HHS Public Access

Author manuscript

Found Data Sci. Author manuscript; available in PMC 2023 January 27.

Published in final edited form as:

Found Data Sci. 2022 June ; 4(2): 165–216. doi:10.3934/fods.2022002.

ASPECTS OF TOPOLOGICAL APPROACHES FOR DATA SCIENCE

Jelena Grbić[†],
School of Mathematical Sciences, University of Southampton, Southampton, UK

Jie Wu[*,†],
School of Mathematical Sciences, Center of Topology and Geometry based Technology, Hebei Normal University, Yuhua District, Shijiazhuang, Hebei, 050024 China

Yanqi Lake Beijing Institute of Mathematica Sciences, Yanqihu, Huairou District, Beijing, 101408 China

Kelin Xia[†],
School of Physical and Mathematical Sciences, Nanyang Technological University, SPMS-MAS-05-18, 21 Nanyang Link, 1, Singapore 63737

Guo-Wei Wei[†]
Department of Mathematics, Department of Computer Science and Engineering, Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA

Abstract

We establish a new theory which unifies various aspects of topological approaches for data science, by being applicable both to point cloud data and to graph data, including networks beyond pairwise interactions. We generalize simplicial complexes and hypergraphs to super-hypergraphs and establish super-hypergraph homology as an extension of simplicial homology. Driven by applications, we also introduce super-persistent homology.

Keywords

Topological data analysis; hypergraph; super-hypergraph; persistent homology; super persistent homology; simplicial complex; Delta set; scoring scheme

1. Introduction.

Topological data analysis (TDA) is a new-born research area mainly stemming from the pioneering works on persistent homology carried in [57, 148] and the landmark paper of Gunnar Carlsson [36] published in 2009. Under its rapid development, TDA has achieved various successful applications in many areas of sciences and technologies such as material science [81, 119, 91, 103], 3D shape analysis [126, 129], multivariate time series analysis [124], biology [29, 32, 31, 90, 140, 105], sensor networks [48], scientific visualization [128],

[*]Corresponding author: Jie Wu. wujie@bimsa.cn.
[†]JG, JW, KX and GW should be considered joint first author.

image analysis [33, 109, 125, 17, 64], dynamics systems [100], etc. A great number of TDA softwares have been developed, including JavaPlex [127], Perseus [104], Dionysus [1], jHoles [21], GUDHI [97], Ripser [13], PHAT [15], DIPHA [14], and R-TDA package [61]. The results from TDA can be visualized by persistent diagram [101] and persistent barcode [66], and further transformed into different representations that are suitable for machine learning models, such as Betti curve [118], persistent landscape [25, 24], persistent image [2], persistent path and signature [43], persistent codebook [147], persistent 2D matrices [29], and others [3, 85, 29, 32, 105]. TDA-based machine learning models [40, 111] have been used in various areas, including image analysis [9, 79, 112, 95, 108, 67], shape analysis [22, 145, 92, 78, 143], time-series data analysis [124, 6, 144, 133, 130], computational biology [109, 50, 29, 32], noise data [107], sphere packing [117], language analysis [146], etc. Other the traditional persistent homology, other TDA models have been proposed for the detailed characterization of the data, such as persistent local homology [18, 62, 16, 19, 5], persistent cohomology [49], multidimensional persistent homology [35, 34, 45, 37], element-specific persistent homology [29, 105], persistent functions [20], persistent spectral [132, 98], persistent Ricci curvatures [134, 135], etc. The wide applications of TDA have made topology as one of the most commonly used mathematical tools in Data Science [120, in Section 1.3]. A summary of TDA and TDA-based learning models can be found in Figure 1.

In a survey paper [41], Chazal and Michel outlined a pipeline that stresses the role of topology and geometry in data science:

- **i.** Input data is given in the form of a finite set of points coming with a notion of distance;
- **ii.** A “continuous shape” is built on top of the input data: this results in a structure over the data;
- **iii.** Topological and geometric information is extracted from the structure;
- **iv.** The topological and geometric information is the output of the analysis and forms the new representation of the data, allowing for an in-depth modeling of the original data.

Such an approach can be naturally applied to point cloud data with a drawback that it can not be immediately or directly applied to non-Euclidean data such as graphs for abstract relationship.

The purpose of this article is to provide a new theory that unifies various aspects of topological approaches for data science that is suitable for both point cloud data and graphic data. In our setting, we explore topological structures on graph data with scoring schemes. The popular persistent homology can be obtained as special cases of our more general theory from a natural transformation from point cloud data to graphic data with scoring schemes.

We start with a graph, which is the working graph for the data analytic purpose. Our approach consists of the following steps:

A. We introduce a homology theory of a collection of subgraphs of the working graph, which is a canonical extension of simplicial homology theory. Briefly, this homology theory will canonically obtain "topological invariants" for collections of subgraphs associated to sample data or experimental data.

B. We assign a *scoring scheme* on the working graph G, where a scoring scheme is a function from the set of subgraphs of G to the set of real numbers. The scoring scheme induces a persistence on homologies in (A), which in this article will be called as *super persistent homology*, as well as its derived *topological features* such as super-persistence diagrams and super-persistence modules for data analytics.

C. The current persistent homology on point cloud data can be deduced from (A) and (B). Hence our approach is suitable for performing topological data analysis on both graphic data and point cloud data.

The pineline of our super persistent homology is as follows:

1. The input is assumed to be a finite (or infinite) graph G with

- **i.** A scoring scheme and
- **ii.** A selection of subgraphs.

The definition of the scoring scheme on the data is usually given as an input or guided by applications. It is however important to notice that the choice of a scoring scheme may be critical to revealing interesting topological and geometric features of the data. The selection of subgraphs on the data is also usually given as an input or guided by the applications at hand. Again it is important to notice that the selection of subgraphs may be critical to revealing interesting topological and geometric features of the data.

2. An "abstract geometry-like" shape is built on top of the data in order to highlight the underlying topological structure. This is a nested family of *super-hypergraphs* filtered by the scoring scheme that reflects the structure of the data at different scales. Super-hypergraphs can be seen as higher dimensional generalizations of neighboring graphs that are classically built on top of data in many standard data analysis or learning algorithms. The challenge here is to define such structures that reflect relevant information about the structure of the data and that can be effectively constructed and manipulated in practice.

3. The extracted topological and geometric information provides new families of features and descriptors of the data. These can be used to better understand the data or they can be combined with other kinds of features for further analysis and machine learning tasks. Demonstrating the added-value and the complementarity (with respect to other features) of the information provided by super persistent homology is an important issue at this step.

4. Adjust the choice of scoring scheme and the selection of subgraphs to get better features and descriptors of the data.

5. One can repeat the procedure on the choices of scoring scheme and selections of subgraphs to obtain the best suitable features and descriptors of the data.

For analyzing both point cloud data and graph data in a unified way, we convert a point cloud data into a graph data by adding exact one edge to each pair of the points in the data to form a fully connected graph. There have been extensive explorations on topological and categorical structures on graphs. In Appendix A, we give a brief review of simplicial complexes constructed from graphs. In addition to the commonly-used construction of a clique complex, here we give a short discussion on a famous construction called *neighborhood complex* introduced by Lovász in his foundational work [94] in the area of topological combinastorics. Let G be a graph. Define the *neighborhood complex* $\mathcal{N}(G)$ to be an abstract simplicial complex whose vertices are the vertices of G and whose simplices are those subsets of the vertex set $V(G)$ which has a common neighbor. Lovász Theorem [94, Theorem 2] states that if $\mathcal{N}(G)$ is $(k-2)$-connected, then G is not k-colorable[1]. As Lovász remarked [94, p.320], in the case $k=2$, the converse statement is also true, and so $\mathcal{N}(G)$ is connected if and only if G is not bipartite. Note that a simplicial complex is connected if and only if its 0-th Betti number is 1. If we choose Lovász' neighborhood complex as the construction filtered by a scoring scheme, then the resulting barcodes of the 0-th super-persistent homology introduced in Section 3 would immediately give the bipartite information for the level subgraphs of the filtration.

The geometric realization of neighborhood complex (of a graph) is quite different from that of the clique complex, see Appendix A for details. This indicates that one could have different topological structures from the same working graph.

It is also possible that simplicial complexes model data from practical problems. According to the review article [12] supported by hundreds of references, extensive research has been recently taken on the networks beyond pairwise interactions. A simplicial complex (hypergraph) can be naturally constructed by taking the nodes having a group-interaction as a simplex (hyperedge).

Hypergraph is a preferred model in some practical problems. Although simplicial complexes overcome some of the problems encountered by other lower dimensional representations, they are still quite limited by the requirement on the existence of all subfaces. In some cases, such as group interactions in social systems, this constraint is too restrictive. In other cases, such as author collaborations in scientific papers and gene pathways, the inclusion constraint can be less easily justified. Hypergraphs provide the most general and unconstrained description of higher-order interactions, see [12, paragraphs 2–4, page 7].

For graph data having higher-order interactions, it is natural to take the collections of subgraphs that have group-interactions. This arises a mathematical question: *Given $\mathcal{H}$ a collection of subgraphs of a working graph, how to introduce topology on $\mathcal{H}$ with as less constraints as possible on $\mathcal{H}$?*

[1]As a consequence of this theorem, he solved the Kneser conjecture in combinatorics [94 Theorem 1].

In this article, we are going to answer this question. Here we give some observations and brief ideas. For graph data, it is likely that two different subgraphs share the same nodes. This observation implies that the notion of simplicial complexes does not work well for the collections of subgraphs due to the fact that any simplex in a simplicial complex must be uniquely determined by its vertices (nodes). The notion of Δ-set in algebraic topology, which is a generalization of simplicial complex, can solve this problem. Roughly, a Δ-set X is a graded (multi-layered) set labeled by X_0, X_1, X_2, ⋯, where X_n can be intuitively viewed as the set of n-dimensional simplices. The structure of a Δ-set X is given by face operators, that is, we assign $n+1$ face operators labeled by d_0, d_1, …, d_n as functions from the set of n-dimensional simplicies to the set of $(n-1)$-dimensional simplices, satisfying the Δ-identity (the matching rule of faces) that $d_i d_j = d_j d_{i+1}$ for $i \geq j$. The geometric realization of a Δ-set is Δ-*complex*, a notion defined in the popular textbook of algebraic topology [80]. A Δ-set can be described in terms of feed-forward neural networks, see Subsection 3.5. Using the notion of Δ-sets, we are allowed to select two or more subgraphs treated as n-simplices sharing the same vertices. For addressing *as less constraints as possible* in the question, we introduce the notion of *super-hypergraph* in Subsection 2.3, which is defined as a graded (multi-layered) subset of a Δ-set. If a Δ-set is given by an oriented simplicial complex, then our definition of super-hypergraph coincides with a hypergraph. Roughly speaking, a super-hypergraph is an extension of a hypergraph that allows hyperedges to form a multiset. An important aspect in the present article is that simplicial homology can be naturally extended to a homology theory on super-hypergraphs as described in Section 2, and so there are topological invariants (in terms of homology) on super-hypergraphs.

To introduce persistence on graph data, we propose to use the notion of a scoring scheme, which is a real valued function on finite subgraphs of the working graph. As discussed in Subsection 3.2, the classical persistent homology can be converted into super-persistent homology under the notion of scoring schemes. Hence super-persistent homology is a novel topological approach that can be applied to broader objects in data science.

Mathematically, the main results of the article are Theorems 2.7, 2.20 and 3.7. Theorem 2.7 states that hypergraph homology does not depend on the choice of orientation, which shows that the hypergraph homology provides the invariants on the intrinsic structure of hypergraph. The Mayer-Vietoris sequence is one of the fundamental tools for computing simplicial homology. Theorem 2.20 gives an analogue of the classical Mayer-Vietoris sequence for super-hypergraphs. Theorem 3.7 gives the structure theorem on super-persistent homology.

