Title
Combinatorial Analysis of Barcodes and Interval Graphs for Applications in Data Science

Author
Jaramillo Rodriguez, Edgar

Publication Date
2023

Peer reviewed|Thesis/dissertation

Combinatorial Analysis of Barcodes and Interval Graphs
for Applications in Data Science

By

EDGAR JARAMILLO RODRIGUEZ
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

MATHEMATICS

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Jesús A. De Loera (Chair)

Thomas Strohmer

Luis Rademacher

Committee in Charge

2023

To my teachers, especially John Montrella and Daniel Willms.

Contents

Edgar Jaramillo Rodriguez, June 2022, Mathematics

Abstract

In this thesis we develop combinatorial methods for studying *barcodes*. A barcode is a collection of closed intervals on the real line. Barcodes arise as key objects in *topological data analysis*, as summaries of the persistent homology groups of a filtration, and in *graph theory* as interval graphs.

We first introduce a map from the space of barcodes to certain equivalence classes of *double occurrence words*, i.e., permutations of a multiset in which every element occurs exactly twice. We call the set of all such words the space of *combinatorial barcodes*. We then define an order relation on this space, based on the weak-Bruhat order, and show that the resulting poset is a graded lattice. We show that this lattice can also be defined using new notions of *inversion multisets*/ *inversion vectors* of double occurrence words. We also use these objects to prove several properties of the lattice. For example, we compute its rank generating function and introduce a natural bijections between combinatorial barcodes, *trapezoidal words*, and *Stirling permutations*. In addition to being of interest from a combinatorial perspective, we also show that the cover relations in this lattice can be used to determine the set of barcode bases of persistence modules.

We then generalize this construction, producing an entire family of multipermutations associated to barcodes. We equip these new multipermutations with a similar order relation and show that the resulting posets also form lattices, which we call the *power-k barcode lattices*. Unfortunately, these lattices do not retain many of the other "nice" combinatorial properties found in our original construction. However, we show that these multipermutations record increasingly detailed information about the arrangement of the bars in a barcode. We prove that for a large class of barcodes these multipermutations can be used to bound two classic, continuous metrics on barcodes: the Wasserstein and bottleneck distances. We also show that these lattices form the face lattices of certain *Bruhat-interval polytopes*.

Finally, we study an original model for generating random interval graphs (or equivalently random barcodes). This model is motivated by scientific sampling problems, where one receives a sequence of time-stamped observations and wants to make conclusions about the start and end of certain events. Although our general model is difficult to analyze, we prove many results about the

expected behaviour of this model in a special case which we call the *stationary case*. For example, we compute the expected number of edges, maximum degree, and maximum clique size. We also study the limit behavior of this model as the number of samples/ observations goes to infinity. In particular, we prove that the special case of this model converges to a complete graph and compute a lower bound on the expected waiting time for this to occur.

Acknowledgments

Thank you to my advisor, Jesús A. De Loera, for your unwavering belief and support. Thank you for counseling me through every hurdle and celebrating every milestone. Thank you for all the long nights checking my proofs or reviewing my slides; you work harder than anyone I know.

Thank you to my committee members, Luis Rademacher and Thomas Strohmer, for reviewing this dissertation and providing helpful feedback. Thank you to my co-authors on this work, Deborah Oliveros and Antonio Torres Hernandez. Thank you to my mentors and collaborators at Los Alamos National Lab, Axel Browne, David Butts, Sara Del Valle, Geoffrey Fairchild, Tim Germann, and Nidhi Parikh. Thank you to my many other collaborators, including Ángel Chávez, Erica Roldan Roa, Nicole Sanderson, and José Simental Rodríguez.

Thank you to my academic siblings, Alex Black, Félix Almendra Hernández, Chengyang Wang, Jack Wesley, and Yue Wu and to my friends and fellow graduate students in the math department for learning and commiserating with me. Thank you to the staff, faculty, and postdocs of the UC Davis math department, especially Tina Denena who keeps the whole thing running. Thank you to my teachers, especially John Montrella and Daniel Willms, for instilling the curiosity that has led me here.

Thank you to Mom, Dad, and Marco for all the love and support you have shown me. Thank you to my friends, especially Esha Datta, Emilia Hernandez, Agatha Scott, Joshua Hartman, and James Hughes for all the times we have shared.

Finally, thank you to my partner, Isabel Johnson. I could write another chapter with all the things I want to thank you for, but I'll just say thank you for sharing your life with me.

I also gratefully acknowledge financial support from NSF grant DMS-1818969, NSF HDR TRIPODS grant CCF-1934568, and the NSF-AGEP supplement. Part of my travel was also supported by the UC Davis Mathematics Travel Awards Program and the UC Davis Graduate Travel Awards program.

CHAPTER 1

Introduction

A *barcode* is a finite multiset of (closed) intervals on the real line, $B = \{[b_i, d_i]^{m_i}\}_{i=1}^{n}$. Recently, barcodes have gained attention because of their role in topological data analysis, where they serve as summaries of the persistent homology groups of a filtration [**ZC05**]. Barcodes also appear in the context of interval orders and interval graphs [**LB62**]. In this thesis we develop combinatorial methods for analyzing barcodes for applications in topological data analysis and random interval graphs.

This thesis consists of two parts. The first part, Chapters 2 and 3, is based on a research project where we developed new combinatorial invariants on the space of barcodes. In addition to being of interest from a combinatorial perspective, these invariants are useful for studying barcodes associated to persistence modules and have connections to classical continuous metrics on the space of barcodes. These contributions are outlined in Sections 1.2.1 and 1.2.2. The second part, Chapter 4, is based on joint work with Jesús A. De Loera, Deborah Oliveros, and Antonio Torres Hernandez. We introduced a new model for generating random interval graphs (equivalently, random barcodes) and study its behavior. This model is motivated by what we call chronological sampling problems, where researchers collect a series of time-stamped observations and wish to learn the shape of the underlying distribution of the data. These contributions are outlined in Section 1.2.3.

1.1. Background

1.1.1. Topological Data Analysis and Persistent Homology. In this section we provide some necessary background on persistent homology to contextualize our work. As our work is motivated by topological data analysis (TDA), where persistent homology is applied to the study of data, we introduce persistence from this viewpoint, following the approach in [**Ghr08**] and [**Car09**]. For a more general introduction to persistence we refer the reader to the survey by Edelsbrunner and Harer [**EH08**] or the foundational paper by Zomorodian and Carlsson [**ZC05**]. We will assume

some familiarity with modern algebra, particularly the theory of rings and modules, as can be found in [**DF04**], for example.

In the context of this thesis, *data* refers to a collection of points $X = \{x_\alpha\}_{\alpha \in A} \subset \mathbb{R}^n$. Thanks to advances in computing, data is currently being generated at an unprecedented rate in a variety of disciplines. Much of this data is high-dimensional, that is to say $X \subset \mathbb{R}^n$ for some very large n. For instance, medical data may represent each patient as a vector x where each component x_i indicates a different characteristic of the patient such as age, weight, blood pressure, or the presence of certain genetic markers [**BHH$^+$21**]. Often it is beneficial to understand the shape of a dataset; for instance, in medicine one might ask how many components are present in a cohort because patients from the same component might have similar health outcomes following some procedure. Naturally, this is more difficult when the data is high dimensional since projections down to 2 or 3 dimensions may remove or distort important features. Persistent homology is a method to measure the shape of data using tools from algebraic topology.

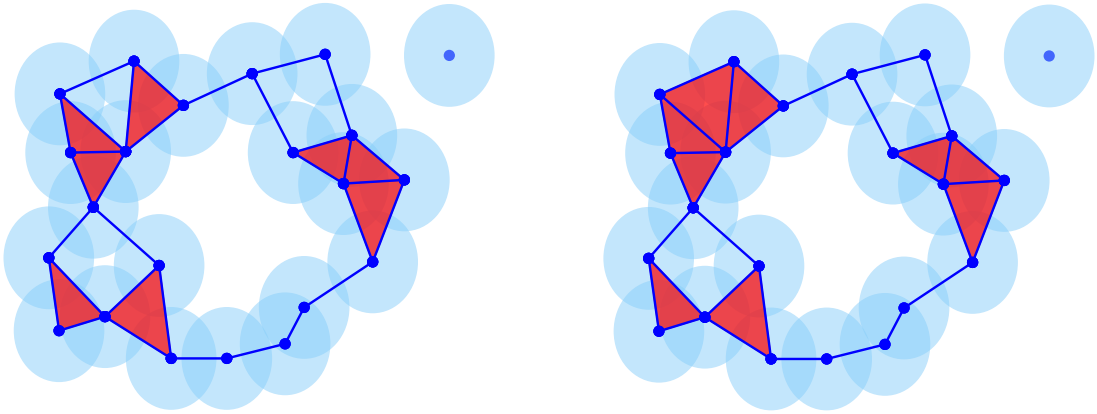
The first step in persistent homology is converting a data set X from a point cloud to a simplicial complex. Recall, a *simplicial complex*, $\mathcal{K}$, is a collection of simplices in $\mathbb{R}$ that satisfy the following conditions: (1) Every face from a simplex in $\mathcal{K}$ is also in $\mathcal{K}$, and (2) For all pairs of simplices $\sigma_1, \sigma_2 \in \mathcal{K}$, $\sigma_1 \cap \sigma_2$ is either empty or else a face of σ_1 and σ_2 [**Hat02**, pg. 102]. The set of k-simplices of a given simplicial complex $\mathcal{K}$ is referred to as the *k-skeleton* of $\mathcal{K}$. Intuitively, a simplicial complex is a topological space formed by gluing simplices (points, edges, triangles, etc.) along their faces. A natural method for constructing a simplicial complex from data is the *Čech complex*.

Definition 1.1.1 ([**Ghr08**]). Given a finite collection of points $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$ and a constant $\varepsilon > 0$, the *Čech complex* is the abstract simplicial complex $C(X, \varepsilon)$ whose k-simplices are the unordered $(k+1)$-tuples of indices $J \subset \{1, \ldots, N\}$ such that $\bigcap_{j \in J} \bar{B}(x_j, \varepsilon) \neq \varnothing$, where $\bar{B}(x_j, \varepsilon)$ denotes the closed ball with center x_j and radius ε.

The *Čech theorem*, also called the *nerve theorem*, states that $C(X, \varepsilon)$ is homotopy equivalent to the union, $\bigcup_{i=1}^n \bar{B}(x_i, \varepsilon)$ [**Bj6**]. Hence, $C(X, \varepsilon)$ is a faithful representation for the topology of X after "thickening" X with balls of radius ε. One issue with the Čech complex is that it is

computationally expensive in the sense that it potentially requires the storage of simplices of all dimensions, from $k = 0$ to $N - 1$. Therefore, in practice, many researchers instead represent their data using the *Vietoris-Rips complex*.

DEFINITION 1.1.2 ([**Ghr08**]). Given a finite collection of points $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^n$ and a constant $\varepsilon > 0$, the *Vietoris-Rips* complex (sometimes called simply the *Rips complex*) is the abstract simplicial complex $R(X, \varepsilon)$ whose k-simplices are the unordered $(k + 1)$-tuples of indices $J \subset \{1, \ldots, N\}$ such that for all $i, j \in J$ we have $\|x_i - x_j\|_2 \leqslant \varepsilon$, where $\| \cdot \|_2$ denotes the ℓ^2 norm.



(a) Čech Complex. (b) Rips Complex.

FIGURE 1.1. For a fixed set of points $X \subset \mathbb{R}^2$ and some $\varepsilon > 0$, we draw circles of radius ε centered at each point and superimpose the resulting Čech complex $C(X, \varepsilon)$ (left) and Rips complex $R(X, 2\varepsilon)$ (right). Note the two complexes differ in the top-left where $C(X, \varepsilon)$ has an open cycle which is a filled 2-simplex in $R(X, 2\varepsilon)$.

Rips complexes belongs to a special class of simplicial complexes known as *flag complexes* or *clique complexes*, which are maximal simplicial complexes (ordered by inclusion) over all simplicial complexes with a given 1-skeleton [**Ghr08**]. Thus, $R(X, \varepsilon)$ is completely determined by its 1-skeleton and, hence, can be stored as a graph and reconstituted as needed, unlike the Čech complex. While Rips complexes are generally not homotopy equivalent to the thickened dataset $\bigcup_{i=1}^n \bar{B}(x_i, \varepsilon)$, Rips complexes are "close" to Čech complexes in that we have the following inclusions:

$$C(X, \varepsilon/2) \subseteq R(X, \varepsilon) \subseteq C(X, \varepsilon), \tag{1.1}$$

noting that the parameter ε refers to radii for Čech complexes and to distances for Rips complexes.

Rips complexes allow us to efficiently transform our data from a point cloud with an effectively trivial topology to a much richer simplicial complex. We can study the topology of this complex using *homology*. We tersely recall the key terms of simplicial homology, below; for a complete introduction to homology see [**Hat02**, Chapter 2].

Let $\mathcal{K}$ be a simplicial complex and let Σ_k denote the set of k-simplices of $\mathcal{K}$. The group of *k-chains* in $\mathcal{K}$ is the free Abelian group on the set Σ_k, denoted $\Delta_k(\mathcal{K})$. By imposing a total order (labeling) on the set of vertices, Σ_0, we can define *boundary maps*, $\delta_k : \Delta_k(\mathcal{K}) \to \Delta_{k-1}(\mathcal{K})$, such that for each k-simplex $\sigma_\alpha = [v_0, v_1, \ldots, v_k]$ we have,

$$\delta_k(\sigma_\alpha) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v_i}, \ldots, v_k], \tag{1.2}$$

where $\hat{v_i}$ indicates that this vertex is deleted from the sequence. Observe that $\delta_k \circ \delta_{k+1} \equiv 0$, hence $\mathrm{Im}(\delta_{k+1}) \subseteq \mathrm{Ker}(\delta_k)$. Thus, we may define the *k-th (simplicial) homology group* of $\mathcal{K}$ by

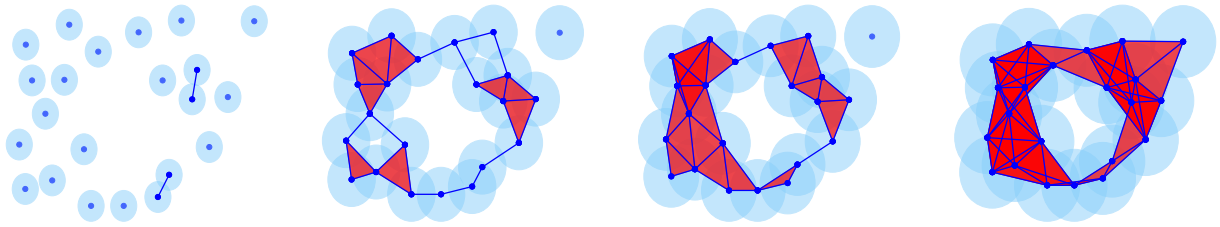
$$H_k(\mathcal{K}, \mathbb{Z}) = \mathrm{Ker}(\delta_k) / \mathrm{Im}(\delta_{k+1}). \tag{1.3}$$

Elements of $\mathrm{Ker}(\delta_k)$ are called *cycles* and elements of $\mathrm{Im}(\delta_{k+1})$ are *boundaries*. The collection of chains and boundary maps, $(\Delta_\bullet, \delta_\bullet)$, is known as a *chain complex*.

Instead of free Abelian groups, the chains $\Delta_*(\mathcal{K})$ can also be defined as free R-modules for an arbitrary ring R, denoted $\Delta_*(\mathcal{K}, R)$ (throughout this thesis, we assume a ring R to be commutative with unity). We denote the resulting homology groups by $H_*(\mathcal{K}, R)$. In practice, we often take R to be a field for computational ease, see [**EH10**, Chapter 4]. For any field $\mathbb{F}$, $H_*(\mathcal{K}, \mathbb{F})$ will be a vector space over $\mathbb{F}$. If $H_k(\mathcal{K}, \mathbb{F})$ is finite dimensional, we let $\beta_k(\mathcal{K}, \mathbb{F}) = \dim(H_k(\mathcal{K}, \mathbb{F}))$, and call $\beta_k(\mathcal{K}, \mathbb{F})$ the *k-th Betti number* of $\mathcal{K}$ with coefficients in $\mathbb{F}$. Informally, the k-th Betti number of $\mathcal{K}$ tells us the number of k-dimensional "holes" in $\mathcal{K}$, where 0-th dimensional holes correspond to connected components. We sometimes omit writing the ring R when discussing chains, homology groups, or Betti numbers if doing so does not create confusion.

Thus, we can describe the topology of the simplicial complex $R(X, \varepsilon)$ in terms of its homology groups, $H_k(R(X, \varepsilon))$ and Betti numbers $\beta_k(R(X, \varepsilon))$. However, the homology of $R(X, \varepsilon)$ depends on the choice of parameter ε. For small values of ε, $R(X, \varepsilon)$ is a disjoint collection of N points.

Meanwhile, when ε is large $R(X,\varepsilon)$ is simply the complete $(N-1)$-simplex. In Figure 1.2 we see how the Betti numbers of $R(X,\varepsilon)$ change for different values of ε between these extremes. This begs the question, what is the correct choice of ε?



(a) $\beta_0 = 17, \beta_1 = 0$. (b) $\beta_0 = 2, \beta_1 = 3$. (c) $\beta_0 = 2, \beta_1 = 1$. (d) $\beta_0 = 1, \beta_1 = 1$.

FIGURE 1.2. We show the evolution of Betti 0 and Betti 1 of $R(X,\varepsilon)$ as ε increases from left to right. Note $\beta_k = 0$ for all $k \geqslant 2$.

The brilliance of persistence homology, as introduced by Edelsbrunner, Letscher, and Zomorodian [**ELZ02**] and refined by Zomorodian and Carlsson [**ZC05**], is that one assumes no perfect choice of ε exists and, instead, analyzes $R(X,\varepsilon)$ at *every* choice of ε. Concretely, let $\mathcal{R} = (R^i)_{i=1}^m$ be a sequence of Rips complexes associated to a fixed dataset X and a strictly increasing sequence of parameters $\varepsilon_1 < \varepsilon_2 < \cdots < \varepsilon_m$ (as in Figure 1.2, for example). Hence, we have natural inclusion maps,

$$R^1 \xhookrightarrow{\iota} R^2 \xhookrightarrow{\iota} \ldots \xhookrightarrow{\iota} R^{m-1} \xhookrightarrow{\iota} R^m. \tag{1.4}$$

Chaining these inclusion maps induces homomorphisms between homology groups, $\iota_* : H_*(R^i) \to H_*(R^j)$, for each pair $i < j$. These homomorphism reveal which topological features *persist* from R^i to R^j; they are exactly those features with non-zero image under the chained inclusion maps ι_* from $H_*(R^i)$ to $H_*(R^j)$. Morally, we say a feature is *significant* if it persists for a "long" range of ε's and say it *noise* if it only persists for a "short" range.

REMARK 1. We note that one can evaluate every choice of ε with only a finite sequence of Rips complexes, as above. Observe that as ε increases new simplices are only added to $R(X,\varepsilon)$ if $\varepsilon = \|x_i - x_j\|$ for some $x_i, x_j \in X$. As X is finite, this leaves only finitely many "critical points" at which $R(X,\varepsilon)$ evolves. Therefore, ordering these pairwise distances gives a sequence of ε's whose corresponding Rips complexes include a copy of $R(X,\varepsilon)$ for every choice of ε.

With this motivation in mind, we now present topological persistence in greater detail and generality. Begin with a *filtration*, which is a sequence of nested simplicial complexes $(\mathcal{K}^i)_{i=0}^m$ such that

$$\varnothing = \mathcal{K}^0 \subseteq \mathcal{K}^1 \subseteq \cdots \subseteq \mathcal{K}^{m-1} \subseteq \mathcal{K}^m. \tag{1.5}$$

From a filtration, we can form a *persistence complex*.

DEFINITION 1.1.3 ([**ZC05**]). A persistence complex is a collection of chain complexes $\{(\Delta_\bullet^i, \delta_\bullet^i)\}_{i\geqslant 0}$ over a ring R together with chain maps, $f^i : (\Delta_\bullet^i, \delta_\bullet^i) \to (\Delta_\bullet^{i+1}, \delta_\bullet^{i+1})$.

The persistence complex associated to a filtration is formed by taking the collection of chain complexes for each $\mathcal{K}^i$ and the chain maps induced by the inclusions $\iota^i : \mathcal{K}^i \to \mathcal{K}^{i+1}$. Below, we illustrate a persistence complex where the filtration index increases from left to right via the maps f^i and each chain complex is descending from top to bottom via the boundary maps δ_*^i.

$$\begin{array}{ccccccc}
\vdots & & \vdots & & \vdots & & \\
\downarrow{\scriptstyle \delta_3^0} & & \downarrow{\scriptstyle \delta_3^1} & & \downarrow{\scriptstyle \delta_3^2} & & \\
\Delta_2^0 & \xrightarrow{f^0} & \Delta_2^1 & \xrightarrow{f^1} & \Delta_2^2 & \xrightarrow{f^2} & \cdots \\
\downarrow{\scriptstyle \delta_2^0} & & \downarrow{\scriptstyle \delta_2^1} & & \downarrow{\scriptstyle \delta_2^2} & & \\
\Delta_1^0 & \xrightarrow{f^0} & \Delta_1^1 & \xrightarrow{f^1} & \Delta_1^2 & \xrightarrow{f^2} & \cdots \\
\downarrow{\scriptstyle \delta_1^0} & & \downarrow{\scriptstyle \delta_2^1} & & \downarrow{\scriptstyle \delta_2^2} & & \\
\Delta_0^0 & \xrightarrow{f^0} & \Delta_0^1 & \xrightarrow{f^1} & \Delta_0^2 & \xrightarrow{f^2} & \cdots
\end{array} \tag{1.6}$$

The maps f^i induce homomorphisms $f_*^i : H_*(\mathcal{K}^i) \to H_*(\mathcal{K}^{i+1})$. For $0 \leqslant i < j$ we define $f_*^{i\to j} : H_*(\mathcal{K}^i) \to H_*(\mathcal{K}^j)$ by,

$$f_*^{i\to j} = f_*^{j-1} \circ f_*^{j-2} \circ \cdots \circ f_*^i. \tag{1.7}$$

DEFINITION 1.1.4 ([**Ghr08**]). Let $\mathcal{C}$ be a persistence complex with chain complexes $\{(\Delta_\bullet^i, \delta_\bullet^i)\}_{i\geqslant 0}$ over a ring R and chain maps f^i. For $0 \leqslant i < j$, the (i, j)-*persistent k-th homology (group)* of $\mathcal{C}$, denoted $H_k^{i\to j}(\mathcal{C})$, is given by

$$H_k^{i\to j}(\mathcal{C}) = \mathrm{Im}(f_k^{i\to j}). \tag{1.8}$$

As an example, consider a filtration consisting of Rips complexes, $\mathcal{R} = (R^i)_{i=1}^m$, where $R^i = R(X, \varepsilon_i)$ as in Equation 1.4. Let γ be a generator of some $H_*(R^i)$ that first appears when $i = i'$. This generator can be mapped to $H_*(R^j)$ via inclusion maps for all $i < j$. Let j' be the greatest j for which $f_*^{i\to j}(\gamma) \neq 0$. Then we say γ persists from i' to j' (equivalently from $\varepsilon_{i'}$ to $\varepsilon_{j'}$) and thus γ is a generator of $H_*^{i\to j}(\mathcal{R})$. In Figure 1.3, below, we display the persistent homology groups for the Rips filtration of a sample dataset in a diagram known as a *persistence barcode*. Each generator of the homology groups corresponds to an interval whose endpoints are the range of epsilon values for which that generator persists. The intervals are stacked vertically in an arbitrary order and colored according to which homology group they correspond to. For simplicity, we do not draw the "essential" 0-th homology generator that persists from zero to infinity. Note that $\beta_k(R^i)$ is given by the number of order k intervals that contain ε_i (plus one in the case of β_0 to account for the missing essential generator). The reason for this is explained in the following sections.

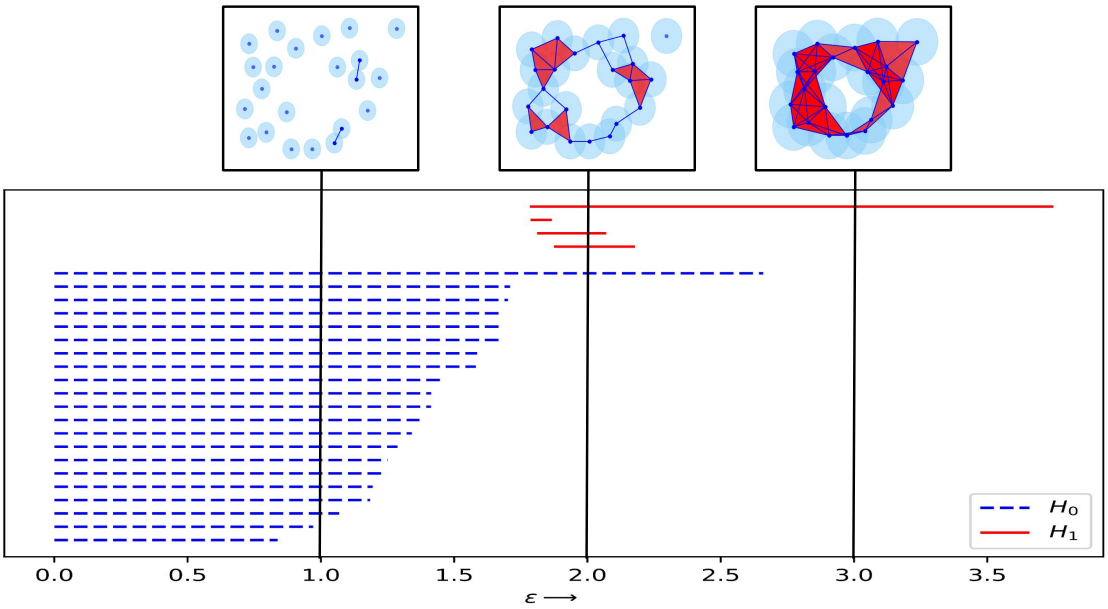


FIGURE 1.3. Persistence barcode of the Rips filtration for the data in Figures 1.1 and 1.2. Above the barcode, we include a representation of the Rips complex at a few values of ε.

1.1.2. Persistence Modules. In this thesis we study the combinatorial structure of the persistent homology groups of filtrations. To do so, we must first develop a classification of these

persistent homology groups. We do this via a classification of a more general class of objects known as *persistence modules*.

DEFINITION 1.1.5 ([**ZC05**]). A *persistence module* $\mathcal{M}$ is a collection of R-modules M^i together with homomorphisms $\varphi^i : M^i \to M^{i+1}$.

DEFINITION 1.1.6. [**ZC05**] A persistence complex $\mathcal{C} = \{(\Delta^i_\bullet, \delta^i_\bullet), f^i\}_{i\geqslant 0}$ (respectively, a persistence module $\mathcal{M} = \{M^i, \varphi^i\}_{i\geqslant 0}$) over a ring R is said to be of *finite type* if each component complex (module) is a finitely generated R-module and there exists some $L \in \mathbb{N}$ such that the maps f^i (φ^i) are isomorphisms for all $i \geqslant L$.

Note that if $\mathcal{K}$ is a finite simplicial complex, then any filtration of $\mathcal{K}$ generates a persistence complex of finite type. Moreover, the homology of this complex forms a persistence module which is also of finite type. Therefore, we would like to classify all persistence modules of finite type over a ring R, as such a classification would allow to classify the persistent homology groups of all filtrations of finite simplicial complexes. Unfortunately, this is an onerous task if we do not make additional assumptions on the ring R.

To see this, consider a persistence module $\mathcal{M} = \{M^i, \varphi^i\}_{i\geqslant 0}$ over a ring R. Equip $R[t]$ with the standard grading and define a graded module over $R[t]$ by,

$$\alpha(\mathcal{M}) = \bigoplus_{i=0}^{\infty} M^i, \tag{1.9}$$

where the R-module structure is the sum of the structures on each component, and where the action by t is given by

$$t \cdot (m^0, m^1, m^2, \dots) = (0, \varphi^0(m^0), \varphi^1(m^1), \varphi^2(m^2), \dots), \tag{1.10}$$

i.e., t shifts elements of the modules up in gradation via the homomorphism φ^i.

THEOREM 1.1.1 ([**ZC05**]). *The correspondence α from Equation 1.9 defines an equivalence of categories between the category of persistence modules of finite type over R and the category of finitely generated non-negatively graded modules over $R[t]$.*

It is well known in commutative algebra that the classification of modules over $\mathbb{Z}[t]$ is extremely complicated. Hence, there is little hope for a classification of persistence modules over an arbitrary ring R. However, if the ground ring is a field $\mathbb{F}$, then the graded ring $\mathbb{F}[t]$ forms a principal ideal domain (PID) and its only graded ideals are of the form (t^n). Thus, by the structure theorem for graded modules over PID's (see [**DF04**, Chapter 12]) we have the following classification theorem, due to Zomorodian and Carlsson.

THEOREM 1.1.2 ([**ZC05**]). *Let $\mathcal{M} = \{M^i, \varphi^i\}_{i\geqslant 0}$ be a persistence module of finite type over a field $\mathbb{F}$ and let α denote the correspondence from Equation 1.9. Then,*

$$\alpha(\mathcal{M}) \cong \left(\bigoplus_{i=1}^{n} t^{a_i}\mathbb{F}[t]\right) \oplus \left(\bigoplus_{j=1}^{m} t^{b_j}\mathbb{F}[t]/(t^{\ell_j})\right), \tag{1.11}$$

for some integers $\{a_1, \ldots, a_n\}, \{b_1, \ldots, b_m\}, \{\ell_1, \ldots, \ell_m\}$.

1.1.3. Interval Decomposition of Persistence Modules. When $\mathcal{M}$ is the persistent homology of a filtration $(\mathcal{K}^i)_{i=1}^m$, Theorem 1.1.2 has a natural interpretation. The free portions of Equation 1.11 are in bijective correspondence with the *essential* homology generators, i.e., those generators that first appear in some $H_*(\mathcal{K}^{a_i})$ and persist in $H_*(\mathcal{K}^r)$ for all $r \geqslant a_i$. Meanwhile, the torsion elements are in bijective correspondence with those homology generators which first appear in $H_*(\mathcal{K}^{b_i})$ but then disappear in $H_*(\mathcal{K}^{b_i+\ell_i})$. Hence, $\mathcal{M}$ is characterized by the following collection of intervals: $\{[a_i, \infty]\}_{i=1}^n \cup \{[b_i, b_j + \ell_j]_{j=1}^m\}$. With this observation, we are finally ready to introduce the main topic of this thesis: barcodes.

DEFINITION 1.1.7 ([**Ghr08**]). A *barcode* is a finite multiset of closed intervals on the real line, $B = \{[b_i, d_i]^{m_i}\}_{i=1}^n$ (here the superscripts m_i indicate multiplicities), where necessarily $b_i < d_i$ for all $i \in [n]$. Each interval is called a *bar*; its left endpoint b_i is called its *birth (time)* and its right endpoint d_i is called its *death (time)*. We denote the set of all barcodes with n distinct bars by $\mathcal{B}^n$.

Now, let $(V^\bullet, f^\bullet)$ be a persistence module of finite type over a field $\mathbb{F}$, so each module V^i is a finite-dimensional vector space over $\mathbb{F}$, the homomorphisms f^i are $\mathbb{F}$-linear maps, and there exists some $L \in \mathbb{N}$ such that f^i is the identity map for all $i \geqslant L$. Hence, we have

$$V^0 \xrightarrow{f^1} V^1 \xrightarrow{f^2} \ldots \xrightarrow{f^{L-1}} V^{L-1} \xrightarrow{f^L} V^L.$$

The *length* of $(V^\bullet, f^\bullet)$ is the number $L+1$.

DEFINITION 1.1.8 ([**ZC05**]). Let $L \in \mathbb{N}$ and let $0 \leqslant i \leqslant j \leqslant L$. The *interval module* over $\mathbb{F}$ of length $L+1$ corresponding to i, j is the persistence module of finite type, $\mathbf{I}[i,j]_\bullet$, given by,

$$0 \longrightarrow \ldots \longrightarrow 0 \longrightarrow \mathbb{F} \longrightarrow \ldots \longrightarrow \mathbb{F} \longrightarrow 0 \longrightarrow \ldots \longrightarrow 0,$$

where $V^i = \mathbb{F}$ for all $i \in [i,j]$, maps between adjacent $\mathbb{F}$'s are identities, and all other vector spaces are trivial.

From Theorem 1.1.2 we have that if $(V^\bullet, f^\bullet)$ is a persistence module of finite type over a field $\mathbb{F}$ and of length $L+1$, then there exists a barcode $B = \{[b_i, d_i]^{m_i}\}_{i=1}^n$ with $b_i, d_i \in \{0, \ldots, L\}$ for all $i \in [n]$ such that $(V^\bullet, f^\bullet)$ is isomorphic to the following direct sum of interval modules:

$$(V^\bullet, f^\bullet) \simeq \bigoplus_{i=1}^{n} \mathbf{I}[b_i, d_i]_\bullet^{m_i}, \tag{1.12}$$

where $\mathbf{I}[b_i, d_i]_\bullet^{m_i}$ denotes the direct sum of m_i copies of the interval module $\mathbf{I}[b_i, d_i]_\bullet$.

Now, let $(V^\bullet, f^\bullet)$ be a persistence module of finite type and of length $L+1$. A *basis family* of $(V^\bullet, f^\bullet)$ is a collection,

$$\mathcal{U} = \{U_i \subset V^i \,:\, 0 \leqslant i \leqslant L\},$$

where U_i is an ordered basis of V^i for each i. Given a fixed basis family, each map f^i can be written as a $\dim(V^{i-1}) \times \dim(V^i)$ matrix, A^i, with entries in $\mathbb{F}$. In order to determine the interval decomposition of $(V^\bullet, f^\bullet)$, we seek basis families which produce matrices of a particularly nice form, known as *barcode form*.

