weights of the vertices in V called eigenvector centrality. Despite the usefulness of the concept of eigenvector centrality, we might be interested in spectral separations that treat all vertices equally, at least a-priori. This means, we want a matrix $\hat{A}$ that has the same off-diagonal elements like A but has $\mathbb{1} \in \mathbb{R}^V$ as an eigenvector of the largest eigenvalue. We define $\hat{A} \coloneqq A - D$, where D is the diagonal matrix containing the degrees of the vertices. Then D contains the sums of A's rows and therefore every row of $\hat{A}$ sums to zero. Hence, $\mathbb{1}$ is an eigenvector of $\hat{A}$ with corresponding eigenvalue 0. We changed only the diagonal, so A' is still essentially positive and by (4.7) eigenvectors corresponding to the maximal eigenvalue are the only ones where all entries are positive. We conclude, that 0 is the biggest eigenvalue of $\hat{A}$, and hence $\hat{A}$ is negative semidefinite.
We have seen the negation of this matrix before, it is the (combinatorial) Laplacian $L = -\hat{A}$ of G. As $|\cdot|_L$ is an order reversing transformation of $|\cdot|_{\hat{A}}$ we found an equivalent matrix where we 'balanced out' the global eigenvector centrality in $\hat{A}$ by adding appropriate negative loops to G. The matrices $\hat{A}$ and $L = -\hat{A}$ have the same eigenvectors, this means, every other eigenvector of L can be chosen orthogonal to $\mathbb{1}$ and the sum of the absolute values of the entries on each side of the induced canonical spectral separation sum to the same value. Only the weights (or the centralities) of the vertices

[16] ... or a graph with non-negatively weighted edges. For simplicity, we will focus on unweighted graphs. All results can be generalised in a straight forward way to graphs with weighted edges.

inside the sides are taken into account.

The Laplacian $L = L(G)$ is historically tightly related to the connectivity of a graph G. It might all have started with Fiedler calling the second smallest eigenvalue λ_2 of L the *algebraic connectivity* of G [12]. He showed that

$$0 \leq \lambda_2 \leq \kappa(G)$$

where $\kappa(G)$ is the vertex connectivity of G and, if $f_i \in \mathbb{R}^V$ denotes an eigenvector corresponding to the i-th smallest eigenvalue, $i \geq 2$ of L, that the subgraph $G[V^+(f_i)]$ with

$$V^+(f_i) \coloneqq \{v \in V \mid f_i(v) \geq 0\}$$

does not have more than $i - 1$ connected components [13] [17]. This leads us to an important notion, the *Nodal Domains* of a vector according to a graph [1].

Definition 4.11. Let $G = (V, E)$ be a connected graph and $f \in \mathbb{R}^V$. A *positive (negative) strong nodal domain* of f is a maximal connected subgraph of G on vertices $v \in V$ with $f(v) > 0$ ($f(v) < 0$). A *positive (negative) weak nodal domain* of f is a maximal connected subgraph of G on vertices $v \in V$ with $f(v) \geq 0$ ($f(v) \leq 0$) that contains at least one nonzero vertex.
The total number of positive and negative strong nodal domains is denoted by $\mathfrak{S}(f)$, the total number of positive and negative weak nodal domains by $\mathfrak{W}(f)$.

Using this definition, Fiedler's result can be stated

$$\mathfrak{W}(f_i) \leq 2(i-1)$$

The inequality was generalised and improved afterwards. An important theorem is known as the *Discrete Nodal Domain Theorem*. A proof can be found in [1].

Theorem 4.12 (Discrete Nodal Theorem)**.** *Let M be a generalised Laplacian of a connected graph $G = (V, E)$ with $|V| = n$. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of M. If λ_i has multiplicity r, then any eigenvector f_i corresponding to λ_i has at most i weak nodal domains and at most $i + r - 1$ strong nodal domains:*

$$\mathfrak{W}(f_i) \leq i \quad \textit{and} \quad \mathfrak{S}(f_i) \leq i + r - 1$$

[17] If f is an eigenvector of λ_2 so is $-f$, hence the same statement applies to $V^-(f_i) \coloneqq \{v \in V \mid f_i(v) \leq 0\}$

□

The generality of the nodal domain theorem allows to write a similar statement about the eigenvectors of an adjacency matrix as a simple corollary. Historically, though, this special case might have been proved earlier [1].