We should point out that it is commonly known (such as the textbook of algebraic topology [80]) that the computation of simplicial homology can be largely simplied using the notion of Δ-complex, if the geometric shape can be homotopically deformed. Also the complexity of computing super-hypergraph homology is essentially the same as that of computing the simplicial homology of Δ-sets.

It should be pointed out that even though this article is a theoretic framework, it has deep roots in application. One of the main motivations is from the drug design, in particular the analysis of binding affinities between proteins and ligands, i.e., how powerful is a

drug (ligand) for a certain disease (protein-related). Figure 2 illustrates the topological representations for the protein-ligand complex (ID: 3E6Y). Biologically, the interactions between the ligand and the protein are key to the drug potency and efficacy. Graph models (as in Figure 2(B)) is widely used to characterize the atomic interactions. A major drawback for graph models is that they only characterize pair-wise interactions and can not be used in many-body interaction characterization. Recently, we have used hypergraph models to describe various types of interactions (beyond pair-wise ones) for biomolecular interactions (Figure 2(C)). Essentially, many-body interactions can be modeled as hyperedges. It has been found that hypergraph models have great advantages over graph models in molecular representation. More details can be found in Section 4. A further generalization of hypergraphs to super-hypergraphs (Figure 2(D)) provides us more flexibility to characterize the complicated topology within each individual hyperedge. We believe that the super-hypergraph model provides an upgraded topological approach to data science, and can help to foster further interactions between topology and data science. The content of the article is organized in the following way.

2. Homology theory on super-hypergraphs.

Recently, homology of hypergaphs has opened new avenues for using topological tools in data analysis. Hypergraphs have been used for data analytics in various areas of sciences from social networks to molecular bioscience. The notion of hypergraph can be generalized as super-hypergraph, see Subsection 2.3. These objects are important for understanding the different explorations of topological structures on spaces of subgraphs. This setting realizes our aim to establish a unified approach to explore data science using topological combinatorics. The purpose of this section is to establish a homology theory of super-hypergraphs as a natural extension of simplicial homology and homology of hypergraphs.

2.1. Algebraic lemmas.

The following algebraic tools will be needed to define a homology theory of super-hypergraphs. Although we will only make use of chain complexes of abelian groups, we note that using simplicial group models, homotopy groups can be combinatorially defined using Moore chain complexes, which are chain complexes of possibly non-abelian groups [47, 139]. There are many studies of the homotopy type of topological structures of subgraphs as indicated in the references in Appendix A. Therefore, we consider chain complexes of possibly non-abelian groups so that the results in this subsection may be relevant for future research.

A graded group $G_* = \{G_n\}_{n \in \mathbb{Z}}$ is a sequence of groups G_n indexed by the integers. A graded subgroup $G'_* = \{G'_n\}_{n \in \mathbb{Z}}$ of $G_* = \{G_n\}_{n \in \mathbb{Z}}$ is a sequence of subgroups G'_n such that $G'_n \leq G_n$ for $n \in \mathbb{Z}$. A chain complex G_* of groups is a graded group G_* with a group homomorphism $\partial_n = \partial_n^{G_*}: G_n \to G_{n-1}$ for $n \in \mathbb{Z}$ such that the composite

$$\partial_{n-1} \circ \partial_n : G_n \to G_{n-2}$$

is the trivial homomorphism. Let us emphasise that in this definition we do not require G_n to be abelian. A subcomplex C_* of G_* is a graded subgroup C_* of G_* such that

$$\partial_n^{G_*}(C_n) \subseteq C_{n-1}$$

for each $n \in \mathbb{Z}$. So C_* together with the restrictions $\partial_n^{G_*}\big|_{C_n} : C_n \to C_{n-1}$ forms a chain complex.

Definition 2.1. Let G_* be a chain complex of groups and let D_* be a graded subgroup of G_*. Define

$$\sup_*^{G_*}(D_*) = \cap \{C_* \mid D_n \leq C_n \text{ for } n \in \mathbb{Z}, \text{ and } C_* \text{ is a subcomplex of } G_*\}$$

$$\inf_*^{G_*}(D_*) = \prod \{E_* \mid E_n \leq D_n \text{ for } n \in \mathbb{Z}, \text{ and } E_* \text{ is a subcomplex of } G_*\}.$$

For simplicity, if the embedding of $D_* \subseteq G_*$ is clear, we denote $\sup_*^{G_*}(D_*)$ by $\sup_*(D_*)$ and $\inf_*^{G_*}(D_*)$ by $\inf_*(D_*)$.

Proposition 2.2. *Let G_* be a chain complex of groups and let D_* be a graded subgroup of G_*. Then*

1. $\sup_*(D_*)$ *is the smallest subcomplex of G_* containing D_*. Moreover,*

$$\sup_n(D_*) = D_n \cdot \partial_{n+1}^{G_*}(D_{n+1})$$

is the product of D_n and $\partial_{n+1}^{G_*}(D_{n+1})$.

2. $\inf_*(D_*)$ *is the largest subcomplex of G_* contained in D_*. Moreover,*

$$\inf_n(D_*) = D_n \cap \partial_n^{-1}(D_{n-1})$$

is the intersection of D_n and $\partial_n^{-1}(D_{n-1})$.

Proof. (1) The first part follows from the definition. Let

$$\widetilde{D}_n = D_n \cdot \partial_{n+1}^{G_*}(D_{n+1})$$

for $n \in \mathbb{Z}$. Let C_* be any subcomplex of G_* such that $D_n \leq C_n$ for each $n \in \mathbb{Z}$. Then

$$\partial_{n+1}^{G_*}(D_n+1) \le \partial_{n+1}^{G_*}(C_n+1) \le C_n$$

and so

$$\widetilde{D}_n = D_n \cdot \partial_{n+1}^{G_*}(D_n+1) \le C_n .$$

Thus $\widetilde{D}_*$ is a graded subgroup of C_* for any subcomplex C_* of G_* with $D_n \le C_n$ for $n \in \mathbb{Z}$, and so $\widetilde{D}_*$ is a graded subgroup of $\sup_*(D_*)$. Notice that

$$\begin{aligned}\partial_n^{G_*}(\widetilde{D}_*) &= \partial_n^{G_*}\left(D_n \cdot \partial_{n+1}^{G_*}(D_n+1)\right)\\ &\le \partial_n^{G_*}(D_n) \cdot \partial_n^{G_*}\left(\partial_{n+1}^{G_*}(D_n+1)\right)\\ &= \partial_n^{G_*}(D_n)\\ &\le \widetilde{D}_{n-1}.\end{aligned}$$

Hence $\widetilde{D}_*$ is a subcomplex of G_* containing D_*, and so $\sup_*(D_*) = \widetilde{D}_*$.

(2) The first part follows from the definition. Let

$$\check{D}_n = D_n \cap \partial_n^{-1}(D_{n-1}).$$

Let $x \in \check{D}_n$. Then $\partial_n(x) \in D_{n-1}$ because $x \in \partial_n^{-1}(D_{n-1})$, and

$$\partial_n(x) \in \partial_{n-1}^{-1}(D_{n-2})$$

because $\partial_{n-1}(\partial_n(x)) = 1 \in D_{n-2}$. Thus $\partial_n(x) \in \check{D}_{n-1}$. It follows that $\check{D}_*$ is a subcomplex of G_* contained in D_*. Hence

$$\check{D}_* \le \inf_*(D_*).$$

Let E_* be any subcomplex of G_* such that $E_n \le D_n$ for $n \in \mathbb{Z}$. Then

$$E_n \le \partial_n^{-1}(E_{n-1}) \le \partial_n^{-1}(D_{n-1}).$$

Thus $E_n \le \check{D}_n$ for $n \in \mathbb{Z}$. It follows that $\inf_*(D_*) \le \check{D}_*$. This finishes the proof.□

Let G_* be a chain complex of groups. The homology of G_* is defined as the right cosets

$$H_n(G_*) = \mathrm{Ker}\left(\partial_n^{G_*}\right) / \partial_{n+1}^{G_*}(G_n+1).$$

Proposition 2.3. *Let G_* be a chain complex of groups and let D_* be a graded subgroup of G_*.*

1. *The inclusion*

$$\inf_*(D_*) \to \sup_*(D_*)$$

induces an injective map on homology.

2. *Suppose that $\partial^{G_*}_{n+1}(D_{n+1})$ is contained in the normalizer of D_n in G_n for each n. Then the inclusion*

$$\inf_*(D_*) \to \sup_*(D_*)$$

induces an isomorphism on homology. In particular, if D_n is normal in G_n for $n \in \mathbb{Z}$, then the inclusion $\inf_*(D_*) \to \sup_*(D_*)$ induces an isomorphism on homology.

Proof. (1) From Proposition 2.2 (2),

$$H_n(\inf_*(D_*)) = \left(D_n \cap \partial_n^{-1}(D_{n-1}) \cap \mathrm{Ker}\left(\partial_n^{G_*}\right)\right) / \partial_{n+1}\left(D_{n+1} \cap \partial_{n+1}^{-1}(D_n)\right)$$

as right cosets. Since $\mathrm{Ker}\left(\partial_n^{G_*}\right) \leq \partial_n^{-1}(D_{n-1})$, we have

$$D_n \cap \partial_n^{-1}(D_{n-1}) \cap \mathrm{Ker}\left(\partial_n^{G_*}\right) = D_n \cap \mathrm{Ker}\left(\partial_n^{G_*}\right).$$

We also claim that

$$\partial_{n+1}\left(D_{n+1} \cap \partial_{n+1}^{-1}(D_n)\right) = D_n \cap \partial_{n+1}(D_{n+1}).$$

Clearly, $\partial_{n+1}\left(D_{n+1} \cap \partial_{n+1}^{-1}(D_n)\right) \leq D_n \cap \partial_{n+1}(D_{n+1})$.

Let $x \in D_n \cap \partial_{n+1}(D_{n+1})$ and let $y \in D_{n+1}$ such that $\partial_{n+1}(y) = x$. Then $y \in D_{n+1} \cap \partial_{n+1}^{-1}(D_n)$. Thus $x \in \partial_{n+1}\left(D_{n+1} \cap \partial_{n+1}^{-1}(D_n)\right)$. Hence $\partial_{n+1}\left(D_{n+1} \cap \partial_{n+1}^{-1}(D_n)\right) = D_n \cap \partial_{n+1}(D_{n+1})$ and so

$$H_n(\inf_*(D_*)) = \left(D_n \cap \mathrm{Ker}\left(\partial_n^{G_*}\right)\right) / (D_n \cap \partial_{n+1}(D_{n+1})).$$

From Proposition 2.2 (1),

$$H_n(\sup_*(D_*)) = \left(\left(D_n \cdot \partial_{n+1}^{G_*}(D_{n+1})\right) \cap \mathrm{Ker}\left(\partial_n^{G_*}\right)\right) / \partial_{n+1}^{G_*}\left(D_{n+1} \cdot \partial_{n+2}^{G_*}(D_{n+2})\right).$$

Since $\partial_{n+1}^{G_*}\left(\partial_{n+2}^{G_*}(D_{n+2})\right) = \left\{1\right\}$, $\partial_{n+1}^{G_*}\left(D_{n+1}\cdot\partial_{n+2}^{G_*}(D_{n+2})\right) = \partial_{n+1}^{G_*}(D_{n+1})$. Thus

$$H_n(\sup_*(D_*)) = \left(\left(D_n\cdot\partial_{n+1}^{G_*}(D_{n+1})\right)\cap \mathrm{Ker}\left(\partial_n^{G_*}\right)\right)/\partial_{n+1}^{G_*}(D_{n+1}).$$

Let $w_1, w_2 \in D_n \cap \mathrm{Ker}\left(\partial_n^{G_*}\right)$ such that $w_1 \equiv w_2$ in $H_n(\sup_*(D_*))$. Then there exists $y = \partial_{n+1}^{G_*}(D_{n+1})$ such that $w_2 = w_1 y$. Note that

$$y = w_1^{-1}w_2 \in D_n \cap \mathrm{Ker}\left(\partial_n^{G_*}\right) \leq D_n.$$

We have $y \in D_n \cap \partial_{n+1}^{G_*}(D_{n+1})$ with $w_2 = w_1 y$. Thus $w_1 \equiv w_2$ in $H_n(\inf_*(D_*))$. So

$$H_n(\inf_*(D_*)) \to H_n(\sup_*(D_*))$$

is injective. This proves (1).