DEFINITION 1.1.9 ([**JNT22**]). An $m \times n$ matrix A of rank r is said to be in *barcode form* if there exists a strictly increasing function $c : [r] \to [n]$ such that

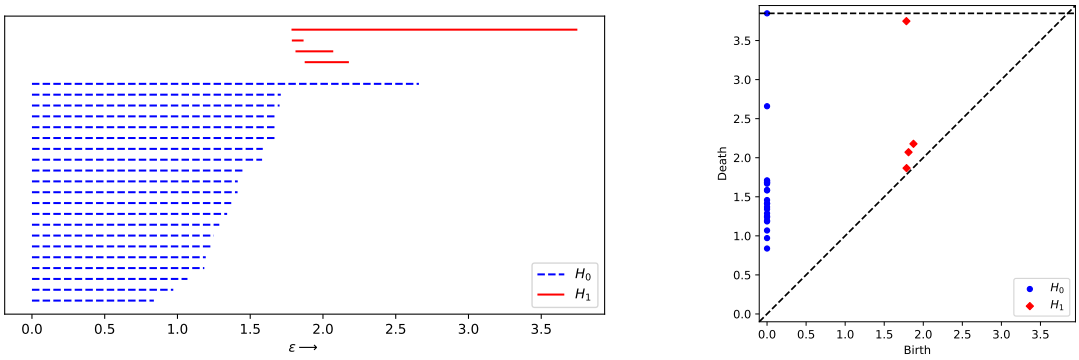
$$A_{ij} = \begin{cases} 1, & \text{if } j = c(i) \\ 0, & \text{otherwise} \end{cases}.$$

Note, A is in barcode form if and only if it has at most a single 1 in each row and column, which appear in the first r rows in strictly increasing column order, and the remaining entries are all 0.

We say that a basis family $\mathcal{U}$ is a *barcode basis* for $(V^\bullet, f^\bullet)$ if the matrix A_i is in barcode form for all $i \in [L]$. Note that given a barcode basis $\mathcal{U}$, one can recover the decomposition from Equation 1.12 and vice-versa; the positions of the 1's in each A_i indicate which basis vectors have non-zero image under f^i. We denote the set of all ordered barcodes bases of $(V^\bullet, f^\bullet)$ by $\mathscr{B}(V^\bullet, f^\bullet)$.

1.1.4. Barcodes. In this section we introduce some additional definitions related to barcodes and introduce some popular metrics used to define a topology on the space of barcodes.

Barcodes are often displayed as a stacked set of intervals above the real line as in Figure 1.3 and Figure 1.4a, below. In these diagrams the heights of the bars are arbitrary, though often bars corresponding to persistent homology groups of the same dimension are grouped together and distinguished from other groups using colors or line-styles. Barcodes are also commonly represented as points $(b_i, d_i) \in \mathbb{R}^2$ in a figure known as a *persistence diagram*. Figure 1.4b shows the persistence diagram for the barcode in Figure 1.4a. Note that the points in Figure 1.4b lie above the diagonal since we require that $b_i < d_i$ for all $i \in [n]$.



(a) Persistence barcode.

(b) Persistence diagram.

FIGURE 1.4. We compute the Rips filtration of the sample data set from the previous figures then display the order 0 and order 1 persistent homology groups as both a barcode (1.4a) and a persistence diagram (1.4b).

DEFINITION 1.1.10. A barcode $B = \{[b_i, d_i]^{m_i}\}_{i=1}^n$ is called *strict* if $m_i = 1$ for all $i \in [n]$ and $d_i \neq d_j$ for all $i \neq j$, i.e., if no bars are repeated and no pair of distinct bars share a death time. We denote the set of strict barcodes with n bars by $\mathcal{B}_{st}^n$. When we refer to a strict barcode, we often simply write $B = \{[b_i, d_i]\}_{i=1}^n$ with the multiplicities m_i omitted.

DEFINITION 1.1.11. We say a persistence module $(V^\bullet, f^\bullet)$ is *strict* if the barcode B associated to its interval decomposition is strict.

Similar definitions of strict barcodes were introduced in [**KGH20**, **CDG$^+$21**]. However, these definitions differs slightly from ours in the following aspects: (1) we do not require that there exists an essential bar $[b_0, d_0]$ that contains all the others, and (2) we only require that no death times are repeated, whereas the definition in [**CDG$^+$21**] requires that no birth times are repeated either, i.e., $b_i \neq b_j$ for all $i \neq j$. We note that our definition is well suited for barcodes arising from the persistent homology groups of Rips/Čech complexes where many generators of the 0-th homology groups may born at time 0, but it is unlikely that a pair of generators will die at the same time (assuming the data is drawn from some continuous distribution).

One can define a topology on the space of all barcodes, regardless of strictness or the number of bars, by way of a distance function. Two popular and well-studied choices are the *bottleneck distance* and the *q-Wasserstein distance*.

DEFINITION 1.1.12 ([**EH08**]). Let $B = \{[b_i, d_i]^{m_i}\}_{i=1}^n$ and $B' = \{[b'_i, d'_i]^{m'_i}\}_{i=1}^m$ be two barcodes. The *bottleneck distance* between B and B' is

$$d_\infty(B, B') = \inf_\gamma \max_{x \in B} \|x - \gamma(x)\|_\infty,$$

where γ runs over all perfect matchings of the points (bars) $x_i = [b_i, d_i]$ in B and the points (bars) in B', allowing bars to be matched to the diagonal $\Delta = \{(x, x) : x \in \mathbb{R}\}$. Here $\|\cdot\|_\infty$ denotes the ℓ^∞-norm on $\mathbb{R}^2$.

Put simply, the bottleneck distance computes the *maximum* distance in ℓ^∞-norm between a pair of matched points on the persistence diagrams of B and B', taking the infimum over all perfect matchings. The q-Wasserstein distance is similar; it can be thought of as the *total* distance between all pairs of matched points, again taking the infimum over all perfect matchings.

DEFINITION 1.1.13 ([**EH08**]). Let $B = \{[b_i, d_i]^{m_i}\}_{i=1}^n$ and $B' = \{[b'_i, d'_i]^{m'_i}\}_{i=1}^m$ be two barcodes. The *q-Wasserstein distance* between B and B' is

$$d_q(B, B') = \inf_\gamma \Big(\sum_{x \in B} \|x - \gamma(x)\|_\infty^q \Big)^{1/q},$$

where γ runs over all perfect matchings of the points (bars) $x_i = [b_i, d_i]$ in B and the points (bars) in B', allowing bars to be matched to the diagonal $\Delta = \{(x, x) : x \in \mathbb{R}\}$. Here $\|\cdot\|_\infty$ denotes the ℓ^∞-norm on $\mathbb{R}^2$.

1.1.5. Barcode Permutations. We note that the space of barcodes does not form a Hilbert space when equipped with either the bottleneck or Wasserstein metrics [**TS20**]. Moreover, the space of barcodes does not even admit a coarse embedding into a Hilbert space when equipped with the bottleneck distance [**BW20**]. Therefore, it is difficult to apply certain statistical methods, such as standard kernel methods, to barcodes; we do note that some progress has been made developing these methods for Banach spaces without an inner-product [**DMW22**]. *In our work, we address this issue by defining new combinatorial invariants for barcodes which record the overlapping arrangements of the bars and can be used to bound the bottleneck and Wasserstein distances between barcodes.*

This work is inspired by a series of papers [**KGH20**, **CDG$^+$21**, **BG22**] which developed similar invariants associated to barcodes. In [**KGH20**, **CDG$^+$21**] the authors introduced a mapping from a special class of strict barcodes with n bars to the symmetric group $\mathfrak{S}_n$, defined as follows. Let $B \in \mathcal{B}^n_{st}$ and assume that the bars in B have distinct endpoints, i.e., $\{b_i, d_i\} \cap \{b_j, d_j\} = \varnothing$ for all $i \neq j$. Begin by ordering the death times increasingly so that $d_{i_1} < d_{i_2} < \cdots < d_{i_n}$. Then the indexing set $[n]$ gives a permutation $\gamma_B \in \mathfrak{S}_n$ defined by $\gamma_B(k) = i_k$, i.e., γ_B is the unique permutation such that $d_{\gamma_B(1)} < d_{\gamma_B(2)} < \ldots < d_{\gamma_B(n)}$. In the same manner, ordering the birth times gives another permutation τ_B. Finally, we set define the *permutation type* of B, π_B, to be the permutation $\pi_B = \tau_B^{-1} \cdot \gamma_B$, which tracks the ordering of the death values with respect to the birth values.

For example, if B is the strict barcode with 3 bars given by $b_2 = 1.0$, $d_2 = 2.0$, $b_1 = 1.5$, $d_1 = 3.0$, $b_3 = 2.5$, $d_3 = 2.75$, then the birth/death times in B_1 are ordered: $b_2 < b_1 < d_2 < b_3 < d_3 < d_1$. So $\tau_B = (2\ 1\ 3)$, $\gamma_B = (2\ 3\ 1)$ and $\pi_B = (1\ 3\ 2)$.

One of the main contributions of [**KGH20**, **CDG$^+$21**] was the discovery that these permutations can be used to describe the fibers of a very different map, known as the topological morphology descriptor (TMD). The TMD, first introduced in [**KDS$^+$18**], is a method for defining a filtration

on *metric trees*, i.e., tree graphs with a length associated to each branch, for applications in neuroscience. By computing the persistent homology groups of this filtration, one can encode the spatial structure of a tree in a barcode.

In [**KGH20**, **CDG$^+$21**] the authors demonstrated that the cardinality of the fibers of TMD, when restricted to a certain class of trees known as *merge trees*, can be computed from π_B alone.

THEOREM 1.1.3 (Curry et al. [**CDG$^+$21**]). *Let* TMD *denote the topological morphology descriptor from [**KDS$^+$18**], and let $B = \{[b_i, d_i]\}_{i=0}^n$ be a strict barcode with n bars. If no birth times in B are repeated, then*

$$|\mathrm{TMD}^{-1}(B)| = \prod_{i=1}^{n} l_i(\pi_B), \tag{1.13}$$

where $l(\pi_B)$ denotes the left inversion vector of π_B (see Section 1.1.6).

Theorem 1.1.3 follows from the observation that inversions in π_B correspond to pairs of *nested* bars in B, i.e., bars $[b_i, d_i], [b_j, d_j]$ satisfying $b_i < b_j < d_j < d_i$. These nested bars allow for choices when attaching branches to tree a T, which in turn are responsible for the non-injectivity of TMD (see [**KGH20**, **CDG$^+$21**] for details).

We note that a pair of non-nested bars may be either *disjoint* or *stepped*, i.e., the bars intersect but are not nested, as defined above. This distinction is important for the study of barcode bases of persistence modules, as demonstrated by the following theorem of Jacquard et al.

THEOREM 1.1.4 ([**JNT22**]). *Let $(V^\bullet, f^\bullet)$ be a persistence module of length $\ell + 1$. Let $B = \{[b_i, d_i]^{m_i}\}_{i=1}^n$ be the barcode associated to the interval decomposition of $(V^\bullet, f^\bullet)$ as in Equation 1.12. Then the set of barcode bases of $(V^\bullet, f^\bullet)$, $\mathscr{B}(V^\bullet, f^\bullet)$, admits a bijection,*

$$\mathscr{B}(V_\bullet, f_\bullet) \cong \prod_{i=1}^{n} \mathrm{GL}(m_i; \mathbb{F}) \times \prod_{\substack{i<j: \\ b_i \leqslant b_j \leqslant d_i \leqslant d_j}} \mathrm{Mat}(m_i \times m_j; \mathbb{F}),$$

where $\mathrm{GL}(m; \mathbb{F})$ denotes the general linear group of $m \times m$ matrices over $\mathbb{F}$.

Note that if the death times in B are distinct, then condition, $b_i \leqslant b_j \leqslant d_i \leqslant d_j$, indicates that a pair of bars are stepped. Although one can identify pairs of *nested* bars in B from their permutation type, π_B, one cannot distinguish non-nested bars as being either *stepped* or *disjoint*

from π_B alone. *Our goal is to develop alternative combinatorial invariants for barcodes which are capable of recording all possible arrangements of bars. We do this by introducing maps from the space of barcodes to multipermutations in the multinomial Newman lattice.* Before summarizing these results, we first recall some necessary concepts regarding posets and permutations.

1.1.6. Partially Ordered Sets and Permutations. We will assume familiarity with the basic definitions regarding partially ordered sets (posets), though we recall several terms below. For a complete introduction to posets see [**Sta11**, Ch.3].

Let $(P, \leqslant)$ be a poset and let $s, t \in P$. An *upper bound* of s and t is an element $u \in P$ such that $s \leqslant u$ and $t \leqslant u$. A *least upper bound* (also called a *join*) of s and t is an upper bound u of s and t such that if v is any upper bound of s and t then $u \leqslant v$. If a least upper bound of s and t exists, then it is necessarily unique and we denote it by $s \vee t$ (read "s join t"). One can define a *lower bound* and the *greatest lower bound* of s and t equivalently. The greatest lower bound, also called a *meet*, is denoted $s \wedge t$ (read "s meet t") when it exists.

A *lattice* is a poset L for which every pair of elements has a least upper bound and greatest lower bound. A *principal ideal* of a lattice L is a subposet of the form $I = \{s \in L : s \leqslant \alpha\}$ for some $\alpha \in L$; we say I is the principal ideal generated by α. It is a well known result that a principal ideal I of a lattice L is a sublattice of L. A *congruence* on a lattice L is an equivalence relation R on L such that if $a \; R \; b$ and $c \; R \; d$ then $(a \wedge c) \; R \; (b \wedge d)$ and $(a \vee c) \; R \; (b \vee d)$ [**Rea16b**].

LEMMA 1.1.1 ([**Rea16b**]). *An equivalence relation R on a finite lattice L is a lattice congruence if and only if it satisfies the the following conditions:*

(1) Each equivalence class is an interval in L.

(2) The map $\pi_{\downarrow}$ which maps each element to the least element of its equivalence class is order-preserving.

(3) The map $\pi_{\uparrow}$ which maps each element to the greatest element of its equivalence class is order-preserving.

Lattice congruences are of interest because they allow us to define the quotient of a lattice in a natural way.

DEFINITION 1.1.14 ([**Rea16b**]). If L is a lattice and R is a congruence on L then the *quotient lattice* L/R is the poset on the R-classes where $C_1 \leqslant C_2$ in L/R if and only if there exists $a \in C_1$ and $b \in C_2$ such that $a \leqslant b$ in L.

One can show that if R is an lattice congruence on a finite lattice L, then L/R is order-isomorphic to the induced subposet of L whose elements are either the minima (or the maxima) of each equivalence class [**Rea16b**].

A lattice of particular interest to us is the *permutohedron* [**CSW16**]. Recall that for $\pi \in \mathfrak{S}_n$ an *inversion* in π is a pair (π_i, π_j) such that $i < j$ and $\pi_i > \pi_j$, i.e., it is a pair of elements that appear out of order. The *inversion set* of π, $\mathrm{inv}(\pi)$, is the set of all inversions in π. The *inversion number* of π is the cardinality $\#\,\mathrm{inv}(\pi)$ of its inversion set. For example, if $\pi = (1\ 2\ 5\ 4\ 3\ 6) \in \mathfrak{S}_6$, written in one-line notation, then $\mathrm{inv}(\pi) = \{(5,4),(5,3),(4,3)\}$ and $\#\,\mathrm{inv}(\pi) = 3$. Notice that $\pi \neq \sigma \implies \mathrm{inv}(\pi) \neq \mathrm{inv}(\sigma)$, so we can think of inv as an injective map from the permutations in $\mathfrak{S}_n$ to subsets of $[n]^2$.

One can also encode the inversions of a permutation $\pi \in \mathfrak{S}_n$ in a vector known as the *inversion vector* of π, denoted $\nu(\pi)$. In particular, $\nu(\pi)$ is defined by $\nu(\pi)_i = \#\{(a,b) \in \mathrm{inv}(\pi) : b = i\}$, i.e., the i-th coordinate of $\nu(\pi)$ is the number of inversions of π in which i is the smaller (right) element. Similarly, the *left inversion vector* of π, denoted $l(\pi)$, is defined $l(\pi)_i = \#\{(a,b) \in \mathrm{inv}(\pi) : a = i\}$, i.e., the i-th coordinate of $\nu(\pi)$ is the number of inversions of π in which i is the larger (left) element. For example, if $\pi = (1\ 2\ 5\ 4\ 3\ 6)$, then $\nu(\pi) = (0,0,2,1,0,0)$ and $l(\pi) = (0,0,0,1,2,0)$.

The *weak Bruhat order* (or *weak order* for short) is the relation $\leqslant_W$ on $\mathfrak{S}_n$ defined by $\pi \leqslant_W \sigma$ if and only if $\mathrm{inv}(\pi) \subseteq \mathrm{inv}(\sigma)$. Note that $\pi \leqslant_W \sigma \implies \#\,\mathrm{inv}(\pi) \leqslant \#\,\mathrm{inv}(\sigma)$, but the converse need not hold. One can show that $\pi \lessdot_W \sigma$ if and only if $\#\,\mathrm{inv}(\pi) + 1 = \#\,\mathrm{inv}(\sigma)$ and $\sigma = (i\ \ i+1)\pi$ for some $i \in [n-1]$, which means that σ equals π after transposing a pair of its *adjacent* entries.

The weak order on $\mathfrak{S}_n$ also forms the face lattice of a well-studied polytope known as the *permutohedron* [**Ber71**, **CSW16**]. Hence, we refer to the poset $(\mathfrak{S}_n, \leqslant_W)$ as the permutohedron as well. Figure 1.5 shows the Cayley graph of the permutohedron, $\mathfrak{S}_4$; the Hasse diagram of $(\mathfrak{S}_4, \leqslant_W)$ can be deduced from this figure since each edge corresponds to a cover relation. The weak Bruhat order can also be defined similarly on arbitrary Coxeter systems (see [**BB05**]), but for this thesis the definition on the symmetric group is sufficient.

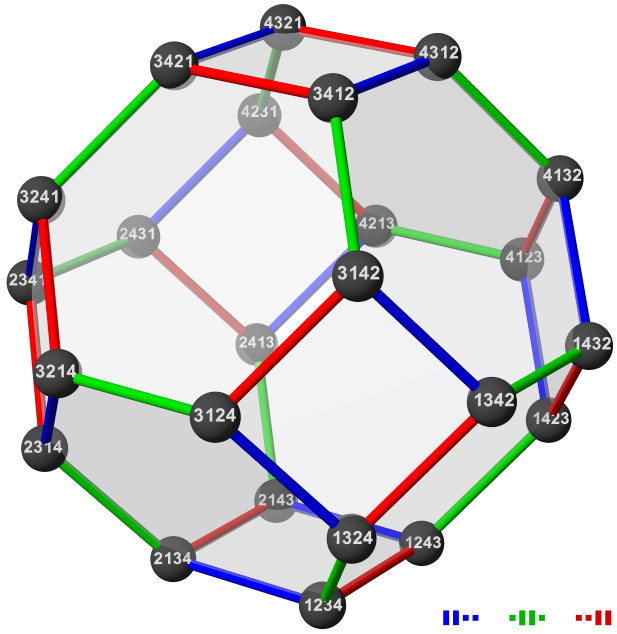


FIGURE 1.5. The permutohedron, $\mathfrak{S}_4$ [**Pie20**]. Edges indicate cover relations. As each cover relation corresponds to swapping a pair of adjacent entries, edges are colored to indicate which entries are swapped.

One can generalize the construction of the permutohedron to multiset permutations. For $\mathbf{m} = (m_1, \ldots, m_n) \in \mathbb{N}^n$, let $L(\mathbf{m})$ denote the set of permutations of the multiset $M = \{1^{m_1}, \ldots, n^{m_n}\}$, here too the exponents m_i indicate multiplicities. The elements of $L(\mathbf{m})$ are called *multipermutations*. An order relation on $L(\mathbf{m})$ can be succinctly defined via the following cover relations. For $s, t \in L(\mathbf{m})$, we say that $s \lessdot t$ if and only if s and t differ only in swapping an adjacent pair of entries, which are in numerical order in s but are reversed in t. For instance, we have that $(1\ 1\ 2\ 3\ 2) \lessdot (1\ 2\ 1\ 3\ 2)$ in $L(2, 2, 1)$. We note that these cover relations are direct analogues of the cover relations in the permutohedron. The poset $(L(\mathbf{m}), \leqslant)$ is called the *multinomial Newman lattice* and was originally introduced by Bennett and Birkhoff in [**BB94**].

The multinomial Newman order can also be defined explicitly as follows. Consider the *set*,

$$S = \{1_1, \ldots, 1_{m_1}, \ldots, n_1, \ldots, n_{m_n}\},$$

which we endow with the lexicographic total ordering $1_1 \lessdot 1_2 \lessdot \cdots \lessdot 1_{m_1} \lessdot \cdots \lessdot n_{m_n}$. We identify a multipermutation $s \in L(\mathbf{m})$ with the unique permutation $\pi \in \mathfrak{S}_S$ which equals s after

removing the labels and where copies of the same elements appear in lexicographic order, that is to say i_j appears before i_k in π for all $i \in [n]$ and all $j < k$. For example, the multipermutation $(1\ 2\ 1\ 3\ 2) \in L(2,2,1)$ is identified with the permutation $(1_1\ 2_1\ 1_2\ 3_1\ 2_2)$. Let $\iota : L(\mathbf{m}) \to \mathfrak{S}_S$ denote this mapping.

One can show that ι is in fact an order-isomorphism from $(L(\mathbf{m}, \leqslant)$ to a principal ideal of the permuhatedron $(\mathfrak{S}_S, \leqslant_W)$; it follows that $(L(\mathbf{m}), \leqslant)$ is also a lattice [**BB94**, **SW16**, **San07**]. Specifically, the multinomial Newman lattice is isomorphic to the principal ideal generated by the permutation $(n_1 \dots n_{m_n} \dots 1_1 \dots 1_{m_1})$. Thus $L(\mathbf{m})$ has (necessarily unique) minimal and maximal elements, denoted $\hat{0}$ and $\hat{1}$ respectively. The minimum, $\hat{0}$, is the identity permutation while the maximum, $\hat{1}$, is the "fully reversed" permutation $(n\ n \dots n\ (n-1)\ (n-1) \dots 1)$.

One can also define $L(\mathbf{m})$ as the set of equivalence classes of $\mathfrak{S}_S$ under the equivalence relation where $\pi_1 \sim \pi_2$ in $\mathfrak{S}_S$ if and only if π_1 and π_2 differ only in permuting the subscripts of each number. For example, $(1_1\ 2_1\ 1_2\ 3_1\ 2_2) \sim (1_2\ 2_2\ 1_1\ 3_1\ 2_1)$ in $\mathfrak{S}_{\{1_1,1_2,2_1,2_2,3_1\}}$. However, we emphasize that $\sim$ is not a lattice congruence on $\mathfrak{S}_S$ except in some trivial cases; note, for example, that the equivalence class of $(1_1\ 2_1\ 1_1) \in \mathfrak{S}_{\{1_1,1_2,2_1\}}$ is $\{(1_1\ 2_1\ 1_1), (1_2\ 2_1\ 1_1)\}$, which does not form an interval in $\mathfrak{S}_{\{1_1,1_2,2_1\}}$. We note that quotient lattices of the permutohedron defined by a lattice congruence are well-studied and include many "famous" posets such as the Tamari and Cambrian lattices [**Rea06**, **Rea12**, **Rea16a**, **HM21**].

In Chapter 2, we will focus on a special multinomial Newman lattice, $L(n,2) = L(2,2,\dots,2)$, whose elements are all permutations of the multiset $\{1^2, 2^2, \dots, n^2\}$. The lattice $L(n,2)$ is closely related to a class of strings known as *double occurrence words*, defined as follows [**Lot02**]. Let $\mathcal{A} = \{a_i\}_{i \in \mathbb{N}}$ be an ordered alphabet, i.e., a countable set of linearly-ordered symbols with a lower bound. A *word* in $\mathcal{A}$ is a sequence of symbols $w = (a_1\ a_2\ \dots\ a_k)$ in $\mathcal{A}$. A *double occurrence word*, sometimes called a *Gauss code*, over $\mathcal{A}$ is a word, w in which every symbol in $\mathcal{A}$ occurs either zero or two times in w. For example, $(1\ 2\ 1\ 3\ 3\ 2)$ and $(4\ 1\ 4\ 3\ 3\ 1)$ are double occurrence words over $\mathbb{N}$ while $(1\ 2\ 4\ 3\ 3\ 2)$ is not. Note that the multipermutations in $L(n,2)$ are the double occurrence words over $[n]$ where each integer occurs exactly twice. Double occurrence words have been studied extensively the context of knot theory [**BJS15**, **Gib11**, **Tur04**], mathematical

logic [**Cou08**], algebraic combinatorics [**STZ09**], enumerative combinatorics [**CFJ+19**, **GKO20**], and genomics [**BNN+18**, **BDJ+13**, **JNS17**].

Often it is beneficial to identify two words w, w' if they are equal after some permutation of the symbols in $\mathcal{A}$. We say w is *combinatorially equivalent* to w', denoted $w \equiv w'$, if there exists a bijection of symbols $f : \mathcal{A} \to \mathcal{A}$ such that $f(w) = w'$, where $f(w)$ indicates that f acts element-wise on w. For example, if $w = (1\ 2\ 1\ 3\ 2\ 3)$ and $w' = (1\ 3\ 1\ 2\ 3\ 2)$ then $w \equiv w'$ because $(2\ 3) \circ (w) = w'$, where $(2\ 3)$ is the permutation map in $\mathfrak{S}_3$, here written in cycle notation. One may verify that $\equiv$ defines an equivalence relation on the set of double occurrence words over $\mathcal{A}$. The equivalence class of w under $\equiv$ is denoted by $[w]$. We emphasize that the equivalence relation $\equiv$ is not a lattice congruence on $L(n, 2)$; note, for example, that the equivalence class of $(1\ 1\ 2\ 2)$ in $L(2, 2)$ is $\{(1\ 1\ 2\ 2), (2\ 2\ 1\ 1)\}$, which does not form an interval in $L(2, 2)$. However, quotients of multinomial lattices given by a lattice congruence have been studied as well [**San07**].

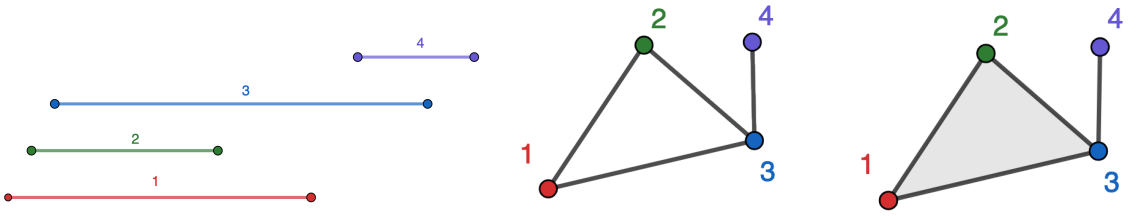
A double occurrence word w is said to be in *ascending order* if $1, 2, \ldots, i-1$ appear before the first instance of i in w, for all $i \in [n]$, i.e., if the first copy of 1 appears before the first copy of 2, which appears before the first copy of 3, and so on. For example, the word $(1\ 2\ 1\ 3\ 2\ 3)$ is in ascending order while the equivalent word $(1\ 3\ 1\ 2\ 3\ 2)$ is not. If $[w]$ is an equivalence class of double occurrence words, we let $\overline{w}$ denote the unique word in $[w]$ which is in ascending order.

1.1.7. Interval Graphs and Random Graphs. In this section we provide some necessary background in graph theory to contextualize our work, particularly Chapter 4. We assume some basic familiarity with graphs as in [**Die17**, Chapter 1], for example, and with elementary probability as in [**Pit93**], for example.

So far, we have discussed how to produce a barcode from a simplicial complex. It is also possible to do the reverse, that is to form a simplicial complex from a barcode, using the notion of a *nerve complex* (see below).

Definition 1.1.15 ([**Mat02**, pg. 197]). Let $\mathcal{F} = \{F_1, \ldots, F_m\}$ be a family of convex sets in $\mathbb{R}^d$. The *nerve complex* $\mathcal{N}(\mathcal{F})$ is the abstract simplicial complex whose k-facets are the $(k+1)$-subsets $I \subset [m]$ such that $\bigcap_{i \in I} F_i \neq \varnothing$.

Figure 1.6c shows the nerve complex constructed from the intervals in Figure 1.6a. Note the presence of a 2-simplex (triangle) with vertices $\{1, 2, 3\}$ because the corresponding intervals mutually intersect. We note that a Čech complex in Definition 1.1.2 is an example of a nerve complex, in particular $C(X, \varepsilon) = \mathcal{N}(\{\bar{B}(x, \varepsilon) : x \in X\})$.



(a) Example barcode, B. (b) Interval graph, $G(B)$. (c) Nerve complex, $\mathcal{N}(B)$.

FIGURE 1.6. A barcode together with its corresponding interval graph and nerve complex.

Given a strict barcode $B = \{[b_i, d_i]\}_{i=1}^n$, the nerve $\mathcal{N}(B)$ records the k-wise intersections between the bars in B for all $k \in [n]$. However, because the intervals are subsets of the real line, the k-wise intersections can actually be inferred from the *pairwise* intersections alone. To see this, suppose we have a collection of intervals $[b_1, d_1], \ldots, [b_k, d_k]$ such that all intervals intersect pairwise, i.e., $[b_i, d_i] \cap [b_j, d_j] \neq \varnothing$ for all $i \neq j$. It follows that $d_i \geqslant b_j$ for all $i, j \in [k]$, and so $[\max\{b_1, \ldots, b_k\}, \min\{d_1, \ldots, d_k\}] \subseteq \bigcap_{i=1}^k [b_i, d_i]$. Hence the whole collection has non-empty intersection. We note that this is a special case of Helly's theorem, which states that if a family of convex sets in $\mathbb{R}^d$ has the property that any $d+1$ of the sets have a non-empty mutual intersection, then the entire collection has a non-empty intersection [**Bar02**].

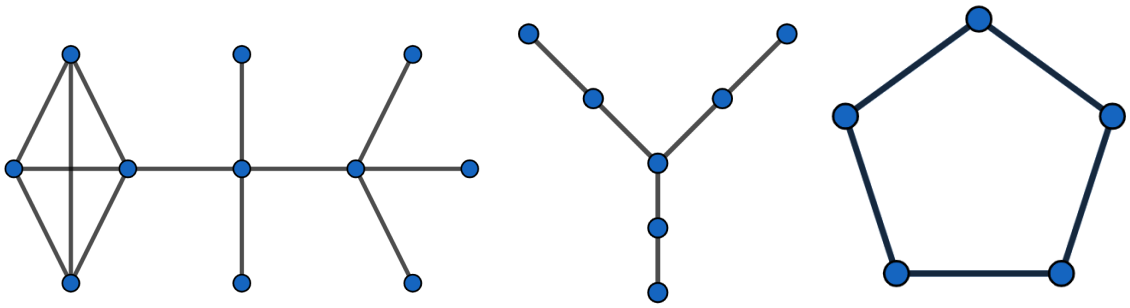
Thus, we see that $\mathcal{N}(B)$ is a *clique complex*, i.e., the k simplices in $\mathcal{N}(B)$ are the $(k+1)$-cliques in its 1-skeleton. Hence, when we refer to $\mathcal{N}(B)$ we may equivalently refer to its 1-skeleton which forms the *interval graph* $G(B)$. Figure 1.6b shows the interval graph for the barcode in Figure 1.6a.

DEFINITION 1.1.16. Given a strict barcode $B = \{[b_i, d_i]\}_{i=1}^n$, the *interval graph* of B is the simple graph $G(B) = G(V, E)$, where $V = [n]$ and $\{i, j\} \in E$ if and only if $[b_i, d_i] \cap [b_j, d_j] \neq \varnothing$.

Interval graphs have been studied extensively due to their wide applicability to topics as diverse as archaeology, genetics, job scheduling, and paleontology [**Fis85**, **Gol04**, **HH07**, **Pip98**]. These graphs have the power to model the overlap of spacial or chronological events and allow for some

inference of structure. Interval graphs also belong to a larger class of graphs known as *intersection graphs*, which are formed analogously from a collection of arbitrary sets [**EGP66**].

There are a number of nice characterizations of interval graphs that have been obtained [**FG65**, **GH64**, **Han82**, **LB62**]. For instance, a graph G is an interval graph if and only if the maximal cliques of G can be linearly ordered in such a way that, for every vertex x of G, the maximal cliques containing x occur consecutively in the list. Another remarkable property of interval graphs is that they are *perfect graphs* and hence the weighted clique and coloring problems are polynomial time solvable [**Gol04**]. Nevertheless, sometimes it is not always obvious whether or not a given graph is an interval graph. For example, of the graphs in Figure 1.7 only 1.7a is an interval graph.



(a) (b) (c)

FIGURE 1.7. Of these three graphs, only 1.7a is an interval graph.

In this thesis we study a new model for generating *random* interval graphs. The most popular model for generating random graphs is the Erdős-Renyi model [**ER59**]. In this model, we construct a graph $G(n,p)$ with vertex set $V = [n]$ by adding edges independently at random, where for all vertices $i \neq j$, $P(ij \in E) = p$ for some constant $p \in (0,1)$. Put simply, for every edge ij in the complete graph K_n, we decide whether to add ij to $G(n,p)$ by flipping a weighted coin independently for each edge.

However, the Erdős-Renyi model is not efficient at generating interval graphs because the probability that $G(n,p)$ is an interval graph rapidly approaches 0 as the number of vertices, n, goes to infinity [**CKM79**]. Therefore, several other authors have proposed their own models for generating random interval graphs. The most famous of these models is due to Scheinerman. Scheinerman's model, introduced in [**Sch88**, **Sch90**] and further analyzed in [**DHJ13**, **JSW90**, **Ili17**], generates a

random interval graphs with n vertices by generating a random barcode with n bars, $B = \{[b_i, d_i]\}$, where each endpoint b_i, d_i is independently chosen from some fixed, continuous probability distribution on the real line. For example, one could generate $2n$-many points, $x_1, \ldots, x_{2n}$ where each x_i is an independent identically distributed (i.i.d.) uniform $[0, 1]$ random variable, then take $B = \{[x_{2k-1}, x_{2k}]\}_{k=1}^{n}$. Note that barcodes generated in this way are almost certainly strict.