Corollary 4.13. *Let $G = (V, E)$ be connected graph on $|V| = n$ vertices and $A \in \mathbb{R}^{V \times V}$ its adjacency matrix. Let $\lambda_1 \leq \cdots \leq \lambda_n$ be the eigenvalues of A and $f_1, \ldots, f_n$ a corresponding orthonormal set of eigenvectors. Let $r_i \in \mathbb{N}$ be the multiplicity of λ_i. Then*

$$\mathfrak{W}(f_i) \leq n - i + 1 \quad \textit{and} \quad \mathfrak{S}(f_i) \leq n - (i + r)$$

Proof. The matrix $-A$ is a generalised laplacian of G. The eigenvalues of $-A$ are $-\lambda_n \leq \cdots \leq -\lambda_1$ and $f_n, \ldots, f_1$ corresponding eigenvectors. □

The bounds in theorem 4.12 are unfortunately not sharp. If a lower bound exist, it only applies to special types of graphs, for example paths [1]. Nevertheless, it helps us to get an idea of the interactions of different eigenvectors and their spectral separations.

We already have mentioned the 'oscillation'-like property of the eigenvectors, the nodal domain theorem now provides something like an upper bound on something that we want loosely describe as something like a 'frequency' of an eigenvector.
Let $G = (V, E)$ be a connected graph and $f \in \mathbb{R}^V$ a vector. Let $D_1, D_2 \subset V$ be different nodal domains of f in G. We call D_1 and D_2 *adjacent* if there is $u \in D_1$, $v \in D_2$ such that u is a neighbour of v in G. Obviously, two adjacent nodal domains must have different sign. If D_1, D_2 are weak nodal domains they do not have to be disjoint. If their intersection is not empty, then for all $v \in D_1 \cap D_2$ we have $f(v) = 0$. We can picture the 'frequency'-like property of the eigenvectors of a laplacian [18] L of G in the following way:

We assume, that all eigenvalues have multiplicity one. Let $\lambda_1 < \cdots < \lambda_n$ be the eigenvalues of L corresponding to orthogonal eigenvectors $f_1, \ldots, f_n$. The first eigenvector f_1 has exactly one nodal domain given by all of G. The second eigenvector f_2 must be orthogonal to the first, so there are exactly two nodal domains, one positive and one negative. Hence, intuitively stretching the analogy to periodic functions, we can imagine f_2 as a function that takes a full period about the full data set, i.e. one with a frequency of 1 over the whole graph. Any eigenvector f_i for $i > 2$ has also at least two nodal domains - for the same reason as f_2 - but it might have more. The picture

[18] or, by the corollary, the adjacency matrix in reversed order

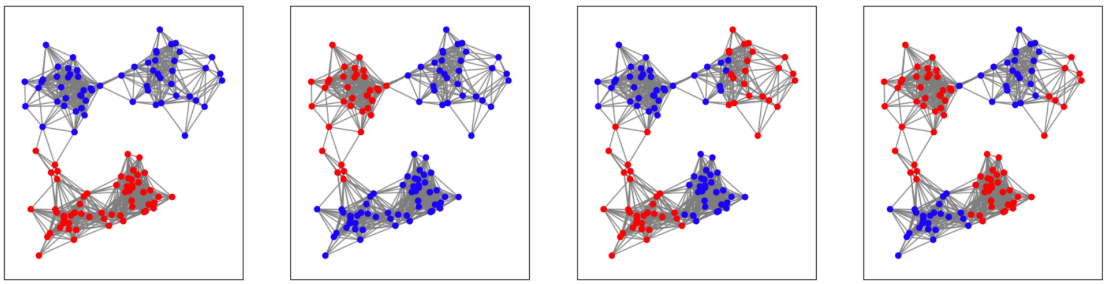

Figure 4: The first four canonical spectral separations of a ϵ-neighborhood graph of points in $\mathbb{R}^2$. Two points are connected if their distance is smaller than a threshold ϵ.

extends to this case naturally, the graph is split by f_i into adjacent nodal domains of different sign. Of course, the analogy to one dimensional periodic functions does usually not work anymore, but the 'oscillation'-like property of higher eigenvectors can be visualised very intuitively for graphs that are embedded into $\mathbb{R}^2$. Figure 4 shows an example.