(2) Let $w \in \left(D_n\cdot\partial_{n+1}^{G_*}(D_{n+1})\right)\cap \mathrm{Ker}\left(\partial_n^{G_*}\right)$. Then $w \in D_n\cdot\partial_{n+1}^{G_*}(D_{n+1})$ and so

$$w = x_1y_1x_2y_2\cdots x_my_m$$

with $x_i \in D_n$ and $y_i \in \partial_{n+1}^{G_*}(D_{n+1})$ for $1 \leq i \leq m$. Since $\partial_{n+1}^{G_*}(D_{n+1})$ is contained in the normalizer of D_n, the product

$$w = x_1\left(y_1x_2y_1^{-1}\right)\left(y_1y_2x_3y_2^{-1}y_1^{-1}\right)\cdots\left(y_1\cdots y_{m-1}x_my_{m-1}^{-1}\cdots y_1^{-1}\right)y_1\cdots y_m = xy$$

with

$$x = x_1\left(y_1x_2y_1^{-1}\right)\left(y_1y_2x_3y_2^{-1}y_1^{-1}\right)\cdots\left(y_1\cdots y_{m-1}x_my_{m-1}^{-1}\cdots y_1^{-1}\right) \in D_n$$

and

$$y = y_1y_2\cdots y_m \in \partial_{n+1}^{G_*}(D_{n+1}).$$

Since $y \in \mathrm{Ker}\left(\partial_n^{G_*}\right)$ and $w \in \mathrm{Ker}\left(\partial_n^{G_*}\right)$,

$$x = wy^{-1} \in \mathrm{Ker}\left(\partial_n^{G_*}\right).$$

It follows that $x \in D_n \cap \mathrm{Ker}\left(\partial_n^{G_*}\right)$, and so

$$H_n(\inf_*(D_*)) \to H_n(\sup_*(D_*))$$

is surjective. From (1), $H_n(\inf_*(D_*)) \to H_n(\sup_*(D_*))$ is injective and so it is an isomorphism. This finishes the proof.□

2.2. Hypergraphs.

Recall that a *hypergraph* $\mathscr{H}$ is a pair $\mathscr{H} = (V_{\mathscr{H}}, \mathscr{E}_{\mathscr{H}})$, where the vertex set $V_{\mathscr{H}}$ is a finite or infinite set and the hyperedge set $\mathscr{E}_{\mathscr{H}}$ is a collection of finite nonempty subsets of $V_{\mathscr{H}}$. Let $\mathscr{P}(V_{\mathscr{H}})$ be the set of all finite subsets of $V_{\mathscr{H}}$. The hypothesis in the definition of hypergraph $\mathscr{H} = (V_{\mathscr{H}}, \mathscr{E}_{\mathscr{H}})$ only requires that $\mathscr{E}_{\mathscr{H}} \subseteq \mathscr{P}(V_{\mathscr{H}}) \setminus \emptyset$. This is different from the notion of an abstract simplicial complex as hypergraphs do not require $\mathscr{E}_{\mathscr{H}}$ to be closed under taking subsets.

The *simplicial closure* (or the *associated simplicial complex* as in [110]) of a hypergraph $\mathscr{H} = (V_{\mathscr{H}}, \mathscr{E}_{\mathscr{H}})$, denoted by $\Delta\mathscr{H}$, is defined as

$$\Delta\mathscr{H} = \{A \neq \emptyset \mid A \subseteq B \text{ for some } B \in \mathscr{E}_{\mathscr{H}}\}.$$

It is straightforward to check that the simplicial closure of $\mathscr{H}$ is the minimal simplicial complex containing $\mathscr{H}$. The homology of $\Delta\mathscr{H}$ has been studied previously in [110]. However, it is desirable for a homology theory of $\mathscr{H}$ to be directly derived from $\mathscr{H}$ itself rather than the simplicial closure $\Delta\mathscr{H}$. Using Proposition 2.3, there is an embedded homology theory of hypergraphs that is an extension of simplicial homology theory.

Definition 2.4. Let $\mathscr{H} = (V_{\mathscr{H}}, \mathscr{E}_{\mathscr{H}})$ be a hypergraph with a total ordering on $V_{\mathscr{H}}$ and let G be an abelian group. Let $C_*(\Delta\mathscr{H}; G)$ be the chain complex with coeffcients in group G. Consider $\mathbb{Z}(\mathscr{H}) \otimes G$ as a graded subgroup of the chain complex of abelian groups $C_*(\Delta\mathscr{H}; G)$. The *embedded homology* $H_*^{\text{emb}}(\mathscr{H}; G)$ *with coefficients in* G is defined by

$$H_*^{\text{emb}}(\mathscr{H}; G) = H_*\left(\inf_*^{C_*(\Delta\mathscr{H}; G)}(\mathbb{Z}(\mathscr{H}) \otimes G)\right) \cong H_*\left(\sup_*^{C_*(\Delta\mathscr{H}; G)}(\mathbb{Z}(\mathscr{H}) \otimes G)\right).$$

The crucial point is that by Proposition 2.3 the inclusion

$$\inf_*^{C_*(\Delta\mathscr{H}; G)}(\mathbb{Z}(\mathscr{H}) \otimes G) \to \sup_*^{C_*(\Delta\mathscr{H}; G)}(\mathbb{Z}(\mathscr{H}) \otimes G)$$

induces an isomorphism on homology. Hence this homology can be considered as a natural topological invariant of $\mathscr{H}$. To detect more subtle information about $\mathscr{H}$, one could explore the acyclic chain complex

$$\sup_*^{C_*(\Delta\mathscr{H};G)}(\mathbb{Z}(\mathscr{H})\otimes G)/\inf_*^{C_*(\Delta\mathscr{H};G)}(\mathbb{Z}(\mathscr{H})\otimes G).$$

For example, when G is a field, one can investigate the Hilbert-Poincaré series

$$\xi^{\mathrm{emb}}(\mathscr{H},t)=\sum_{n=0}^{\infty}\left(\dim\left(\sup_*^{C_*(\Delta\mathscr{H};G)}(\mathbb{Z}(\mathscr{H})\otimes G)/\inf_*^{C_*(\Delta\mathscr{H};G)}(\mathbb{Z}(\mathscr{H})\otimes G)\right)\right)t^n$$

to detect gaps and get more robust information.

Let $\delta(\mathscr{H})$ denote the maximal simplicial complex contained in $\mathscr{H}$. In general, $H_*^{\mathrm{emb}}(\mathscr{H};G)$ is different from $H_*(\delta(\mathscr{H});G)$ and $H_*(\Delta(\mathscr{H});G)$ as shown in the following example.

Example 2.5. Let $\mathscr{H}$ be the boundary of a 2-simplex with all vertices removed, $V_{\mathscr{H}}=\{0,1,2\}$ and $\mathscr{E}_{\mathscr{H}}=\{\{0,1\},\{0,2\},\{1,2\}\}$ as depicted in Figure 4. Then $\delta(\mathscr{H})$ is the empty set, and $\Delta(\mathscr{H})$ is the boundary of the 2-simplex. By definition, $H_1^{\mathrm{emb}}(\mathscr{H};\mathbb{Z})=\mathbb{Z}$ and $H_0^{\mathrm{emb}}(\mathscr{H};\mathbb{Z})=0$. Thus $H_*^{\mathrm{emb}}(\mathscr{H};\mathbb{Z})$ is different from $H_*(\delta(\mathscr{H});\mathbb{Z})$ and $H_*(\Delta(\mathscr{H});\mathbb{Z})$.

This example shows that $H_*^{\mathrm{emb}}(\mathscr{H};G)$ may not be the homology of any simplicial complex as $H_0^{\mathrm{emb}}(\mathscr{H};\mathbb{Z})=0$, which is not the case for any nonempty simplicial complex. Let us consider another example.

Example 2.6. Let $\mathscr{H}=(V_{\mathscr{H}},\mathscr{E}_{\mathscr{H}})$ with $V_{\mathscr{H}}=\{0,1,2\}$ ordered by $0<1<2$, and

$$\mathscr{E}_{\mathscr{H}}=\{\{0,1,2\},\{0,1\},\{0,2\},\{0\},\{1\},\{2\}\}$$

see Figure 4. Then $\Delta\mathscr{H}$ is the abstract simplicial complex of a 2-simplex with vertices labeled by 0, 1, 2. The 1-face {1, 2} is not in $\mathscr{H}$. Let $G=\mathbb{Z}$. Then the chain complex $C_*(\Delta\mathscr{H})$ is given by $C_0(\Delta\mathscr{H})=\mathbb{Z}^{\oplus 3}=\mathbb{Z}\{\{0\},\{1\},\{2\}\}$, $C_1(\Delta\mathscr{H})=\mathbb{Z}^{\oplus 3}=\mathbb{Z}\{\{0,1\},\{0,2\},\{1,2\}\}$, and $C_2(\Delta\mathscr{H})=\mathbb{Z}=\mathbb{Z}\{\{0,1,2\}\}$.

We have $\inf_0=C_0(\Delta\mathscr{H})=\mathbb{Z}\{\{0\},\{1\},\{2\}\}$,

$$\inf_1=\mathbb{Z}(\mathscr{E}_1)\cap\partial_1^{-1}(\mathbb{Z}(\mathscr{E}_0))=\mathbb{Z}(\mathscr{E}_1)\cap C_1(\Delta\mathscr{H})=\mathbb{Z}(\mathscr{E}_1)=\mathbb{Z}\{\{0,1\},\{0,2\}\}$$

$$\inf_2=\mathbb{Z}(\mathscr{E}_2)\cap\partial_2^{-1}(\mathbb{Z}(\mathscr{E}_1))=0$$

with $\partial_1(\inf_1)=\mathbb{Z}\{\{1\}-\{0\},\{2\}-\{0\}\}$. Thus $H_0^{\mathrm{emb}}(\mathscr{H})=\mathbb{Z}$ and $H_i^{\mathrm{emb}}(\mathscr{H})=0$ for $i\geq 1$.

Let $\mathscr{H}'=(V_{\mathscr{H}'},\mathscr{E}_{\mathscr{H}'})$ with $V_{\mathscr{H}'}=V_{\mathscr{H}}=\{0,1,2\}$ ordered by $0<1<2$, and

$$\mathscr{E}_{\mathscr{H}'} = \{\{0,1,2\},\{0,1\},\{0\},\{1\},\{2\}\}.$$

Then $H_0^{\mathrm{emb}}(\mathscr{H}') = \mathbb{Z} \oplus \mathbb{Z}$ and $H_i^{\mathrm{emb}}(\mathscr{H}) = 0$ for $i \geq 1$. Thus the embedded homology of $\mathscr{H}'$ can not be realized as the homology of a path-connected topological space.

These examples indicate that embedded homology is a new homology theory with unusual properties and that poses its own questions and challenges.

The definition of embedded homology of a hypergraph $\mathscr{H}$ depends on the orientation of its simplicial closure $\Delta\mathscr{H}$. It is well-known that simplicial homology is independent on the choice of orientation. The following theorem shows that this is also true for the embedded homology of hypergraphs.