1.2. Outline of Our Contributions

1.2.1. The Combinatorial Barcode Lattice. We begin Chapter 2 by introducing a mapping from the space of strict barcodes with n bars, $\mathcal{B}_{st}^n$, to certain double occurrence words. This map is defined as follows. Given a strict barcode $B = \{[b_i, d_i]\}_{i=1}^{n}$, linearly order the birth and death times then record the sequence of labels. This produces a double occurrence word $\sigma_B \in L(n, 2)$. To account for potential redundancy from different initial labelings, consider the equivalence class of σ_B under combinatorial equivalence, which we denote by $[\sigma_B]$. We let $L(n, 2)/\mathfrak{S}_{\mathfrak{n}}$ denote the set of all such equivalence classes and define the map $g : \mathcal{B}_{st}^n \to L(n, 2)/\mathfrak{S}_{\mathfrak{n}}$ so that $g(B) = [\sigma_B]$. The map g provides a new combinatorial invariant on the space of strict barcodes. Our goal in this chapter is to study the combinatorial structure of this invariant and to uncover what connections it has to persistence modules and topological data analysis more broadly.

To that end, recall that an equivalence class $[w] \in L(n, 2)/\mathfrak{S}_{\mathfrak{n}}$ has a unique representative $\overline{w}$ which is in ascending order. Hence, there is a bijection of sets from $L(n, 2)/\mathfrak{S}_{\mathfrak{n}}$ to the set of all words in $L(n, 2)$ that are in ascending order, which we denote by $\overline{L}(n, 2)$. We call the set $\overline{L}(n, 2)$ the space of *combinatorial barcodes* with n bars and the elements $s \in \overline{L}(n, 2)$ *combinatorial barcodes*.

For example, if B is the strict barcode with 3 bars given by $b_1 = 1.5$, $d_1 = 3.0$, $b_2 = 1.0$, $d_2 = 2.0$, $b_3 = 2.5$, $d_3 = 2.75$, then the birth/death times are ordered, $b_2 < b_1 < d_2 < b_3 < d_3 < b_1$, and $\sigma_B = (2\ 1\ 2\ 3\ 3\ 1) \in L(3, 2)$; we note that some care must be taken when the birth/death times are not all distinct, see Chapter 2 for details). If the permutation $(1\ 2) \in \mathfrak{S}_3$, written in cycle notation, acts on σ_B the resulting multipermutation is $(1\ 2\ 1\ 3\ 3\ 2)$. Hence, both permutations belong to the same equivalence class, $g(B) = [\sigma_B]$ and $\overline{\sigma_B} = (1\ 2\ 1\ 3\ 3\ 2)$.

Although combinatorial equivalence is not a lattice congruence on $L(n, 2)$, one can still define a partial order on $\overline{L}(n, 2)$ by simply taking the induced order from $L(n, 2)$. In our first main result,

Theorem 2.2.1, we show that $(\overline{L}(n,2),\leqslant)$ is order-isomorphic to a principal ideal of $L(n,2)$ and, hence, is also a lattice. Figure 1.8, below, depicts a Hasse diagram of $(\overline{L}(3,2),\leqslant)$.

THEOREM 2.2.1. *The combinatorial barcode poset $(\overline{L}(n,2),\leqslant)$ is order-isomorphic to the principal ideal of the multinomial Newman lattice, $L(n,2)$ generated by the "fully nested" permutation: $(1\ 2\ \ldots\ (n-1)\ n\ n\ (n-1)\ \ldots\ 2\ 1)$. Consequently, $(\overline{L}(n,2),\leqslant)$ is a lattice.*

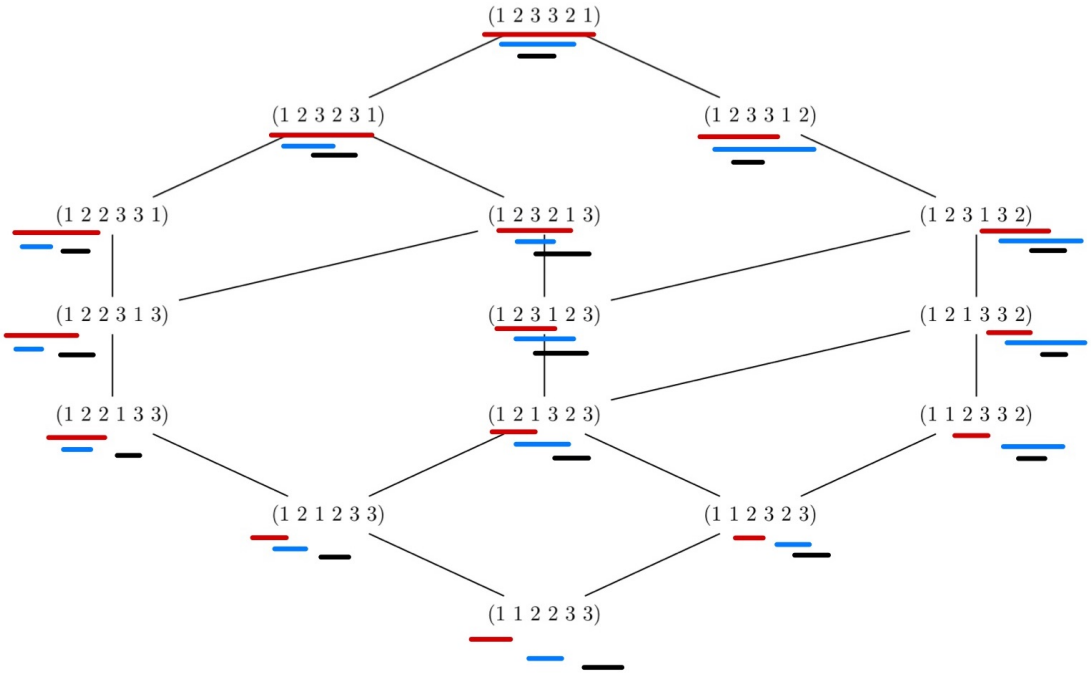


FIGURE 1.8. Hasse diagram of $(\overline{L}(3,2),\leqslant)$. Below each element, $\overline{w}$, a barcode is depicted for which $\overline{\sigma_B} = \overline{w}$, illustrating Theorem 2.3.2.

We then introduce a notion of inversion sets for combinatorial barcodes, which we call *inversion multisets*. The *inversion multiset* of a combinatorial barcode s is the multiset of pairs $\mathrm{invm}(s) = \{(j,i)^{a_{ij}} : 1 \leqslant i < j \leqslant n\}$ where a_{ij} is equal to the number of pairs of indices (k,ℓ) such that $s_k = i, s_\ell = j$ and $k > \ell$.

In Proposition 2.3.1 we show that $s \leqslant t$ in $\overline{L}(n,2)$ if and only if $\mathrm{invm}(s) \subseteq \mathrm{invm}(t)$. Hence, one can define the order $\leqslant$ on $\overline{L}(n,2)$ in terms of the inversion multisets alone, in a manner analogous to the classic permutohedron. Motivated by this, we further define the notion of the crossing number of a combinatorial barcode. For $s \in \overline{L}(n,2)$ and $(j,i) \in [n]^2$ with $j > i$, the *crossing number* of

i and j in s as the multiplicity of $(j,i) \in \text{invm}(s)$, which we denote by $\text{cross}\#(s,j,i)$, or simply $\text{cross}\#(j,i)$ when no confusion may occur. In our next major result, we show that the rank of $s \in \overline{L}(n,2)$ can be computed from the crossing numbers of s.

THEOREM 2.3.2. *If $s \in \overline{L}(n,2)$ is a combinatorial barcode and $\rho(s)$ denotes the rank of s in $\overline{L}(n,2)$, then $\rho(s) = \sum_{i<j} \text{cross}\#(j,i)$.*

From the inversion multisets/ crossing numbers, we also introduce a notion of *inversion vectors* for combinatorial barcodes (see Definition 2.3.2). Our next major result, Theorem 2.3.1, shows that the map which sends a combinatorial barcode to its inversion vector is a bijection of sets. As a corollary, we are able to compute the rank generating function of $\overline{L}(n,2)$.

THEOREM 2.3.1. *Let $T_n = \prod_{i=1}^{n}[0, 2(n-i)]$. Then the map $J : \overline{L}(n,2) \to T_n$ which sends each combinatorial barcode to its inversion vector is a bijection.*

COROLLARY 2.3.1. *For $k \in [0, 2n]$, let $c_k = \#\{s \in \overline{L}(n,2) : \rho(s) = k\}$, i.e., the number of combinatorial barcodes in $\overline{L}(n,2)$ of rank k. Then,*

$$\sum_{k=0}^{2n} c_k q^k = \prod_{i=1}^{n} \left(1 + q + \cdots + q^{2(n-i)}\right).$$

We also show that the number of combinatorial barcodes with k distinct elements in their inversion vectors is given by the second-order Eulerian numbers, which are also known to enumerate trapezoidal words, Stirling permutations, and plane-recursive trees with certain statistics [**Rio76**, **GS78**, **Jan08**].

COROLLARY 2.4.3. *Let $C_{n,k}$ denote the second-order Eulerian numbers defined recursively as follows:*

$$C_{n,k} = kC_{n-1,k} + (2n-k)C_{n-1,k-1}.$$

Then, $C_{n,k}$ is equal to the number of combinatorial barcodes with n bars, i.e., ascending-order double occurrence words over $[n]$, with k distinct elements in its inversion vector.

Finally, we show how the cover relations and crossing numbers in $\overline{L}(n,2)$ relate back to persistence modules and topological data analysis. In Propositions 2.5.1 and 2.5.2 we show that the

cardinality of the fibers of TMB^{-1} (Equation 1.13) and the set of barcode bases of a persistence module (Theorem 1.1.4) can also be expressed in terms the crossing numbers of the barcode B. We also prove the following connection between the cover relations in $(\overline{L}(n,2),\leqslant)$ and the set of barcode bases of strict persistence modules.

THEOREM 2.5.1. *Consider two strict persistence modules $(V_\bullet, f_\bullet), (W_\bullet, h_\bullet)$. Let $B = \{[b_i, d_i]\}_{i=1}^n$, $B' = \{[b'_i, d'_i]\}_{i=1}^n$ be the barcodes associated to the interval decomposition of $(V_\bullet, f_\bullet)$ and $(W_\bullet, h_\bullet)$, respectively. Assume without loss of generality that B are labeled so that $\sigma_B = \overline{\sigma_B}$ (and likewise for B'). Suppose $\overline{\sigma_B} \lessdot \overline{\sigma_{B'}}$, so $\overline{\sigma_B}$ and $\overline{\sigma_{B'}}$ differ only in swapping an adjacent pair of entries $i < j$, which are in ascending order in $\overline{\sigma_B}$ but inverted in $\overline{\sigma_{B'}}$. Then,*

$$\mathscr{B}(V_\bullet, f_\bullet) \cong \mathscr{B}(W_\bullet, h_\bullet) \times \mathbb{F}, \text{ if } \operatorname{cross}\#(\overline{\sigma_B}, j, i) = 0$$

$$\mathscr{B}(W_\bullet, h_\bullet) \cong \mathscr{B}(V_\bullet, f_\bullet) \times \mathbb{F}, \text{ otherwise.}$$

1.2.2. The Power-k Barcode Lattices. In Chapter 3, we generalize the construction of the combinatorial barcode lattice by considering a family of maps g_k from the space of strict barcodes with n to equivalence classes of the multinomial Newman lattice $L(n, 2^k + 1)$, whose elements are the multipermutations of $\{1^{2^k+1}, 2^{2^k+1}, \ldots, n^{2^k+1}\}$, where $k \in \mathbb{Z}_{\geqslant 0}$. These maps are defined as follows. Given a strict barcode B and a non-negative integer k, begin by bisecting each bar k-many times so as to produce 2^k sub-intervals. Ordering the endpoints of these sub-intervals produce a multipermutation $f_k(B) \in L(n, 2^k + 1)$, see Figure 1.9, below.

As before, consider the classes of equivalent words to account for redundancy in the initial labelings. We denote the set of all such classes by $L(n, 2^k + 1)/\mathfrak{S}_n$ and let $\overline{L}(n, 2^k + 1)$ denote the set of ascending order representatives of each class. We let $\overline{\sigma}_k(B)$ denote the ascending order representative of the equivalence class of $f_k(B)$.

We note that some care must be taken to ensure that bisecting two distinct intervals does not produce a repeated endpoint, for example, the bars $[-2, 2]$ and $[-1, 1]$ share no endpoints but share a midpoint at zero. We handle such cases by splitting the bars into not quite equal halves, as discussed in Chapter 3.

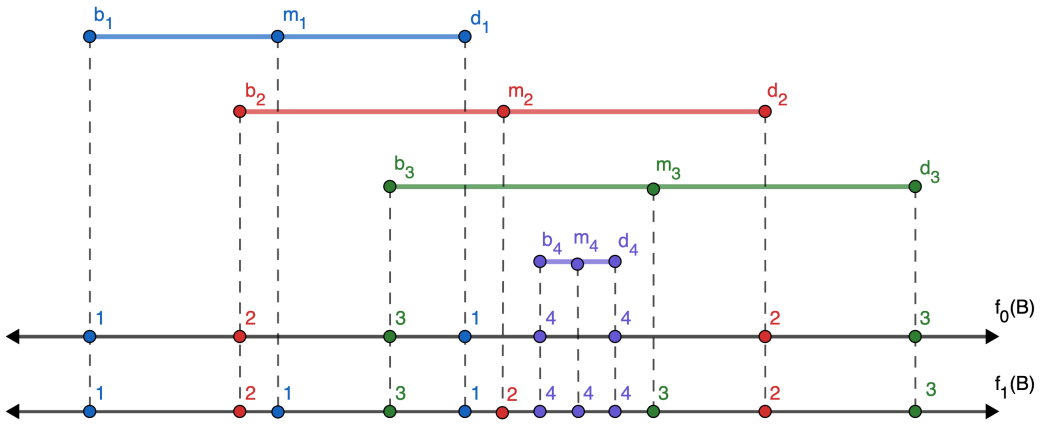


FIGURE 1.9. An example barcode B with its associated multipermutations $f_0(B) = (1\ 2\ 3\ 1\ 4\ 4\ 2\ 3)$ and $f_1(B) = (1\ 2\ 1\ 3\ 1\ 2\ 4\ 4\ 4\ 3\ 2\ 3)$ displayed below.

Thus, we have an entire collection of invariants $\overline{\sigma}_k(B)$ associated to barcodes, where $\overline{\sigma}_0(B)$ is exactly the combinatorial barcode invariant discussed above. We can define an order relation on $\overline{L}(n, 2^k + 1)$ in the same manner, which we denote by $\leqslant_k$. In Theorem 3.2.1 we show that $(\overline{L}(n, 2^k + 1), \leqslant_k)$ retains the graded lattice structure found in the $k = 0$ case, and so we call this poset the *power-k barcode lattice*.

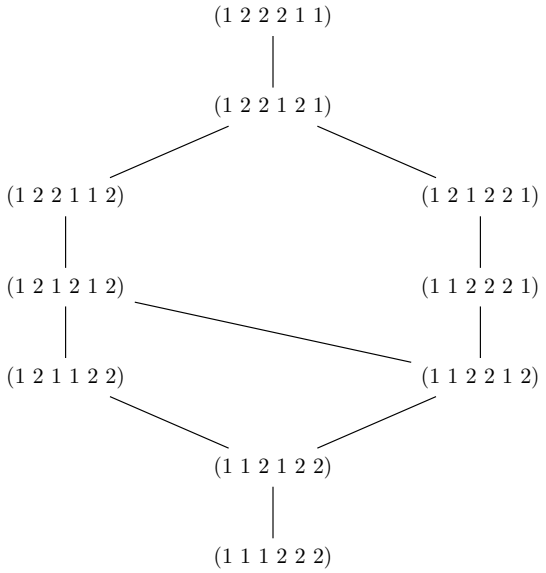


FIGURE 1.10. Hasse diagram of $(\overline{L}(n, 2^1 + 1), \leqslant_k)$.

THEOREM 3.2.1. *The power k barcode poset $(\overline{L}(n, 2^k+1), \leqslant_k)$ is order-isomorphic to a principal ideal of the multinomial Newman lattice, $L(n, 2^k+1)$. Consequently, $(\overline{L}(n, 2^k+1), \leqslant_k)$ is a lattice.*

Note that the invariants $\overline{\sigma_k(B)}$ are nested, in the sense that $\overline{\sigma_k(B)}$ is a substring of $\overline{\sigma_k(B)}$ for all $j \leqslant k$. The reason for this is that increasing k adds new endpoints for consideration, but does not change the relative positions of existing endpoints. Hence, we have the following lemma.

LEMMA 3.3.1. *Let B_1, B_2 be strict barcodes with n bars. If $\overline{\sigma_k(B_1)} = \overline{\sigma_k(B_1)}$, then $\overline{\sigma_k(B_1)} = \overline{\sigma_k(B_2)}$ for all $j \leqslant k$.*

It follows that increasing k amounts to producing ever more sensitive invariants $g_k(B)$ that capture increasingly nuanced information about the overlaps of pairs of bars. In fact, as k goes to infinity, the invariants $\overline{\sigma_k(B)}$ completely determine a large class of barcodes up to an affine transformation.

THEOREM 3.3.1. *Let B, B' be strict barcodes with n bars, where $B = \{[b_i, d_i]\}_{i=1}^n$ and $B' = \{[b'_i, d'_i]\}_{i=1}^n$ such that $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$ for all $k \in \mathbb{N}$. If the interval graph G_B (equivalently $G_{B'}$) is connected, then there exist constants $\alpha > 0$ and $\delta \in \mathbb{R}$ such that $B = \alpha B' + \delta$, where $\alpha B' + \delta := \{(\alpha b'_i + \delta, \alpha d'_i + \delta) : i \in [n]\}$.*

Even when $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$ for only finitely many k, these invariants can still be used to bound the bottleneck and q-Wasserstein distances between B and B'. Thus, our discrete invariants can be used to approximate continuous metrics on the space of barcodes.

THEOREM 3.3.2. *Let B, B' be strict barcodes with n bars such that $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$. Suppose there exists a bar $[b_*, d_*] \in B$ (or equivalently in B') which strictly contains all others, that is to say $b_* \leqslant b_i$ and $d_* \geqslant d_i$ for all $i \in [n]$. Then there exist constants $\alpha > 0$ and $\delta \in \mathbb{R}$ such that*

$$d_\infty(B, \alpha B' + \delta) \leqslant \frac{|d_* - b_*|}{2^k}, \;\; d_q(B, \alpha B' + \delta) \leqslant (n-1)^{\frac{1}{q}} \frac{|d_* - b_*|}{2^k}.$$

Now, recall that the permutohedron $\mathfrak{S}_n$ is also the face lattice of the polytope

$$P_{\mathfrak{S}_n} = \text{conv}\{(\pi_1, \ldots, \pi_n) \in \mathbb{R}^n : \pi \in \mathfrak{S}_\mathfrak{n}\}. \tag{1.14}$$

We show that one may embed $\overline{L}(n, 2^k + 1)$ into $P_{\mathfrak{S}_{n(2^k+1)}}$ by identifying multipermutations in $\overline{L}(n, 2^k + 1)$ with permutations of the totally ordered set $1_1 < 1_2 < \ldots n_1 < \ldots < n_{2^k+1}$. This gives us a new polytope, $P_{n,k} = \text{conv}\{(\pi_1, \ldots, \pi_{n(2^k+1)}) \in \mathbb{R}^{n(2^k+1)} : \pi \in \text{Im}(\iota \circ g_k)\}$. We call $P_{n,k}$ the *power-k barcode polytope*.

Because this embedding sends $\overline{L}(n, 2^k + 1)$ to a prime-ideal, the polytope $P_{n,k}$ is an example of a *Bruhat interval polytope*.

DEFINITION 1.2.1 ([**TW15**]). Let $u \leqslant v$ be permutations in $\mathfrak{S}_n$. The Bruhat interval polytope $Q_{u,v}$ is the convex hull of all permutation vectors $(z_1, z_2, \ldots, z_n)$ with $u \leqslant z \leqslant v$.

Note that $P_{n,k}$ is equal to $Q_{u,v}$ for $u = e \in \mathfrak{S}_{n(2^k+1)}$ and v the "fully nested" permutation $(1_1\ 2_1\ \ldots\ n_1\ n_2\ \ldots\ n_{2k+1}\ (n-1)_2\ \ldots\ 1_{2^k}\ 1_{2^k+1})$.

In [**TW15**], the authors prove, among other things, the following formula for computing the dimension of a Bruhat interval polytope. Let $u \leqslant v$ be permutations in $\mathfrak{S}_n$, and let $C : u = x_0 \lessdot x_1 \lessdot \ldots \lessdot x_\ell = v$ be any maximal chain from u to v. Define a labeled graph G^C on $[n]$ having an edge between vertices a and b if and only if $x_i(ab) = x_{i+1}$ for some $0 \leqslant i \leqslant \ell - 1$. Define $\Pi_C = V_1, V_2, \ldots, V_r$ to be the partition of $[n]$ whose blocks V_j are the connected components of G^C. One can show that the number of blocks does not depend on the choice of maximal chain C, so we let $\#\Pi_{u,v}$ denote the number of blocks, r. The authors then prove the following theorem.

THEOREM 1.2.1 ([**TW15**]). *The dimension of the Bruhat interval polytope $Q_{u,v}$ is $(n - \#\Pi_{u,v})$.*

From this result, one can easily compute the dimension of the barcode polytopes $P_{n,k}$.

COROLLARY 3.4.1. *The dimension of the power-k barcode polytope, $P_{n,k}$ is $n(2^k + 1) - 2$.*

1.2.3. Random Interval Graphs for Chronological Sampling Problems. In Chapter 4 we study a new model for generating random interval graphs, or equivalently random barcodes. The model is motivated by what we call "chronological sampling problems", where scientists are recording a series of time-stamped random observations and wish to deduce the support of different events over a given time period (see Chapter 4 for details).

Our model can be tersely described as follows. Let $Y = \{Y_t : t \in [0, T]\}$ be a stochastic process with state space $[m]$ and let $\mathcal{P} = \{t_1, t_2, ..., t_n\}$ be a set of n distinct points in $[0, T]$ with

$t_1 < t_2 < \ldots < t_n$. Then let $Y = (Y_1, Y_2, \ldots, Y_n)$ be a random vector whose components Y_i are samples from Y where $Y_i = Y_{t_i}$, so each Y_i takes values $[m]$. For each label $i \in [m]$ we define the (possibly empty) interval $I_n(i)$ as the convex hull of the points t_j for which $Y_j = i$, i.e., the interval defined by points colored i. Explicitly $I_n(i) = \text{Conv}(\{t_j \in \mathcal{P} : Y_j = i\})$, and we refer to $I_n(i)$ as the *empirical support* of label i. Furthermore, because it comes from the n observations or samples, we call the nerve complex, $\mathcal{N}(\{I_n(i) : i = 1, \ldots m\})$, the *empirical nerve* of Y. As we have shown, the empirical nerve of Y is fully determined by its 1-skeleton, which is a random interval graph, so we denote it by $G(Y, n)$.

In this thesis we mainly focus on a special case of this more general model, where we assume that the observations $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ are i.i.d. variables. In this case, there exist constants $p_1, p_2, \ldots, p_m \geqslant 0$ such that $\sum_{i=1}^{m} p_i = 1$ and $P(Y_{t_k} = 1) = p_i$ for all observations Y_{t_k} and all labels $i \in [m]$. We refer to this special case as the *stationary case* and all other cases as *non-stationary.*

We prove many results regarding the behaviour of this model in the stationary case, such as the probability of a particular edge being present and the expected number of edges.

THEOREM 4.3.1. *Let $G(Y, n)$ be the random interval graph generated by the stationary model described above. Then, for any pair $\{i, j\}$, $1 \leqslant i < j \leqslant m$, the probability of event $A_{ij} = \{\{i, j\} \in G(Y, n)\}$, i.e., that the edge $\{i, j\}$ is present in the graph $G(Y, n)$, is given by*

$$P(A_{ij}) = 1 - q_{ij}^n - \sum_{k=1}^{n} \binom{n}{k} \left[\left(2 \sum_{r=1}^{k-1} p_i^r p_j^{k-r} \right) + p_i^k + p_j^k \right] q_{ij}^{n-k},$$

where $q_{ij} = 1 - (p_i + p_j)$.

When $p_i = \frac{1}{m}$ for all $i \in [m]$, then $P(A_{ij}) = 1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n}$.

COROLLARY 4.3.1. *Let $G(Y, n)$ be the random interval graph generated by the stationary model described above. Additionally, let X be the random variable equal to the number of edges in the random interval graph $G(Y, n)$. Then,*

$$\mathbb{E}X = \sum_{1 \leqslant i < j \leqslant m} 1 - q_{ij}^n - \sum_{k=1}^{n} \left[\binom{n}{k} \left(2 \sum_{r=1}^{k-1} p_i^r p_j^{k-r} \right) + p_i^k + p_j^k \right] q_{ij}^{n-k},$$

where $q_{ij} = 1 - (p_i + p_j)$. *In the uniform case where* $p_i = \frac{1}{m}$ *for all* $i \in [m]$, *this expectation equals*

$$\binom{m}{2}\left(1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n}\right).$$

The also prove the following lower bound on the probability of finding an interval intersecting all others, i.e., that the maximum degree $\mathrm{Deg}(G(Y,n))$ of $G(Y,n)$ is $m-1$. In the following theorem we let $\mathcal{X}^n_{m,k}$ denote the set of weak-compositions of n with length m containing exactly k-many non-zero parts [**Sta11**, p. 25]. Formally, $\mathcal{X}^n_{m,k} = \{(x_1, ..., x_m) \in \mathbb{Z}^m_{\geqslant 0} : \sum_{i=1}^m x_i = n, |\{x_i : x_i \neq 0\}| = k\}$. Also let $M(x) = \frac{(x_1+x_2+...+x_m)!}{x_1!x_2!...x_m!}\prod_{i=1}^m p_i^{x_i}$ denote the multinomial distribution applied to the vector $x \in \mathcal{X}^n_{m,k}$ considering the associated probabilities $p_1, p_2, ..., p_m$. Finally, let S^k_n denotes the *Stirling numbers* of the second kind [**Sta11**, p. 81].

THEOREM 4.3.2. *Let* $G(Y,n)$ *be the random interval graph generated by the stationary model described above. Then, maximum degree of* $G(Y,n)$ *satisfies*

$$P(\mathrm{Deg}(G(Y,n)) = m-1) \geqslant \max_r\{[1 - \sum_{k=1}^{m-1}\frac{k^r}{m^r}\binom{m}{k}\sum_{x\in\mathcal{X}^r_{m,k}} M(x)(m-k)^r p_*^r][\sum_{x\in\mathcal{X}^{n-2r}_{m,m}} M(x)]\}.$$

Moreover, in the uniform case where $p_i = \frac{1}{m}$ *for all* $i \in [m]$, *we have that*

$$P(\mathrm{Deg}(\mathcal{N}_n) = m-1) \geqslant \max_r\{[1 - \frac{m!}{m^{2r}}\sum_{k=1}^{m-1}\frac{(m-k)^r}{(m-k)!}S^k_r], [\frac{m!}{m^{n-2r}}S^m_{n-2r}]\}.$$

We also derive the following lower bound on the expected *clique number* of $\mathcal{N}_n$, i.e., the size of the largest clique in the graph.

THEOREM 4.3.3. *Let* $G(Y,n)$ *be the random interval graph generated by the stationary model described above. Additionally, let* ω *be the random variable equal to the clique number of* $G(Y,n)$. *Then,*

$$\mathbb{E}\ \omega \geqslant \sum_{i=1}^m (1 - q_i^{\lceil\frac{n}{2}\rceil} - q_i^{n-\lceil\frac{n}{2}\rceil+1} + q_i^n)$$

where $q_i = 1 - p_i$. *Moreover, in the uniform case where* $p_i = \frac{1}{m}$ *for all* $i \in [m]$, *we have that*

$$\mathbb{E}\ \omega \geqslant m - \left(\tfrac{m-1}{m}\right)^{\lceil\frac{n}{2}\rceil} - \left(\tfrac{m-1}{m}\right)^{n-\lceil\frac{n}{2}\rceil+1} + \left(\tfrac{m-1}{m}\right)^n.$$

Note that as the number of samples n grows large, Theorem 4.3.3 implies that the expected clique number $\mathbb{E}\ \omega \to m$. Since ω only takes values in $\{1, \ldots, m\}$ it follows that the clique number

also converges to m in probability. Thus, as n goes to infinity, the probability that the nerve of the observations is the $(m-1)$-simplex denoted by Δ_{m-1}, i.e., a complete graph, goes to 1. The following theorem provides a lower bound on this convergence.

THEOREM 4.4.1. *Let $G(Y,n)$ be the random interval graph generated by the stationary model described above. Then, the probability that $G(Y,n)$ is isomorphic to the complete graph $\mathcal{K}_m$ satisfies*

$$P(G(Y,n) \cong \mathcal{K}_{m-1}) \geqslant (\sum_{x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}} M(x))^2$$

where $\mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor} = \{(x_1, x_2, ..., x_m) \in \mathbb{N}^m : \sum_{i=1}^m x_i = \lfloor \frac{n}{2} \rfloor\}$.
In the uniform case where $p_i = \frac{1}{m}$ for every $i \in [m]$, this gives that

$$P(G(Y,n) \cong \mathcal{K}_m) \geqslant \left(\frac{m!}{m^{\lfloor \frac{n}{2} \rfloor}} S^m_{\lfloor \frac{n}{2} \rfloor} \right)^2$$

where, again, S^k_n denotes the Stirling numbers of the second kind.

Theorem 4.4.1 tells us how likely it is for the empirical nerve of n samples to form the $(m-1)$-simplex for fixed n. A related question asks what is the *first* observation n for which this occurs, i.e., if we have a sequence of observations $Y_1, Y_2, \ldots$ what is the least n such that $G(Y,n) \cong \mathcal{K}_m$? We call this quantity the *waiting time* to form the $(m-1)$-simplex and provide an upper bound on its expectation, below.

THEOREM 4.4.2. *Let $Y = Y_1, Y_2, \ldots$ be a sequence i.i.d. random variables such that $P(Y_j = i) = p_i > 0$ for all $i \in [m]$. For $n \in \mathbb{N}$, let $G(Y,n)$ denote the empirical nerve produced by the first n variables, $Y_1, \ldots, Y_n$, as in the stationary model described above. Let X be the random variable for the waiting time until $G(Y,n) \cong \mathcal{K}_m$, explicitly $X = \inf\{n \in \mathbb{N} : G(Y,n) \cong \mathcal{K}_m\}$. Then,*

$$\mathbb{E}X \leqslant 2 \int_0^\infty \Big(1 - \prod_{i=1}^m (1 - e^{-p_i x})\Big) dx.$$

Moreover, in the uniform case, where $p_i = \frac{1}{m}$ for all $i \in [m]$, we have that

$$\mathbb{E}X \leqslant 2m \sum_{i=1}^m \frac{1}{i}.$$

CHAPTER 2

The Combinatorial Barcode Lattice

Our goal in this chapter is to develop a new combinatorial invariant for barcodes which is capable of recording all possible arrangements of bars. We accomplish this by defining a map from the space of strict barcodes with n bars, $\mathcal{B}^n_{st}$, to a set of equivalence classes of certain multipermutations. We note that chapter contains proofs of the main contributions outlined in Section 1.2.1, as well as additional results.

2.1. The Space of Combinatorial Barcodes

DEFINITION 2.1.1. Let $B = \{[b_i, d_i]\}_{i=1}^n$ be a strict barcode and let T_n be the set of symbols $\{x_1, y_1, \ldots, x_n, y_n\}$. Then, the relation $\preccurlyeq_B$ on T_n is given by,

$$y_i \preccurlyeq_B y_j \iff d_i \leqslant d_j, \tag{2.1}$$

$$y_i \preccurlyeq_B x_j \iff d_i < b_j, \tag{2.2}$$

$$x_i \preccurlyeq_B y_j \iff b_i \leqslant d_j, \tag{2.3}$$

$$x_i \preccurlyeq_B x_j \iff (b_i < b_j) \text{ or } (b_i = b_j \text{ and } d_i \leqslant d_j). \tag{2.4}$$

Put simply, the relation $\preccurlyeq_B$ orders a set of label T_n according to the ordering of the endpoints in B, with some rules governing ties between birth and death times (2.3) or two birth times (2.4); note that since B is assumed to be strict, it is not possible for two bars to share a death time. In fact, this relation defines a total ordering of T_n.