We mentioned above that the orthogonality of eigenvectors can have the effect that spectral separations contain nonlocal information. This means, canonical laplacian separations of higher rank tend to contain global information about the structure of a graph. The unions of nodal domains with equal sign, i.e. a side of the canonical spectral separation, might be scattered all over the graph. This is most probably not what we want if we search guseful bipartitions of a graph. Nevertheless, they contain a lot of information that can be used to derive more local separations.

Nodal domains provide a way to extract local information from canonical spectral separations, we therefore define a special type of spectral separation.

Definition 4.14. Let $G = (V, E)$ be a graph and $f \in \mathbb{R}^V$ a vector with (weak or strong) nodal domains $D_1, \ldots, D_k$ on G. Then the bipartitions

$$(D_1^c, D_1), \ldots, (D_k^c, D_k) \in \vec{\mathcal{S}}_V$$

are called *nodal domain separations* of G induced by f.

Using this definition, we can describe the non-locality of a higher rank canonical spectral separation induced from a (laplacian) eigenvector more exactly: The orientation pointing to the positive side arises as the infimum of the nodal domain separations induced by the positive strong nodal domains of f_k. A canonical spectral separation of a graph

can therefore be seen as a composition of appropriate local separations that point to strongly connected subgraphs of G.

4.6 Principal Separations

So far we have mainly focused on sub- or supermodular matrix order functions, this means order functions given by offdiagonal-nonpositive or offdiagonal-nonnegative matrices. For matrices with mixed signs outside the diagonal the situation is a bit more involved because we cannot rely on the Perron-Frobenius theorem or one of its variations. In this section we want to have a look at another matrix type that arises very naturally if our data points are given as elements of the euclidean vector space $\mathbb{R}^k$. Then a canonical measure of similarity between two vectors $u, v \in \mathbb{R}^k$ is the inner product $\langle u, v\rangle = u^t v$ and the matrix containing the dot products of all pairs serves as a very natural similarity matrix.

Let $V = \{v_1, \dots, v_n\}$ be a set of objects and $Q = \{q_1, \dots, q_k\}$ a set of *attributes* or *features*. We think of the attributes as functions $q_i : V \to \mathbb{R}$ that map each object to its attribute value and combine all these attribute values in a *data matrix*

$$M = \begin{pmatrix} q_1(v_1) & \dots & q_k(v_1) \\ \vdots & \ddots & \\ q_1(v_n) & & q_k(v_n) \end{pmatrix} \in \mathbb{R}^{V \times Q}$$

We assume that the data set is centred around zero, this means $\sum_{v \in V} q(v) = 0$ for all $q \in Q$. It will simplify the notation if we identify $V \subset \mathbb{R}^k$ by its embedding in $\mathbb{R}^k$ induced by Q, in particular we associate $v \in V$ with its feature vector $\big(q_1(v), \dots, q_k(v)\big)^t$. Then the attributes in Q are projections onto the coordinate axes, $q_i : (q_1(v), \dots, q_k(v) \mapsto q_i(v)$ for $1 \leq i \leq k$. In this section we are interested in the order function $|\cdot|_K$ defined by the matrix

$$K = MM^t \in \mathbb{R}^{V \times V}$$

Then $K(u, v) = u^t v$ is the dot product between the feature vectors of $u \in V$ and $v \in V$ and we can interpret it as the correlation of the feature values of u and v. The order function $|\cdot|_K$ take a large value for a bipartition $s \in \mathcal{B}_V$ if data points on the same side of s are positively correlated and points on different sides are negatively correlated. The order function induced by K is very similar to the cosine similarity but without