Theorem 2.7. *The embedded homology $H_*^{\mathrm{emb}}(\mathscr{H}; G)$ of a hypergraph $\mathscr{H}$ does not depend on a choice of orientation on $\Delta\mathscr{H}$.*

Proof. Let $H_*^{\mathrm{emb}}(\mathscr{H})$ and $G(\mathscr{H})$ denote $H_*^{\mathrm{emb}}(\mathscr{H}, G)$ and $\mathbb{Z}(\mathscr{H}) \otimes G$, respectively. We assume that $V_{\mathscr{H}}$ is a finite set $\{v_1, v_2, \ldots, v_m\}$. Take a linear ordering on $V_{\mathscr{H}}$ so that $v_1 < v_2 < \cdots < v_n$ as a fixed choice of total order and let $C_* = C_*(\Delta\mathscr{H}; G)$ denote the oriented chain complex. It suffices to show that the homology stays the same up to isomorphism under the transpositions $(i, i+1)$ of the ordering on $V(\mathscr{H})$ for $1 \leq i \leq m-1$.

Let $\partial'_n : C_n \to C_{n-1}$, $n \geq 1$ be the boundary homomorphism defined using the new order on $V_{\mathscr{H}}$, that is, $v_1 < v_2 < \cdots < v_{i-1} < v_{i+1} < v_i < v_{i+2} < \cdots < v_n$. For $n \geq 1$, the abelian group C_n admits a direct sum decomposition

$$C_n = C_n^{v_i v_i+1} \oplus C_n^{\widehat{v_i v_i+1}} \tag{1}$$

where $C_n^{v_i v_i+1}$ is the subgroup of C_n given by linear combinations with coefficients in group G of the n-simplicies $\sigma \in \Delta\mathscr{H}$ whose vertex set contains both v_i and v_{i+1}, and $C_n^{\widehat{v_i v_i+1}}$ is the subgroup of C_n given by linear combinations with coefficients in group G of the remaining n-simplicies in $\Delta\mathscr{H}$. For any chain $\alpha \in C_n$, there is a corresponding unique decomposition

$$\alpha = \alpha^{v_i v_i+1} + \alpha^{\widehat{v_i v_i+1}}. \tag{2}$$

Since v_i and v_{i+1} are neighbored vertices in the order, we have $\partial'(\sigma) = \partial(\sigma)$ if σ does not contain both v_i and v_{i+1} in its vertex set.

Therefore

$$\partial'_n| = \partial_n| : C_n^{\widehat{v_i v_i+1}} \to C_{n-1}. \tag{3}$$

Let $\sigma = [a_1 \cdots a_t v_i v_{i+1} b_1 \cdots b_s]$ be an oriented simplex in $\Delta\mathscr{K}$ with $a_1 < \cdots < a_t < v_i < v_{i+1} < b_1 < \cdots < b_s$. By the definition of $\partial(\sigma)$, we have

$$\partial(\sigma)^{v_i v_i+1} = \sum_{j=1}^{t} (-1)^{j-1}\left[a_1 \cdots \hat{a}_j \cdots a_t v_i v_j+1 b_1 \cdots b_s\right] + \sum_{k=1}^{s} (-1)^{t+k+1}\left[a_1 \cdots a_t v_i v_i+1 b_1 \cdots \hat{b}_k \cdots b_s\right]$$

where $\cdots \hat{x} \cdots$ means that x is deleted, and

$$\partial(\sigma)^{\widehat{v_i v_i+1}} = (-1)^t [a_1 \cdots a_t v_i+1 b_1 \cdots b_s] + (-1)^{t+1} [a_1 \cdots a_t v_i b_1 \cdots b_s].$$

By switching the order of v_i and v_{i+1}, we have

$$\partial'(\sigma) = (\partial(\sigma))^{v_i v_i+1} - (\partial(\sigma))^{\widehat{v_i v_i+1}}.$$

Extending this formula linearly with coefficients in group G, we obtain the formula

$$\partial'(\alpha) = (\partial(\alpha))^{v_i v_i+1} - (\partial(\alpha))^{\widehat{v_i v_i+1}} \text{ for } \alpha \in C_*^{v_i v_i+1}. \tag{4}$$

Define the group homomorphism

$$\phi_n : C_n = C_n^{v_i v_i+1} \oplus C_n^{\widehat{v_i v_i+1}} \to C_n = C_n^{v_i v_i+1} \oplus C_n^{\widehat{v_i v_i+1}}$$

by setting

$$\phi_n\left(z^{v_i v_i+1} + z^{\widehat{v_i v_i+1}}\right) = z^{v_i v_i+1} - z^{\widehat{v_i v_i+1}}.$$

Clearly, ϕ_n is an isomorphism. Let $z = z^{v_i v_i+1} + z^{\widehat{v_i v_i+1}} \in C_n$ be a chain. Then

$$\partial(z) = \partial\left(z^{v_i v_i+1}\right) + \partial\left(z^{\widehat{v_i v_i+1}}\right) = \left(\partial\left(z^{v_i v_i+1}\right)\right)^{v_i v_i+1} + \left(\partial\left(z^{v_i v_i+1}\right)\right)^{\widehat{v_i v_i+1}} + \partial\left(z^{\widehat{v_i v_i+1}}\right)$$

so

$$(\partial(z))^{v_i v_i+1} = \left(\partial\left(z^{v_i v_i+1}\right)\right)^{v_i v_i+1}$$

and

$$(\partial(z))^{\widehat{v_i v_i+1}} = \left(\partial\left(z^{v_i v_i+1}\right)\right)^{\widehat{v_i v_i+1}} + \partial\left(z^{\widehat{v_i v_i+1}}\right).$$

On the other hand, by direct computation

$$\begin{aligned}\partial'(\phi_n(z)) &= \partial'\left(z^{v_i v_i+1} - z^{\widehat{v_i v_i+1}}\right)\\ &= \partial\left(z^{v_i v_i+1}\right))^{v_i v_i+1} - \left(\left(\partial\left(z^{v_i v_i+1}\right)\right)^{\widehat{v_i v_i+1}} + \partial\left(z^{\widehat{v_i v_i+1}}\right)\right)\\ &= (\partial(z))^{v_i v_i+1} - (\partial(z))^{\widehat{v_i v_i+1}}.\end{aligned}$$

This gives a commutative diagram

$$\begin{array}{ccc} C_n & \xrightarrow[\cong]{\phi_n} & C_n \\ \downarrow{\scriptstyle \partial_n} & & \downarrow{\scriptstyle \partial'_n} \\ C_{n-1} & \xrightarrow[\cong]{\phi_{n-1}} & C_{n-1}. \end{array}$$

Note that the decomposition (1) restricted to $G(\mathscr{H}_n)$ gives the decomposition

$$G(\mathscr{H}) = G(\mathscr{H})^{v_i v_i+1} \oplus G(\mathscr{H})^{\widehat{v_i v_i+1}}$$

with the same rule on simplices. The subgroup $G(H_n)$ is invariant under ϕ_n. Moreover, there is a commutative diagram

$$\begin{array}{ccc} \inf_n^{C_*(\Delta\mathcal{H};G)}(G(\mathcal{H})) & \xrightarrow[\cong]{\phi_n} & \inf_n^{C_*(\Delta\mathcal{H};G)}(G(\mathcal{H})) \\ \downarrow{\scriptstyle \partial_n|} & & \downarrow{\scriptstyle \partial'_n|} \\ \inf_{n-1}^{C_*(\Delta\mathcal{H};G)}(G(\mathcal{H})) & \xrightarrow[\cong]{\phi_{n-1}} & \inf_{n-1}^{C_*(\Delta\mathcal{H};G)}(G(\mathcal{H})). \end{array}$$

The assertion then follows by taking homology of this commutative diagram. □

The embedded homology of hypergraphs was introduced in 2019 in [23]. Previously, cohomological aspects on k-uniform hypergraphs have been studied, see [28, 27, 42, 44, 96, 99, 122, 123, 138, 142], using cohomology introduced in a combinatorial way. Also

Emtander [59] studied the homology of the *independence complex* $\Delta^c\mathscr{H}$ of a hypergraph $\mathscr{H} = (V_{\mathscr{H}}, \mathscr{E}_{\mathscr{H}})$ in 2009, where $\Delta^c\mathscr{H} = \{F \subseteq V_{\mathscr{H}} \mid E \nsubseteq F \text{ for any } E \in \mathscr{E}_{\mathscr{H}}\}$. The approach of embedded homology is different from the classical research on topological structures related to hypergraphs as it is directly define on the hypergraph.

Although the embedded homology of hypergraphs is a new topic with surprising properties, it inherits many characteristics of simplicial homology. The following theorem is an example of this, the proof of this theorem is similar to that of [102, Theorem 8.2, p.45].

Recall that the *cone* CK of a simplicial complex K is defined as a join $CK = w*K$ with w a vertex not in K. Analogously, we can define the join of hypergraphs and the cone $C\mathscr{H} = w * \mathscr{H}$.

Theorem 2.8. *Let* $\mathscr{H}$ *be a hypergraph and let* G *be an abelian group. Then*

$$H_n^{\text{emb}}(C\mathscr{H}; G) = \begin{cases} 0 & if\ n > 0 \\ G & if\ n = 0. \end{cases}$$

□

2.3. Super-hypergraphs.

Recall [47, 139] that a Δ-*set* X_* is a sequence of sets $X_* = (X_n)_{n\geq 0}$ with maps d_i: $X_n \to X_{n-1}$, for $0 \leq i \leq n$ and $n \geq 1$, called *face operations*, satisfying the following Δ-identity

$$d_i d_j = d_j d_{i+1} \text{ for } i \geq j. \tag{5}$$

Definition 2.9. A *super-hypergraph* is a pair $(\mathscr{H}, X)$, where X is a Δ-set and $\mathscr{H}$ is a graded subset of X. We call $\mathscr{H}$ a *super-hypergraph born from* X, and X is called *a parental* Δ-*set* of $\mathscr{H}$. The Δ-*closure* of $\mathscr{H}$ in X is defined by

$$\Delta^X(\mathscr{H}) = \bigcap \{Y \mid \mathscr{H} \subseteq Y \text{ as a graded subset and } Y \subseteq X \text{ as a } \Delta\text{-subset}\}.$$

A *morphism* ϕ: $(\mathscr{H}, X) \to (\mathscr{H}', Y)$ of super-hypergraphs is a Δ-map ϕ: $X \to Y$ such that $\phi(\mathscr{H}) \subseteq \mathscr{H}'$.

2.3.1. Homology of super-hypergraphs.—

Using Proposition 2.3, there is an embedded homology on super-hypergraphs.

Definition 2.10. Let $(\mathscr{H}, X)$ be a super-hypergraph and let G be an abelian group. The *embedded homology* $H_*^{\text{emb}, X}(\mathscr{H}; G)$ *with coefficients in* G of $(\mathscr{H}, X)$ is defined by

$$H_*^{\text{emb}, X}(\mathscr{H}; G) = H_*\left(\inf_*^{C_*(X;G)}(\mathbb{Z}(\mathscr{H}) \otimes G)\right) \cong H_*\left(\sup_*^{C_*(X;G)}(X; G)(\mathbb{Z}(\mathscr{H}) \otimes G)\right)$$

where $\mathbb{Z}(\mathscr{K}) \otimes G$ is a graded subgroup of the chain complex of abelian groups $C_*(X;G)$.

We want to show that this definition is an extension of the embedded homology of hypergraphs. An *oriented hypergraph* is a hypergraph $\mathscr{K}$ with a partial order on its vertex set so that the restriction to the vertices of each hyperedge of $\mathscr{K}$ is linear. If the vertices of a simplex are totally ordered, then the restricted order on the vertices of any of its faces is linear. Thus the simplicial closure $\Delta\mathscr{K}$ can be oriented with its orientation induced by the order on $\mathscr{K}$. From Definition 2.4,

$$H_*^{\mathrm{emb}}(\mathscr{K};G) = H_*^{\mathrm{emb},\,\Delta\mathscr{K}}(\mathscr{K};G)$$

by considering the oriented simplicial complex $\Delta\mathscr{K}$ as a Δ-set. From Theorem 2.7, this definition is independent on the choice of orientation.