LEMMA 2.1.1. *Let $B = \{[b_i, d_i]\}_{i=1}^n$ be a strict barcode with n bars and let T_n be the set of symbols $\{x_1, y_1, \ldots, x_n, y_n\}$. Then the poset $(T_n, \preccurlyeq_B)$ is totally ordered.*

PROOF. One can easily verify that $\preccurlyeq_B$ is reflexive and anti-symmetric. It is also clear that $(T, \preccurlyeq_B)$ is strongly-connected, i.e., any two elements are comparable. Thus, we only have left to show that $\preccurlyeq_B$ is transitive. Let $a, b, c \in T$ with $a \preccurlyeq_B b$ and $b \preccurlyeq_B c$. Now, proceed by cases:

(1) If $a = y_i, b = y_j, c = y_k$, then $d_i \leqslant d_j \leqslant d_k$. Hence, $y_i \preccurlyeq_B y_k$ and $a \preccurlyeq_B c$.
(2) If $a = y_i, b = y_j, c = x_k$, then $b_i \leqslant d_j < b_k$. Hence, $y_i \preccurlyeq_B x_k$ and $a \preccurlyeq_B c$. By the same logic, if exactly one of a, b, c is of the form x_i, then again $a \preccurlyeq_B c$.
(3) If $a = y_i, b = x_j, c = x_k$, then $d_i < b_j \leqslant b_k$. Hence, $y_i \preccurlyeq_B x_k$ and $a \preccurlyeq_B c$. By the same logic, if exactly one of a, b, c is of the form x_i, then again $a \preccurlyeq_B c$.
(4) If $a = x_i, b = x_j, c = x_k$, then
    i If $b_i < b_j$ or $b_j < b_k$ then $b_i < b_k$. Hence, $x_i \preccurlyeq_B x_k$ and $a \preccurlyeq_B c$.
    ii Otherwise, $b_i = b_j = b_k$ and $d_i \leqslant d_j \leqslant d_k$. Hence, $x_i \preccurlyeq_B x_k$ and $a \preccurlyeq_B c$.

□

Now, let $B = \{[b_i, d_i]\}_{i=1}^n$ be a strict barcode. If we linearly order the elements in T_n with respect to $\preccurlyeq_B$, then the indices of the symbols produce a multipermutation (equivalently, a double occurrence word) $\sigma_B \in L(n, 2)$. Recall, $L(n, 2)$ denotes the the multinomial Newman lattice $L(2, \ldots, 2)$ whose elements are the multipermutations of the multiset $\{1^2, \ldots, n^2\}$, or equivalently, the set of double occurrence words over $[n]$ where each symbol $i \in [n]$ appears exactly twice. For example, if B is the strict barcode with 3 bars given by $b_1 = 1.0$, $d_1 = 2.0$, $b_2 = 1.0$, $d_2 = 3.0$, $b_3 = 2.5$, $d_3 = 2.75$, then T_3 is ordered,

$$x_1 \lhd_B x_2 \lhd_B y_1 \lhd_B x_3 \lhd_B y_3 \lhd_B y_2,$$

and hence $\sigma_B = (1\ 2\ 1\ 3\ 3\ 2) \in L(3, 2)$.

Now, let $f : \mathcal{B}_{st}^n \to L(n, 2)$ denote the map given by $f(B) = \sigma_B$. The map f provides a new combinatorial invariant on the space of strict barcodes with n bars by associating a double occurrence word to each barcode. However, one issue with this invariant is that the map f is highly dependent on the given labeling of the bars. For example, consider the strict barcode B_2 given by: $b_2 = 1.0$, $d_2 = 2.0$, $b_1 = 1.0$, $d_1 = 3.0$, $b_3 = 2.5$, $d_3 = 2.75$. Clearly, B_2 is the same barcode as B_1 from the prior example, except that the labels of bars 1 and 2 have been swapped. As a result,

$f(B_1) = (1\ 2\ 1\ 3\ 3\ 2)$ while $f(B_2) = (2\ 1\ 2\ 3\ 3\ 1)$. We would like our invariant to be the same for any two barcodes that are equal up to some such relabeling.

To that end, consider the combinatorial equivalence class of σ_B, denoted $[\sigma_B]$. Recall, we say two multipermutations $\sigma_1, \sigma_2 \in L(n,2)$ are *combinatorially equivalent* if there exists a permutation $\tau \in \mathfrak{S}_n$ such that $\tau \circ \sigma_1 = \sigma_2$, where $\tau \circ \sigma_1$ indicates that τ acts element-wise on σ_1. When this is the case we write $\sigma_1 \equiv \sigma_2$. For example, $(1\ 2) \circ \sigma_{B_1} = \sigma_{B_2}$, for B_1, B_2 as above and $(1\ 2)$ the permutation in $\mathfrak{S}_3$ written in cycle notation. Hence, $\sigma_{B_1} \equiv \sigma_{B_2}$. We let $[\sigma]$ denote the equivalence class of σ and let $L(n,2)/\mathfrak{S}_n$ denote the set of equivalence classes, $\{[\sigma] : \sigma \in L(n,2)\}$. These equivalence classes allow us to define an equivalence relation on B^n_{st} which we call combinatorial equivalence for barcodes. One can verify that the following relation is reflexive, symmetric, and transitive.

Definition 2.1.2. Let $B_1, B_2 \in \mathcal{B}^n_{st}$. Let $g : \mathcal{B}^n_{st} \to L(n,2)/\mathfrak{S}_n$ denote the map given by $g(B) = [\sigma_B]$. We say B_1, B_2 are *combinatorially equivalent* if and only if $g(B_1) = g(B_2)$.

The equivalence class $g(B) = [\sigma_B]$ defines a new combinatorial invariant on the space of strict barcode with n bars. However, unlike the permutation invariant π_B studied in [**KGH20**, **CDG$^+$21**, **BG22**], the set of equivalence classes $L(n,2)/\mathfrak{S}_n$ does not have an easily interpretable algebraic structure. Moreover, because the relation $\equiv$ does not define a lattice congruence on $L(n,2)$, this set is not a quotient lattice as defined in Definition 1.1.14. Therefore, we must ask the following questions:

(1) What is the algebraic or combinatorial structure of $L(n,2)/\mathfrak{S}_n$?

(2) How does the structure of this space relate back to barcodes and/or persistence modules?

The remainder of this chapter is devoted to answering these questions.

2.2. Ordering Combinatorial Barcodes

Although combinatorial equivalence is not a lattice congruence, $L(n,2)/\mathfrak{S}_n$ can still be endowed with a partial order, which is inherited from $L(n,2)$, by selecting representatives of each equivalence class in a suitable manner. Recall that a multipermutation $\sigma \in L(n,2)$ is in *ascending order* if $1, 2, \ldots, i-1$ appear before the first instance of i in σ, for all $i \in [n]$, i.e., if the first copy of 1

appears before the first copy of 2, which appears before the first copy of 3, and so on. For example, the word (1 2 1 3 2 3) is in ascending order while the equivalent word (1 3 1 2 3 2) is not.

Now, let $w \in L(n,2)$. One can find an ascending order word $\overline{w} \in [w]$ as follows. Let τ_w be the substring of w given by the first occurrence of each element; when $w = \sigma_B$ for some barcode B, τ_w this is the string of the labels of the birth times. Now consider τ_w as a permutation in $\mathfrak{S}_n$. Then, the action of τ_w^{-1} on w relabels w so that the birth times now appear in ascending order and we may take $\overline{w} = \tau_w^{-1} \circ w$. For example, if $w = (2\ 1\ 4\ 1\ 3\ 3\ 2\ 4) \in L(n,2)$ we have that $\tau_w = (2\ 1\ 4\ 3)$ and so $\tau_w^{-1} \circ w = (1\ 2\ 3\ 2\ 4\ 4\ 1\ 3)$.

Note that if $s, t \in [w]$, then $\tau_s^{-1} \circ s = \tau_t^{-1} \circ t$. Hence, each equivalence class has a unique member which is in ascending order. Therefore, the map $\psi : L(n,2)/\mathfrak{S}_n \to L(n,2)$ given by $\psi([s]) = \overline{s}$, which sends each equivalence class $[s]$ to its ascending order representative $\overline{s}$, is well defined. In particular, ψ is a bijection of sets between $L(n,2)/\mathfrak{S}_n$ and $\overline{L}(n,2)$, the set of all words in $L(n,2)$ which are in ascending order. Hence, two strict barcodes $B_1, B_2 \in B_{st}^n$ are combinatorially equivalent if and only if $\overline{\sigma_{B_1}} = \overline{\sigma_{B_2}}$. For that reason, we may treat the sets $L(n,2)/\mathfrak{S}_n$ and $\overline{L}(n,2)$ interchangeably.

Definition 2.2.1. The *combinatorial barcode poset* is the induced subposet $(\overline{L}(n,2), \leqslant)$ of the multinomial Newman lattice $(L(n,2), \leqslant)$, where $\leqslant$ denotes the weak order. A double occurrence word $\sigma \in \overline{L}(n,2)$ is called a *combinatorial barcode*.

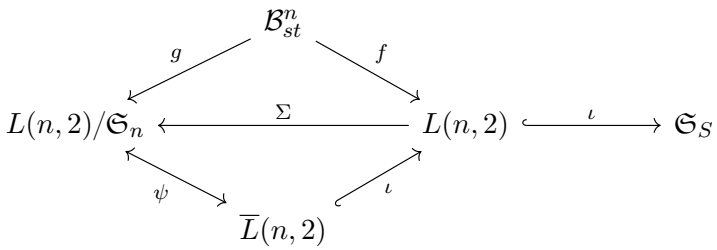
Now, recall that the relation $\leqslant$ on the multinomial Newman lattice is itself defined using the embedding $\iota : L(\mathbf{m}) \hookrightarrow \mathfrak{S}_S$. Hence, for all $s, t \in \overline{L}(n,2)$ we have that ,

$$s \leqslant t \iff \mathrm{inv}(\iota(s)) \subseteq \mathrm{inv}(\iota(t)). \tag{2.5}$$

The roles of f, g, ψ, ι and the equivalence relation $\equiv$ are summarized in the diagram (2.6), below. Here, we let Σ denote the quotient map that sends each word $w \in L(n,2)$ to its equivalence class $[w]$. By abuse of notation we also let ι denote both the inclusion map from $\overline{L}(n,2)$ to $L(n,2)$ and the inclusion map from $L(n,2)$ to $\mathfrak{S}_S$. We emphasize that this diagram is *not* commutative,

although it is the case that $\Sigma \circ f = g$.

(2.6)



Our first main result is that the combinatorial barcode poset, $\overline{L}(n,2)$, is order-isomorphic to a principal ideal of the multinomial Newman lattice $L(n,2)$ and, hence, is a lattice. Therefore, we also refer to $\overline{L}(n,2)$ as the *combinatorial barcode lattice*.

THEOREM 2.2.1. *The combinatorial barcode poset* $(\overline{L}(n,2) \leqslant)$ *is order-isomorphic to the principal ideal of the multinomial Newman lattice,* $L(n,2)$ *generated by the "fully nested" permutation:* $(1\ 2\ \ldots\ (n-1)\ n\ n\ (n-1)\ \ldots\ 2\ 1)$. *Consequently,* $(\overline{L}(n,2), \leqslant)$ *is a lattice.*

PROOF. Let $\alpha = (1\ 2\ \ldots\ n\ n\ \ldots\ 2\ 1) \in \overline{L}(n,2)$ and let $I(\alpha)$ denote the principal ideal generated by α in $L(n,2)$. We wish to show that $\overline{L}(n,2) = I(\alpha)$. To begin, note that α is in ascending order so $\alpha \in \overline{L}(n,2)$. We claim α is maximal in $(\overline{L}(n,2), \leqslant)$. Indeed, observe that every pair of distinct integers in α are inverted with the exception of the first occurrences of each integer, which are required to be appear in ascending order for all words $\overline{L}(n,2)$. Since the relation $\leqslant$ is induced by inversions in $\iota(\alpha)$, it follows that α is maximal. Thus, $\overline{L}(n,2) \subseteq I(\alpha)$.

To prove the reverse inclusion, let $s \in I(\alpha)$ and let $\tau_s \in \mathfrak{S}_n$ be the permutation given by the string of the first occurrences of each integer in s. Recall that the map $\psi : L(n,2)/\mathfrak{S}_n \to \overline{L}(n,2)$ which sends each equivalence class to its unique ascending order representative is defined as $\psi([s]) = \tau_s^{-1} \circ s$. Assume for the sake of contradiction that τ_s is not the identity permutation, then it follows that there exists a pair $i < j$ for which the first copy of j appears before the first copy of i in s. Hence, $(j_1, i_1) \in \mathrm{inv}(\iota(s))$. However, $s \leqslant \alpha$ implies that $\mathrm{inv}(\iota(s)) \subseteq \mathrm{inv}(\iota(\alpha))$ and $(k_1, \ell_1) \notin \mathrm{inv}(\iota(\alpha))$ for any $k > \ell$. Hence, we have a contradiction. Therefore it must be the case that $\tau_s = \mathrm{Id}_n$ and, hence, s is in ascending order and so $s \in \overline{L}(n,2)$. Thus, $I(\alpha) \subseteq \overline{L}(n,2)$, as desired. □

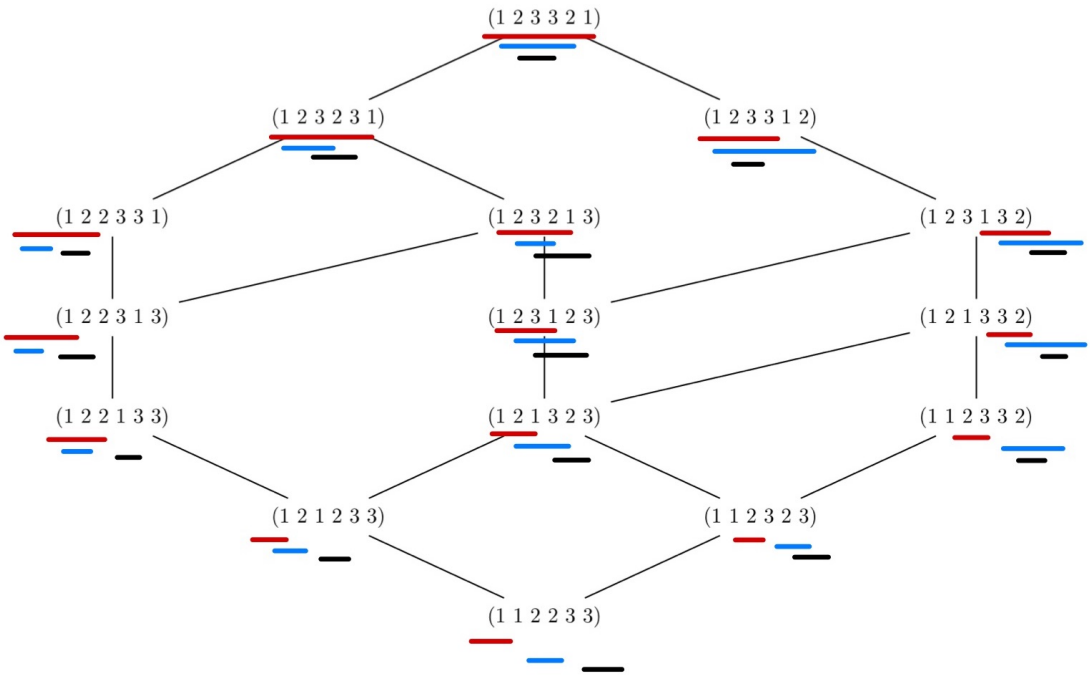


FIGURE 2.1. Hasse diagram of $(\overline{L}(3,2), \leqslant)$. Below each element, s, a barcode is depicted for which $g(B) = s$, illustrating Proposition 2.3.2.

2.3. Inversion Multisets and Barcode Crossing Numbers

A remarkable property of the combinatorial barcode lattice $\overline{L}(n,2)$ is that it admits an elegant, alternate construction based directly on inversions in the double occurrence words that does not require first "translating" to the symmetric group via the embedding ι.

DEFINITION 2.3.1. Let $s \in L(n,2)$ be a double occurrence word. Then, the *inversion multiset* of s is the multiset of pairs $\text{invm}(s) = \{(j,i)^{a_{ij}} : 1 \leqslant i < j \leqslant n\}$ where a_{ij} is equal to the number of pairs of indices (k,ℓ) such that $s_k = i, s_\ell = j$ and $k > \ell$.

Put simply, the inversion multiset has as elements the pairs (j,i), $i < j$, with multiplicity equal to the number of pairs of i's and j's that appear out of order in s. For example, $\text{invm}((1\ 2\ 3\ 2\ 4\ 4\ 1\ 3)) = \{(2,1)^2, (3,1)^1, (4,1)^2, (3,2)^1, (4,3)^2\}$.

Now, for $s,t \in \overline{L}(n,2)$, write $s \preceq t$ if $\text{invm}(s) \subseteq \text{invm}(t)$; recall, given multisets $A = \{x_1^{a_1}, \ldots, x_n^{a_n}\}$, $B = \{x_1^{b_1}, \ldots, x_n^{b_n}\}$ we say that $A \subseteq B$ if and only if $a_i \leqslant b_i$ for all $i \in [n]$.

PROPOSITION 2.3.1. *For $s,t \in \overline{L}(n,2)$, we have that $s \preceq t \iff s \leqslant t$.*

PROOF. Let $s \in L(n,2)/\mathfrak{S}_n$ and let (j,i) be an element in $\mathrm{invm}(s)$ with multiplicity k. Recall that since s is in ascending order, the first copy of i appears before the first copy of j for all $i < j$. Therefore, the copies of i, j must appear according to one of three patterns:

$$(1)\, i \ldots i \ldots j \ldots j,\ (2)\, i \ldots j \ldots i \ldots j,\ (3)\, i \ldots j \ldots j \ldots i\,.$$

It follows that $k \in \{0,1,2\}$, specifically, $k = 0$ when s contains pattern (1), $k = 1$ when s contains pattern (2), and $k = 2$ when s contains pattern (3).

Now, let $A_{ij} = \{(j_1,i_1),(j_2,i_1),(j_1,i_2),(j_2,i_2)\}$ be the set of all possible inversions involving i and j in $\iota(s)$, which is a permutation of the set $S = \{1_1, 1_2, 2_1, 2_2, \ldots, n_1, n_2\}$. It follows that,

$$\mathrm{inv}(\iota(s)) \cap A_{ij} = \begin{cases} \varnothing & , k = 0 \\ \{(j_1,i_2)\} & , k = 1 \\ \{(j_1,i_2),(j_2,i_2)\} & , k = 2 \end{cases}.$$

Now suppose $s \preceq t$, for some $t \in \overline{L}(n,2)$. Then, $(j,i) \in \mathrm{invm}(t)$ with multiplicity ℓ and necessarily $\ell \geqslant k$. As a result, $(\mathrm{inv}(\iota(s)) \cap A_{ij}) \subseteq (\mathrm{inv}(\iota(t)) \cap A_{ij})$. Applying this argument to all elements in $\mathrm{invm}(s)$, it follows that $\mathrm{inv}(\iota(s)) \subseteq \mathrm{inv}(\iota(t))$. Thus, $\iota(s) \leqslant_W \iota(s)$ and hence $s \leqslant t$.

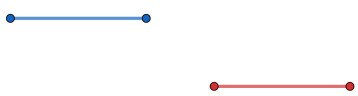
Moreover, this argument is reversible in the sense that we can deduce the multiplicity of $(j,i) \in \mathrm{invm}(s)$ from $\mathrm{inv}(\iota(s)) \cap A_{ij}$. Thus, we also have that if $s \leqslant t$, then $s \preceq t$, as desired. □

REMARK 2. We note that the key to Proposition 2.3.1, above, is the fact that the inversion *multisets* of elements in $\overline{L}(n,2)$ are in one-to-one correspondence with the inversion *sets* of those elements after embedding them in $\mathfrak{S}_S$. This is not true for general multipermutations in $L(n,2)$. For example, (1 2 2 1) and (2 1 1 2) both have the inversion multiset $\{(2,1)^2\}$ but their inversion sets are $\{(2_1,1_2),(2_2,1_2)\}$ and $\{(2_1,1_1),(2_1,1_2)\}$, respectively. Hence, we cannot use inversion multisets to define the ordering of $L(n,2)$; this construction is only valid on the ascending order words in $\overline{L}(n,2)$.

Now, let $s \in \overline{L}(n,2)$ and let $(j,i) \in [n]^2$ with $j > i$. We define the *crossing number* of i and j in s as the multiplicity of $(j,i) \in \mathrm{invm}(s)$ and denote it $\mathrm{cross}\#(s,j,i)$, or simply $\mathrm{cross}\#(j,i)$ when no confusion may occur. The name come from the fact that if $s = \overline{\sigma_B}$ for some strict barcode

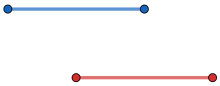
$B = \{[b_i, d_i]\}_{i=1}^n$, then the crossing numbers of s have a natural interpretation in terms of the bars in B: cross$\#(\overline{\sigma_B}, j, i)$ equals 0 if the bars $[b_i, d_i]$ and $[b_j, d_j]$ (labeled by ascending birth time) are disjoint, 1 if they are stepped, or 2 if they are nested, i.e.,

$$\text{cross}\#(j,i) = \begin{cases} 0, & \text{if } d_i < b_j & \text{(disjoint)} \\ 1, & \text{if } b_i \leqslant b_j \leqslant d_i < d_j & \text{(stepped)} \\ 2, & \text{if } b_i < b_j < d_j < d_i & \text{(nested)} \end{cases} . \tag{2.7}$$



(a) Disjoint bars.

(b) Nested bars.



(c) Two pairs of stepped bars.

FIGURE 2.2. A representation of all possible arrangements of a pair of bars in a strict barcode. The arrangement type determines the crossing number of their corresponding symbols in $\overline{\sigma_B}$.

PROPOSITION 2.3.2. *Let $s \in \overline{L}(n,2)$ be a combinatorial barcode and let $\rho(s)$ denote the rank of s in $\overline{L}(n,2)$. Then,*

$$\rho(s) = \sum_{i<j} \text{cross}\#(j,i).$$

PROOF. The result follows immediately from the observation that $|\text{invm}(s)| = |\text{inv}(\iota(s))| = \rho(s)$, where elements are counted with multiplicity in the inversion multiset. □

Recall that the inversions in a permutation $\pi \in \mathfrak{S}_n$ can be recorded as a vector $\nu(\pi)$ where $\nu(\pi)_i = \#\{(a,i) \in \text{inv}(\pi)\}$. We can generalize this construction to combinatorial barcodes by defining the *inversion vector* of $s \in \overline{L}(n,2)$.

DEFINITION 2.3.2. Let $s \in L(n,2)$ be a double occurrence word. The *inversion vector* of s is the vector $\mu(s) \in \mathbb{Z}^n$ where $\mu(s)_i = \sum_{j>i} a_{ij}$, where a_{ij} denotes the multiplicity of (j,i) in invm(s).

Put simply, the i-th component of $\mu(s)$ is the number of inversions in s where i is the smaller (right) element (counting multiplicity). It is a well-known result that the map which sends permutations $\pi \in \mathfrak{S}_n$ to their inversion vectors $\nu(\pi) \in \prod_{i=1}^{n}[0, n-i]$, where $[0,k] = \{0, 1, \ldots, k\}$, is a bijection of sets. Hence, a permutation $\pi \in \mathfrak{S}_n$ is completely determined by its inversion vector. In the following theorem, we show this is also the case for combinatorial barcodes.

THEOREM 2.3.1. *Let $V_n = \prod_{i=1}^{n}[0, 2(n-i)]$. Then the map $J : \overline{L}(n,2) \to V_n$ which sends each combinatorial barcode to its inversion vector is a bijection.*

PROOF. Let $x = (x_1, \ldots, x_n) \in V_n$. The following algorithm describes how to construct $J^{-1}(x)$ by building it "backwards" from n to 1. **Step 0**: Begin by writing the word with just two copies of n. **Step 1**: Insert two copies of $n-1$ by placing the first copy to the left of the current word and the second copy so that x_{n-1}-many terms of the current word are to the left of it. For example, if $x_{n-1} = 2$, then the second copy of $n-1$ would go on the far-right, so both copies of n are to the left of it. **Step k**: Repeat the process above until termination. At each step insert one copy of k to the left of the current word and insert the second copy so that x_{n-k} terms of the current word are to its left.

For example, if $x = (2, 0, 12, 5, 2, 3, 0, 1, 0)$ then we construct $J^{-1}(x)$ in the following manner:

9 9

8 9 8 9

7 7 8 9 8 9

6 7 7 8 6 9 8 9

5 6 7 7 5 8 6 9 8 9

4 5 6 7 4 7 5 8 6 9 8 9

3 4 5 6 7 4 7 3 5 8 6 9 8 9

2 3 4 5 6 7 4 7 3 5 8 6 9 2 8 9

1 2 3 1 4 5 6 7 4 7 3 5 8 6 9 2 8 9.

Note that by construction the number of terms $j > i$ appearing between the two copies of i in $J^{-1}(x)$ is exactly x_i for each $i \in [n]$. Therefore, this process is the inverse of J and, hence, J is a bijection. □

COROLLARY 2.3.1. *Let $c_k = \#\{s \in \overline{L}(n,2) : \rho(s) = k\}$, i.e., the number of combinatorial barcodes in $\overline{L}(n,2)$ of rank k, where k is a non-negative integer. Then,*

$$\sum_{k=0}^{\infty} c_k q^k = \prod_{i=1}^{n} \left(1 + q + \cdots + q^{2(n-i)}\right).$$

PROOF. Note that from Proposition 2.3.2 and Theorem 2.3.1 $\rho(s) = \sum_{i=1}^{n} \mu(s)_i$. Therefore,

$$\begin{aligned}
\sum_{k=0}^{\infty} c_k q^k = \sum_{s \in \overline{L}(n,2)} q^{\rho(s)} &= \sum_{a_1=0}^{2(n-1)} \sum_{a_1=0}^{2(n-2)} \cdots \sum_{a_n=0}^{0} q^{a_1 + a_2 \cdots + a_n} \\
&= \left(\sum_{a_1=0}^{2(n-1)} q^{a_1} \right) \left(\sum_{a_1=0}^{2(n-2)} q^{a_2} \right) \cdots \left(\sum_{a_n=0}^{0} q^{a_n} \right) \\
&= \prod_{i=1}^{n} \left(1 + q + \cdots + q^{2(n-i)}\right).
\end{aligned}$$

□

From Theorem 2.3.1, it is clear that we can identify combinatorial barcodes in $\overline{L}(n,2)$ with the integer lattice points in the polytope $A_n = \{x \in \mathbb{R}^{n-1} : 0 \leqslant x_i \leqslant 2(n-i)\, \forall i \in [n-1]\}$ by associating each multipermutation with its corresponding inversion vector (see Figures 2.3 and 2.4, below). Upon examination, we see that $\overline{L}(n,2)$ is a sublattice of the integer lattice in A_n formed by removing a few edges.

2.4. Connections to Trapezoidal Words, Stirling Permutations, & Second-Order Eulerian Numbers

In this section, we take a digression to discuss connections between the space of combinatorial barcodes and several related objects in combinatorics.

Recall that the q-analogue of the non-negative integer n is the polynomial $[n]_q = \frac{1-q^n}{1-q} = 1 + q + \cdots + q^{n-1}$. From this, one can form the q-analogues of other quantities; for example, the q-analogue of $n!$ is defined to be $[n!]_q = [1]_q \cdot [2]_q \cdot \cdots \cdot [n]_q$. We note that the right hand side

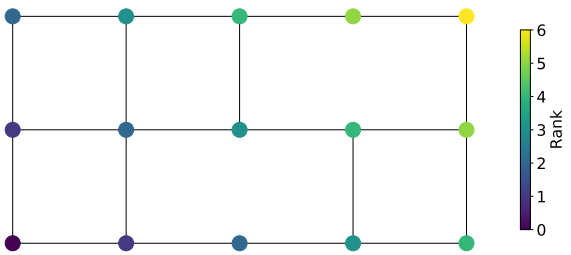


FIGURE 2.3. The graph of integer lattice points in P_3 with edges corresponding to cover relations in $(\overline{L}(3,2)\}, \leqslant)$ and node colors corresponding the rank of their associated multipermutation. Note that this graph is isomorphic to the Hasse diagram of $(\overline{L}(3,2)\}, \leqslant)$ in Figure 2.1.

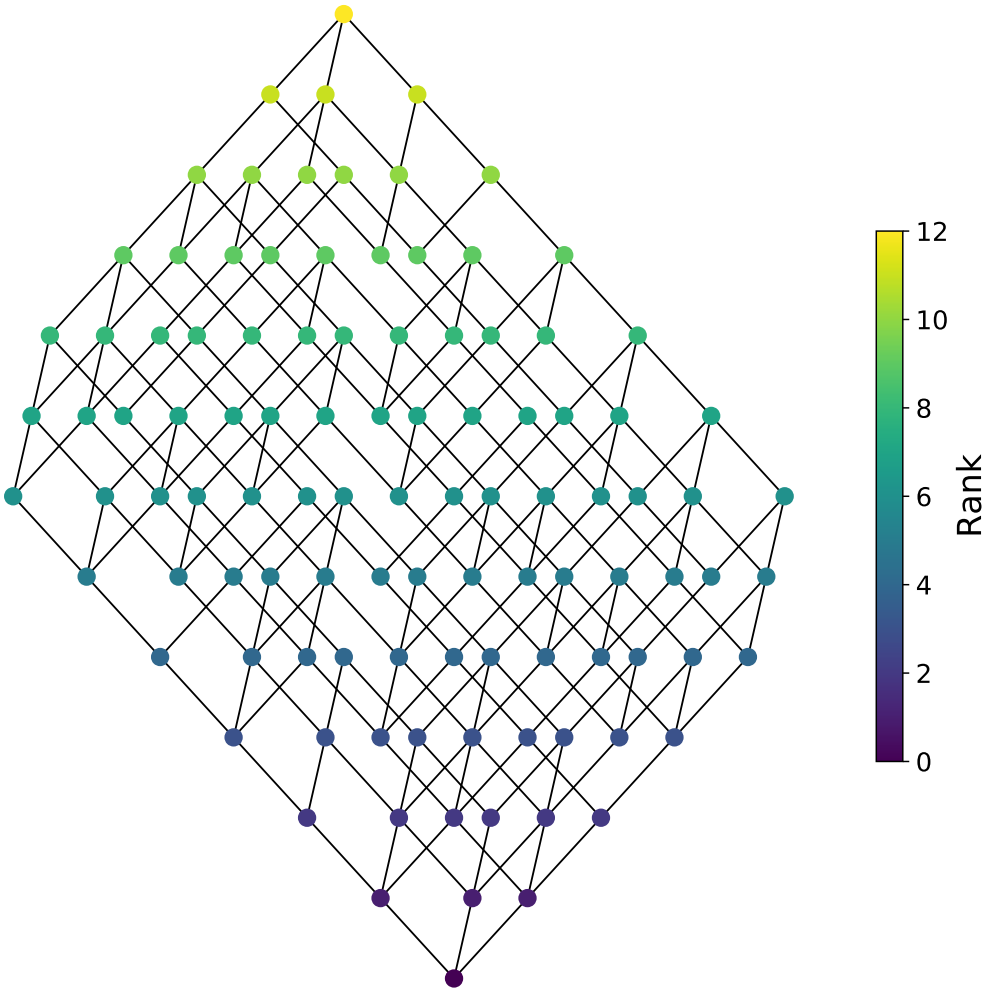


FIGURE 2.4. Hasse diagram of $(\overline{L}(4,2)\}, \leqslant)$. We note that this diagram is an "almost" isometric projection of the integer lattice points in A_4.

in Corollary 2.3.1 is the q-analogue of the product of the first n odd numbers, denoted $(2n-1)!!$, which is precisely the cardinality of $\overline{L}(n,2)$ (this follows from Theorem 2.3.1 by computing $|V_n|$).

It is a well known result that the quantity $(2n-1)!!$ also counts the number of *trapezoidal words* of length n [**Rio76**], the number of *Stirling permutations* over $[n]$ [**GS78**], and the number of plane recursive trees with $n+1$ vertices [**Jan08**].

DEFINITION 2.4.1 ([**Rio76**]). A *trapezoidal word* is an element of the Cartesian product $T_n = [1] \times [3] \times \cdots \times [2n-1]$, or equivalently, a word $w = (w_1\ w_2\ \ldots\ w_n)$ over $[2n-1]$ with the property that $1 \leqslant w_i \leqslant 2i-1$ for all $i \in [n]$.

DEFINITION 2.4.2 ([**GS78**]). A *Stirling permutation* is a double occurrence word $w \in L(n,2)$ with the property that if j appears between the two copies i in w, then $j > i$, for all $i \in [n]$. We denote the set of Stirling permutations over n by $\mathcal{Q}_n$.

DEFINITION 2.4.3 ([**Jan08**]). A *plane recursive tree* is a rooted, ordered tree obtained by starting with the root and recursively adding leaves to the tree. The root is labeled 0 and the vertices are labeled $1, 2, \ldots$ in the order in which they are added. Thus, a plane recursive tree with $n+1$ vertices is a labeled, rooted plane tree (with labels $0, 1, \ldots, n$) where the labels increase along each branch as we travel from the root.

For example, $T_2 = \{(1,1),(1,2),(1,3)\}$, $\mathcal{Q}_n = \{(1\ 1\ 2\ 2),(1\ 2\ 2\ 1),(2\ 2\ 1\ 1)\}$. and the three plane recursive trees with 3 vertices are depicted in Figure 2.5, below.

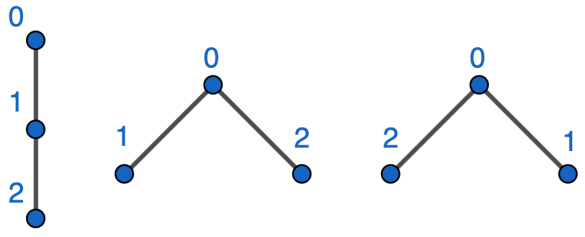


FIGURE 2.5. The three plane recursive trees with 3 vertices. We note that the second and third trees are distinct because plane recursive trees are oriented, i.e., the left-to-right order of vertices matters.

All three objects have connections to the *second-order Eulerian numbers*, $C_{n,k}$, defined recursively as follows:

$$C_{n,k} = kC_{n-1,k} + (2n-k)C_{n-1,k-1}. \tag{2.8}$$

The numbers $C_{n,k}$ count the number of words of length n with k distinct elements [**Rio76**], the number of Stirling permutations over $[n]$ with k descents [**GS78**], and the number of plane recursive trees with $n+1$ vertices and k leaves [**Jan08**]. Recently, there has been much interest in developing bijections between these objects that preserve these and other statistics [**Jan08**,**MQYY23**,**Liu23**].