normalising the length of the feature vectors of u and v.
The matrix K is symmetric, hence it has real eigenvalues and according to Rayleigh's inequality (4.2) the solution of the relaxed problem (7) is the largest eigenvalue of K achieved by a corresponding eigenvector.
Rayleigh's inequality applies to arbitrary real symmetric matrices but here we are more interested in the special case where the matrix K comes as a product of a data matrix and its transpose. In this case, we can say a bit more about the solution of the relaxed optimisation problem by decomposing the matrix M into a product of the matrices MM^t, D and M^tM where D is a diagonal matrix. This decomposition is called *Singular Value Decomposition (SVD)* [2] and one of its important uses in data analysis is a technique called *Principal Component Analysis (PCA)*. We call canonical spectral separations derived from eigenvectors of the matrix $K = MM^t$ the *principal separations* of the data matrix M.

To better understand PCA and principal separations, we informally model each of our features in Q as a random variable and interpret the rows of M as results of an i.i.d sampling process from the unknown joint distribution of the attributes $q_1, \ldots, q_k$. Each one dimensional subspace of $\mathbb{R}^k$ can be associated with an attribute of our data points. In particular, we can associate the one dimensional subspaces spanned by the standard unit basis vectors with the *original attributes* $q_1, \ldots, q_k$ and the projection of the data to these subspaces reveal the columns of the matrix M. The subspaces spanned by other unit vectors can be interpreted as *derived attributes*. Each derived attribute arises as a linear combination of the original attributes. The goal of PCA is to find a new basis of this space and thereby a new representation of the data with some favourable properties. The first property we want our new basis to have is that the associated derived attributes are uncorrelated, in the sense that there is no linear (or monotone) relationship between values of different attributes, the second is that most of the information the data set contains should be concentrated in the subspace spanned by the first few of the new basis vectors. Both of these properties can be very useful. The first allows us to 'whiten' the data, i.e. to remove linear dependencies between the attributes, the second can be used to reduce the dimensionality of the data while preserving most of the information. We will look at these properties in more detail, but let us first go back to where we started and find out, what this has to do with the solution to our relaxed optimisation problem stated in (7).

The descition of principal component analysis in this section primarily follows [2].

4.6.1 The first principal component

Suppose we want to reduce the dimensionality of our data most aggressively by projecting all data points to a single one dimensional subspace of $\mathbb{R}^k$ while keeping as much information as possible. This means, we want to introduce a single new derived attribute that describes our data as good as possible.
There are a lot of ways to quantify the information that is kept or lost in such a dimensionality reduction process, in PCA we focus on two very obvious notions that turn out to be two sides of the same coin: the *variance* of the derived attribute and the *least square error* of the projection.

Let $q \in \mathbb{R}^n$ be a vector of observed attribute values, for example a column of M. Then an estimate of the attribute's variance is $\mathrm{Var}(q) = \frac{1}{n} q^t q = \frac{1}{n}\|q\|^2$, assuming the attribute has zero mean[19]. If $w \in \mathbb{R}^k$ is a derived attribute, its variance is $\mathrm{Var}(Mw) \propto w^t M^t M w$ and maximal, by Rayleigh's Inequality, if w is an eigenvector corresponding to the largest eigenvalue of the symmetric matrix $M^t M$. Interestingly, maximising the variance is equivalent to minimising the least squared distance of the data to the subspace spanned by w. This follows directly from Pythagoras' Theorem: For a single point $v \in V$, i.e. a row of M, we have

$$|\langle w, v\rangle|^2 = \|v\|^2 - \|v - \langle w, v\rangle w\|^2$$

and as $\|v\|^2$ is constant, the left side is maximal if the minuend on the right side, that is the squared distance from v to its projection on the space spanned by w, is minimal. Summing over all rows of M shows the correspondence:

$$\mathrm{Var}(Mw) \propto (Mw)^t Mw = \sum_{v \in V} |\langle w, v\rangle|^2 = \sum_{v \in V} \|v\|^2 - \sum_{v \in V} \|v - \langle w, v\rangle w\|^2$$

4.6.2 The eigenvectors of MM^t and M^tM

We are actually interested in the vector $f \in \mathbb{R}^n$ that solves the relaxed optimisation problem (7), that is the unit vector that maximises $f^t K f = f^t M M^t f$. We already know, that f must be an eigenvector of MM^t corresponding to the largest eigenvalue.