We now consider how morphisms of super-hypergraphs, recall Definition 2.9, induce maps on the infimum and supremum chain complexes as well as embedded homology of super-hypergraphs.

Proposition 2.11. *Let $\phi\colon(\mathscr{K},X)\to(\mathscr{K}',Y)$ be a morphism of super-hypergraphs. Then there is a commutative diagram*

$$\begin{array}{ccccccc}
\inf_*^{C_*(X;G)}(\mathbb{Z}(\mathcal{H})\otimes G) & \subseteq & \mathbb{Z}(\mathcal{H})\otimes G & \subseteq & \sup_*^{C_*(X;G)}((\mathbb{Z}(\mathcal{H})\otimes G) & \hookrightarrow & C_*(X;G) \\
\downarrow{\scriptstyle \phi_\#|} & & \downarrow{\scriptstyle \phi_\#|} & & \downarrow{\scriptstyle \phi_\#|} & & \downarrow{\scriptstyle \phi_\#} \\
\inf_*^{C_*(Y;G)}(\mathbb{Z}(\mathcal{H}')\otimes G) & \subseteq & \mathbb{Z}(\mathcal{H}')\otimes G & \subseteq & \sup_*^{C_*(Y;G)}((\mathbb{Z}(\mathcal{H}')\otimes G) & \hookrightarrow & C_*(Y;G)
\end{array} \tag{6}$$

which induces a map $\phi_*\colon H_*^{\mathrm{emb},\,X}(\mathscr{K};G)\to H_*^{\mathrm{emb},\,Y}(\mathscr{K}';G)$. Moreover, if $\phi(\mathscr{K})=\mathscr{K}'$, then there is a short exact sequence of chain complexes

$$\begin{aligned}
&\sup_*^{C_*(X;G)}((\mathbb{Z}(\mathscr{K})\otimes G)\cap \mathrm{Ker}(\phi_\#) \hookrightarrow \sup_*^{C_*}(X;G)((\mathbb{Z}(\mathscr{K})\otimes G \\
&) \overset{\phi_\#}{\twoheadrightarrow} \sup_*^{C_*(Y;G)}((\mathbb{Z}(\mathscr{K}')\otimes G)\,.
\end{aligned} \tag{7}$$

Proof. Since $\sup_*^{C_*(Y;G)}((\mathbb{Z}(\mathscr{K}')\otimes G)$ is a subcomplex of $C_*(Y;G)$ containing $\mathbb{Z}(\mathscr{K}')\otimes G$, its preimage

$$\phi_\#^{-1}\Big(\sup_*^{C_*(Y;G)}((\mathbb{Z}(\mathscr{K}')\otimes G))\Big)$$

is a subcomplex of $C_*(X;G)$ containing $\mathbb{Z}(\mathscr{K})\otimes G$ because $\phi(\mathscr{K})\subseteq\mathscr{K}'$. Thus

$$\sup_*^{C_*(X;G)}\Big((\mathbb{Z}(\mathscr{K})\otimes G)\subseteq \phi_\#^{-1}\Big(\sup_*^{C_*(Y;G)}((\mathbb{Z}(\mathscr{K}')\otimes G)\Big) \tag{8}$$

as a subcomplex. Similarly

$$\phi_{\#}\left(\inf_{*}^{C_*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)\right)\subseteq \inf_{*}^{C_*(Y;G)}(\mathbb{Z}(\mathscr{H}')\otimes G) \tag{9}$$

as a subcomplex. Therefore, there is a commutative diagram 6.

Now, we assume that $\phi(\mathscr{H})=\mathscr{H}'$. Then

$$\phi_{\#}\left(\sup_{*}^{C_*(X;G)}((\mathbb{Z}(\mathscr{H})\otimes G))\supseteq \phi_{\#}((\mathbb{Z}(\mathscr{H})\otimes G))=\mathbb{Z}(\mathscr{H}')\otimes G\right.$$

is a subcomplex of $C_*(Y;G)$ containing $\mathbb{Z}(\mathscr{H}')\otimes G$. Hence

$$\phi_{\#}\left(\sup_{*}^{C_*(X;G)}((\mathbb{Z}(\mathscr{H})\otimes G))\supseteq \sup_{*}^{C_*(Y;G)}((\mathbb{Z}(\mathscr{H}')\otimes G)\right.\,.$$

Together with the containment 8, we have

$$\sup_{*}^{C_*(Y;G)}\left((\mathbb{Z}(\mathscr{H}')\otimes G)\supseteq \phi_{\#}\left(\sup_{*}^{C_*(X;G)}((\mathbb{Z}(\mathscr{H})\otimes G)\right)\right.$$

and hence the short exact sequence 7. □

Corollary 2.12. *Let* $\phi:(\mathscr{H},X)\to(\mathscr{H}',Y)$ *be a morphism of super-hypergraphs. Suppose that*

1. ϕ: $X\to Y$ *is an injective* Δ*-map and*
2. $\phi(\mathscr{H})=\mathscr{H}'$.

Then

$$\phi_{\#}:\sup_{*}^{C_*(X;G)}\left((\mathbb{Z}(\mathscr{H})\otimes G)\to \sup_{*}^{C_*(Y;G)}((\mathbb{Z}(\mathscr{H}')\otimes G)\right.$$

is an isomorphism. In particular, $\phi_*:H_*^{\mathrm{emb},X}(\mathscr{H};G)\to H_*^{\mathrm{emb},Y}(\mathscr{H}';G)$ *is an isomorphism.*

Proof. By the assumption (1),

$$\sup_{*}^{C_*(X;G)}((\mathbb{Z}(\mathscr{H})\otimes G)\cap \mathrm{Ker}(\phi_{\#})=0$$

and so the assertion follows by Proposition 2.11. □

2.3.2. Variations of parental Δ-sets.—In recent topological applications to data analytics and machine learning, one of the most common approaches is to use discrete Hodge-Laplacian theory. Mathematically, combinatorial Laplacian operators defined on linear transformations on cochains of simplicial complexes have been studied, for example in [55, 82]. Therefore, in addition to simplicial homology, research on (co)chains of simplicial complexes such as spectral analysis on combinatorial Laplacian operators is also important for potential applications in data science. Similarly, the research on chains

$\inf_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)$ and $\sup_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)$ could be useful for the applications of super-hypergraphs. Next, we describe some basic properties related to the chain complexes arising from super-hypergraphs.

A super-hypergraph $\mathscr{H}$ is assumed to have a parental Δ-set X that carries geometric structural information about $\mathscr{H}$. The embedded homology $H_*^{\text{emb},X}(\mathscr{H};G)$ is defined using the geometric information inherited from X. On level of chains, there are inclusions of graded groups

$$\inf_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)\hookrightarrow\mathbb{Z}(\mathscr{H})\otimes G\hookrightarrow\sup_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)\hookrightarrow C*(X;G)$$

where the right most inclusion is a chain map. By Proposition 2.3, the inclusion

$$\inf_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)\hookrightarrow\sup_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)$$

is a chain homotopy equivalence that defines the embedded homology. The gap complex

$$\sup_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G)/\inf_*^{C*(X;G)}(\mathbb{Z}(\mathscr{H})\otimes G) \tag{10}$$

which is an acyclic chain complex, gives more robust information about the graded set $\mathscr{H}$.

Let $\mathscr{H}$ be a fixed graded data set. Our aim is to vary the parental Δ-set X such that the corresponding infimum and supremum chain complexes reveal different aspects of the topological structure of $\mathscr{H}$. One natural way to vary the parental Δ-set would be to consider morphisms $\phi:(\mathscr{H},X)\to(\mathscr{H},Y)$ that fix $\mathscr{H}$ and investigate how these affect the embedded homology. Another important question is whether one can vary the parental Δ-set so that the gap complex (10) is as small as possible. We consider the following example.

Example 2.13. Let n be an odd positive integer. Let $X=\Delta^+[n]$ be the Δ-set induced by an n-simplex with vertices labelled 0, 1, …, n. Let Y be the Δ-set with $Y_k=\{a_k\}$, $0\le k\le n$, $Y_k=\varnothing$ for $k>n$ and $d_i(a_k)=a_{k-1}$ for $0\le i\le k\le n$. Let $\mathscr{H}$ be the graded set given by $\mathscr{H}_n=\{x_n\}$ and $\mathscr{H}_k=\varnothing$ for $k\neq n$. Consider the super-hypergraphs $(\mathscr{H},X)$, $x_n=[0,1,2,\ldots,n]$, and $(\mathscr{H},Y)$, $x_n=a_n$. There is a unique morphism $\phi:(\mathscr{H},X)\to(\mathscr{H},Y)$ such that $\phi|_{\mathscr{H}}=\mathrm{id}_{\mathscr{H}}$. Then, we have $H_k^{\text{emb},X}(\mathscr{H};\mathbb{Z})=0$ for $k\ge 0$ and

$$H_k^{\text{emb},Y}(\mathscr{H};\mathbb{Z})=\begin{cases}\mathbb{Z} & \text{if } k=n,\\ 0 & \text{otherwise.}\end{cases}$$

This shows the embedded homology or super-hypergraphs depends on the parentl Δ-set. The gap complex for $(\mathscr{H},X)$ is the same as the acyclic complex

$$\sup_k^{C_*(X;\mathbb{Z})}(\mathbb{Z}(\mathscr{H})) = \begin{cases} \mathbb{Z} & \text{if } k = n, n-1, \\ 0 & \text{otherwise} \end{cases}$$

with $\partial : \sup_n^{C_*(X;\mathbb{Z})}(\mathbb{Z}(\mathscr{H})) \to \sup_{n-1}^{C_*(X;\mathbb{Z})}(\mathbb{Z}(\mathscr{H}))$.an isomorphism, and the gap complex for $(\mathscr{H}, Y)$ is 0.

Now we consider the effect of morphisms on super-hypergraphs with a fixed hypergraph more generally. Let $\phi : (\mathscr{H}, X) \to (\mathscr{H}, Y)$ be a morphism of super-hypergraphs so that $\phi|_{\mathscr{H}} = \mathrm{id}_{\mathscr{H}}$. By 9, we have

$$\inf_*^{C_*(X;G)}(\mathbb{Z}(\mathscr{H}) \otimes G) \hookrightarrow \inf_*^{C_*(Y;G)}(\mathbb{Z}(\mathscr{H}) \otimes G)$$

namely the chain complex of Y is closer to the graded group $\mathbb{Z}(\mathscr{H}) \otimes G$ than the chain complex of X. Together with the inclusion 8, it follows that the gap complex for $(\mathscr{H}, Y)$ is smaller, as shown in the above example. If ϕ is injective, then both

$$\inf_*^{C_*(X;G)}(\mathbb{Z}(\mathscr{H}) \otimes G) = \inf_*^{C_*(Y;G)}(\mathbb{Z}(\mathscr{H}) \otimes G) \text{ and}$$

$$\sup_*^{C_*(X;G)}(\mathbb{Z}(\mathscr{H}) \otimes G) = \sup_*^{C_*(Y;G)}(\mathbb{Z}(\mathscr{H}) \otimes G)$$

which means that we can replace X by any of its Δ-subsets that contain $\mathscr{H}$ without changing the infimum and supremum chain complexes. In particular, we have

$$H_*^{\mathrm{emb},X}(\mathscr{H}; G) = H_*^{\mathrm{emb},\Delta^X(\mathscr{H})}(\mathscr{H}; G) \tag{11}$$

where $\Delta^X(\mathscr{H})$ is the Δ-closure of $\mathscr{H}$ in X, which is the minimal Δ-subset of X containing $\mathscr{H}$ defined in Definition 2.9.

Definition 2.14. A super-hypergraph $(\mathscr{H}, X)$ is called *regular* if $X = \Delta^X(\mathscr{H})$.