Theorem 2.3.1 provides a natural bijection between combinatorial barcodes (ascending order double occurrence words) and trapezoidal words; note that each inversion vector $\mu(s)$ corresponds to a unique trapezoidal words after adding 1 to each entry. We can also link combinatorial barcodes to Stirling permutations via a similar bijection. We first require the following definition.

DEFINITION 2.4.4. Let $s \in L(n,2)$ be a double occurrence word. The *left-inversion vector* of s is the vector $\mu_L(s) \in \mathbb{Z}^n_{\geqslant 0}$ where $\mu_L(s)_j = \sum_{i<j} a_{ij}$, where a_{ij} denotes the multiplicity of (j,i) in $\mathrm{invm}(s)$.

Put simply, the j-th component of $\mu(s)$ is the number of inversions in s where j is the larger (left) element (counting multiplicity). The following theorem shows that there is a one-to-one correspondence between Stirling permutations and their left-inversion vectors.

THEOREM 2.4.1. *Let $U_n = \prod_{i=1}^{n}\{0,2,4,\ldots,4(i-1)\}$. Then the map $H : \mathcal{Q}_n \to U_n$ which sends each Stirling permutation to its left-inversion vector is a bijection.*

PROOF. Let $x = (x_1,\ldots,x_n) \in U_n$. The following algorithm describes how to construct $H^{-1}(x)$ by building it "forwards" from 1 to n. **Step 1**: Begin by writing the word with just two copies of 1. **Step k**: Insert kk into the current word so that $\frac{x_k}{2}$ elements of the current word are to its right. Repeat this process until termination.

For example, if $x = (0, 2, 4, 0, 12, 8, 2, 24, 6)$ then we construct $H^{-1}(x)$ in the following manner:

1 1

1 2 2 1

1 2 3 3 2 1

1 2 3 3 2 1 4 4

1 2 5 5 3 3 2 1 4 4

1 2 5 5 3 3 6 6 2 1 4 4

1 2 5 5 3 3 6 6 2 1 4 7 7 4

1 2 8 8 5 5 3 3 6 6 2 1 4 7 7 4

1 2 8 8 5 5 3 3 6 6 2 1 4 9 9 7 7 4

Note that by construction the number of inversions in $H^{-1}(x)$ where j is larger element is precisely x_j for all $j \in [n]$. Hence, this process is the inverse of H and thus H is a bijection. □

REMARK 3. We note that the mapping H^{-1} described in Theorem 2.4.1 is similar but distinct from the bijection between Stirling permutations and inversion sequences introduced in [**Liu23**]. The main advantage of Liu's bijection is that it generalizes to k-Stirling permutations (Stirling permutations with more than 2 copies of each integer) and it preserves some desired statistics that ours does not. The main advantage of our bijection is that it allows us to define another bijection between Stirling permutations and combinatorial barcodes which preserves the number of inversions up to a factor of 2 (see Corollary 2.4.1, below).

Taking Theorems 2.3.1 and 2.4.1 together, we can produce a natural bijection between combinatorial barcodes in $\overline{L}(n, 2)$ and the Stirling permutations $\mathcal{Q}_n$.

COROLLARY 2.4.1. *Let J, H be the bijections from Theorems 2.3.1 and 2.4.1, respectively. Then the map $\varphi : \overline{L}(n,2) \to \mathcal{Q}_n$ given by,*

$$\varphi(s) = H^{-1}(2(J(s))^R),$$

is a bijection, where if $x = (x_1, \ldots, x_n)$ is a vector then x^R the reversal of x, $(x_n, x_{n-1}, \ldots, x_1)$. Moreover, $|\operatorname{invm}(\varphi(s))| = 2|\operatorname{invm}(s)|$.

PROOF. Note that if $x \in \prod_{i=1}^{n}[0, 2(n-i)]$ then $2x^R \in \prod_{i=1}^{n}\{0, 2, 4 \ldots, 4(i-1)\}$ and this mapping is a bijection. Hence φ is a bijection. Finally, recall that the cardinality of the inversion multiset of a double occurrence is equal to the sum of the components of its inversion vector, or equivalently, its left-inversion vector. Thus, $|\operatorname{invm}(\varphi(s))| = 2|\operatorname{invm}(s)|$. □

The map φ from Corollary 2.4.1 can be described in a simple manner as follows. For each combinatorial barcode with $[n]$ bars, compute its inversion vector then scale it by 2 and reverse it. This produces an inversion vector which corresponds uniquely to a Stirling permutation over $[n]$. Moreover, this map is particularly natural because it preserves the cardinality of the inversion multisets up to a factor of 2.

We note that although φ preserves the cardinality of the inversion multisets up to a factor of 2, it does not necessarily preserve which pairs are inverted. For example, if $s = (1\ 2\ 1\ 3\ 2\ 3) \in \overline{L}(n,2)$, then $\operatorname{invm}(s) = \{(2,1)^1, (3,2)^1\}$ and $\mu(s) = (1,1,0)$. Reversing this vector and scaling by 2 yields $(0,2,2)$ and then taking the inverse of this vector under H as in Theorem 2.4.1 produces the Stirling permutation $\varphi(s) = (1\ 2\ 2\ 3\ 3\ 1)$. Although $\mu_L(\varphi(s)) = (0,2,2)$, the pairs of inverted elements do not match since $\operatorname{invm}(\varphi(s)) = \{(2,1)^2, (3,1)^2\}$.

The map φ allows us to prove the following combinatorial identities involving the generating function of the number of Stirling permutations over $[n]$ with k-many inversions.

COROLLARY 2.4.2. *Let $b_k = \#\{w \in \mathcal{Q}_n : |\operatorname{invm}(w)| = k\}$, i.e., the number of Stirling permutations over $[n]$ with k inversions, where k is a non-negative integer. Then,*

$$\sum_{k=0}^{\infty} b_k q^k = \prod_{i=1}^{n} \left(1 + q^2 + q^4 + \cdots + q^{4(i-1)}\right), \tag{2.9}$$

and equivalently,

$$\sum_{k=0}^{\infty} b_{2k} q^k = \prod_{i=1}^{n} \left(1 + q^1 + q^2 + \cdots + q^{2(n-i)}\right). \tag{2.10}$$

PROOF. Equation 2.9 follows from Theorem 2.4.1 by observing that $|\operatorname{invm}(w)|$ is equal to the sum of the components of its left-inversion vector and following the approach in Corollary 2.3.1. Equation 2.10 follows directly from Corollaries 2.3.1 and 2.4.1. □

From Corollary 2.4.1, we also have a new interpretation of $C_{n,k}$ in terms of inversion vectors of combinatorial barcodes.

COROLLARY 2.4.3. *Let $C_{n,k}$ denote the second-order Eulerian numbers defined recursively as follows:*

$$C_{n,k} = kC_{n-1,k} + (2n-k)C_{n-1,k-1}.$$

Then, $C_{n,k}$ is equal to the number of combinatorial barcodes with n bars, i.e., ascending-order double occurrence words over $[n]$, with k distinct elements in its inversion vector.

PROOF. The result follows immediately from Theorem 2.3.1 after noting that the set $\prod_{i=1}^{n}[0, 2(n-i)]$ is in bijection with the set of trapezoidal words T_n after adding 1 to each component (note that this mapping preserves the number of distinct elements). □

Now, recall that if $w = (w_1\ w_2\ \ldots\ w_l)$ is a word over $[n]$, an index $i \in [l]$ is called a *descent* if $w_i > w_{i+1}$, where we set $w_{l+1} = 0$ so that w_l is always a descent [**GS78**]. We let $\operatorname{des}(w)$ denote the set of descents of w. For example, if $w = (1\ 3\ 3\ 2\ 2\ 1)$ then the indices $\operatorname{des} w = \{3, 5, 6\}$ are descents. As we mentioned, the second-order Eulerian numbers $C_{n,k}$ count the number of trapezoidal words with k distinct elements and the number of Stirling permutations with k descents. Naturally, we wonder whether the number of distinct elements in the left-inversion vector of a Stirling permutation w is equal to the number of descents in w, i.e., if the relevant statistics for Stirling permutations and trapezoidal words are preserved under the bijection H from Theorem 2.4.1. Unfortunately, this is not the case. Consider $w = (2\ 2\ 1\ 3\ 4\ 4\ 3\ 1) \in \mathcal{Q}_4$. Then $\operatorname{des}(w) = \{2, 6, 7, 8\}$ while $\mu_L(w) = (0, 4, 2, 4)$, which has only 3 distinct components.

2.5. Connections to Barcode Bases and TMD

Recall our two motivating questions from earlier in the chapter:

(1) What is the algebraic or combinatorial structure of $L(n,2)/\mathfrak{S}_n$ (equivalently, $\overline{L}(n,2)$)?

(2) How does the structure of this space relate back to barcodes and/or persistence modules?

So far, we have mainly focused on the first question. In particular, we showed that $L(n,2)/\mathfrak{S}_n$ has a natural bijection to the set of combinatorial barcodes $\overline{L}(n,2)$ and then studied this space as an induced subposet of $L(n,2)$. We found that $\overline{L}(n,2)$ is a principal ideal of $L(n,2)$ and, hence, is a lattice. We also developed an alternate construction for $\overline{L}(n,2)$ based on a new notion of inversion multisets and used this construction to prove, among other things, a formula for the rank generating function of $\overline{L}(n,2)$.

We now turn our attention to the second question. We have already shown that the crossing numbers of a combinatorial barcode $\overline{\sigma_B}$ have a natural interpretation in terms of the geometric arrangements of the bars in B (see Equation 2.7 and Figure 2.2).

In fact, we can also use these crossing numbers to study the fibers of the TMD and the set of barcodes bases of persistence modules. Observe that we can restate Theorem 1.1.3 and Theorem 1.1.4 in terms of the crossing numbers of a barcode.

PROPOSITION 2.5.1. *Let* TMD *denote the topological morphology descriptor from [**KDS$^+$18**], and let $B = \{[b_i, d_i]\}_{i=1}^n$ be a strict barcode with n bars. Suppose no birth times in B are repeated and without loss of generality assume that the bars in B are labeled so that $\sigma_B = \overline{\sigma_B}$. Then,*

$$|\mathrm{TMD}^{-1}(B)| = \prod_{j=1}^{n} \#\{0 \leqslant i < j : \mathrm{cross}\#(j,i) = 2\}. \tag{2.11}$$

PROPOSITION 2.5.2. *Let $(V_\bullet, f_\bullet)$ be a persistence module of length $\ell+1$. Let $B = \{[b_i, d_i]^{m_i}\}_{i=1}^n$ be the barcode associated to the interval decomposition of $(V_\bullet, f_\bullet)$ as in Equation 1.12. Suppose $d_i \neq d_j$ for all $i \neq j$ and assume without loss of generality assume that B is labeled so that $\sigma_B = \overline{\sigma_B}$. Then the set of barcode bases of $(V_\bullet, f_\bullet)$, $\mathscr{B}(V_\bullet, f_\bullet)$, admits a bijection,*

$$\mathscr{B}(V_\bullet, f_\bullet) \cong \prod_{i=1}^{n} \mathrm{GL}(m_i; \mathbb{F}) \times \prod_{\substack{i<j: \\ \mathrm{cross}\#(j,i)=1}} \mathrm{Mat}(m_i \times m_j; \mathbb{F}), \tag{2.12}$$

where $\mathrm{GL}(m;\mathbb{F})$ *denotes the general linear group of* $m \times m$ *matrices over* $\mathbb{F}$.

Now, consider a strict persistence module $(V_\bullet, f_\bullet)$ with associated barcode $B = \{[b_i, d_i]\}_{i=1}^n$. Note that the first term on right hand side of Equation (2.12) reduces to $\mathrm{GL}(1,\mathbb{F})^n \cong (\mathbb{F} - \{0\})^n$, while the second reduces to,

$$\prod_{\substack{i<j: \\ \mathrm{cross}\#(j,i)=1}} \mathrm{Mat}(1;\mathbb{F}) \cong \prod_{\substack{i<j: \\ \mathrm{cross}\#(j,i)=1}} \mathbb{F}.$$

From this observation, we see how the cover relations in $\overline{L}(n,2)$ relate to the space of barcode bases for strict persistence modules.

THEOREM 2.5.1. *Consider two strict persistence modules* $(V_\bullet, f_\bullet), (W_\bullet, h_\bullet)$. *Let* $B = \{[b_i, d_i]\}_{i=1}^n$, $B' = \{[b'_i, d'_i]\}_{i=1}^n$ *be the barcodes associated to the interval decomposition of* $(V_\bullet, f_\bullet)$ *and* $(W_\bullet, h_\bullet)$, *respectively. Assume without loss of generality that* B *are labeled so that* $\sigma_B = \overline{\sigma_B}$ *(and likewise for* B'*). Suppose* $\overline{\sigma_B} \lessdot \overline{\sigma_{B'}}$, *so* $\overline{\sigma_B}$ *and* $\overline{\sigma_{B'}}$ *differ only in swapping an adjacent pair of entries* $i < j$, *which are in ascending order in* $\overline{\sigma_B}$ *but inverted in* $\overline{\sigma_{B'}}$. *Then,*

$$\mathscr{B}(V_\bullet, f_\bullet) \cong \mathscr{B}(W_\bullet, h_\bullet) \times \mathbb{F}, \text{ if } \mathrm{cross}\#(\overline{\sigma_B}, j, i) = 0,$$

$$\mathscr{B}(W_\bullet, h_\bullet) \cong \mathscr{B}(V_\bullet, f_\bullet) \times \mathbb{F}, \text{ otherwise.}$$

PROOF. Swapping i, j produces a new inversion. If the bars i and j were disjoint (had crossing number 0 in B), this produces a new pair of stepped bars. Hence, the product in Equation (2.12) gains an extra term, $\mathbb{F}$. Otherwise, the bars i and j were already stepped, in which case swapping i and j forms a pair of nested bars from the stepped bars. Hence, the product in Equation (2.12) loses one copy of $\mathbb{F}$. □

2.6. Conclusion and Further Questions

In this chapter we introduced a new combinatorial invariant on the space of barcodes by associating to each barcode an ascending-order double occurrence word $\overline{\sigma_B}$. We then studied the set of all such words, $\overline{L}(n,2)$ and studied the structure of this space as an induced subposet of the multinomial Newman lattice $L(n,2)$. In particular, we showed that $\overline{L}(n,2)$ is isomorphic to a principal ideal

of $L(n,2)$ and hence is a lattice (Theorem 2.2.1). We then developed an alternate construction of $\overline{L}(n,2)$ using the notion of inversion multisets (Proposition 2.3.1). We also introduced the notion of inversion vectors of multipermutations and used this construction to compute the rank generating function of $\overline{L}(n,2)$ (Corollary 2.3.1) and establish bijections between $\overline{L}(n,2)$ and trapezoidal words, Stirling words, and the second-order Eulerian numbers (Theorem 2.3.1, Theorem 2.4.1, Corollary 2.4.1, and Corollary 2.4.3). Finally, we showed how the cover relations of $\overline{L}(n,2)$, which we define via crossing numbers for barcodes, relate to the enumeration of barcode bases of strict persistence modules and to the fibers of the topological morphology descriptor. These results have inspired many new questions which still remain open. We outline a few of these, below.

Firstly, Corollary 2.4.1 establishes a bijection between $\overline{L}(n,2)$ and the Stirling permutations $\mathcal{Q}_n$ via a mapping between the inversion vectors of the former and the left-inversion vectors of the latter. Moreover, this mapping preserves the cardinality of the respective inversion multisets up to a factor of 2. Separately, it is known that the second-order Eulerian numbers $C_{n,k}$ count the number of trapezoidal words with k distinct components, the number of Stirling permutations with k descents, and the number of plane recursive trees with $n+1$ vertices and k leaves. In Corollary 2.4.3 we deduce that $C_{n,k}$ also enumerates the number of combinatorial barcodes whose inversion vectors (which are trapezoidal words after shifting by 1) have k distinct parts. Thus, we ask the following questions

(1) Is there another permutation statistic (such as descents, maj, etc.) on $\overline{L}(n,2)$ which is enumerated by $C_{n,k}$? Is there a different bijection from $\overline{L}(n,2)$ to Stirling permutations, trapezoidal words, or plane recursive trees that preserves this value and the relevant statistic for each set?

(2) We have already seen that the rank of a combinatorial barcode w is related to the inversion number of a Stirling permutation associated with w and the sum of the trapezoidal word associated to w. Is the rank also related to some statistic on plane recursive trees? What is the bijection that provides this relationship?

Secondly, Theorem 2.3.1 demonstrates that $\overline{L}(n,2)$ is a sublattice of the integer lattice in the polytope $A_n = \{x \in \mathbb{R}^{n-1} : 0 \leqslant x_i \leqslant 2(n-i),\ \forall i \in [n-1]\}$.

(1) *From inspecting the Hasse Diagram of* $\overline{L}(3,2)$ *and* $\overline{L}(4,2)$ *it seems as though* $\overline{L}(n,2)$ *can be obtained from the integer lattice in* A_n *by removing a few edges, thus producing chord-less cycles of length* 6. *Is this true in general, i.e., for arbitrary for* n*, are the longest chord-less cycles in* $\overline{L}(n,2)$ *also of length 6?*

(2) *Can we say what proportion of the edges are removed as* n *grows? That is, if we regard the Hasse diagram of* $\overline{L}(n,2)$ *as a graph, how does the cardinality of its edge set compare to the number of edges in the integer lattice in* A_n*?*

Thirdly, recall that Proposition 2.5.2 and Theorem 2.5.1 allows us to compute the space of barcode bases of strict persistence modules based on the crossing in in their associated barcodes. Therefore, we naturally ask,

(1) *How many combinatorial barcodes correspond to isomorphic spaces of barcode bases? That is to say, if* $(V_\bullet, f_\bullet)$ *and* $(W_\bullet, g_\bullet)$ *are strict persistence modules with associated barcodes* B, B'*, what is the maximal set of combinatorial barcodes* $E \subset \overline{L}(n,2)$ *for which,*

$$\{\overline{\sigma_B}, \overline{\sigma_{B'}}\} \subset E \implies \mathscr{B}(V_\bullet, f_\bullet) \cong \mathscr{B}(W_\bullet, g_\bullet)?$$

We note that this is equivalent to asking how many combinatorial barcodes have k *elements of multiplicity 1 in their inversion multisets, where* $k \in \binom{n}{2}$ *is arbitrary.*

CHAPTER 3

The Power-k Barcode Lattices

In the previous chapter we introduced a new combinatorial invariant on the space of barcodes by associating to each barcode an ascending-order double occurrence word $\overline{\sigma_B}$. In this chapter we generalize the construction from Chapter 2, thus producing an entire family of multipermutations associated to barcodes. We will show that these new multipermutations also form lattices, although they do not retain many of the other "nice" combinatorial properties found in $\overline{L}(n,2)$. However, we will show that these multipermutations still provide value by recording increasingly detailed information about the arrangement of the bars in a barcode. Ultimately, we prove that for a large class of barcodes these multipermutations can be to used bound two classic, continuous metrics on barcodes: the Wasserstein and bottleneck distances. A summary of these results can be found in Section 1.2.2.

We first recall a mapping, introduced in [**KGH20**] and further studied in [**CDG$^+$21**, **BG22**], from the space of strict barcodes with n bars to the symmetric group $\mathfrak{S}_n$, defined as follow. Let $B \in \mathcal{B}_{st}^n$ and suppose that $b_i \neq b - j$ for all $i \neq j$. Begin by ordering the death times in ascending order so that, $d_{i_1} < d_{i_2} < \cdots < d_{i_n}$. Then the indexing set $[n]$ gives a permutation $\gamma_B \in \mathfrak{S}_n$ defined by $\gamma_B(k) = i_k$, i.e., γ_B is the unique permutation such that $d_{\gamma_B(1)} < d_{\gamma_B(2)} < \ldots < d_{\gamma_B(n)}$. In the same manner, ordering the birth times gives another permutation τ_B. Thus, to each strict barcode with distinct birth times we can associate a permutation π_B given by $\pi_B = \tau_B^{-1} \cdot \gamma_B$ which tracks the ordering of the death values with respect to the birth values.

For example, if B_1 is the strict barcode with 3 bars given by $b_2 = 1.0$, $d_2 = 2.0$, $b_1 = 1.5$, $d_1 = 3.0$, $b_3 = 2.5$, $d_3 = 2.75$, then the birth/death times in B_1 are ordered: $b_2 < b_1 < d_2 < b_3 < d_3 < d_1$. So $\tau_B = (2\ 1\ 3)$, $\gamma_B = (2\ 3\ 1)$ and $\pi_B = (1\ 3\ 2)$, here all permutations are written in one-line notation.

We note that the double occurrence word $\overline{\sigma_B}$, defined in Chapter 2, actually contains π_B as a sub-word; π_B is exactly the sub-word formed by deleting the first occurrence of each integer in $\overline{\sigma_B}$.

PROPOSITION 3.0.1. *Let $B \in \mathcal{B}^n_{st}$ with distinct birth times and let π_B denote the barcode associated to B under the mapping defined in [**KGH20**]. Then π_B is a sub-permutation of the multipermutation $\overline{\sigma_B}$.*

It follows that the invariant $\overline{\sigma}_B$ is more sensitive than the invariant π_B because it captures the relative positions of the birth times along with the death times. This motivates the following questions: *(1) Is there a further generalization of this construction where we consider more points in each bar rather than just the birth and death times? (2) What additional information could be gleaned from such a construction?*

To that end, we must first determine a sensible way of selecting more points from each bar. A natural choice is to take the endpoints of all the intervals we get when splitting each bar into 2^k subintervals of equal length, where k is a non-negative integer. For instance, if $k = 0$ we consider just the endpoints of each bar as before, whereas if $k = 1$ then we consider both the endpoints and midpoint of each bar. For general k, we obtain the $(2^k + 1)$-many points $\left\{b_i + \ell\frac{d_i - b_i}{2^k} : \ell = 0, \ldots, 2^k\right\}$ for each bar $[b_i, d_i]$. We consider this choice natural because the points obtained by larger values of k contain all points obtained by smaller values of k. One can then linearly order these points to produce a multipermutation in $L(n, 2^k + 1)$, the multinomial Newman lattice consisting of permutations of the multiset $\{1^{2^k+1}, \ldots, n^{2^k+1}\}$. We let $f_k : \mathcal{B}^n_{st} \to L(n, 2^k + 1)$ denote this mapping.

For example, let B be the strict barcode with two bars, $b_1 = 0, d_1 = 2, b_2 = 1.5, d_2 = 3$. When $k = 1$, we bisect each bar and obtain the collection of points two new points (the midpoints) $m_1 = 1, m_2 = 2.25$ and we may order them, $b_1 < m_1 < b_2 < d_1 < m_2 < d_2$. The sub-scripts produce the multipermutation, $f_1(B) = (1\ 1\ 2\ 1\ 2\ 2) \in L(2, 2^1 + 1)$. Figure 3.1, below, contains an example where B has four bars.

One issue with this approach is that even strict barcodes might produce repeated points after bisecting, which we call *collisions*, that complicate the linear ordering. For example, if $B = \{[-1, 1], [-2, 2]\}$, then although all birth/death times are distinct, the two bars share the midpoint 0, i.e., the bars produce a collision at 0. It is also possible for bars to collide at an endpoint, for example consider $B = \{[-1, 1], [0, 1]\}$. One approach to resolving these collisions is to define a new order relation, an analogue of $\preccurlyeq_B$, for each value of k with rules to determine the ordering of collisions. However, because the number of points we add grows exponentially in k, this task

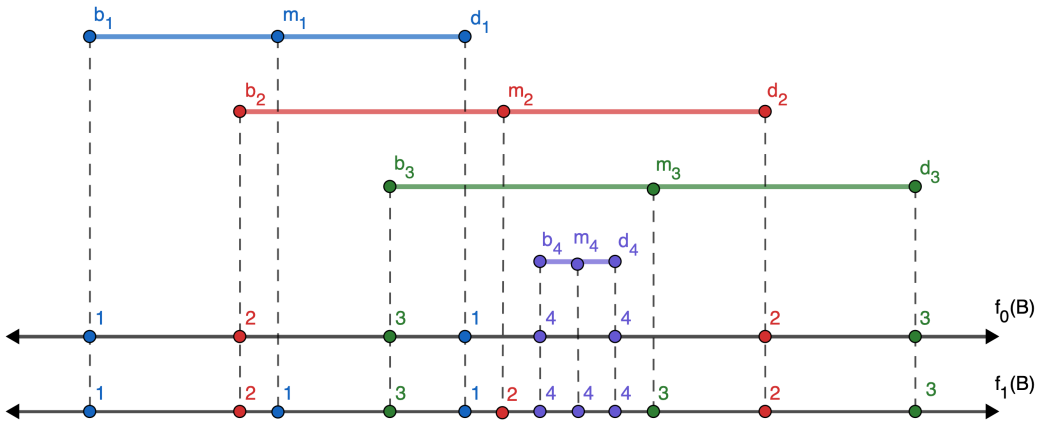


FIGURE 3.1. An example barcode B with its associated multipermutations $f_0(B) = (1\ 2\ 3\ 1\ 4\ 4\ 2\ 3)$ and $f_1(B) = (1\ 2\ 1\ 3\ 1\ 2\ 4\ 4\ 4\ 3\ 2\ 3)$ displayed below.

quickly becomes cumbersome. Instead, we take a different approach: rather than bisecting each bar into two equal pieces of length $\frac{1}{2}(d_i - b_i)$, we bisect them into pieces of almost equal length in such a way that avoids collisions. In the following section we explain this approach in greater details. The results therein are technical in nature and are not necessary for understanding the remainder of this thesis.

3.1. Technical Results for Collision-Free Bisections

Let $B = \{[b_i, d_i]\}_{i=1}^n$ be a strict barcode. Then for each $i \in [n]$ and $0 < t < 1$, let $E_B(i, t, k)$ be the set defined recursively as follows. When $k = 0$, $E_B(i, t, 0) = \{b_i, d_i\}$. For $k > 0$, let $E_B(i, t, k-1) = \{x_1, x_2, \ldots, x_{2^{k-1}+1}\}$ be linearly ordered so $x_1 < x_2 < \cdots < x_{2^{k-1}+1}$, and define $E_B(i, t, k) = \{x_1, x_{2^{k-1}+1}\} \cup \{tx_i + (1-t)x_{i+1} : i \in [n-1]\}$. Note that $E_B(i, t, k)$ is the collection of endpoints one obtains by splitting the bars in a fractal manner; for example, when $t = \frac{1}{2}$, this process is exactly the bisection of bars discussed prior. We claim that for an arbitrary strict barcode and non-negative integer k, we can find a t, arbitrarily close to $\frac{1}{2}$, such that these sets are collision-free.

LEMMA 3.1.1. *Let $B = \{[b_i, d_i]\}_{i=1}^n$ be a strict barcode. For each $i \in [n]$ and arbitrary $t \in [0, 1]$, let*

$$m_i(t) = tb_i + (1-t)d_i.$$

Then, for all $\varepsilon > 0$ there exists a $0 \leqslant \delta < \varepsilon$ such that $m_i(\frac{1}{2} - \delta) \notin \{b_j, m_j(\frac{1}{2} - \delta), d_j\}$ for all $i \neq j$.

PROOF. Note that $m_i(t) \in \{b_j, m_j(\frac{1}{2} - \delta), d_j\}$ if and only if t is a solution to one of the following linear equations:

$$tb_i + (1-t)d_i = b_j,$$

$$tb_i + (1-t)d_i = d_j,$$

$$tb_i + (1-t)d_i = tb_j + (1-t)d_j.$$

One can verify that each of these equations is non-degenerate, i.e., it has at most a single solution. Hence, the set,

$$\bigcup_{1 \leqslant i < j \leqslant n} \{t \in [0,1] : m_i(t) = b_j \vee m_i(t) = d_j \vee m_i(t) = m_j(t)\}$$

is finite. Therefore, we can always find a non-negative δ, arbitrarily close to zero, such that $m_i(\frac{1}{2} - \delta) \notin \{b_j, m_j(\frac{1}{2} - \delta), d_j\}$ for all $i \neq j$. □

THEOREM 3.1.1. *Let $B = \{[b_i, d_i]\}_{i=1}^n$ be a strict barcode. Then, for all $\varepsilon > 0$ and all non-negative integers k, there exists a $0 \leqslant \delta < \varepsilon$ such that $E_B(i, \frac{1}{2} - \delta, k) - \{b_i, d_i\} \cap E(j, \frac{1}{2} - \delta, k) = \varnothing$ for all $i \neq j$.*

PROOF. For simplicity let $E_B^c(i,t,k) = E_B(i,t,k) - \{b_i, d_i\}$. Fix some $\varepsilon > 0$ and proceed by induction on k. The $k = 1$ case is proved by Lemma 3.1.1, above. Now, suppose we have found some $0 \leqslant \delta_k < \varepsilon$ such that $E_B^c(i, \frac{1}{2} - \delta_k, k) \cap E(j, \frac{1}{2} - \delta_k, k) = \varnothing$ for all $i \neq j$. If $E_B^c(i, \frac{1}{2} - \delta_k, k+1) \cap E(j, \frac{1}{2} - \delta_k, k+1) = \varnothing$ for all $i \neq j$, then we are done. Otherwise, let $z \in E_B^c(i, \frac{1}{2} - \delta_k, k+1) \cap E(j, \frac{1}{2} - \delta_k, k+1)$ for some $i \neq j$. It follows that $(\frac{1}{2} - \delta)$ is a solution to one of the following linear equations:

$$tx_s + (1-t)x_{s+1} = b_j,$$

$$tx_s + (1-t)x_{s+1} = d_j,$$

$$tx_s + (1-t)x_{s+1} = ty_r + (1-t)y_{r+1},$$

for some $x_s, x_{s+1} \in E(i, \frac{1}{2} - \delta_k, k)$ and $y_r, yr + 1 \in E(j, \frac{1}{2} - \delta_k, k)$. As in the proof of Lemma 3.1.1, one can verify that each of these equations is non-degenerate. Hence, the union of the solution sets

of these equations, over all pairs $i < j$ and all indices s, r is finite. Let M denote the minimum of this set. Finally, by the inductive hypothesis we can find a $0 \leqslant \delta^* < \min(M, \varepsilon)$ such that $E^c_B(i, \frac{1}{2} - \delta^*, k) \cap E(j, \frac{1}{2} - \delta^*, k) = \varnothing$ for all $i \neq j$, and by construction $E^c_B(i, \frac{1}{2} - \delta^*, k+1) \cap E(j, \frac{1}{2} - \delta^*, k+1) = \varnothing$ for all $i \neq j$ as well. □

Let $f_k(B, t)$ denote the multipermutation obtained by linearly ordering the points in the set $\bigcup_{i \in [n]} E_B(i, t, k)$, where we assume t is chosen suitably to avoid collisions, and taking the sequence of the subscripts, as described above. Observe that if t_m is a sequence converging to $\frac{1}{2}$ from below, then there exists some $N \in \mathbb{N}$ such that $f_k(B, t_m) = f_k(B, t_l)$ for all $m, l > N$. Indeed, by construction the points in $E_B(i, t, k)$ are continuous functions of t which intersect at only finitely many points. Hence we may always find a sufficiently small $\delta > 0$ such such that $f_k(B, t_m) = f_k(B, t_l)$ for all $t_m, t_l \in [\frac{1}{2} - \delta, \frac{1}{2})$.

Thus, we formally define,

$$f_k(B) = \lim_{m = \to \infty} f_k(B, t_m), \tag{3.1}$$

where t_m is an arbitrary sequence converging to $\frac{1}{2}$ from below. However, in the remainder of this thesis, we will ignore these technical details for simplicity and assume that taking $t = \frac{1}{2}$ does not produce collisions so $f_k(B) = f_k(B, \frac{1}{2})$. It is not hard to verify that the results that follow do not depend on this assumption.

3.2. The Power-k Barcode Lattice

Now, for each non-negative integer k let $f_k \in \mathcal{B}^n_{st} \to L(n, 2^k + 1)$ described above. As in the $k = 0$, we find that these maps are highly dependent on the initial labeling of the bars in a barcode B. Therefore, we consider the maps $g_k : \mathcal{B}^n_{st} \to L(n, 2^k + 1)/\mathfrak{S}_n$ given by $g_k(B) = [f_k(B)]$, where $L(n, 2^k + 1)/\mathfrak{S}_n$ denotes the set of equivalence classes of $L(n, 2^k + 1)$ under combinatorial equivalence and $[\sigma]$ denotes the equivalence class of σ. As before, there is a bijection of sets $\psi_k : L(n, 2^k + 1)/\mathfrak{S}_n \to \overline{L}(n, 2^k + 1)$, where $\overline{L}(n, 2^k + 1)$ denotes the set of ascending order words in $L(n, 2^k + 1)$, which sends each class to its unique representative which is in ascending order. If $B \in \mathcal{B}^n_{st}$, we let $\overline{\sigma_k(B)} = \overline{f_k(B)}$. In the remainder of this chapter, we study $\overline{L}(n, 2^k + 1)$ as an induced subposet of $L(n, 2^k + 1)$.

DEFINITION 3.2.1. Let k be a non-negative integer. The *power-k barcode poset* is the induced subposet $(\overline{L}(n, 2^k + 1), \leqslant)$ of the multinomial Newman lattice $(L(n, 2^k + 1), \leqslant)$, where $\leqslant$ denotes the weak order.

Recall that there exists a map ι which embeds $L(n, 2^k+1)$ into $\mathfrak{S}_{n(2^k+1)}$ by identifying $\mathfrak{S}_{n(2^k+1)}$ with permutations of the set $S = \{1_1, 1_2, \ldots, 1_{2^k+1}, \ldots, n_{2^k+1}\}$. As in the $k = 0$ case from Chapter 2, we have that for all $s, t \in \overline{L}(n, 2^k + 1)$,

$$s \leqslant t \iff \mathrm{inv}(\iota(s)) \subseteq \mathrm{inv}(\iota(t)). \tag{3.2}$$

Our first main result in this chapter is that the power-k barcode poset is a lattice. We note that the proof of Theorem 3.2.1, below, is a direct generalization of proof of Theorem 2.2.1.