[19] If we explicitely centred the data, an unbiased estimate would be $\frac{1}{n-1} q^t q$ but this does not make a difference for our purpose. We are interested in maximising the estimated variance and can happily ignore any constant factors.

The spectra of MM^t and M^tM are closely related: Let $w \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}$, then

$$M^tMw = \lambda v \Rightarrow MM^t(Mw) = \lambda(Mw)$$

and in the other direction, if $f \in \mathbb{R}^n$ we have

$$MM^tf = \lambda f \Rightarrow M^tM(M^tf) = \lambda(M^tf)$$

Hence, MM^t and M^tM have the same nonzero eigenvalues and there is a one to one correspondence between the associated eigenvectors[20].

Summing up, we can interpret a solution to the relaxed optimisation problem as follows. If $f \in \mathbb{R}^n$ is a unit vector maximising f^tMM^tf then there is a vector $w \in \mathbb{R}^k$ such that $f = Mw$ and w^tM^tMw is maximal. Please note that different to the discussion above, here w is usually not a unit vector.[21] The vector w spans a one dimensional subspace of $\mathbb{R}^k$ that can be associated with a derived attribute containing the most information about the data a single linearly derived attribute can. The components of the vector f are the values the new attribute takes on the elements of V and the rows of the matrix $fw^t \in \mathbb{R}^{n\times k}$ are least square approximations to the data points within the subspace spanned by w. The vector $\frac{w}{\|w\|}$ is called the *first principal component* of M.
The first *principal separation* then splits the data according to the sign of f. In other words, we project the data onto the first principal component w and split the data set at the mean of this new derived attribute. This semantically makes sense: We found the best way to characterise the data by a linear combination of all attributes and use this information to decide where to split the data set in roughly balanced subsets.

4.6.3 Principal Component Analysis

The one dimensional case can be extended to higher dimensional subspaces. In the previous section we found the direction in $\mathbb{R}^k$ where the data has the largest variance and called it the *first principal component*. To extend the definition, we write the first

[20] One of the matrices M^tM and MM^t might be bigger than the other. The spectrum of the bigger matrix is filled up with zeros

[21] w already contains a scaling factor needed to reconstruct the data, we will come back to this point later in the section about SVD

principal component in the following form

$$w_1 \coloneqq \underset{\|w\|=1,}{\arg\max} \|Mw\|^2$$

and define the remaining principal components inductively as the directions of maximal variance in the orthogonal complement of the subspace spanned by the previous components

$$w_{j+1} \coloneqq \underset{\substack{\|w\|=1,\\ w\perp w_1,\ldots,w_j}}{\arg\max} \|Mw\|^2$$

for $1 \leq j < k$. Again, Rayleigh's inequality confirms that $w_1, \ldots, w_k$ are eigenvectors corresponding to the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ of the matrix M^tM.
Despite a generally pleasing aspect and a nice match with properties of the eigenvectors, the orthogonality constraint might look somehow arbitrary on the first view. That's not the case, in contrary, this emerges quite naturally if we construct the principal components one after the other: Assuming we have already found the first $j-1$ principal components, the j-th principal component w_j is the direction of largest variance of the matrix $M_j \coloneqq M - \sum_{i=1}^{j-1} Mw_iw_i^t$. This matrix contains the remaining information that is not already described by the first $j-1$ principal components. It sends the whole subspace spanned by $w_1, \ldots, w_{j-1}$ to zero and behaves exactly like M on the *complement*, so w_j becomes the largest eigenvector of $M_j^tM_j$, i.e. the first principal component of M_j.