It is straightforward to see that a super-hypergraph $(\mathscr{H}, X)$ is regular if and only if all elements in X are obtained from the elements in $\mathscr{H}$ together with their iterated faces in X. The following proposition may be useful for analysing for variations of parental Δ-sets.

Proposition 2.15. *Let $\mathscr{H} = \{\mathscr{H}_n\}_{n \geq 0}$ be a graded set such that the cardinality of the set $\coprod_{n \geq 0} H_n$ is finite. Then there are finitely many regular super-hypergraphs $(\mathscr{H}, X)$ up to isomorphisms.*

Proof. Consider the collection of all possible Δ-sets X such that $(\mathscr{H}, X)$ is a regular super-hypergraph. If $(\mathscr{H}, X)$ is regular, then all elements in X are given by the elements in the

graded subset $\mathscr{H}$ together with their iterated faces in X. Therefore, as a Δ-set X is a finite extension of $\mathscr{H}$ together with finitely many face operations.□

Definition 2.16. A super-hypergraph $(\mathscr{H}, X)$ is *complete* if it is regular, and for any morphism of super-hypergraphs $\phi:(\mathscr{H}, X) \to (\mathscr{H}, Y)$ with $\phi|_{\mathscr{H}} = \mathrm{id}_{\mathscr{H}}$, the Δ-map $\phi: X \to Y$ is injective.

Therefore, for a complete super-hypergragh $(\mathscr{H}, X)$, the infimum complex

$$\inf_{*}^{C*(X;G)}(\mathbb{Z}(\mathscr{H}) \otimes G)$$

reaches a maximum and the gap complex (10) reaches a minimum.

By definition, a complete super-hypergraph is regular. However, the converse may not be true. For instance, the super-hypergraph $(\mathscr{H}, X)$ in Example 2.13 is regular but not complete as the morphism to $(\mathscr{H}, Y)$ is not an injective Δ-set map. Therefore completeness provides a notion of maximality in the set of super-hypergraphs related to a given hypergraph.

Theorem 2.17 (Completeness Criterion). *A regular super-hypergraph* $(\mathscr{H}, X)$ *with* $\mathscr{H} \neq \varnothing$ *is complete if and only if it has the following properties:*

1. *(Vertex Property) If* $\mathscr{H}_0 \neq \varnothing$, *then* $X_0 = \mathscr{H}_0$. *If* $\mathscr{H}_0 = \varnothing$, *then* X_0 *is a one-point set.*
2. *(Matching Face Property) Let* $z_1, z_2 \in X_n$ *with* $n > 0$. *Suppose that*
   i. $d_i z_1 = d_i z_2$, $0 \leq i \leq n$, *and*
   ii. $\{z_1, z_2\} \not\subseteq \mathscr{H}_n$.

   Then $z_1 = z_2$.

Proof. Suppose that $(\mathscr{H}, X)$ is complete. We first show that properties (1) and (2) hold.

1. Assume that $\mathscr{H}_0 \neq \varnothing$. Suppose that there exists $z \in \Delta^X(\mathscr{H})_0 \backslash \mathscr{H}_0$. Choose an element $x \in \mathscr{H}_0$. Let Z be the Δ-quotient of X by identifying z and x. Let $q: X \to Z$ be the quotient map. Then $q|_{\mathscr{H}}$ is injective, but $q: X \to Z$ is not injective, which is a contradiction. Hence $X_0 = \mathscr{H}_0$.

   If $\mathscr{H}_0 \neq \varnothing$, then the same argument shows that X_0 must be a one-point set.

2. Suppose $z_1 \neq z_2$. Then, similar to the proof of (1), we can construct the Δ-quotient Z obtained from X by identifying z_1 with z_2 in dimension n. Since all faces of z_1 and z_2 match, the equivalence relation $z_1 \sim z_2$ does not induce nontrivial identifications in X_m for $m \neq n$. Since $\{z_1, z_2\} \not\subseteq \mathscr{H}_n$, the equivalence relation $z_1 \sim z_2$ does not effect elements in $\mathscr{H}$. This proves property (2).

Now let $\phi:(\mathscr{H}, X) \to (\mathscr{H}, Y)$ be a morphism of super-hypergraphs with $\phi|_{\mathscr{H}} = \mathrm{id}_{\mathscr{H}}$ such that $(\mathscr{H}, X)$ is a regular super-hypergraph satisfying properties (1) and (2).

We show that the Δ-map ϕ: $X \to Y$ is injective. By the vertex property, ϕ: $X_0 \to Y_0$ is injective. Suppose that ϕ: $X \to Y$ is not injective. Then there exists $n > 0$ such that ϕ: $X_k \to Y_k$ is injective for $k < n$ and ϕ: $X_n \to Y_n$ is not injective. It follows that there exists $z_1, z_2 \in X_n$ with $z_1 \neq z_2$ such that $\phi(z_1) = \phi(z_2)$. Then $\{z_1, z_2\} \not\subseteq \mathscr{H}_n$ because $\phi|_{\mathscr{H}}$ is injective. Since ϕ is a Δ-set map, we have

$$\phi(d_i z_1) = d_i \phi(z_1) = d_i \phi(z_2) = \phi(d_i z_2)$$

for $0 \leq i \leq n$ and ϕ: $X_{n-1} \to Y_{n-1}$ is injective. Therefore $d_i z_1 = d_i z_2$ for $0 \leq i \leq n$. However, by the matching face property, $z_1 = z_2$, which contradicts the assumption that $z_1 \neq z_2$. □

For a given regular super-hypergraph $(\mathscr{H}, X)$, the above proof gives a way to construct a complete super-hypergraph $(\mathscr{H}, Y)$ with Y as a Δ-quotient of X. The following example shows that $(\mathscr{H}, X)$ may have non-isomorphic complete quotients.

Example 2.18. Let

$$\mathscr{H} = \{\{0, 1, 2\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{1\}, \{2\}\}$$

be the graded subset of the 2-simplex $X = \Delta^+[2]$ without the 0 vertex. Let Y be the Δ-quotient of X by identifying vertices $\{1\}$ and $\{0\}$, and let Z be the Δ-quotient of X by identifying vertices $\{2\}$ and $\{0\}$. Then

1. $(\mathscr{H}, X)$, $(\mathscr{H}, Y)$ and $(\mathscr{H}, Z)$ are regular super-hypergraph.
2. $(\mathscr{H}, Y)$ and $(\mathscr{H}, Z)$ are complete, but $(\mathscr{H}, X)$ is not complete.
3. There are non-injective Δ-quotients $X \twoheadrightarrow Y$ and $X \twoheadrightarrow Z$ with $(\mathscr{H}, Y) \ncong (\mathscr{H}, Z)$.

2.3.3. Mayer-Vietoris sequence.—

The Mayer-Vietoris sequence (MV sequence) is one of the fundamental tools in topology for inductively computing homology. In general, the MV sequence fails for embedded homology of super-hypergraphs.

Example 2.19. Let $X = \{f_1, f_2, e_1, e_2, v\}$ be a Δ-set with face operations given by

1. $d_0 f_i = d_2 f_i = e_1$, $d_1 f_i = e_2$ where $j = 1, 2$, and
2. $d_i e_j = v$ for $i = 0, 1$ and $j = 1, 2$.

Let

$$\mathscr{H} = X \backslash \{e_1, e_2\}$$

be the graded subset of X missing the edges e_1, e_2. Let $A = \{f_1, e_1, e_2, v\}$ and $B = \{f_2, e_1, e_2, v\}$ with face operations induced from X. Then

$$(\mathscr{H}, X) = (\mathscr{H} \cap A, A) \cup (\mathscr{H} \cap B, B)$$

with $A \cap B = \{e_1, e_2, v\}$ and $\mathscr{H} \cap A \cap B = \{v\}$. Then there is no exact sequence

$$\cdots \longrightarrow H_2^{\mathrm{emb},A}(\mathcal{H} \cap A) \oplus H_2^{\mathrm{emb},B}(\mathcal{H} \cap B) \longrightarrow H_2^{\mathrm{emb},X}(\mathcal{H})$$

$$\longrightarrow H_1^{\mathrm{emb},A\cap B}(\mathcal{H} \cap A \cap B) \longrightarrow \cdots$$

Since $\mathscr{H}_1 = \varnothing$ and $\mathscr{H}_2 = X_2$,

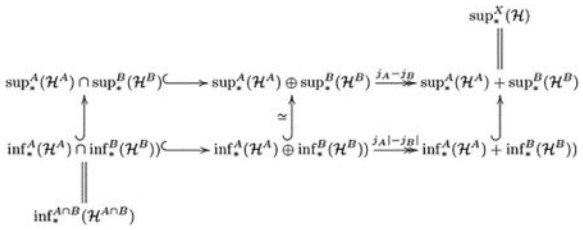
$$\begin{aligned}
\inf_2^{C_*(A)}(\mathbb{Z}(\mathscr{H} \cap A)) &= \mathbb{Z}(\mathscr{H}_2 \cap A) \cap \mathrm{Ker}\left(\partial_2^A\right) = 0,\\
\inf_2^{C_*(B)}(\mathbb{Z}(\mathscr{H} \cap B)) &= \mathbb{Z}(\mathscr{H}_2 \cap B) \cap \mathrm{Ker}\left(\partial_2^B\right) = 0,\\
\inf_2^{C_*(X)}(\mathbb{Z}(\mathscr{H})) &= \mathbb{Z}(\mathscr{H}_2) \cap \mathrm{Ker}\left(\partial_2^X\right) = \mathbb{Z}\{f_1 - f_2\} = \mathbb{Z},\\
\inf_1^{C_*(X)}(\mathbb{Z}(\mathscr{H})) &= \mathbb{Z}(\mathscr{H}_1) \cap \left(\partial_1^X\right)^{-1}(\mathbb{Z}(\mathscr{H}_0)) = 0,\\
\inf_1^{C_*(A\cap B)}(\mathbb{Z}(\mathscr{H} \cap A \cap B)) &= 0\,.
\end{aligned}$$

Then $H_2^{\mathrm{emb},A}(\mathscr{H} \cap A) = H_2^{\mathrm{emb},B}(\mathscr{H} \cap B) = H_1^{\mathrm{emb},A\cap B}(\mathscr{H} \cap A \cap B) = 0$ and $H_2^{\mathrm{emb},X}(\mathscr{H}) = \mathbb{Z}$. Hence the above sequence cannot be exact.

An analogue of the classical Mayer-Vietoris sequence for super-hypergraphs is a multi-exact sequence derived from the following theorem.

Theorem 2.20. *Let* $(\mathscr{H}, X)$ *be a super-hypergraph and let* A *and* B *be* Δ*-subsets of* X *such that* $A \cup B = X$. *Let* $\mathscr{H}^A = \mathscr{H} \cap A$, $\mathscr{H}^B = \mathscr{H} \cap B$ *and* $\mathscr{H}^{A\cap B} = \mathscr{H} \cap A \cap B$. *Let* G *be an abelian group, and denote* $\sup_*^{C_*(X;G)}(\mathscr{H})$ *and* $\inf_*^{C_*(X;G)}(\mathscr{H})$ *by* $\sup_*^X(\mathscr{H})$ *and* $\inf_*^X(\mathscr{H})$, *respectively.*

Then there is a commutative diagram



where the middle two rows are short exact sequences of chain complexes, the maps j_A *and* j_B *are canonical inclusions and the vertical arrows are inclusions.*

Example 2.19 shows that the left and right vertical arrows in the above diagram are not chain homotopy equivalences. However, the middle vertical arrow is always a chain homotopy equivalence.