THEOREM 3.2.1. *The power k barcode poset $(\overline{L}(n, 2^k + 1), \leqslant)$ is isomorphic to a principal ideal of the multinomial Newman lattice, $L(n, 2^k + 1)$. Consequently, $(\overline{L}(n, 2^k + 1), \leqslant)$ is a lattice.*

PROOF. Let $\alpha \in L(n, 2^k + 1)$ be the word given by writing the integers 1 to n, followed by the remaining 2^k copies of n, followed by the remaining 2^k copies of $(n-1)$, and so on, terminating with the remaining 2^k copies of 1. For example, when $k = 1$ and $n = 3$, we have 3 copies of each integer and $\alpha = (1\ 2\ 3\ 3\ 3\ 2\ 2\ 1\ 1)$. Let $I(\alpha)$ denote the principal ideal generated by α in $L(n, 2^k + 1)$. We wish to show that $\overline{L}(n, 2^k + 1) = I(\alpha)$.

To begin, note that α is in ascending order so $\alpha \in \overline{L}(n, 2^k + 1)$. We claim α is maximal in $(\overline{L}(n, 2^k + 1), \leqslant)$. Indeed, observe that every pair of distinct integers in α are inverted with the exception of the first occurrences of each integer, which are required to be appear in ascending order for all words $\overline{L}(n, 2^k + 1)$. Since the relation $\leqslant$ is induced by inversions in $\iota(\alpha)$, it follows that α is maximal. Thus, $\overline{L}(n, 2^k + 1) \subseteq I(\alpha)$.

To prove the reverse inclusion, let $s \in I(\alpha)$ and let $\tau_s \in \mathfrak{S}_n$ be the permutation given by the string of the first occurrences of each integer in s. Assume for the sake of contradiction that τ_s is not the identity permutation, then it follows that there exists a pair $i < j$ for which the first copy of j appears before the first copy of i in s. Hence, $(j_1, i_1) \in \mathrm{inv}(\iota(s))$. However, $s \leqslant \alpha$ implies that $\mathrm{inv}(\iota(s)) \subseteq \mathrm{inv}(\iota(\alpha))$ and $(k_1, \ell_1) \notin \mathrm{inv}(\iota(\alpha))$ for any $k > \ell$. Hence, we have a contradiction.

Therefore it must be the case that $\tau_s = \mathrm{Id}_n$ and, hence, s is in ascending order and so $s \in \overline{L}(n,2)$. Thus, $I(\alpha) \subseteq \overline{L}(n,2)$, as desired. □

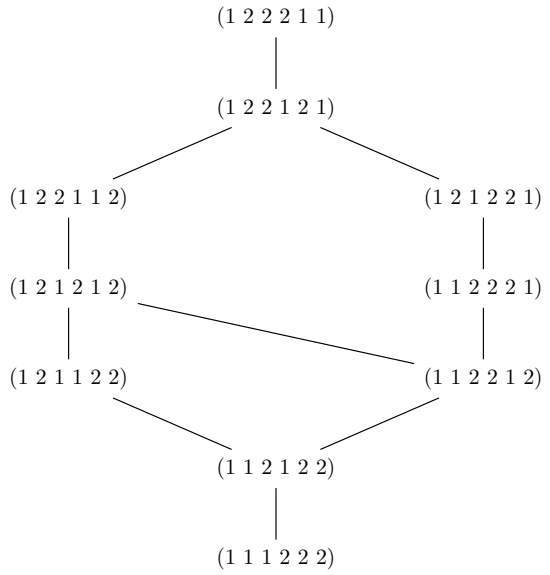


FIGURE 3.2. Hasse diagram of $\overline{L}(2, 2^1 + 1)$.

Unfortunately, many of the combinatorial results we proved for $\overline{L}(n,2)$ in Chapter 2 do *not* generalize to $\overline{L}(n, 2^k + 1)$ when $k \geqslant 1$. Notably, although we can define inversion multisets for the multipermutations in $\overline{L}(n, 2^k + 1)$ analogously, this correspondence is no longer one-to-one. For instance, $\mathrm{invm}(1\ 2\ 1\ 2\ 1\ 2) = \mathrm{invm}(1\ 1\ 2\ 2\ 2\ 1) = \{(2,1)^3\}$. Therefore, although one could still use inversion multisets to compute the rank of an element, it is not clear how to define inversion vectors in a way that admits a bijection as in Theorem 2.3.1. Moreover, without this correspondence it is not obvious how to compute the rank generating function of $\overline{L}(n, 2^k + 1)$ as in Corollary 2.3.1.

3.3. Connections to Bottleneck and Wasserstein Distances

Note that $\overline{\sigma_j(B)}$ is a sub-word of $\overline{\sigma_j(B)}$ for all $k > j$, just as the permutation π_B from [**KGH20**] is a sub-word of $\overline{\sigma_0(B)}$. Specifically, observe that if we delete every other occurrence (beginning with the second) of i in $\overline{\sigma_{k+1}(B)}$, for each $i \in [n]$, then the resulting word is precisely $\overline{\sigma_k(B)}$. Hence, we have a map $\delta_k : \overline{L}(n, 2^{k+1} + 1) \to \overline{L}(n, 2^k + 1)$ such that $\delta_k \circ \overline{\sigma_{k+1}(B)} = \overline{\sigma_k(B)}$. For instance, consider the barcode B given by $b_1 = 1.0$, $d_1 = 2.5$, $b_2 = 1.5$, $d_2 = 4.0$, $b_3 = 3.0$, $d_3 = 3.5$. Taking $k = 1$, we add the points $m_1 = 1.75$, $m_2 = 2.75$, $m_3 = 3.25$ so $\overline{\sigma_1(B)} = (1\ 2\ 1\ 1\ 2\ 3\ 3\ 3\ 2)$. The

map δ_0 deletes the points m_i, which gives $\delta_0(\overline{\sigma_1(B)}) = (1\ 2\ 1\ 3\ 3\ 2) = \overline{\sigma_0(B)}$. Hence, we have the following lemma.

LEMMA 3.3.1. *Let $B_1, B_2 \in \mathcal{B}^n_{st}$. If $\overline{\sigma_k(B_1)} = \overline{\sigma_k(B_2)}$ for some non-negative integer k, then $\overline{\sigma_j(B_1)\sigma_j(B_2)}$ for all $0 \leqslant j < k$.*

Thus, we see that increasing k amounts to producing ever more sensitive discrete invariants $\overline{\sigma_k(B)}$. These higher order invariants capture more nuanced information about the overlaps of pairs of bars. For instance, we have seen that if a barcode B contains two nested bars then $\overline{\sigma_0(B)}$ will contain the pattern (1 2 2 1). Going up a level, $\overline{\sigma_1(B)}$ confirms that the bars are nested but also tells us whether bar 2 is contained in the left half of bar 1, in the right half of bar 1, or whether it straddles the midpoint of 1 (see Figure 3.3). By the same logic, higher values of k provide even more granular intersection data.

(a) Left-nested bars, $\overline{\sigma_1(B)} = (1\ 2\ 2\ 2\ 1\ 1)$.

(b) Right-nested bars, $\overline{\sigma_1(B)} = (1\ 1\ 2\ 2\ 2\ 1)$.

FIGURE 3.3. Two barcodes with the same power 0 invariant, $\overline{\sigma_0(B)} = (1\ 2\ 2\ 1)$, but different power 1 invariants.

We will show that as k increases, these invariants provide enough information that they can be used to bound both the bottleneck and Wasserstein distances between a large class of barcodes. However, before proceeding we require the following lemma.

LEMMA 3.3.2. *Let B, B' be strict barcodes and let $G_B, G_{B'}$ denote the interval graphs of B and B', respectively. If $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$ for some $k \geqslant 0$ then $G_B \cong G_{B'}$.*

PROOF. It is clear that the intersection of a pair of bars $i < j$ can be determined from cross$\#(j, i)$. Recall one can deduce cross$\#(j, i)$ from the power 0 invariant of a barcode. Therefore, the power 0 invariant, and by Lemma 3.3.1 any power k invariant, completely determines the interval graph of its associated barcode. □

THEOREM 3.3.1. *Let B, B' be strict barcodes with n bars, where $B = \{[b_i, d_i]\}_{i=1}^n$ and $B' = \{[b'_i, d'_i]\}_{i=1}^n$, such that $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$ for all $k \in \mathbb{N}$. If the bars in B and B' have distinct endpoints, i.e., $\{b_i, d_i\} \cap \{b_j, d_j\} = \varnothing$ for all $i \neq j$ (and likewise for B'), and the interval graph G_B (equivalently $G_{B'}$) is connected, then there exist constants $\alpha > 0$ and $\delta \in \mathbb{R}$ such that $B = \alpha B' + \delta$, where $\alpha B' + \delta := \{(\alpha b'_i + \delta, \alpha d'_i + \delta) : i \in [n]\}$.*

PROOF. Without loss of generality assume that B, B' are suitably labelled so that $f_k(B) = \overline{\sigma_k(B)}$ and $b_1 < b_2 < \cdots < b_n$, and likewise for B'. Now, let $\alpha = \frac{d_1 - b_1}{d'_1 - b'_1}$, $\delta = b_1 - \alpha b'_1$, and define the map $T : \mathbb{R} \to \mathbb{R}$ such that $T(x) = \alpha x + \delta$. Observe that $T(b'_1) = b_1$, and that

$$T(d'_1) = \frac{d'_1(d_1 - b_1)}{d'_1 - b'_1} + b_1 - \frac{b'_1(d_1 - b_1)}{d'_1 - b'_1} = \frac{(d'_1 - b'_1)(d_1 - b_1)}{d'_1 - b'_1} + b_1 = d_1.$$

Now, let i be a neighbor of 1 in G_B (and hence in $G_{B'}$, by Lemma 3.3.2). We know at least one such i exists since G_B is connected. Take $k > 0$ to be arbitrary and let m denote the number of 1's that appear before the first i in $\overline{\sigma_k(B)}$, or equivalently in $\overline{\sigma_k(B')}$. Note $0 < m < 2^k$ since necessarily $b_1 < b_i < d_1$. Thus, we have that

$$\begin{aligned} b_1 + (m-1)\frac{d_1 - b_1}{2^k} < b_i < b_1 + m\frac{(d_1 - b_1)}{2^k}, \text{ and} \\ b'_1 + (m-1)\frac{d'_1 - b'_1}{2^k} < b'_i < b'_1 + m\frac{(d'_1 - b'_1)}{2^k}. \end{aligned}$$

Since $\alpha > 0$, it follows that

$$\begin{aligned} &\alpha\Big(b'_1 + (m-1)\frac{d'_1 - b'_1}{2^k}\Big) + \delta < T(b'_i) < \alpha\Big(b'_1 + m\frac{(d'_1 - b'_1)}{2^k}\Big) + \delta \\ \implies &b_1 + \frac{(m-1)(d_1 - b_1)}{2^k} < T(b'_i) < b_1 + \frac{m(d_1 - b_1)}{2^k}. \end{aligned}$$

Therefore, $|b_i - T(b'_i)| < \frac{d_1 - b_1}{2^k}$. Sending $k \to \infty$, it follows that $T(b'_i) = b_i$.

Now, if $b_1 < d_i < d_1$, then repeating the argument above gives $T(d'_i) = d_i$. Otherwise, we have that $b_1 < b_i < d_1 < d_i$. Let ℓ be the number of i's that appear before the last 1 in $\overline{\sigma_k(B)}$. Note

that $0 < \ell < 2^k + 1$, then we have

$$\begin{aligned}\frac{\ell(d_i - b_i)}{2^k} < d_1 - b_i < \frac{(\ell+1)(d_i - b_i)}{2^k}, \text{ and}\\ \frac{\ell(d'_i - b'_i)}{2^k} < d'_1 - b'_i < \frac{(\ell+1)(d'_i - b'_i)}{2^k}.\end{aligned}$$

Again, as $\alpha > 0$, it follows that,

$$\begin{aligned}&\frac{\ell(\alpha d'_i - \alpha b'_i)}{2^k} < \alpha d'_1 - \alpha b'_i < \frac{(\ell+1)(\alpha d'_i - \alpha b'_i)}{2^k},\\ \implies &\frac{\ell(T(d'_i) - T(b'_i))}{2^k} < T(d'_1) - T(b'_i) < \frac{(\ell+1)(T(d'_i) - T(b'_i))}{2^k},\\ \implies &\frac{\ell(T(d'_i) - b_i)}{2^k} < d_1 - b_i < \frac{(\ell+1)(T(d'_i) - b_i)}{2^k}.\end{aligned}$$

Rearranging the inequalities above gives,

$$\begin{aligned}\frac{\ell}{2^k} < \frac{d_1 - b_i}{d_i - b_i} < \frac{(\ell+1)}{2^k}, \text{ and}\\ \frac{\ell}{2^k} < \frac{d_1 - b_i}{T(d'_i) - b_i} < \frac{(\ell+1)}{2^k}.\end{aligned}$$

Sending $k \to \infty$ it follows that $\frac{d_1 - b_i}{d_i - b_i} = \frac{d_1 - b_i}{T(d'_i) - b_i}$, from which we see that $T(d'_i) = d_i$. Hence, if i is a neighbor of 1, then $(T(b'_i), T(d'_i)) = (b_i, d_i)$. Finally, note that we can repeat the arguments above, now considering i and one of its neighbors j, to show that this holds for all vertices connected to 1 by some finite path. Since G_B is connected and finite, this gives the desired result. □

We note that the assumption of distinct endpoints in Theorem 3.3.1 is not necessary if B (equivalently B') has an "essential" bar $[b_*, d_*]$ which contains all others.

COROLLARY 3.3.1. *Let B, B' be strict barcodes with n bars such that $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$ for all $k \in \mathbb{N}$. Suppose there exists a bar $[b_*, d_*] \in B$ (or equivalently in B') which strictly contains all others, that is to say $b_* < b_i$ and $d_* > d_i$ for all $i \neq *$. Then there exist constants $\alpha > 0$ and $\delta \in \mathbb{R}$ such that $B = \alpha B' + \delta$, where $\alpha B' + \delta := \{(\alpha b'_i + \delta, \alpha d'_i + \delta) : i \in [n]\}$.*

PROOF. This follows from the proof of Theorem 3.3.1, since i is a neighbor of 1 in G_B and $b_1 < b_i$ for all $i \in \{2, \ldots, n\}$. □

Thus, we see that the entire family of power k invariants completely determines a barcode B up to an affine transformation. If the barcodes contain an essential bar strictly containing all others as in Corollary 3.3.1, we can even bound continuous metrics on barcodes when only finitely many of the invariants $\overline{\sigma_k(B)}$ agree.

THEOREM 3.3.2. *Let B, B' be strict barcodes with n bars such that $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$ for some non-negative integer k. Suppose there exists a bar $[b_*, d_*] \in B$ (or equivalently in B') which strictly contains all others, that is to say $b_* < b_i$ and $d_* > d_i$ for all $i \neq *$. Then there exist constants $\alpha > 0$ and $\delta \in \mathbb{R}$ such that*

$$d_\infty(B, \alpha B' + \delta) \leqslant \frac{|d_* - b_*|}{2^k}, \text{ and}$$
$$d_q(B, \alpha B' + \delta) \leqslant (n-1)^{\frac{1}{q}} \frac{|d_* - b_*|}{2^k}.$$

PROOF. Without loss of generality assume that B, B' are labeled appropriately so that $f_k(B) = \overline{\sigma_k(B)}$, and likewise for B'. Note that this implies that $[b_*, d_*] = [b_1, d_1]$. Now, let $\alpha = \frac{d_1 - b_1}{d_1' - b_1'}$, $\delta = b_1 - \alpha b_1'$, and define $T : \mathbb{R} \to \mathbb{R}$ such that $T(x) = \alpha x + \delta$. Observe that $T(b_1') = b_1$, and that $T(d_1') = d_1$. Now let (b_i, d_i) be another bar and let m denote the number of 1's that appear before the first i in $\overline{\sigma_k(B)}$, or equivalently in $\psi_k(g_k(B'))$. Note $0 < m < 2^k$ since $b_1 < b_i < d_1$. Then we have that,

$$b_1 + (m-1)\frac{d_1 - b_1}{2^k} < b_i < b_1 + m\frac{(d_1 - b_1)}{2^k}, \text{ and}$$
$$b_1' + (m-1)\frac{d_1' - b_1'}{2^k} < b_i' < b_1' + m\frac{(d_1' - b_1')}{2^k},$$

and since $\alpha > 0$, it follows that

$$b_1 + \frac{(m-1)(d_1 - b_1)}{2^k} < T(b_i') < b_1 + \frac{m(d_1 - b_1)}{2^k}.$$

Therefore, $|b_i - T(b_i')| \leqslant \frac{d_1 - b_1}{2^k}$. Recall that $d_1 \geqslant d_i$ for all $i \in [n]$, so by repeating the argument we get that $|d_i - T(d_i')| \leqslant \frac{d_1 - b_1}{2^k}$. Thus, $d_\infty(B, \alpha B' + \delta) \leqslant \frac{|d_* - b_*|}{2^k}$. For the bound on the q-Wasserstein

distance, observe that:

$$\Big(\sum_{i=1}^{n} \|(b_i, d_i) - (T(b_i'), T(d_i'))\|_\infty^q\Big)^{\frac{1}{q}} \leqslant \Big(\sum_{i=2}^{n} (\frac{d_1 - b_1}{2^k})^q\Big)^{\frac{1}{q}} = \Big((n-1)(\frac{d_1 - b_1}{2^k})^q\Big)^{\frac{1}{q}}$$
$$= (n-1)^{\frac{1}{q}} \frac{d_1 - b_1}{2^k},$$

from which the result follows. □

We note that the assumption in Theorem 3.3.2 requiring the existence of an essential bar containing all others is necessary. To see this, fix some non-negative integer and consider the barcodes $B = \{(0,1), (1-\varepsilon, \frac{3}{2})\}$ and $B' = \{(0,1), (1-\varepsilon, 2)\}$, where $\varepsilon << \frac{1}{2^k}$. Observe that $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$. Let $T(x) = \alpha x + \delta$, with $\alpha = \frac{d_1 - b_1}{d_1' - b_1'}$ and $\delta = b_1 - \alpha b_1'$ as in Theorems 3.3.1 and 3.3.2. Note T is simply the identity map, so $|T(2) - \frac{3}{2}| = \frac{1}{2}$. Hence, $d_\infty(B, \alpha B' + \delta) = \frac{1}{2}$. Thus, for arbitrary k we can find a pair of barcodes B, B' satisfying the following conditions:

(1) $\overline{\sigma_k(B)} = \overline{\sigma_k(B')}$,

(2) The length of the longest bar in B is bounded above by a fixed constant,

(3) $d_\infty(B, \alpha B' + \delta) = \frac{1}{2}$.

Thus, the convergence from Theorem 3.3.2 is not possible in this case.

3.4. Barcode Polytopes

It is a well known result that the permutohedron $\mathfrak{S}_n$ is also the face lattice of the polytope $P_{\mathfrak{S}_n} = \text{conv}\{(\pi_1, \ldots, \pi_n) \in \mathbb{R}^n : \pi \in \mathfrak{S}_\mathfrak{n}\}$. Recall that one can embed $\overline{L}(n, 2^k + 1)$ into $\mathfrak{S}_{n(2^k+1)}$, the set of permutations of the totally ordered set $1_1 < 1_2 < \ldots n_1 < \ldots < n_{2^k+1}$, via the map ι discussed earlier. Thus, we may view $\overline{L}(n, 2^k + 1)$ as a subset of the vertices of its corresponding permutohedron which gives rise to a new polytope, $P_{n,k} = \text{conv}\{(\pi_1, \ldots, \pi_{n(2^k+1)}) \in \mathbb{R}^{n(2^k+1)} : \pi \in \iota \circ \overline{L}(n, 2^k + 1)\}$. We call $P_{n,k}$ the *power-k barcode polytope*.

Because the map ι sends $\overline{L}(n, 2^k + 1)$ to a prime-ideal, the polytope $P_{n,k}$ is an example of a *Bruhat interval polytope*.

Definition 3.4.1 ([**TW15**]). Let $u \leqslant v$ be permutations in $\mathfrak{S}_n$. The Bruhat interval polytope $Q_{u,v}$ is the convex hull of all permutation vectors $(z_1, z_2, \ldots, z_n)$ with $u \leqslant z \leqslant v$.

Note that $P_{n,k}$ is equal to $Q_{u,v}$ for $u = e \in \mathfrak{S}_{n(2^k+1)}$ and v the "fully nested" permutation $(1_1\ 2_1\ \ldots\ n_1\ n_2\ \ldots\ n_{2k+1}\ (n-1)_2\ \ldots\ 1_{2^k}\ 1_{2^k+1})$.

In [**TW15**], the authors prove, among other things, the following formula for computing the dimension of a Bruhat interval polytope. Let $u \leqslant v$ be permutations in $\mathfrak{S}_n$, and let $C : u = x_0 \lessdot x_1 \lessdot \ldots \lessdot x_\ell = v$ be any maximal chain from u to v. Define a labeled graph G^C on $[n]$ having an edge between vertices a and b if and only if $x_i(ab) = x_{i+1}$ for some $0 \leqslant i \leqslant \ell - 1$. Define $\Pi_C = V_1, V_2, \ldots, V_r$ to be the partition of $[n]$ whose blocks V_j are the connected components of G^C. The authors show that the number of blocks does not depend on the choice of maximal chain C, so we let $\#\Pi_{u,v}$ denote the number of blocks, r. The authors then prove the following theorem.

THEOREM 3.4.1 ([**TW15**]). *The dimension of the Bruhat interval polytope $Q_{u,v}$ is $(n - \#\Pi_{u,v})$.*

From this result, it is easy to compute the dimension of the barcode polytopes $P_{n,k}$.

COROLLARY 3.4.1. *The dimension of the power-k barcode polytope, $P_{n,k}$ is $n(2^k + 1) - 2$.*

PROOF. Recall, $P_{n,k} = Q_{u,v}$ for $u = e \in \mathfrak{S}_{n(2k+1)}$ and v the "fully nested" permutation $(1_1\ 2_1\ \ldots\ n_1\ n_2\ \ldots\ n_{2k+1}\ (n-1)_2\ \ldots\ 1_{2^k}\ 1_{2^k+1})$. Consider the maximal chain C from u to v given by moving each element into position one by one, starting with the 1's in descending lexicographic order, then the 2's is descending order, and so on. Note that traversing this chain requires that we use all adjacent transpositions except for transposing the first and second elements, since the 1_1 term does not move. Hence, $\Pi_C = \{\{1_1\}, \{1_2, \ldots, 1_{2k+1}, 2_1, \ldots, n_{2k+1}\}$. Therefore $\Pi_{u,v} = 2$, from which the result follows. □

There is reason to believe that these barcode posets may be useful in performing a stratification of the space of barcodes. In [**BG22**] the authors adapted the mapping from [**KGH20**, **CDG$^+$21**] in order to stratify the space of barcodes with n bars. In essence, this stratification can be performed as follows. For each strict barcode B, consider the permutation $\pi_B \in \mathfrak{S}_n$ as defined in [**KGH20**]. Recall that the elements of $\mathfrak{S}_n$ can be identified with vertices of the permutohedron $P_{\mathfrak{S}_n}$, which is a dimension $(n-2)$ polytope in $\mathbb{R}^n$. Let $P^*_{\mathfrak{S}_n}$ denote the dual of $P_{\mathfrak{S}_n}$. $P^*_{\mathfrak{S}_n}$ induces a simplicial decomposition of the $(n-2)$-sphere. Then each permutation π_B can be associated to a top-dimensional simplex of this decomposition. By construction, crossing from one top-dimensional

simplex to another corresponds to changing the permutation type π_B by swapping a pair of adjacent entries. Therefore, the lower dimensional simplices correspond to non-strict barcodes where birth and death times may be repeated. Thus, the simplices in this decomposition induce a stratification on the space of barcodes. Finally, because $P^*_{\mathfrak{S}_n}$ comes with an embedding in $\mathbb{R}^n$, the two "extra" coordinates of $\mathbb{R}$ can be used to encode additional metrics describing the lengths of the bars.

It is possible to mirror this construction by considering $P^*_{n,k}$, the polar dual of $P_{n,k}$. Like the permutohedron, $P^*_{n,k}$ induces a simplicial decomposition of the $(n-2)$-sphere whose simplices have a natural interpretation in terms of the multipermutations in $\overline{L}(n, 2^k + 1)$. While we have not performed an in-depth investigation of this approach, we believe it is a ripe area for further study.

3.5. Conclusion and Further Questions

In this chapter, we generalized the construction of combinatorial barcodes in Chapter 2 to produce an entire family of discrete invariants associate to barcodes. These invariants are multipermutations in the power-k barcode lattices $\overline{L}(n, (2^k+1)^n+1)$, or equivalently, they are ascending order words over $[n]$ with exactly k occurrences of each integer, where $k \geqslant 2$. These invariants retain the lattice structure of $\overline{L}(n, 2)$ (Theorem 3.2.1) while recording increasingly granular information about the arrangement of the bars (Lemma 3.3.1). In Theorems 3.3.1 and 3.3.2, we showed that these multipermutations can provide bounds on the bottleneck and Wasserstein distances for a large class for barcodes. Finally, we showed how the barcode lattices can be interpreted as polytopes and computed some basic statistics of these polytopes (Corollary 3.4.1). These results have inspired many new questions which remain open. We outline a few of these below.

Firstly, one of the main limitations of our methods is that we are only ever comparing barcodes with the same number of bars. We ask,

(1) Is there some way to extend these methods in order to compare barcodes with different numbers of bars? We note one approach could be to add dummy bars of length 0 to the smaller barcode, but it is unclear what consequences this might have.

Secondly, from Lemma 3.3.1, we have surjective maps $\delta_k : \overline{L}(n, (2^k+1)^n+1) \to \overline{L}(n, (2^k)^n+1)$ such that $\delta_k \circ \overline{\sigma_{k+1}(B)} = \overline{\sigma_k(B)}$ (recall, this map deletes every other occurrence of each integer i). We ask,

(1) *What are the fibers of δ_k? We note this is equivalent to asking: if we only know $\overline{\sigma_k(B)}$, how many different words could $\overline{\sigma_{k+1}(B)}$ be?*

Thirdly, Theorem 3.3.2 allows us to bound the bottleneck and q-Wasserstein distances between barcodes B, B' based on their associated permutations $\overline{\sigma_k(B)}, \overline{\sigma_k(B')}$. We ask,

(1) *Can these bounds be improved by considering the distance between $\sigma_k(B)$ and $\overline{\sigma_k(B')}$ on $\overline{L}(n, 2^{k+1}+1)$?*

Fourthly, as we discussed in Section 3.4, there is reason to believe that the duals of the barcode polytopes could be used to perform a stratification of the space of barcodes by following the approach in [**BG22**]. We ask,

(1) *Is such a stratification possible and, if so, what benefits might it provide? Does it allow us to weaken our strictness assumption by allowing us to consider points on the boundaries of regions?*

CHAPTER 4

Random Interval Graphs for Chronological Sampling Problems

This final chapter takes a departure from the previous two. Rather that studying the relationship between barcodes and multipermutations, we will study the relationship between barcodes and their interval graphs. In particular, we study the behaviour of a random interval graph model of our own design. Though the complete model is quite generalizable, we will mainly focus on a special case which is very combinatorial in nature. We begin with some motivation.

Suppose you are an avid birdwatcher and you are interested in the migratory patterns of different birds passing through your area this winter. You know from prior knowledge that there are m different species of birds that pass over your home every year. Each day you go out and make a note whenever you see a bird, recording the species, day, and time you observed it. You hope that after enough time you will have observed at least one representative of each species. Naturally, you begin to wonder: *after n observations, how likely is it that I have seen every type of bird?* If we make the assumption that each observation is an independent, identically-distributed (i.i.d.) random variable , we recognize this situation as an example of the famous *coupon collector's problem* (for a comprehensive review of this problem see [**FS14**] and references therein). In this old problem a person is trying to collect m types of objects, called coupons, which are labeled $1, 2, \ldots, m$. The coupons arrive one by one as an ordered sequence $X_1, X_2, \ldots$ of i.i.d. random variables taking values in $[m] = \{1, \ldots, m\}$. The collector is interested in the expected *waiting time*, which is the random variable $W = \inf\{n \in \mathbb{N} : \{X_1, \ldots, X_n\} = [m]\}\}$, i.e., the least value of n for which the set $\{X_1, \ldots, X_n\}$ forms a complete collection.

The coupon collector problem dates back to 1708 when it first appeared in De Moivre's *De Mensura Sortis (On the Measurement of Chance)* [**FS14**]. The answer for the coupon collector problem depends on the assumptions we make about the distributions of the X_i. Euler and Laplace proved that if the coupons are equally likely, that is if $P(X_i = k) = \frac{1}{m}$ for every $k \in [m]$, then $\mathbb{E}(W) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{k}$, so $\mathbb{E}(W) = O(n \log n)$. The problem lay dormant until 1954 when H. Von

Schelling obtained the expected waiting time when the coupons are not equally likely [**Sch54**]. More recently, Flajolet et. al. introduced a unified framework relating the coupon collector problem to many other random allocation processes [**FGT92**].

Returning to our bird watcher, one might ask several other related questions about the arrival of the birds or coupons. For example, the birdwatcher might also ask:

- *What are the chances that the visits of k types of birds do not overlap at all?*
- *What are the chances that a pair of birds is present in my area at the same time?*
- *What are the chances of one bird type was in my area at the same time as k-many others?*
- *What are the chances that all the bird types were in my area at the same time?*

We note that very similar situations, where scientists collect or sample time-stamped data that comes in m types or classes and wish to predict overlaps, appear in applications as diverse as archaeology, genetics, job scheduling, and paleontology [**Gol04**, **Fis85**, **Pip98**, **HH07**]. The goal of this chapter is to present a *random graph model* to answer the four questions above.

4.1. Establishing a General Random Interval Graph Model

In order to answer any of the questions above we need to determine how to estimate the time(s) each species of bird might be present from a finite number of observations. This requires some modeling choices which we outline below.

The first modeling choice is that our observations are samples from a stochastic process indexed by a real interval $[0, T]$ and taking values in $[m]$. We tersely recall the definition of a stochastic process, see [**Ros96**] for a complete introduction. Let I be a set and let $(\Omega, \mathcal{F}, P)$ be a probability space. Suppose that for each $\alpha \in I$, there is a random variable $Y_\alpha : \Omega \to S \subset \mathbb{R}$ defined on $(\Omega, \mathcal{F}, P)$. Then the function $Y : I \times \Omega \to S$ defined by $Y(\alpha, \omega) = Y_\alpha(\omega)$ is called a *stochastic process* with *indexing set* I and *state space* S, and is written $Y = \{Y_\alpha : \alpha \in I\}$. When we conduct an observation at some time $t_0 \in [0, T]$, we are taking a sample of the random variable Y_{t_0}.

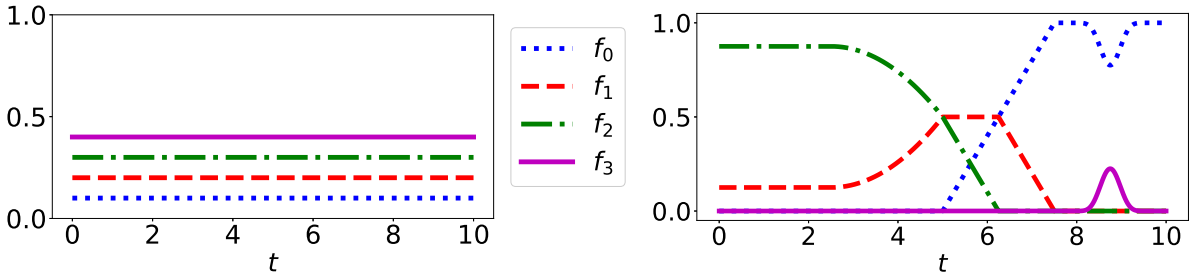
For each $i \in [m]$, the probabilities, $P(Y_t = i)$ determine a function from $[0, T] \to [0, 1]$, which we call the *rate function* of Y corresponding to i; the name is inspired by the language of Poisson point processes where the density of points in an interval is determined by a *rate* parameter (see [**Ros96**]).

DEFINITION 4.1.1 (Rate function). Let $Y = \{Y_t : t \in [0,T]\}$ be a stochastic process with indexing set $I = [0,T]$ and state space $S = [m]$. The *rate function* corresponding to label $i \in S$ in this process is the function $f_i : I \to [0,1]$ given by

$$f_i(t) = P(Y_t = i) = P(\{\omega : Y(t,\omega) = i\}).$$

Figure 4.2 gives two examples of the rate functions of some hypothetical stochastic processes (we will clarify the meaning of stationary and non-stationary in the next section). Note for each time $t_0 \in [0,T]$, the values $f_i(t_0)$ sum to 1 and define the probability density function of Y_{t_0}. Thus, we see that the rate functions describe the evolution of the probability density functions of the variables Y_t with respect to the indexing (time) variable t.

Note that in our bird watcher story the *support* of the rate function f_i corresponds to the interval in $[0,T]$ where species i might be present. Therefore, *our key problem is to estimate the support of the rate functions from finitely many samples.*



(a) Stationary rate functions.

(b) Non-stationary rate functions.

FIGURE 4.1. Two examples of hypothetical rate functions.

The state space, $[m]$, of Y is another, subtler modeling choice of ours. Note that one could alternatively have Y take values in the power set $2^{[m]}$, so as to allow for multiple species of birds to be observed at the same time. However, choosing $[m]$ rather than $2^{[m]}$ simplifies some calculations and, moreover, is quite reasonable. Rather than registering "three birds at 6 o'clock," our birdwatcher can instead register three sightings: one bird at 6:00:00, a second at 6:00:01, and a third a 6:00:02, for example.