If V is a linear vector space and $U \subseteq V$ is a linear subspace, we write $U \leq V$ and denote the projection of $v \in V$ to U as

$$pr_U(v) \coloneqq \sum_{i=1}^{l} \langle v, b_i \rangle b_i,$$

if $b_1, \ldots, b_l \in V$ is a basis of U.
We call U a *least squares fit subspace* to a set of points $p_1, \ldots, p_n \subseteq \mathbb{R}^k$, if the average squared distance between the points and their projections to U is minimal, i.e. if

$$\frac{1}{n} \sum_{i=1}^{n} \|p_i - \mathrm{pr}_U(p_i)\|^2$$

is minimal. With these definitions we are ready to state the most important fact about principal components: The least square property of the first principal component also

extends to higher dimensional subspaces.

Theorem 4.15. *Let $M \in \mathbb{R}^{n\times k}$ a matrix containing data points $v_1, \dots, v_n \in \mathbb{R}^k$ as rows. Let $w_1, \dots, w_k$ an orthonormal family of eigenvectors of M^tM corresponding to eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$. Then for each $1 \leq l \leq k$ the subspace $W_l \leq \mathbb{R}^k$ spanned by $w_1, \dots, w_l$ is a least square fit subspace to $v_1, \dots, v_n$.* □

A proof can be found in [2].

Principal component analysis might serve the intuition about how a spectral separation of rank $k > 1$ refines the sides of spectral separations of rank $< k$. The first principal separation provides the best split of the data along the axis of maximal variance. Every additional principal separation then refines the previous splits utilising a derived attribute that is uncorrelated to all derived attributes before and contains the most information, in the sense of variance, that *was not already used for an earlier split.* Hence, every principal separations should be used in combination with the lower rank separations, for example by building corners or by using a more specialised approach like nodal domains for graphs.

5 Summary

We discussed *Spectral Separations*, an important method to generate bipartitions and set separations to analyse the internal structure of data sets. The method comes with a somewhat heuristic flavour but we have analysed the approach rigorously using arguments from optimisation, matrix theory and graph connectivity analysis.

We first encountered a very broad class of order functions defined as quadratic forms involving matrices representing pairwise information between elements of a set. The class of these *matrix order functions* covers the standard evaluation criterium for a cut in a graph, namely its *cut weight*. Furthermore it allows us to transfer general characterisations of a data set's pairwise structure to the theory and practice of separation systems, like for example covariances and distances.
We have seen a fundamental way to influence qualities of a matrix order function through *matrix shifting*. This operation allows us to change a matrix order's preference for balanced separations. We have shown that the balancing behaviour of a matrix order function is tightly connected to sub- or supermodularity, an important property of an order function in the theory of separation systems and tangles.
Another important matrix modification, the *clamping* of matrix elements, allowed us to draw an interesting comparison to a well known class of partitioning methods in supervised machine learning. By restricting the matrix entries to a certain range we can make an order function focus more on *local* relationships between data points. This allowed us to intuitively link similarity graph clustering methods to *margin*-classifiers.

Matrix order functions provide a way to evaluate the usefulness of a separation in describing the internal structure of a set. The task to find separations that optimise such an order function can be stated as a *quadratic unconstrained binary optimisation* problem. A relaxation of such problems naturally motivated the definition of *spectral separations*, that is separations derived from eigenvectors of the order function's defining matrix.
For the special class of *essentially positive* matrices, or their negations, we justified this method by resorting to the Perron-Frobenius Theory of nonnegative matrices. The theory allowed us to verify basic properties of spectral separations and lead to the important concept of *eigenvector centrality*. A slight generalisation allowed us to interpret the absolute values of an eigenvector's entries as a measure of the corresponding data points' importance within their side of a separation. According to this, a spectral separation, derived from an eigenvector corresponding to a large eigenvalue, cuts a data set at a

bottleneck if we take the importance weighting into account.
The class of nonnegative matrices in particular includes adjacency matrices of graphs. Analysing the connectivity of graphs using spectral methods is well established in the literature. Spectral separations fit nicely into the theory, especially the notion of a *nodal domain* provides a way to extract local connectivity information from eigenvectors. We further observed that spectral separations of different rank *refine* each other.
This refinement process finally was illustrated from a different point of view by a discussion of *principal component analysis*. Associating spectral separations with principal components intuitively explained the interaction of multiple spectral separations as an iterative refinement process.