Proof. We need to show that

$$\sup_*^A\left(\mathscr{H}^A\right)+\sup_*^B\left(\mathscr{H}^B\right)=\sup_*^X(\mathscr{H}) \text{ and}$$

$$\inf_*^{A\cap B}\left(\mathscr{H}^{A\cap B}\right)=\inf_*^A\left(\mathscr{H}^A\right)\cap\inf_*^B\left(\mathscr{H}^B\right).$$

First we prove that $\sup_*^A\left(\mathscr{H}^A\right)+\sup_*^B\left(\mathscr{H}^B\right)=\sup_*^X(\mathscr{H})$. Since $\sup_*^A\left(\mathscr{H}^A\right)+\sup_*^B\left(\mathscr{H}^B\right)$ is a sub complex of $C_*(X;G)$ containing $\mathbb{Z}(\mathscr{H})\otimes G=\mathbb{Z}\left(\mathscr{H}^A\right)\otimes G+\mathbb{Z}\left(\mathscr{H}^B\right)\otimes G$, we have $\sup_*^X(\mathscr{H})\subseteq\sup_*^A\left(\mathscr{H}^A\right)+\sup_*^B\left(\mathscr{H}^B\right)$. On the other hand, $\sup_*^A\left(\mathscr{H}^A\right),\sup_*^B\left(\mathscr{H}^B\right)\subseteq\sup_*^X(\mathscr{H})$. Thus $\sup_*^A\left(\mathscr{H}^A\right)+\sup_*^B\left(\mathscr{H}^B\right)\subseteq\sup_*^X(\mathscr{H})$, and so $\sup_*^X(\mathscr{H})=\sup_*^A\left(\mathscr{H}^A\right)+\sup_*^B\left(\mathscr{H}^B\right)$

Now we show that $\inf_*^{A\cap B}\left(\mathscr{H}^{A\cap B}\right)=\inf_*^A\left(\mathscr{H}^A\right)\cap\inf_*^B\left(\mathscr{H}^B\right)$. Clearly

$$\inf_*^{A\cap B}\left(\mathscr{H}^{A\cap B}\right)\subseteq\inf_*^A\left(\mathscr{H}^A\right)\cap\inf_*^B\left(\mathscr{H}^B\right)$$

Conversely, note that

$$\inf_*^A\left(\mathscr{H}^A\right)\cap\inf_*^B\left(\mathscr{H}^B\right)\subseteq(\mathbb{Z}(\mathscr{H}\cap A)\otimes G)\cap(\mathbb{Z}(\mathscr{H}\cap B)\otimes G)=\mathbb{Z}(\mathscr{H}\cap A\cap B)\otimes G$$

where $(\mathbb{Z}(\mathscr{H}\cap A)\otimes G)\cap(\mathbb{Z}(\mathscr{H}\cap B)\otimes G)=\mathbb{Z}(\mathscr{H}\cap A\cap B)\otimes G$ because $C_*(X;G)$ is the direct sum of the copies of G with its coordinates labeled by the graded set X. Thus $\inf_*^A\left(\mathscr{H}^A\right)\cap\inf_*^B\left(\mathscr{H}^B\right)$ is a subcomplex of $C_*(X;G)$ contained in $\mathbb{Z}(\mathscr{H}\cap A\cap B)\otimes G$, and so it is contained in $\inf_*^{A\cap B}\left(\mathscr{H}^{A\cap B}\right)$. □

Corollary 2.21. *Using the notation as in the theorem above, the inclusion*

$$\inf_*^A\left(\mathscr{H}^A\right)\cap\inf_*^B\left(\mathscr{H}^B\right)\hookrightarrow\sup_*^A\left(\mathscr{H}^A\right)\cap\sup_*^B\left(\mathscr{H}^B\right)$$

is a chain homotopy equivalence if and only if so is the inclusion

$$\inf_*^A\left(\mathscr{H}^A\right)+\inf_*^B\left(\mathscr{H}^B\right)\hookrightarrow\sup_*^A\left(\mathscr{H}^A\right)+\sup_*^B\left(\mathscr{H}^B\right).$$

Proof. The statement follows by applying the Five Lemma to the long exact sequence obtained from Theorem 2.20. □

2.3.4. Gap complexes.—

Let $(\mathscr{H}, X)$ be a super-hypergraph and define

$$\delta^X(\mathscr{H})=\bigcup\{Y\subseteq\mathscr{H}\mid Y\text{ is a }\Delta-\text{subset of }X\} \tag{12}$$

to be the largest Δ-subset of X contained in $\mathscr{H}$. Then $\delta^X(\mathscr{H})$ consists of the elements in $\mathscr{H}$ whose all iterated faces lie in $\mathscr{H}$. The gap between $\delta^X(\mathscr{H}) \subseteq \Delta^X(\mathscr{H})$ measures how far $\mathscr{H}$ is from being a Δ-set. For a finite hypergraph $\mathscr{H}$, the differences can be expressed as

$$\#\left(\Delta^X(\mathscr{H}) \setminus \delta^X(\mathscr{H})\right) = \#\left(\Delta^X(\mathscr{H})\right) - \#\left(\delta^X(\mathscr{H})\right).$$

Topological invariants of the geometric gap complex

$$\left|\Delta^X(\mathscr{H})\right| / \left|\delta^X(\mathscr{H})\right| \tag{13}$$

such as homology groups and homotopy groups, provide different means of measuring how far $\mathscr{H}$ is from being a Δ-set.

Algebraically, at the chain level

$$\begin{aligned} &C_*\left(\delta^X(\mathscr{H}); G\right) \hookrightarrow \inf_*^{C_*\left(\Delta^X; G\right)}(G(\mathscr{H})) \subseteq G(\mathscr{H}) \subseteq \sup_*^{C_*\left(\Delta^X; G\right)}(G(\mathscr{H})) \\ &\hookrightarrow C_*\left(\Delta^X(\mathscr{H}); G\right) \end{aligned} \tag{14}$$

where $G(\mathscr{H}) = \mathbb{Z}(\mathscr{H}) \otimes G$. An important consequence of Proposition 2.3 is that the inclusion

$$\inf_*^{C_*\left(\Delta^X; G\right)}{}_{(G(\mathscr{H}))} \to \sup_*^{C_*\left(\Delta^X; G\right)}{}_{(G(\mathscr{H}))}$$

is a chain homotopy equivalence, which implies that the algebraic gap complex (10) is acyclic. However, the geometric gap complex (13) is not contractible in general. For example, let X be any Δ-set and let $\mathscr{H}$ be the graded subset of X by removing the vertex set X_0. Then $\delta^X(\mathscr{H}) = \varnothing$ and so its geometric gap complex is $|X|^+$, the space $|X|$ disjoint union with a one-point set. The homology of the chain complexes

$$\begin{aligned} &\inf_*^{C_*\left(\Delta^X; G\right)}{}_{(G(\mathscr{H}))} / C_*\left(\delta^X(\mathscr{H}); G\right), \\ &C_*\left(\Delta^X(\mathscr{H}); G\right) / \sup_*^{C} * \left(\Delta^X; G\right)(G(\mathscr{H})) \end{aligned}$$

could give extra information in addition to the topology of the geometric gap complex (13).

The proof of the following proposition is immediate.

Proposition 2.22. *Let* $(\mathscr{H}, X)$ *be a super-hypergraph. Then*

$$\begin{aligned} &\inf_*^{C_*\left(\Delta^X(\mathscr{H}); G\right)}{}_{(G(\mathscr{H}))} / C_*\left(\delta^X(\mathscr{H}); G\right) = \inf_*^{C_*\left(\Delta^X(\mathscr{H}); G\right)} / C_*\left(\delta^X(\mathscr{H}); G\right)\left(G(\mathscr{H}) / C_*\left(\delta^X(\mathscr{H}); G\right)\right), \\ &\sup_*^{C_*\left(\Delta^X(\mathscr{H}); G\right)}{}_{(G(\mathscr{H}))} / C_*\left(\delta^X(\mathscr{H}); G\right) = \sup_*^{C_*\left(\Delta^X(\mathscr{H}); G\right)} / C_*\left(\delta^X(\mathscr{H}); G\right)\left(G(\mathscr{H}) / C_*\left(\delta^X(\mathscr{H}); G\right)\right). \end{aligned}$$

□

2.3.5. Computations.—With the potential applications in mind, it is important to consider the computability of these topological constructions. There have been various algorithms developed for computing simplicial homology that have led to the applications of topology in data analytics. The computations of embedded homology are quite similar to those of simplicial homology. Below we detail a procedure for computing embedded homology.

For a super-hypergraph $(\mathscr{H}, X)$, let us consider the computations for $H_*^{\mathrm{emb},X}(\mathscr{H}, F)$ using the chain complex $\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})$ with coefficients in a field $\mathbb{F}$.

By definition,

$$H_n^{\mathrm{emb},X}(\mathscr{H}, F) = Z_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right) / B_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right)$$

where

$$Z_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right) = \mathbb{F}(\mathscr{H}_n) \cap \mathrm{Ker}(\partial_n : C_n(X;\mathbb{F}) \to C_{n-1}(X;\mathbb{F})),$$
$$B_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right) = \mathbb{F}(\mathscr{H}_n) \cap \partial_{n+1}(C_{n+1}(X;\mathbb{F})).$$

The *Betti number* $b_n(\mathscr{H}, X)$ is defined as

$$\begin{aligned} b_n(\mathscr{H}, X) &= \dim H_n^{\mathrm{emb},X}(\mathscr{H};\mathbb{F}) \\ &= \dim\left(Z_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right)\right) - \dim\left(B_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right)\right). \end{aligned} \tag{15}$$

To compute $Z_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right)$, we can consider the restriction of the linear transformation $\partial_n : C_n(X;\mathbb{F}) \to C_{n-1}(X;\mathbb{F})$ to $\mathbb{F}(\mathscr{H}_n)$. Namely, consider $\mathbb{F}(\mathscr{H}_n)$ as a vector spaces over $\mathbb{F}$ with a basis given by the elements in $\mathscr{H}_n$. For each element x in $\mathscr{H}_n$, express $\partial_n(x)$ as an element in $C_{n-1}(X;\mathbb{F}) = \mathbb{F}(X_{n-1})$. This defines a linear transformation

$$\partial_n | : \mathbb{F}(\mathscr{H}_n) \to \mathbb{F}(X_{n-1})$$

whose kernel is $Z_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right)$.

For computing $B_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right)$, we can first consider $\partial_{n+1}(C_{n+1}(X;\mathbb{F}))$ as a subspace of the vector space $C_n(X;\mathbb{F}) = \mathbb{F}(X_n)$, which is spanned by linear combinations of $\partial_{n+1}(\sigma)$ for $\sigma \in X_{n+1}$. Then consider the decomposition

$$\mathbb{F}(X_n) = \mathbb{F}(\mathscr{H}_n) \oplus \mathbb{F}(X_n \backslash \mathscr{H}_n).$$

Let

$$p: \mathbb{F}(X_n) \to \mathbb{F}(X_n \backslash \mathscr{H}_n)$$

be the projection. Then $B_n\left(\inf_*^{C_*(X;\mathbb{F})}(\mathscr{H};\mathbb{F})\right)$ is the kernel of the restriction

$$p \mid : \partial_{n+1}(C_{n+1}(X;\mathbb{F})) \to \mathbb{F}(X_n \backslash \mathscr{H}_n).$$

If the data $(\mathscr{H}, X)$ is large, the complexity of direct computation of $H_*^{\mathrm{emb}, X}(\mathscr{H}, F)$ increases. The existing computational methods for chain complexes aim at reducing this complexity.

3. Super-persistent homology.

The general idea of super-persistent homology is to use the geometry of Δ-sets and super-hypergraphs as tools to investigate collections of subgraphs in a given graph and to get topological features from for example graphic data and various networks. We will see that a Δ-set model performs much better than models using simplicial complexes, particularly for exploring topological features arising from the structures related to clustering.

3.1. General theory.

Let G be a directed/undirected (multi)-graph. Let $\mathscr{FP}(G)$ denote the set of all finite subgraphs of G.