This brings us to our next modeling choice: we assume that the rate function f_i has connected support for all $i \in [m]$. This is reasonable for our motivation; after all, a bird species first seen

on a Monday and last seen on a Friday is not likely to suddenly be out of town on Wednesday. The main benefit of this assumption is that now the support of the rate function f_i, $\text{supp}(f_i)$, is a sub-interval of $[0, T]$. This fact provides a natural way of approximating the support of f_i: given a sequence of observations $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ with $0 \leqslant t_1 < t_2 < \ldots < t_n \leqslant T$, let $I_n(i)$ denote the sub-interval of $[0, T]$ whose endpoints are the first and last times t_k for which $Y_{t_k} = i$ (note that it is possible for $I_n(i)$ to be empty or a singleton). It follows that $I_n(i) \subset \text{supp}(f_i)$ so we can use it to approximate $\text{supp}(f_i)$. We call the interval $I_n(i)$ the *empirical support* of f_i, as it is an approximation of $\text{supp}(f_i)$ taken from a random sample.

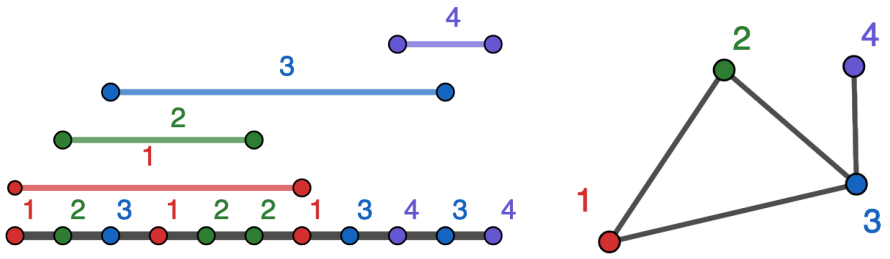
In summary, our model is actually quite simple: given a sequence of observations $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ from a stochastic process Y, we construct a random barcode $B = \{I_n(i) : i \in [m]\}$, where $I_n(i)$ is the empirical support of f_i, i.e., the interval whose endpoints are the least and greatest values t_k for which $Y_{t_k} = i$. These empirical supports are approximations of the supports of the rate functions, f_i and satisfy $I_n(i) \subset \text{supp}(f_i)$. With this model, we can now formulate our four bird watching questions as follows:

- *What is the probability that none of the empirical supports $I_n(i)$ intersect?*
- *What is the probability that a pair of empirical supports $I_n(i)$ and $I_n(j)$ intersect?*
- *What is the probability that one empirical support, $I_n(i)$, intersects with k-many others?*
- *What is the probability that the empirical supports all mutually intersect?*

Now, recall that given a family of convex sets $\mathcal{F} = \{F_1, \ldots, F_m\} \subseteq \mathbb{R}^d$. The *nerve complex* $\mathcal{N}(\mathcal{F})$ is the abstract simplicial complex whose k-facets are the $(k+1)$-subsets of indices $I \subset [m]$ such that $\bigcap_{i \in I} F_i \neq \varnothing$. Recall further that, as a consequence of Helly's theorem, the nerve complex of a barcode is a clique complex, so it is completely determined by its 1-skeleton (see Section 1.1.7 for details). Hence, given a barcode B, we will refer to the nerve complex $\mathcal{N}(B)$ and the graph G_B interchangeably depending on the context.

Figure 4.2, below, outlines how one can construct an interval graph from a set of observations $Y_{t_1}, \ldots, Y_{t_n}$. Figure 4.2a shows a sequence of $n = 11$ points on the real line, which corresponds to some observation times $t_1, \ldots t_n$. Above each point t_k we have a label, representing the value of Y_{t_k}. Displayed above the data are the empirical supports $I_n(i)$ for each $i \in [m] = [4]$. Figure 4.2b shows the interval graph constructed from these four intervals where each vertex is labeled with

the interval it corresponds to. In this example there are no times shared by the species $\{1, 2\}$ and the species $\{4\}$, so there are no edges drawn between those nodes. We emphasize that the interval graph constructed in this way will contain *up to* m-many vertices, but may contain fewer if some of the intervals $I_n(i)$ are empty, i.e., if we never see species i in our observations.



(a) Observations and empirical supports. (b) Interval graph.

FIGURE 4.2. Example observations with their corresponding empirical supports and interval graph.

Thus, we now have a model for generating an interval graph from a set of random observations $Y_{t_1}, \ldots, Y_{t_n}$ which records the overlaps of the empirical supports $I_n(i)$ in its edge sets. We summarize this model below, then re-frame the four motivating questions in terms of our model.

DEFINITION 4.1.2 (The Random Interval Graph Model). Let $Y = \{Y_t : t \in [0, T]\}$ be a stochastic process with state space $[m]$ and let $\mathcal{P} = \{t_1, t_2, ..., t_n\}$ be a set of n points in the indexing set $[0, T]$, satisfying $t_1 < t_2 < \ldots < t_n$. Then, let $Y_{t_1}, \ldots, Y_{t_n}$ be samples of from Y. For each state i define the (possibly empty) interval $I_n(i)$ as the convex hull of the points t_j for which $Y_{t_j} = i$. We refer to $I_n(i)$ as the *empirical support* of label i and let $G(Y, t_1, \ldots, t_n)$ denote the interval graph of $\{I_n(i) : i \in [m]\}$.

Under this random interval graph model our four questions can be formulated as follow:

- *What is the probability that* $G(Y, t_1, \ldots, t_n)$ *has no edges?*
- *What is the probability that a particular edge* (i, j) *is present in* $G(Y, t_1, \ldots, t_n)$*?*
- *What is the probability that* $G(Y, t_1, \ldots, t_n)$ *has a vertex of degree* $\geqslant k$*?*
- *What is the probability that* $G(Y, t_1, \ldots, t_n)$ *is isomorphic to the complete graph* K_m*?*

4.2. The Stationary Case

The model we present above is quite general in order to capture the potential nuances of our motivating problem. However, without some additional assumptions on the distribution of Y, the prevalence of pathological cases makes answering the motivating questions above very difficult. Therefore, our analysis here will focus on a special case of this problem where we make two additional assumptions on Y so that our analysis only requires discrete probability.

Our first assumption is that our observations $Y_{t_1}, Y_{t_2}, \ldots, Y_{t_n}$ are mutually independent random variables. Our second assumption is that the rate functions f_i are constant throughout the interval $[0, T]$. In this case, there exist constants $p_1, p_2, \ldots, p_m \geqslant 0$ such that $\sum_{i=1}^{m} p_i = 1$ and $f_i(t) = p_i$ for all $t \in [0, T]$ and all $i \in [m]$. We call the special case of our model where both of these assumptions are satisfied the *stationary case* and all other cases as *non-stationary*. Figure 4.2 shows examples of a stationary case, 4.1a, and a non-stationary case, 4.1b. We will also refer to the *uniform case*, which is the extra-special situation where $p_i = \frac{1}{m}$ for all $i \in [m]$. Note Figure 4.1a is stationary but not uniform.

The stationary case assumptions directly lead to two important consequences that greatly simplify our analysis. The first is that now the random variables $Y_{t_1}, \ldots, Y_{t_n}$ are independent and identically distributed (i.i.d.) such that $P(Y_{t_k} = i) = p_i > 0$. Note that this is true for any set of distinct observation times $\mathcal{P} = \{t_1, \ldots, t_n\}$. The second consequence simplifies things further still: though the points $\mathcal{P}$ corresponding to our sampling times have thus far been treated as arbitrary, one can assume without loss of generality that $\mathcal{P} = [n] = \{1, 2, \ldots, n\}$ since all sets of n points in $\mathbb{R}$ are combinatorially equivalent, as explained in the following lemma.

LEMMA 4.2.1. *Let $Y = \{Y_t : t \in [0, T]\}$ be stationary a stochastic process with state space $[m]$ and let $\mathcal{P} = \{t_1, \ldots, t_n\}$ and $\mathcal{P}' = \{t'_1, \ldots, t'_n\}$ be two sets of n distinct points in the state space $[0, T]$. Without loss of generality assume $t_1 < \ldots < t_n$ and $t'_1 < \ldots < t'_n$. Then for an arbitrary graph G_0, $P(G(Y, t_1, \ldots, t_n) \cong G_0) = P(G(Y, t'_1, \ldots, t'_n) \cong G_0)$.*

PROOF. Let w be an arbitrary word of length n over the alphaber $[m]$. Because Y_t, Y_s are i.i.d. for all $t, s \in [0, T]$ we have that $P(Y_{t_k} = w_k, k \in [n]) = P(Y_{t'_k} = w_k, k \in [n])$. Hence, the random

words $(Y_{t_1} \ \dots \ Y_{t_n})$ and $(Y_{t'_1} \ \dots \ Y_{t'_n})$ are i.i.d. Finally, note that the interval graph $G(Y, t_1, \dots, t_n)$ is completely determined by the word $(Y_{t_1} \ \dots \ Y_{t_n})$. □

Of course, the stationary case is less realistic and applicable in many situations. For example, it is not unreasonable to suppose that the presence of a dove at 10 o'clock should influence the presence of another at 10:01, or that the presence of doves might fluctuate according to the season and time of day. However, the stationary case is still rich in content and connections. Note that the stationary case of our model has the same assumptions as this famous problem: an observer receives a sequence of i.i.d. random variables taking values in $[m]$. In the language of our model, the coupon collector problem could be posed as, *How many samples does one need before they can expect that the graph $G(Y, t_1, \dots, t_n)$ contains m vertices?* Thus, we can consider the stationary model a generalization of the coupon collector problem which seeks to answer more nuanced questions about the arrival of different coupons. We summarize the key assumptions of the stationary case, below; we will reference these assumptions throughout this chapter.

Summary of stationary case assumptions: *In all results that follow let $Y_1, \dots, Y_n$ be a sequence of i.i.d. random variables such that $P(Y_j = i) = p_i > 0$ for all $i \in [m]$. When we refer to the uniform case this means the special situation when $p_i = \frac{1}{m}$ for all $i = 1, \dots, m$. In some problems it is useful to think of $Y_1, \dots, Y_n$ as a random coloring of the points $1, 2, \dots, n$ using m-many colors. Now, for each label/color $i \in [m]$, let $I_n(i)$ denote the empirical support of i, i.e., the interval $[b_i, d_i]$ whose left and right endpoints are the first and last indices k for which $Y_k = i$ (note these intervals may be empty or singletons). Finally, let $G(Y, n)$ denote the interval graph of the barcode $\{I_n(i) : i \in [m]\}$.*

4.3. Behavior with a Fixed Number of Samples

4.3.1. Elementary results. In this section we prove several results about the expected behaviour of the random interval graph $G(Y, n)$. We begin with a few elementary results as a warm-up.

PROPOSITION 4.3.1. *Under the key assumptions in Section 4.2, the probability that the random graph $G(Y, n)$ is isomorphic to the empty graph with $0 \leqslant k \leqslant m$ vertices but no edges, denoted K_k^c,*

satisfies

$$P(G(Y,n) \cong K_k^c) \geqslant p_*^n k! \binom{m}{k} \binom{n-1}{k-1},$$

where $p_* = \min\{p_1, p_2, ..., p_m\}$. *Moreover, in the uniform case where* $p_i = \frac{1}{m}$ *for all* $i \in [m]$, *we have that*

$$P(G(Y,n) \cong K_k^c) = \frac{k!}{m^n} \binom{m}{k} \binom{n-1}{k-1}.$$

PROOF. Note that for $G(Y,n)$ to form a disjoint collection of k points, the intervals induced by the $Y_1, \ldots, Y_n$ must also be disjoint. This occurs if and only if all points of the same color are grouped together. Given k fixed colors it is well known that the disjoint groupings are counted by the number of compositions of n into k parts, $\binom{n-1}{k-1}$. Each composition occurs with probability at least p_*^n. Finally, considering the $\binom{m}{k}$ different ways to choose these k colors and the $k!$ ways to order the colors, we have that,

$$P(G(Y,n) \cong K_k^c) \geqslant p_*^n k! \binom{m}{k} \binom{n-1}{k-1}.$$

The uniform case follows from the fact that every $k-$coloring of the n points is equally likely and occurs with probability $\frac{1}{m^n}$. □

In the next result we compute the probability that a particular edge is present in $G(Y,n)$.

THEOREM 4.3.1. *Under the key assumptions in Section 4.2 and for any* $1 \leqslant i < j \leqslant m$, *the probability of the event* $A_{ij} = \{\{i,j\} \in G(Y,n)\}$, *i.e., that the edge* $\{i,j\}$ *is present in the graph* $\mathcal{N}_n$, *is given by*

$$P(A_{ij}) = 1 - q_{ij}^n - \sum_{k=1}^{n} \binom{n}{k} \left[\left(2 \sum_{r=1}^{k-1} p_i^r p_j^{k-r} \right) + p_i^k + p_j^k \right] q_{ij}^{n-k},$$

where $q_{ij} = 1 - (p_i + p_j)$. *Moreover, in the uniform case where* $p_i = \frac{1}{m}$ *for all* $i \in [m]$, *we have that* $P(A_{ij}) = 1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n}$.

PROOF. We will find the probability of the complement, A_{ij}^c, which is the event where the two empirical supports do not intersect, i.e., $A_{ij}^c = \{I_n(i) \cap I_n(j)\} = \varnothing$. Let $C_i = \{\ell : Y_\ell = i, 1 \leqslant \ell \leqslant n\}$ and define C_j analogously. Note that A_{ij}^c can be expressed as the disjoint union of three events:

(1) $\{C_i$ and C_j are both empty$\}$,

(2) {Exactly one of C_i or C_j is empty},

(3) $\{C_i$ and C_j are both non-empty but $I_n(i)$ and $I_n(j)$ do not intersect$\}$.

The probability of the first event is simply q_{ij}^n. For the second event, assume for now that C_i will be the non-empty set and let $k \in [n]$ be the desired size of C_i. There are $\binom{n}{k}$ ways of choosing the locations of the k points in C_i. Once these points are chosen, the probability that these points receive label i and no others receive label i nor label j is exactly $p_i^k q_{ij}^{n-k}$. Summing over all values of k and noting that the argument where C_j is non-empty is analogous, we get that the probability of the second event is exactly $\sum_{k=1}^n \binom{n}{k}(p_i^k + p_j^k)q_{ij}^{n-k}$.

Now, note that the third event only occurs if all the points in C_i are to the left of all points in C_j or vice versa; for now assume C_i is to the left. Let $k \in [n]$ be the desired size of $C_i \cup C_j$ and let $r \in [k-1]$ be the desired size of C_i. As before there are $\binom{n}{k}$ ways of choosing the locations of the k points in $C_i \cup C_j$. Once these points are fixed, we know C_i has to be the first r many points, C_j has to be the remaining $k-r$ points, and all other points cannot have label i nor label j. This occurs with probability $p_i^r p_j^{k-r} q_{ij}^{n-k}$. Finally, summing over all values of k and r and adding a factor of 2 to account for flipping the sides of C_i and C_j we get that the third event occurs with probability $2\sum_{k=1}^n \binom{n}{k} \sum_{r=1}^{k-1} p_i^r p_j^{k-r} q_{ij}^{n-k}$.

Since A_{ij}^c is the disjoint union of these three events, $P(A_{ij}^c)$ is equal to the sum of these three probabilities, which gives the desired result. For the uniform case, simply set $p_i = p_j = 1/m$ in the general formula and note,

$$\begin{aligned}
P(A_{ij}) =& 1 - (\frac{m-2}{m})^n - \sum_{k=1}^n \binom{n}{k}\left[\left(2\sum_{r=1}^{k-1}\frac{1}{m^k}\right) + \frac{2}{m^k}\right](\frac{m-2}{m})^{n-k} \\
=& 1 - (\frac{m-2}{m})^n - \frac{1}{m^n}\sum_{k=1}^n \binom{n}{k} 2k(m-2)^{n-k} \\
=& 1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n}.
\end{aligned}$$

□

Corollary 4.3.1. *Let X be the random variable equal to the number of edges in $G(Y, n)$. Under the key assumptions in Section 4.2,*

$$\mathbb{E}X = \sum_{1 \leqslant i < j \leqslant m} 1 - q_{ij}^n - \sum_{k=1}^n \left[\binom{n}{k}\left(2\sum_{r=1}^{k-1} p_i^r p_j^{k-r}\right) + p_i^k + p_j^k\right] q_{ij}^{n-k},$$

where $q_{ij} = 1 - (p_i + p_j)$. *Moreover, in the uniform case where* $p_i = \frac{1}{m}$ *for all* $i \in [m]$, *we have*

$$\mathbb{E}X = \binom{m}{2}\left(1 - \frac{2n(m-1)^{n-1} + (m-2)^n}{m^n}\right).$$

PROOF. This follows immediately from Theorem 4.3.1 and the linearity of expectation. □

4.3.2. Connectivity. We now turn our attention to answering several questions regarding the connectivity of the graph $G(y, n)$. We begin by proving a lower bound on the probability of finding an interval intersecting all others, i.e., that the maximum degree of $G(Y, n)$ is $m-1$. In our bird watching story this can be interpreted as the probability of finding a species which overlaps in time with all others.

In the following theorem we let $\mathcal{X}_{m,k}^n$ denote the set of weak-compositions of n with length m containing exactly k-many non-zero parts. Recall, a weak composition of $n \in \mathbb{N}$ of length m is a sequence of non-negative integers $c_1, \ldots, c_m$ such that $\sum_{i=1}^m c_i = n$ [**Sta11**, p. 25]. Formally,

$$\mathcal{X}_{m,k}^n = \{(x_1, ..., x_m) \in \mathbb{Z}_{\geqslant 0}^m : \sum_{i=1}^m x_i = n, |\{x_i : x_i \neq 0\}| = k\} \tag{4.1}$$

Additionally, we let $M(x) = \frac{(x_1+x_2+...+x_m)!}{x_1!x_2!...x_m!}\prod_{i=1}^m p_i^{x_i}$ denote the multinomial distribution applied to the vector $x \in \mathcal{X}_{m,k}^n$, with associated probabilities $p_1, p_2, ..., p_m$ [**Pit93**, p. 155],

$$M(x) = \frac{(x_1 + x_2 + ... + x_m)!}{x_1!x_2!...x_m!}\prod_{i=1}^m p_i^{x_i}. \tag{4.2}$$

Finally, we let S_n^k denote the *Stirling numbers* of the second kind, which count the number of partitions of $[n]$ into k many parts [**Sta11**, p. 81]. Specifically,

$$S_n^k = \frac{1}{k!}\sum_{i=0}^k (-1)^{k-i}\binom{k}{i} i^n. \tag{4.3}$$

THEOREM 4.3.2. *Under the key assumptions in Section 4.2, the maximum degree of* $G(Y, n)$ *satisfies*

$$P(\mathrm{Deg}(G(Y,n)=m-1)\geqslant \max_{1\leqslant r\leqslant \frac{n-m}{2}}\{a(r)\cdot b(r)\},\ \textit{where,}$$

$$a(r)=1-\sum_{k=1}^{m-1}\frac{k^r}{m^r}\binom{m}{k}\sum_{x\in\mathcal{X}^r_{m,k}}M(x)(m-k)^r p_*^r,\ \textit{and}$$

$$b(r)=\sum_{x\in\mathcal{X}^{n-2r}_{m,m}}M(x),$$

and $p_*=\max\{p_i : i\in[m]\}$. *Moreover, in the uniform case where* $p_i=\frac{1}{m}$ *for all* $i\in[m]$, *we have that,*

$$a(r)=1-\frac{m!}{m^{2r}}\sum_{k=1}^{m-1}\frac{(m-k)^r}{(m-k)!}S^k_r,\ \textit{and}$$

$$b(r)=\frac{m!}{m^{n-2r}}S^m_{n-2r}.$$

Proof. For each $i\in[m]$ we let $S(i)=\{j\in[n] : Y_j=i\}$. Fix some $1\leqslant r\leqslant\frac{n-m}{2}$ and consider the sets $L=\{1,2,...,r\}$, $C=\{r,r+1,...,n-(r+1)\}$ and $R=\{n-r,n-(r-1),...,n\}$. Note that $\mathrm{Deg}(G(Y,n))=m-1$ if the following events occur,

$$A(r)=\{\text{There exists a label } i^* \text{ for which } S(i^*)\cap L\neq\varnothing \text{ and } S(i^*)\cap R\neq\varnothing\},$$

$$B(r)=\{S(i)\cap C\neq\varnothing \text{ for all } i\in[m]\}.$$

In order to calculate $P(A(r))$, we will compute $P(A(r)^c)$. Recall we may think of the stationary case as a random coloring of the points $1,\ldots,n$ with m-many colors. Thus, $A(r)^c$ corresponds to the event where no color appears in both L and R. First we calculate the probability of L being colored with exactly k colors with $1\leqslant k\leqslant m-1$. Observe that there are $\binom{m}{k}$ ways to choose these colors and $k^r\sum_{x\in\mathcal{X}^r_{m,k}}M(x)$ ways to color L with them. As there exist m^r different colorings with all the m colors, we have that for a fixed k the probability that L is colored with exactly k colors is given by

$$\frac{1}{m^r}k^r\binom{m}{k}\sum_{x\in\mathcal{X}^r_{m,k}}M(x).$$

In order for A^c to occur, we need that R be colored with only the $(m-k)$ remaining colors. Note that this event is independent from the coloring of L as the two sets are disjoint. There are $(m-k)^r$ different ways of coloring R, and each occurs with probability at most p_*^r, where $p_* = \max\{p_i : 1 \leqslant i \leqslant m\}$. Thus, for a fixed k we have that the probability that no color appears in both L and R is at most

$$\left[\frac{1}{m^r}k^r\binom{m}{k}\sum_{x\in\mathcal{X}_{m,k}^r}M(x)\right]\left[(m-k)^r p_*^r\right].$$

Then, by summing over all k we have that

$$P(A^c) \leqslant \sum_{k=1}^{m-1}\left[\frac{1}{m^r}k^r\binom{m}{k}\sum_{x\in\mathcal{X}_{m,k}^r}M(x)\right]\left[(m-k)^r p_*^r\right],$$

which implies that

$$P(A) \geqslant 1 - \sum_{k=1}^{m-1}\frac{1}{m^r}k^r\binom{m}{k}\sum_{x\in\mathcal{X}_{m,k}^r}M(x)(m-k)^r p_*^r.$$

To compute $P(B)$, note that the probability of coloring C with m colors is exactly

$$\sum_{x\in\mathcal{X}_{m,m}^{n-2r}}M(x).$$

Finally, as A and B are independent events, we have $P(\mathrm{Deg}(G(Y,n)) = m-1)$ is greater than

$$\left[1 - \sum_{k=1}^{m-1}\frac{1}{m^r}k^r\binom{m}{k}\sum_{x\in\mathcal{X}_{m,k}^r}M(x)(m-k)^r p_*^r\right]\left[\sum_{x\in\mathcal{X}_{m,m}^{n-2r}}M(x)\right].$$

Maximizing over r gives the desired result. For the case uniform, we just apply $p_* = 1/m$ and use the former equality together with the fact that $k!/k^n S_n^k = \sum_{x\in\mathcal{X}_{m,k}^n} M(x)$. □

Next, we compute a lower bound on the expected clique number of $G(Y,n)$. This quantity is of special interest to us since this corresponds to the maximal subset of birds whose time intervals all intersect in our bird watching story.

LEMMA 4.3.1. *Under the key assumptions in Section 4.2, the probability that an arbitrary point $x \in [n]$ lies inside the interval $I_n(i)$, is exactly $1 - q_i^x - q_i^{n-x+1} + q_i^n$, where $q_i = 1 - p_i$.*

PROOF. Fix an arbitrary $x \in [n]$ and define the event $A = \{x \in I_n(i)\}$. Note that in order for A to occur either x lies between two points with label i or x itself is labeled i. Now consider the complementary probability event, $A^c = \{x \notin I_n(i)\}$. Next define the events L, R where $L = \{$none of the points *less than or equal* to x have label $i\}$ and $R = \{$none of the points *greater than or equal to* x have label $i\}$. Note $A^c = L \cup R$ and $P(L) = q_i^x$, $P(R) = q_i^{n-x+1}$ and $P(L \cap R) = p_i^n$. Therefore, by the inclusion-exclusion principle we have,

$$P(A^c) = P(L) + P(R) - P(L \cap R) = q_i^x + q_i^{n-x+1} - q_i^n,$$

and hence $P(A) = 1 - q_i^x - q_i^{n-x+1} + q_i^n$. □

THEOREM 4.3.3. *Let ω be the random variable equal to the clique number of $G(Y, n)$, i.e., the size of the largest clique. Under the key assumptions in Section 4.2 we have that,*

$$\mathbb{E}\ \omega \geqslant \sum_{i=1}^{m} (1 - q_i^{\lceil \frac{n}{2} \rceil} - q_i^{n-\lceil \frac{n}{2} \rceil + 1} + q_i^n)$$

where $q_i = 1 - p_i$. Moreover, in the uniform case where $p_i = \frac{1}{m}$ for all $i \in [m]$, we have that

$$\mathbb{E}\ \omega \geqslant m - \left(\tfrac{m-1}{m}\right)^{\lceil \frac{n}{2} \rceil} - \left(\tfrac{m-1}{m}\right)^{n-\lceil \frac{n}{2} \rceil + 1} + \left(\tfrac{m-1}{m}\right)^n.$$

PROOF. By the preceding lemma we know that the probability that $x \in I_n(i)$ for some $x \in [n]$ is exactly $1 - q_i^x - q_i^{n-x+1} + q_i^n$. To maximize this quantity over $x \in [n]$ we will first minimize $f(x) = q_i^x + q_i^{n-x+1} - q_i^n$ over all x. Note f is convex so a simple calculus exercise shows that f is minimized at $x^* = \frac{n+1}{2}$. This can also be seen directly from the fact that f is convex and symmetric about $\frac{n+1}{2}$. When n is odd the minimizer x^* is an integer and lies in $[n]$. To handle the case when n is even, note that f is symmetric about the minimizer x^*. Therefore, when n is even, f is minimized over $[n]$ at the integers closest to x^*, which are $\frac{n}{2}$ and $\frac{n}{2} + 1$. We see then that f is minimized over $[n]$ at the point $x = \lceil \frac{n}{2} \rceil$, which holds whether n is even or odd.

Now, for $i = 1, \ldots, m$ let X_i be the indicator random variable which equals 1 if $\lceil \frac{n}{2} \rceil \in I_n(i)$ and is 0 otherwise and set $X = \sum_{i=1}^m X_i$, so X counts the number of intervals containing the point $\lceil \frac{n}{2} \rceil$. Note that the clique number $\omega \geqslant X$, so

$$\mathbb{E}\ \omega \geqslant \mathbb{E}X = \sum_{i=1}^{m} \mathbb{E}X_i = \sum_{i=1}^{m} P(X_i) = \sum_{i=1}^{m} (1 - q_i^{\lceil \frac{n}{2} \rceil} - q_i^{n-\lceil \frac{n}{2} \rceil + 1} + q_i^n).$$

The result for the uniform case follows directly by setting $p_i = \frac{1}{m}$ for every i. □

4.3.3. Occurrence of Tree/Caterpillar Graphs. In this subsection we study the probability that the graph $G(Y, n)$ forms a *tree* (recall, a tree is a a simple graph with no cycles). In [**Eck93**] Eckhoff showed that the the only interval graphs which are also trees are the so-called *caterpillar graphs.*

DEFINITION 4.3.1 ([**Eck93**]). A *caterpillar graph* is a tree T in which the removal of all leaves results in a path P_l on l vertices, called the *central path.* The vertices connecting the leaves with the central path are called *support vertices.*

Note that a caterpillar with central path $P_l = \{v_1, v_2, \ldots, v_l\}$ can be defined uniquely with the following code $\text{Cat}(k_1, k_2, ..., k_s)$, where k_i is the number of leaf vertices adjacent to v_i (see [**HS73**]); figure 4.3 illustrates some examples. Hence, we will often refer to a caterpillar graph simply by its corresponding code.

(a) $\text{Cat}(3, 2)$ (b) $\text{Cat}(3, 2, 3)$

FIGURE 4.3. Two examples of caterpillar graphs.

THEOREM 4.3.4. *Let* $\text{Cat}(k_1, k_2, ..., k_l)$ *denote a caterpillar and let* $v = \sum_{j=0}^{l}(k_j + 1)$ *denote the number of vertices in* $\text{Cat}(k_1, k_2, ..., k_l)$*. If* $v \geqslant 3$*, then nder the key assumptions in Section 4.2 and in the uniform case where* $p_i = \frac{1}{m}$ *for all* $i \in [m]$*, we have that,*

$$P(G(Y, n) \cong \text{Cat}(k_1, k_2 \ldots, k_l)) = \frac{v!}{m^n}\binom{m}{v}\sum_{x \in X} 2^{x_1 + x_2 + \ldots + x_{v-1}},$$

where $X = \{(x_1, x_2, ..., x_{v+l+1}) \in Z^{v+l+1} : x_i \geqslant 0$ *and* $\sum_{i=1}^{v+s+1} x_i = n - (v + l)\}$.

PROOF. For each $i \in [m]$ let $C_i = \{j \in [n] : Y_j = i\}$. We will refer to C_i as the *chromatic class* of i, since it is the set of points that are colored i when we view our model as a random coloring. Without loss of generality, let $C_1, C_2, ..., C_l$ denote the chromatic classes corresponding to the vertices of the central path P_l, and let $C_{i,1}, C_{i,2}, ..., C_{i,k_i}$ denote the chromatic classes corresponding to the leaves leaves adjacent to the support vertex C_i.

Now, consider the empirical supports $I_n(i) = [b_i, d_i]$ and $I_n(i,j) = [b_{i,j}, d_{i,j}]$ of the chromatic classes $C(i)$ and $C_{i,j}$, respectively. Observe that,

(i) $I_n(i) \cap I_n(j) \neq \varnothing \iff |i-j| = 1$,
(ii) $I_n(i) \cap I_n(j,k) \neq \varnothing \iff i = j$,
(iii) $I_n(i,r) \cap I_n(j,s) \neq \varnothing \iff i = j$ and $r = s$.

These restrictions imply that the birth times and death times of $I_n(i)$ and $I_n(i,j)$ will appear in the following pattern:

$$\cdots < b_i < d_{i-1} < b_{i,1} \leqslant d_{i,1} < b_{i,2} \leqslant d_{i,2} < \cdots < b_{i,k_i} \leqslant d_{i,k_i} < b_{i+1} < d_i < \ldots .$$

Thus we see that the collection of $(v+l+1)$-many intervals,

$$(-\infty, b_1),\ (b_1, d_{1,1}),\ (d_{1,1}, d_{1,2}),\ \ldots, (d_{1,k_1-1}, d_{1,k_1}),\ (d_{1,k_1}, b_2),\ (b_2, d_1),\ (d_1, d_{2,1}),\ (d_{2,1}, d_{2,2}),\ \ldots,$$

partition the $n-(v+l)$ points in $[n]$ which are not an endpoint of one of the intervals above. Observe that the intervals $(-\infty, b_1), (d_{1,k_1}, b_2), (d_{2,k_2}, b_3), \ldots, (d_{l-1,k_{l-1}}, b_l), (d_{l,k_l}, d_l)$ only intersect a single empirical support while all other intervals intersect exactly two. Therefore, the number of possible ways to place and color the remaining $n-(v+l)$ points in the $v+l+1$ intervals is

$$\sum_{x \in X} (2^{x_1+x_2+\ldots+x_{v-1}})(1^{x_v,\ldots x_{v+l+1}}) = \sum_{x \in X} 2^{x_1+x_2+\ldots+x_{v-1}},$$

where $X = \{(x_1, x_2, ..., x_{v+l+1}) \in Z^{v+l+1} : x_i \geqslant 0 \text{ and } \sum_{i=1}^{v+s+1} x_i = n-(v+l)\}$. □

Corollary 4.3.2. *Under the key assumptions in Section 4.2 and in the uniform case where $p_i = \frac{1}{m}$ for all $i \in [m]$, the probability that $G(Y,n)$ is a tree is given by,*

$$\frac{1}{m^n}\left[m + \binom{m}{2}(2^n - 2n) + \sum_{v=3}^{m}\sum_{s=1}^{v-2}\binom{v-3}{s-1}\binom{m}{v}(v)! \sum_{x \in X} 2^{x_1+x_2+\ldots+x_{v-1}}\right],$$

where $X = \{(x_1, x_2, ..., x_{v+s+1}) : \sum_{i=1}^{v+s+1} x_i = n-(v+s)\}$.

PROOF. By Theorem 4.3.4, the function $f(n, m, v, l)$ counts all the possible m-colorings of the set $[n]$ such that $G(Y, n)$ is isomorphic to the caterpillar $\text{Cat}(k_1, k_2, ..., k_l)$,

$$f(n, m, v, l) = \binom{m}{v}(v)! \sum_{x \in X} 2^{x_1 + x_2 + ... + x_{v-1}},$$

when $v = \sum_{i=1}^{l}(k_i + 1) \geqslant 3$, and $X = \{(x_1, x_2, ..., x_{v+s+1}) : \sum_{i=0}^{v+s+1} x_i = n - (v + s)\}$.

As there exist $\binom{v-3}{l-1}$ different ways to assign the k_i values so that k_1 and k_l are not zero , the equation,

$$\binom{v-3}{l-1} f(n, m, v, l),$$

counts all the m-colorings that induce caterpillars whose central path length is l and with $v - l$ leaves. Summing over all possible values of v and l, we see that the number of ways to color $[n]$ so so the $G(Y, n)$ is a caterpillar with 3 or more vertices is given by:

$$\sum_{v=3}^{m} \sum_{s=1}^{v-2} \binom{v-3}{s-1} f(n, m, v, s).$$

Finally, note that there are m different m-colorings for which $G(Y, n)$ is contains a single vertex, and $\binom{m}{2}(2^n - 2n)$ different m-colorings such that $G(Y, n)$ contains exactly two vertices connected by a single edge. Since these are all the possible interval trees, we may add all these quantities and divide by all the possible m-colorings, which gives the desired result. □

4.4. Behavior as the Number of Samples goes to Infinity.

Note that Theorem 4.3.3 implies that the expected clique number $\mathbb{E}\ \omega \to m$ as the number of samples, n, goes to infinity. Since ω takes values in $[m]$, it follows that the clique number also converges to m in probability, i.e., the probability that $G(Y, n)$ is isomorphic to the complete graph, $\mathcal{K}_m$, goes to 1 as n goes to infinity. In our bird watching story, this means that with sufficiently many observations one is almost sure to find an interval where all m species can be observed. The following theorem provides a lower bound on this convergence.