This thesis focused mainly on theoretical aspects but spectral separations could provide a promising tool in the practical data analyst's toolbox. This could especially apply for its use as a preprocessing step in tangle analysis. A deeper examination of the interplay between tangles and spectral separations could be a very interesting future research direction.

References

[1] T. Biyikoglu, J. Leydold, and P. F. Stadler, editors. *Laplacian eigenvectors of graphs : Perron-Frobenius and Faber-Krahn type theorems*, volume 1915 of *Lecture notes in mathematics*. Springer, Berlin [u.a.], 2007.

[2] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of Data Science*. Cambridge University Press, Cambridge, 2020.

[3] W. Cunningham. On submodular function minimization. *Combinatorica*, 5:185–192, 07 1985.

[4] R. Diestel. Abstract separation systems. *Order*, 35(1):157–170, apr 2017.

[5] R. Diestel. *Graph Theory: 5th edition*. Springer Graduate Texts in Mathematics. Springer-Verlag, © Reinhard Diestel, 2017.

[6] R. Diestel. Tangles in the social sciences. *arXiv preprint arXiv:1907.07341*, 2019.

[7] R. Diestel, F. Hundertmark, and S. Lemanczyk. Profiles of separations: in graphs, matroids and beyond, 2011.

[8] R. Diestel and S.-i. Oum. Tangle-tree duality: In graphs, matroids and beyond. *Combinatorica*, 39, 01 2017.

[9] R. Diestel and G. Whittle. Tangles and the mona lisa. *CoRR*, abs/1603.06652, 2016.

[10] V. Ekambaram. *Graph Structured Data Viewed Through a Fourier Lens*. PhD thesis, EECS Department, University of California, Berkeley, Dec 2013.

[11] C. Elbracht, D. Fioravanti, S. Klepper, J. Kneip, L. Rendsburg, M. Teegen, and U. von Luxburg. Tangles: From weak to strong clustering. *arXiv preprint arXiv:2006.14444*, 2020.

[12] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[13] M. Fiedler. A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory. *Czechoslovak Mathematical Journal*, 25(4):619–633, 1975.

[14] C. Godsil and G. Royle. *Algebraic Graph Theory*, volume 207 of *Graduate Texts in Mathematics.* volume 207 of Graduate Texts in Mathematics. Springer, 2001.

[15] R. v. d. Hofstad. *Random Graphs and Complex Networks*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2016.

[16] G. Kochenberger, J.-K. Hao, F. Glover, M. Lewis, Z. Lü, H. Wang, and Y. Wang. The unconstrained binary quadratic programming problem: a survey. *Journal of Combinatorial Optimization*, 28(1):58–81, 2014.

[17] M. W. Lewis and F. W. Glover. Quadratic unconstrained binary optimization problem preprocessing: Theory and empirical analysis. *CoRR*, abs/1705.09844, 2017.

[18] C. D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, 2001.

[19] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):85778582, May 2006.

[20] M. E. J. Newman. *Mathematics of Networks*, pages 1–8. Palgrave Macmillan UK, London, 2016.

[21] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

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

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit im Bachelorstudiengang Mathematik selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel - insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen - benutzt habe. Alle Stellen, die wörtlich oder sinngemäß aus Veröffentlichungen entnommen wurden, sind als solche kenntlich gemacht. Ich versichere weiterhin, dass ich die Arbeit vorher nicht in einem anderen Prüfungsverfahren eingereicht habe und die eingereichte schriftliche Fassung der auf dem elektronischen Speichermedium entspricht.

Ort, Datum	Unterschrift