THEOREM 4.4.1. *Under the key assumptions in section 4.2, the probability that $G(Y,n)$ is the complete graph $\mathcal{K}_m$ satisfies*

$$P(G(Y,n) \cong \mathcal{K}_m) \geqslant \left(\sum_{x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}} M(x) \right)^2$$

where $\mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor} = \{(x_1, x_2, ..., x_m) \in \mathbb{N}^m : \sum_{i=1}^m x_i = \lfloor \frac{n}{2} \rfloor\}$. In the uniform case where $p_i = \frac{1}{m}$ for every $i \in [m]$, this gives that

$$P(G(Y,n) \cong \mathcal{K}_m) \geqslant \left(\frac{m!}{m^{\lfloor \frac{n}{2} \rfloor}} S_{\lfloor \frac{n}{2} \rfloor}^m \right)^2$$

where, S_n^k denotes the Stirling numbers of the second kind.

PROOF. Consider the events $L = \{$the first $\lfloor \frac{n}{2} \rfloor$ points are colored with exactly m colors$\}$ and $R = \{$the last $\lfloor \frac{n}{2} \rfloor$ points are colored with exactly m colors$\}$. For each vector $x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}$ the multinomial $M(x)$ computes the probability that there exist exactly x_i points with color i for every $1 \leqslant i \leqslant m$. Therefore, the sum over all the vectors of $\mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}$ gives us the probability of having at least one point of each color. Thus, we have that,

$$P(L) = P(R) = \sum_{x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}} M(x).$$

Finally, since $P(G(Y,n) \cong \mathcal{K}_m) \geqslant P(L \cap R)$ and since L and R are independent events, we conclude

$$P(G(Y,n) \cong \mathcal{K}_m) \geqslant P(L \cap R) = P(L)P(R) = \left(\sum_{x \in \mathcal{X}_m^{\lfloor \frac{n}{2} \rfloor}} M(x) \right)^2.$$

The result for the uniform case follows from the fact that $k!/k^n S_n^k = \sum_{x \in \mathcal{X}_{m,k}^n} M(x)$. □

Theorem 4.4.1 tells us how likely it is for the empirical nerve of n samples to form the $(m-1)$-simplex for fixed n. A related question asks what is the *first* observation n for which this occurs, i.e., if we have a sequence of observations $Y_1, Y_2, \ldots$ what is the least n such that $G(Y,n) \cong \mathcal{K}_m$? We call this quantity the *waiting time* to form the complete graph. The following theorem provides a lower bound on its expectation below.

THEOREM 4.4.2. *Let X be the random variable for the waiting time until $G(Y,n) \cong \mathcal{K}_m$, explicitly $X = \inf\{n \in \mathbb{N} : G(Y,n) \cong \mathcal{K}_m\}$. Then, under the key assumptions in Section 4.2, we have*

that,

$$\mathbb{E}X \leqslant 2\int_0^\infty \Big(1 - \prod_{i=1}^{m}(1 - e^{-p_i x})\Big)dx.$$

Moreover, in the uniform case, where $p_i = \frac{1}{m}$ *for all* $i \in [m]$*, we have that,*

$$\mathbb{E}X \leqslant 2m\sum_{i=1}^{m}\frac{1}{i}.$$

Proof. The results follow directly from the expected waiting time of the classical coupon collector problem. Let Z denote the waiting time until we have observed every label, i.e., Z is the waiting time until we have completed a collection of coupons if each coupon is an i.i.d. random variable that takes value i with probability p_i. It is known that $\mathbb{E}Z = 2\int_0^\infty \big(1 - \prod_{i=1}^{m}(1 - e^{-p_i x})\big)dx$, and in the uniform case where $p_i = \frac{1}{m}$ for all $i \in [m]$, $\mathbb{E}Z = m\sum_{i=1}^{m}\frac{1}{i}$ (see [**FS14**] for several detailed proofs). Now, note that $G(Y, n) \cong \mathcal{K}_m$ if we complete a collection, then complete a second collection, disjoint from the first. Let Z_1 denote the waiting time to complete the first collection, and let Z_2 be the additional waiting time to complete a second collection. Then $X \leqslant Z_1 + Z_2$ and Z_1, Z_2 are equal in distribution to Z, so $\mathbb{E}X \leqslant \mathbb{E}(Z_1 + Z_2) = 2\mathbb{E}Z$. □

4.4.1. Non-Stationary Results. So far we have focused on a stationary case of our model where the rate functions $f_i : [0, T] \to [0, 1]$ are constant and our samples are i.i.d. random variables. In this section we study the convergence of our model in the non-stationary case. We begin with the following definition.

Definition 4.4.1. Let $Y = \{Y_t : t \in [0, T]\}$ be a stochastic process with indexing set $[0, T]$, state space $[m]$, and rate functions $\{f_i : i \in [m]\}$. The *support graph* of Y is the interval graph of the supports the rate functions, $G_s(Y) = G(\{\text{supp}(f_i)^\circ : i \in [m]\})$.

We note that unlike the random interval graphs $G(Y, n)$, $G_s(Y)$ is non-random, as it is completely determined by Y and not a random sample. The following theorem shows that if we densely sample in $[0, T]$, then the empirical support graphs converge to the support graph of Y.

Theorem 4.4.3. *Let* $\{Y_t : t \in [0, T]\}$ *be a stochastic process with indexing set* $I = [0, T]$*, state space* $S = \{1, \ldots, m\}$*, and rate functions* $\{f_i : i \in [m]\}$*. Let* $(t_n)_{n\in\mathbb{N}}$ *be a dense sequence in* $[0, T]$

with the property that Y_{t_j} and Y_{t_k} are independent for all $j \neq k$. Then,

$$\lim_{n\to\infty} P(G(Y,n) \cong G_s(Y)) = 1$$

.

PROOF. Without loss of generality we may assume that $T = 1$. Note that in order for $G(Y,n)$ to be isomorphic to $G_s(Y)$ it is necessary and sufficient that for all edges $\{i,j\} \in G_s(Y)$ we have that $\{i,j\} \in G(Y,n)$. Let E_s denote the edge set of $G_s(Y)$, then we wish to show that

$$\lim_{n\to\infty} P\big(\bigcap_{e\in E_s} A_n^e\big) = 1,$$

where A_n^e denotes the even $\{e \in G(Y,n)\}$. Since the sequence $G(Y,n)$ is a filtration, it follows that $A_n^e \subset A_{n+1}^e$, so it is sufficient to prove that

$$P\bigg(\bigcap_{e\in G(Y,n)} \big(\bigcup_{n=1}^{\infty} A_n^e\big)\bigg) = 1.$$

And, since $G_s(Y)$ is a finite graph, it is sufficient to show that $P(\cup_{n=1}^{\infty} A_n^e) = 1$ for each $e \in E_s$ because a finite intersection of almost sure events also occurs almost surely.

Thus let $e \in E_s$ be arbitrary and without loss of generality say $e = \{i,j\}$. It follows that $\text{supp}(f_i)^\circ \cap \text{supp}(f_j)^\circ$ is non-empty and also open, being a finite intersection of open sets. Therefore, there exists an $x_0 \in [0,1]$ and an $\varepsilon > 0$ such that $[x_0 - \varepsilon, x_0 + \varepsilon] \subset \text{supp}(f_i)^\circ \cap \text{supp}(f_j)^\circ$. Therefore for all $x \in [x_0 - \varepsilon, x_0 + \varepsilon]$ we have that $f_i(x) = P(Y_x = i) > 0$, and likewise for j. Since the functions f_i, f_j are continuous we may apply the Extreme Value Theorem which gives us that $p_i := \min_{x\in[x_0-\varepsilon,x_0+\varepsilon]} f_i(x)$ exists and is strictly greater than 0, and likewise for j. Thus we can define $p^* := \min\{p_i, p_j\} > 0$.

By assumption of the density of $(t_n)_{n\in\mathbb{N}}$ we also have that there exist subsequences (t_{k_n}) and (t_{l_n}) such that for all $n \in \mathbb{N}$ we have that $t_{k_n} \in [x_0, x_0 + \varepsilon]$ and $t_{l_n} \in [x_0 - \varepsilon, x_0]$. Then by the independence of the Y_t we have that for all $n \in \mathbb{N}$,

$$P(Y_{t_{k_n}} \neq i) \leqslant (1 - p^*) \implies P\big(\bigcap_{n=1}^{\infty} (Y_{t_{k_n}} \neq i)\big) \leqslant \prod_{n=1}^{\infty} (1 - p^*) = 0,$$

and likewise for j. As a result,

$$P\Big(\bigcup_{n=1}^{\infty}(Y_{t_{k_n}}=i)\Big)=1 \implies P\bigg(\Big(\bigcup_{n=1}^{\infty}(Y_{t_{k_n}}=i)\Big)\bigcap\bigcup_{n=1}^{\infty}(Y_{t_{k_n}}=i)\Big)\bigg)=1,$$

i.e., that as $n\to\infty$ the probability that $Y_{t_n}=i$ at least once when $t_n\in[x_0,x_0+\varepsilon]$ goes to 1 as $n\to\infty$. Analogously, we get the same result for the interval $[x_0-\varepsilon,x_0]$, and so we have that x_0 lies in the intersection of the empirical supports of i and j with probability tending to 1 as n tends to infinity. This event is contained in A_n^e so we have that $\lim_{n\to\infty}P(A_n^e)=1$, as desired. □

4.5. Experimental Results

In Theorems 4.3.2, 4.3.3, and 4.4.1 we provided bounds on the likelihood of various events occurring in $G(Y,n)$ when we have a fixed number of points n and colors m. To study the usefulness of these bounds we ran simulations. For each pair (m,n) we randomly colored n points on the real line using m colors with uniform probability (each color was equally likely) then constructed the induced interval graph. We repeated this process 100 times for each pair (m,n) and plotted the percentage of the simulations where the desired event occurred. We also plotted our lower bounds from the theorems above and found that, in general, our bounds perform well for most values of m and n.

Figure 4.4 compares the bound on $P(\mathrm{Deg}(G(Y,n)=m-1)$ in Theorem 4.3.2 and the empirical approximation generated by our simulations. Figure 4.5 compares the bound on the expected clique number obtained in Theorem 4.3.3 and the empirical approximation generated by our simulations. Figure 4.6 compares the bound on the probability of the nerve being the $(m-1)$ simplex obtained in Theorem 4.4.1 and the empirical approximation generated by our simulations.

4.6. Conclusion and Further Questions

In this chapter we introduced a random interval graphs model which is well suited for applications involving the overlap patterns of chronological observations. We then proved many results about the expected behaviour of this model in a special case we call stationary. We proved how this model behaves in the general case as our number of samples points goes to infinity. This work has inspired many additional questions which remain open. We outline a few of these below.

Firstly, it is clear that the non-stationary case is better for applications where many factors may affect the distribution of our observations. We ask, *which results from stationary case can be extended to non-stationary examples? Although the general non-stationary model is intractably general, can we say something if we assume the rate functions have some nice, non-constant distributions, such as Gaussian mixtures, Poisson, or similar?*

Secondly, the story we told involving bird watching is about data samples indexed by a single parameter, say time. *But what happens when other variables such as geographical coordinates, temperature, humidity, are considered?* This would involve extending our model to higher dimensions where our empirical supports are now convex sets in $\mathbb{R}^n$. This poses many new challenges; for example, the random interval graphs are no longer sufficient to capture all the information. Instead, one needs to investigate random simplicial complexes (see [**DLH20**]).

Thirdly, Hanlon presented in [**Han82**] a characterization of all interval graphs using a unique interval representation and used characterization to enumerate all interval graphs. The analysis we presented in Theorem 4.4.1 indicates that when we use our stochastic process to generate random intervals on the line, the probability of getting an interval graph other than the complete graph goes to 0 as the number of samples n goes to infinity. *A natural challenge is to understand the decay of probabilities for different classes of graphs.* For instance, Theorem 4.3.4 describes the decaying likelihood of producing a tree.

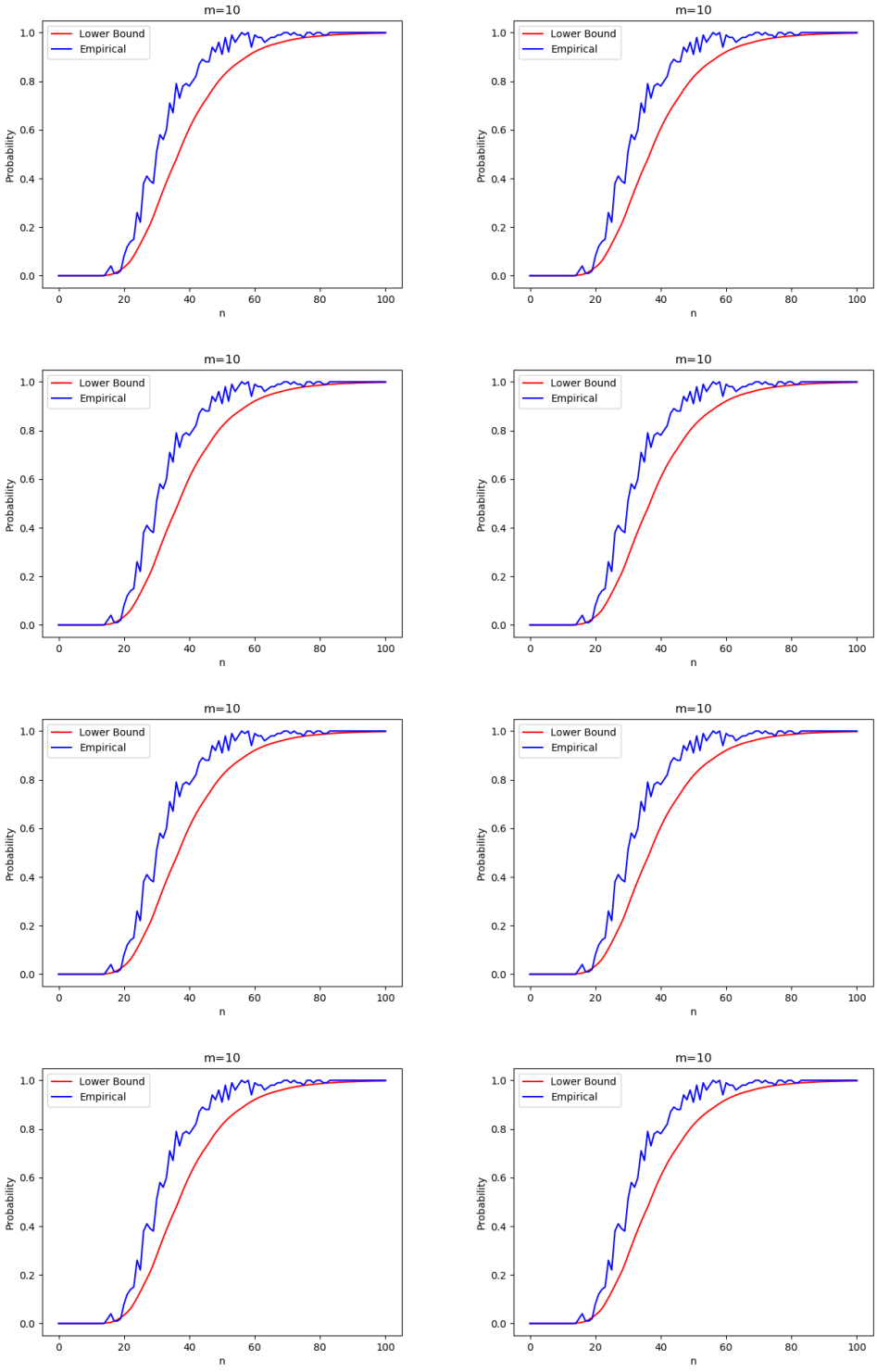


FIGURE 4.4. Probability that the maximum degree of $G(Y, n) = m - 1$.

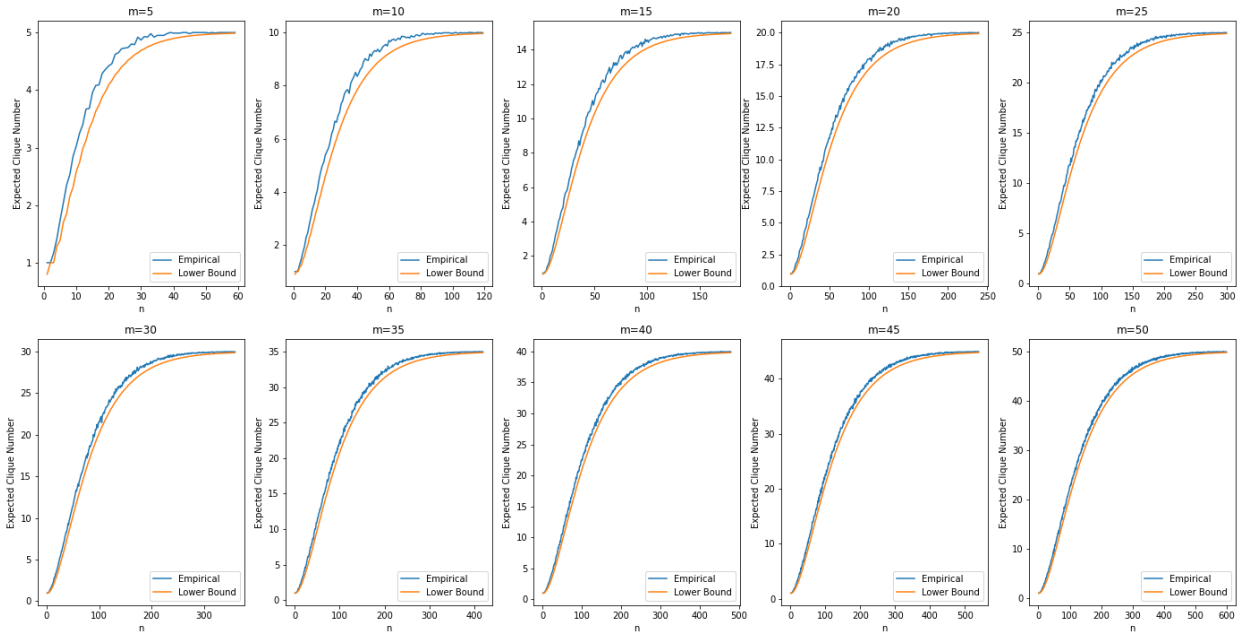


FIGURE 4.5. Expected clique number of $G(y, n)$ in uniform case.

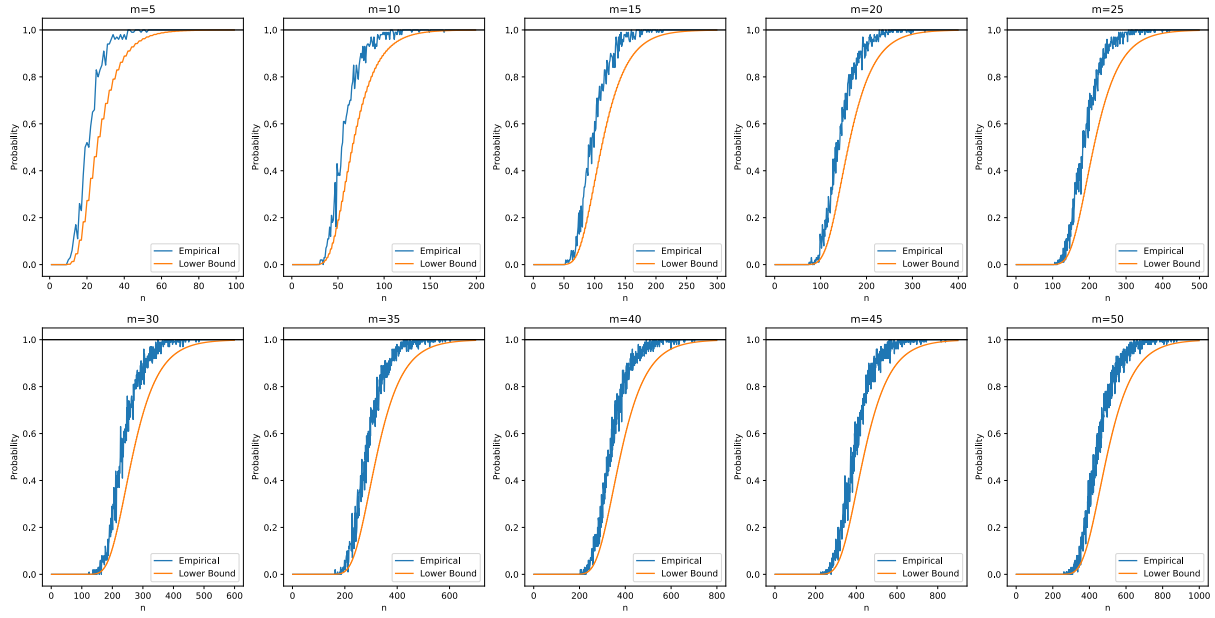


FIGURE 4.6. Probability that $G(Y, n) \cong K_m$.

Bibliography

[Bar02] A. Barvinok, *A course in convexity*, Graduate Studies in Mathematics, vol. 54, American Mathematical Society, Providence, R.I., 2002.

[BB94] M. Bennett and G. Birkhoff, *Two families of Newman lattices*, Algebra Universalis **32** (1994), 115–144.

[BB05] A. Björner and F. Brenti, *Combinatorics of coxeter groups*, Springer, Berlin and Heildelberg, Germany, 2005.

[BDJ+13] J. Burns, E. Dolzhenko, N. Jonoska, T. Muche, and M. Saito, *Four-regular graphs with rigid vertices associated to dna recombination*, Discrete Applied Mathematics **161** (2013), no. 10, 1378–1394.

[Ber71] C. Berge, *Principles of combinatorics*, 1st ed., Academic Press, New York, NY and London, UK, 1971.

[BG22] B. Brück and A. Garin, *Stratifying the space of barcodes using Coxeter complexes*, Journal of Applied and Computational Topology (2022).

[BHH+21] V. Berisha, P. Hahn, S. Hahn, G. Dasarathy, P. Turaga, and J. Liss, *Digital medicine and the curse of dimensionality*, npj Digital Medicine **4** (2021), no. 153.

[Bj6] A. Björner, *Topological methods*, p. 1819–1872, MIT Press, Cambridge, MA, USA, 1996.

[BJS15] J. Burns, N. Jonoska, and M. Saito, *Genus ranges of chord diagrams*, Journal of Knot Theory and its Ramifications **24** (2015), no. 4.

[BNN+18] J. Braun, L. Nabergall, R. Neme, L. F. Landweber, M. Saito, and N. Jonoska, *Russian doll genes and complex chromosome rearrangements in oxytricha trifallax*, G3 Genes—Genomes—Genetics **8** (2018), no. 5, 1669–1674.

[BW20] P. Bubenik and A. Wagner, *Embeddings of persistence diagrams into hilbert spaces*, Journal of Applied and Computational Topology (2020), 339–351.

[Car09] G. E. Carlsson, *Topology and data*, Bull. Amer. Math. Soc. **46** (2009), 255–308.

[CDG+21] J. Curry, J. DeSha, A. Garin, K. Hess, L. Kanari, and B. Mallery, *From trees to barcodes and back again ii: combinatorial and probabilistic aspects of a topological inverse problem*, 2021, 2107.11212, arXiv:2107.11212.

[CFJ+19] D. Cruz, M. Ferrari, N. Jonoska, L. Nabergall, and M. Saito, *Insertions yielding equivalent double occurrence words*, Fundamenta Informaticae **171** (2019), no. 1-4, pp. 113–132.

[CKM79] J. E. Cohen, J. Komlós, and T. Mueller, *The probability of an interval graph, and why it matters*, Proc. of the Symposia in Pure Math., vol. 34, American Mathematical Society, 1979, pp. 97–115.

[Cou08] B. Courcelle, *Circle graphs and monadic second-order logic*, Journal of Applied Logic **6** (2008), no. 3, 416–442.

[CSW16] N. Caspard, L. Santocanale, , and F. Wehrung, *Permutohedra and associahedra*, Lattice Theory: Special Topics and Applications (G. Grätzer and F. Wehrung, eds.), vol. 2, Springer International Publishing, Cham, Switzerland, 2016, pp. 215–286.

[DF04] D. S. Dummit and R. M. Foote, *Abstract algebra*, 3rd ed ed., Wiley, New York, 2004.

[DHJ13] P. Diaconis, S. Holmes, and S. Janson, *Interval graph limits*, Ann. of Comb. **17** (2013), no. 1, 27–52.

[Die17] R. Diestel, *Graph theory*, 5 ed., Springer, 2017.

[DLH20] J. De Loera and T. Hogan, *Stochastic Tverberg theorems with applications in multiclass logistic regression, separability, and centerpoints of data*, SIAM Journal on Mathematics of Data Science **2** (2020), 1151–1166.

[DMW22] A. Dedieu, R. Mazumder, and H. Wang, *Solving l1-regularized svms and related linear programs: Revisiting the effectiveness of column and constraint generation*, J. Mach. Learn. Res. **23** (2022), no. 1.

[Eck93] J. Eckhoff, *Extremal interval graphs*, J. Graph Theory **17** (1993), no. 1, 117–127.

[EGP66] P. Erdös, A. W. Goodman, and L. Pósa, *The representation of a graph by set intersections*, Canadian Journal of Mathematics **18** (1966), 106–112.

[EH08] H. Edelsbrunner and J. Harer, *Persistent homology - a survey*, Surveys on discrete and computational geometry: Twenty years later. AMS-IMS-SIAM Joint Summer Research Conference (Providence, RI) (J. Goodman, J. Pach, , and R. Pollack, eds.), vol. 453, American Mathematical Society, 2008.

[EH10] H. Edelsbrunner and J. Harer, *Computational topology: An introduction*, 01 2010.

[ELZ02] H. Edelsbrunner, D. Letscher, and A. Zomorodian, *Topological persistence and simplification*, Discrete Computational Geometry (2002), 511–533.

[ER59] P. Erdős and A. Renyi, *On random graphs i.*, Publ. Math. Debrecen **6** (1959), no. 290-297, 18.

[FG65] D. R. Fulkerson and O. A. Gross, *Incidence matrices and interval graphs.*, Pacific Jour. Math. **15** (1965), no. 3, 835–855.

[FGT92] P. Flajolet, D. Gardy, and L. Thimonier, *Birthday paradox, coupon collectors, caching algorithms and self-organizing search*, Discrete Appl. Math. **39** (1992), no. 3, 207 – 229.

[Fis85] P. C. Fishburn, *Interval orders and interval graphs : a study of partially ordered sets*, New York : Wiley, 1985 (English).

[FS14] M. Ferrante and M. Saltalamacchia, *The coupon collector's problem*, MATerials MATematics **2014** (2014), no. 2, 35.

[GH64] P. C. Gilmore and A. J. Hoffman, *A characterization of comparability graphs and of interval graphs*, Canadian Jour. of Math. **16** (1964), 539–548.

[Ghr08] R. Ghrist, *Barcodes: The persistent topology of data*, Bull. Amer. Math. Soc. **45** (2008), 61–75.

[Gib11] A. Gibson, *Homotopy invariants of gauss words*, Mathematische Annalen **349** (2011), no. 4, 871–887.

[GKO20] A. Guterman, E. Kreines, and N. Ostroukhova, *Double occurrence words: Their graphs and matrices*, Journal of Mathematical Sciences **249** (2020), no. 2.

[Gol04] M. C. Golumbic, *Chapter 8 - interval graphs*, Algorithmic Graph Theory and Perfect Graphs (M. Golumbic, ed.), Ann. of Discrete Math., vol. 57, Elsevier, 2004, pp. 171 – 202.

[GS78] I. Gessel and R. P. Stanley, *Stirling polynomials*, Journal of Combinatorial Theory, Series A **24** (1978), no. 1, 24–33.

[Han82] P. Hanlon, *Counting interval graphs*, Trans. of the Amer. Math. Soc. **272** (1982), no. 2, 383–426.

[Hat02] A. Hatcher, *Algebraic topology*, Cambridge University Press, 2002.

[HH07] O. Hammer and D. A. T. Harper, *Paleontological data analysis*, John Wiley and Sons, Ltd, 2007.

[HM21] H. Hoang and T. Mütze, *Combinatorial generation via permutation languages. ii. lattice congruences*, Israel Journal of Mathematics **244** (2021), 359–417.

[HS73] F. Harary and A. Schwenk, *The number of caterpillars*, Discrete Mathematics **6** (1973), no. 4, 359–365.

[Ili17] V. Iliopoulos, *A study on properties of random interval graphs and Erdös Renyi graph* $\#(n, 2/3)$, Jour. of Discrete Math. Sci. and Cryptography **20** (2017), no. 8, 1697–1720.

[Jan08] S. Janson, *Plane recursive trees, Stirling permutations and an urn model*, Discrete Mathematics & Theoretical Computer Science Proceedings of Fifth Colloquium on Mathematics and Computer Science **AI** (2008).

[JNS17] N. Jonoska, L. Nabergall, and M. Saito, *Patterns and distances in words related to dna rearrangement*, Fundamenta Informaticae **154** (2017), no. 1-4, 225–238.

[JNT22] E. Jacquard, V. Nanda, and U. Tillmann, *The space of barcode bases for persistence modules*, Journal of Applied and Computational Topology (2022).

[JSW90] J. Justicz, E. Scheinerman, and P. Winkler, *Random intervals*, The Amer. Math. Monthly **97** (1990), no. 10, 881–889.

[KDS+18] L. Kanari, P. Dłotko, M. Scolamiero, R. Levi, J. Shillcock, K. Hess, and H. Markram, *A topological representation of branching neuronal morphologies*, Neuroinformatics **16** (2018), 3–13.

[KGH20] L. Kanari, A. Garin, and K. Hess, *From trees to barcodes and back again: theoretical and statistical perspectives*, Algorithms **13** (2020), no. 12.

[LB62] C. Lekkeikerker and J. Boland, *Representation of a finite graph by a set of intervals on the real line*, Fundamenta Mathematicae **51** (1962), 45–64.

[Liu23] S. Liu, *Statistics on trapezoidal words and k-inversion sequences*, Discrete Applied Mathematics **325** (2023), 1–8.

[Lot02] M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2002.

[Mat02] J. Matousek, *Lectures on discrete geometry*, Graduate Texts in Mathematics, Springer New York, 2002.

[MQYY23] S. Ma, H. Qi, J. Yeh, and Y. Yeh, *Stirling permutation codes*, 2023, 2210.11372, arXiv:2210.11372.

[Pie20] T. Piesk, *Symmetric group 4; cayley graph 1,2,6 (1-based).png*, 2020, File: Symmetric group 4; Cayley graph 1,2,6 (1-based).png.

[Pip98] N. Pippenger, *Random interval graphs.*, Random Struct. Algorithms **12** (1998), no. 4, 361–380.

[Pit93] J. Pitman, *Probability*, Springer New York, 1993.

[Rea06] N. Reading, *Cambrian lattices*, Advances in Mathematics **205** (2006), no. 2, 313–353.

[Rea12] N. Reading, *From the tamari lattice to cambrian lattices and beyond*, pp. 293–322, Springer Basel, Basel, 2012.

[Rea16a] N. Reading, *Finite coxeter groups and the weak order*, Lattice Theory: Special Topics and Applications (G. Grätzer and F. Wehrung, eds.), vol. 2, Springer International Publishing, Cham, 2016, pp. 489–561.

[Rea16b] ———, *Lattice theory of the poset of regions*, Lattice Theory: Special Topics and Applications (G. Grätzer and F. Wehrung, eds.), vol. 2, Springer International Publishing, Cham, 2016, pp. 399–487.

[Rio76] J. Riordan, *The blossoming of schröder's fourth problem*, Acta Mathematica **137** (1976), 1–16.

[Ros96] S. Ross, *Stochastic processes*, 5 ed., Wiley, 1996.

[San07] L. Santocanale, *On the join dependency relation in multinomial lattices*, Order **24** (2007), 155–179.

[Sch54] H. V. Schelling, *Coupon collecting for unequal probabilities*, The Amer. Math. Monthly **61** (1954), no. 5, 306–311.

[Sch88] E. R. Scheinerman, *Random interval graphs*, Combinatorica **8** (1988), no. 4, 357–371.

[Sch90] ———, *An evolution of interval graphs.*, Discrete Math. **82** (1990), no. 3, 287–302.

[Sta11] R. P. Stanley, *Enumerative combinatorics: Volume 1*, 2nd ed., Cambridge University Press, USA, 2011.

[STZ09] B. Shtylla, L. Traldi, and L. Zulli, *On the realization of double occurrence words*, Discrete Mathematics **309** (2009), no. 6, 1769–1773.

[SW16] L. Santocanale, , and F. Wehrung, *Generalizations of the permutohedron*, Lattice Theory: Special Topics and Applications (G. Grätzer and F. Wehrung, eds.), vol. 2, Springer International Publishing, Cham, Switzerland, 2016, pp. 287–397.

[TS20] K. Turner and G. Spreemann, *Same but different: Distance correlations between topological summaries*, Topological Data Analysis (Cham) (N. A. Baas, G. E. Carlsson, G. Quick, M. Szymik, and M. Thaule, eds.), Springer, 2020, pp. 459–490.

[Tur04] V. Turaev, *Virtual strings*, Annales de l'Institut Fourier **54** (2004), no. 7, 2455–2525.

[TW15] E. Tsukerman and L. Williams, *Bruhat interval polytopes*, Advances in Mathematics **285** (2015), 766–810.

[ZC05] A. Zomorodian and G. Carlsson, *Computing persistent homology*, Discrete Comput. Geom. **33** (2005), 249–274.