Definition 3.1. A Δ-set X is said to be *dominated by* G if there exists an injective map $\phi: X \to \mathscr{FP}(G)$ such that $\phi(d_i(\sigma))$ is a subgraph of $\phi(\sigma)$ for any $0 \le i \le n$ and any element $\sigma \in X_n$.

A super-hypergraph $(\mathscr{H}, X)$ is said to be *dominated by* G if its parental Δ-set X is dominated by G.

For a Δ-set X dominated by G, we identify the elements σ in X with its image $\phi(\sigma)$, a finite subgraph of G, and so we consider X as a collection of finite subgraphs of G. A super-hypergraph $(\mathscr{H}, X)$ dominated by G can be described as a multi-layered collection of families of finite subgraphs $X = \{X_0, X_1, \ldots\}$, where each X_i is a family of finite subgraphs, with face operations $d_i: X_n \to X_{n-1}$, $0 \le i \le n$, satisfying the Δ-identity, and a marked graded subset $\mathscr{H}$ of X.

To introduce persistence we need a scoring scheme on G. For graphs P and Q, denote by $P \le Q$ if P is a subgraph of Q.

Definition 3.2. A *scoring scheme* on a directed/undirected (multi-)graph G is a function

$$\mathfrak{M}:\mathscr{FP}(G)\to\mathbb{R}$$

from the set of finite subgraphs of G to the real numbers.

A scoring scheme $\mathfrak{M}$ on G is called *regular* if for every $P, Q\in\mathscr{FP}(G)$ such that $P\preceq Q$,

$$\mathfrak{M}(P)\leq\mathfrak{M}(Q).$$

Definition 3.3. A *persistent* Δ-*filtration* of a Δ-set X over $\mathbb{R}$ is a family of Δ-subsets $X(t)$ of X, indexed by $t\in\mathbb{R}$, such that

1. $X(s)$ is a Δ-subset of $X(t)$ for $s\leq t$, and
2. $X=\bigcup_{t\in\mathbb{R}}X(t)$.

A *persistent super-hypergraph filtration* of a super-hypergraph $(\mathscr{H}, X)$ over $\mathbb{R}$ is a family of super-hypergraphs $(\mathscr{H}(t), X(t))$, indexed by $t\in\mathbb{R}$, such that

1. The indexed family $X(t)$, $t\in\mathbb{R}$, is a persistent Δ-filtration of X over $\mathbb{R}$,
2. $\mathscr{H}(t)=\mathscr{H}\cap X(t)$.

Proposition 3.4. *Let G be a directed/undirected (multi-)graph with a regular scoring scheme $\mathfrak{M}:\mathscr{FP}(G)\to\mathbb{R}$. Let $(\mathscr{H}, X)$ be a super-hypergraph dominated by G. Then*

$$X(t)=\mathfrak{M}^{-1}((-\infty,t])\cap X,\quad t\in\mathbb{R}$$

is a persistent Δ-*filtration of X, and the pair*

$$(\mathscr{H}(t),X(t))=\left(\mathfrak{M}^{-1}((-\infty,t])\cap\mathscr{H},\mathfrak{M}^{-1}((-\infty,t])\cap X\right),\quad a\in\mathbb{R}$$

is a persistent super-hypergraph filtration of $(\mathscr{H}, X)$.

Proof. The proof follows from the definitions. □

Let $\mathbb{F}$ be a fixed choice of a ground field. A (*graded/ungraded*) *persistence module* over $\mathbb{F}$, denoted by $\mathbb{V}$, is defined to be an indexed family of (graded/ungraded) $\mathbb{F}$ vector spaces

$$\mathbb{V}=(V(t)\mid t\in\mathbb{R})$$

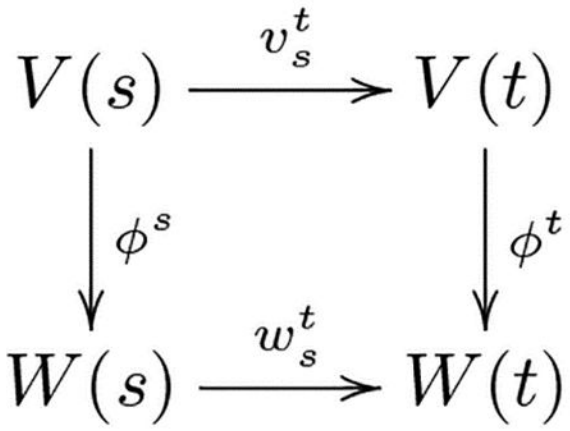
and a bi-indexed family of (graded/ungraded) linear maps

$$\left(v_s^t:V(s)\to V(t)\mid s\leq t\right)$$

which satisfy the composition law

$$v_s^t \circ v_r^s = v_r^t$$

whenever $r \le s \le t$, where v_t^t is the identity map on $V(t)$. For graded vector spaces V and W, a *graded linear map* of degree q is a collection of linear maps $\phi = (\phi_n)_{n \in \mathbb{Z}}$ with $\phi_n\colon V_n \to W_{n+q}$. A *persistence morphism* Φ *of dregree* q between two graded persistence modules $\mathbb{V}$ and $\mathbb{W}$ is a collection of graded linear maps of degree q, $\left(\phi^t : V(t) \to W(t) \mid t \in \mathbb{R}\right)$, such that the diagram



commutes for $s \le t$. If $q = 0$, then Φ is called a *persistence morphism*. A persistence morphism between ungraded persistence modules is defined in the same way.

Definition 3.5. Let $(\mathscr{H}, X)$ be a super-hypergraph. Let A be an abelian group. The *relative embedded homology* $H_*^{\text{emb}, X}(X, \mathscr{H}; A)$ *with coefficients in* A of $(\mathscr{H}, X)$ is defined by

$$\begin{aligned} H_*^{\text{emb}, X}(X, \mathscr{H}; A) &= H_*\left(C_*(X; A)/\inf_*^{C_*(X;A)}(\mathbb{Z}(\mathscr{H}) \otimes A)\right) \\ &\cong H_*\left(C_*(X; A)/\sup_*^{C_*(X;A)}(\mathbb{Z}(\mathscr{H}) \otimes A)\right). \end{aligned}$$

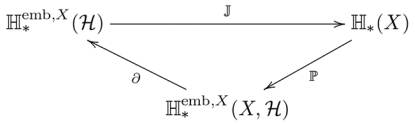
Now we assume that all homology is taken with coefficients in a ground field $\mathbb{F}$. Therefore, we can simplify our notation of homology groups $H_*(-;\mathbb{F})$ to $H_*(-)$. Let $(\mathscr{H}, X)$ be a super-hypergraph dominated by a directed/undirected (multi-)graph G with a scoring scheme $\mathfrak{M}$. Let $X(t) = \mathfrak{M}^{-1}((-\infty, t]) \cap X$ and $\mathscr{H}(t) = \mathfrak{M}^{-1}((-\infty, t]) \cap \mathscr{H}$. Then

$$\begin{aligned} &\mathbb{H}_*(X) = \left(H_*(X(t)) \mid t \in \mathbb{R}\right) \\ &\mathbb{H}_*^{\text{emb}, X}(\mathscr{H}) = \left(H_*^{\text{emb}, X}(\mathscr{H}(t)) \mid t \in \mathbb{R}\right) \quad \text{and} \\ &\mathbb{H}_*^{emb, X}(X, \mathscr{H}) = \left(H_*^{\text{emb}, X}(X, \mathscr{H}(t)) \mid t \in \mathbb{R}\right) \end{aligned} \tag{16}$$

are graded persistence modules. The short exact sequence of chain complexes

$$\inf_*^{C_*(X(t))}(\mathbb{F}(\mathscr{H}(t))) \xrightarrow{j^t} C_*(X(t)) \xrightarrow{p^t} C_*(X(t))/\inf_*^{C_*(X(t))}(\mathbb{F}(\mathscr{H}(t)))$$

induces a long exact sequence on homology, which can be written as an exact triangle of graded persistence modules



(17)

where $\mathbb{J}$ and $\mathbb{P}$ are persistence morphisms and the boundary homomorphism ∂ is persistence morphism of degree -1.

Definition 3.6. Let $(\mathscr{H}, X)$ be a super-hypergraph dominated by a directed/undirected (multi-)graph G with a scoring scheme. Then the three graded persistence modules listed in (16) together with the exact triangle (17) give a *super-persistent homology* of $(\mathscr{H}, X)$ with coefficients in a field $\mathbb{F}$.

Equally, we could call $\mathbb{H}_*^{emb,X}(\mathscr{H})$ super-persistent homology. However, the definition given above can carry more information than the embedded homology of the super-hypergraph, as we will now illustrate.

Let $J \subseteq \mathbb{R}$ be an interval. The *interval persistence module* $\mathbb{F}^J = (J(t) \mid t \in \mathbb{R})$ is defined by

$$J(t) = \begin{cases} \mathbb{F} & \text{if } t \in J, \\ 0 & \text{otherwise} \end{cases}$$

with double indexed linear maps

$$j_s^t = \begin{cases} \mathrm{id} & \text{if } s,t \in J, \\ 0 & \text{otherwise.} \end{cases}$$

For an ungraded persistence module $\mathbb{V}$, the q-th *suspension* $\Sigma^q\mathbb{V}$ is a graded persistence module with

$$\Sigma^q\mathbb{V}_n = \begin{cases} \mathbb{V} & \text{if } n = q \\ 0 & \text{otherwise.} \end{cases}$$

A *graded interval persistence module* is a q-th suspension of the ungraded interval persistence module for some q.

Recall that a Δ-set X is called of *finite type* if X_n is finite for each n.

Theorem 3.7 (Structure Theorem). *Let* $(\mathscr{H}, X)$ *be a super-hypergraph dominated by a directed/undirected (multi-)graph G with a scoring scheme. Suppose that X is of finite type. Then the graded persistence modules* $\mathbb{H}_*(X)$, $\mathbb{H}_*^{\mathrm{emb},X}(\mathscr{H})$ *and* $\mathbb{H}_*^{\mathrm{emb},X}(X,\mathscr{H})$ *admit direct sum decompositions in terms of graded interval persistence modules and these decompositions are unique up to the order of factors in the category of graded persistence modules.*

Proof. For any graded persistence module $\mathbb{V}$, there is a canonical decomposition

$$\mathbb{V} \cong \bigoplus_{n \in \mathbb{Z}} \Sigma^n \mathbb{V}_n$$

in the category of graded persistence modules, where $\mathbb{V}_n$ is considered as an ungraded persistence module. It suffices to show that $\mathbb{H}_n(X)$, $\mathbb{H}_n^{\mathrm{emb},X}(\mathscr{H})$ and $\mathbb{H}_n^{\mathrm{emb},X}(X,\mathscr{H})$ admit unique factorization as ungraded persistence modules for each n.

Since X is of finite type, the chain complex $C_*(X)$ is of finite type and so is any subcomplex or quotient complex. The assertion follows from the structure theorem on persistence modules [46, Theorem 1.1] derived from the classical Gabriel Theorem in representation theory [65]. □

We now briefly summarise persistence diagrams/persistent barcodes; for which this structure theorem is prominent in the calculations, for more details see [39].

Let $\mathbb{V}$ be an ungraded persistence module with a unique decomposition (up to the order of factors)

$$\mathbb{V} = \bigoplus_{\alpha \in I} \mathbb{F}^{J_\alpha}$$

in terms of interval persistence modules, where I is the index set. Then the multiset given by $\inf(J_\alpha), \sup(J_\alpha)) \subset \mathbb{R}^2$, $\alpha \in I$, is the *persistence diagram* (or *barcode*) of $\mathbb{V}$, denoted by $\mathrm{dgm}(\mathbb{V})$. In our case, under the hypothesis that X is of finite type, we have three persistence diagrams.

Corollary 3.8. *Let* $(\mathscr{H}, X)$ *be a super-hypergraph dominated by a directed/undirected (multi-)graph G with a scoring scheme and let X be of finite type. Then there are three multi-layer persistence diagrams* $\mathrm{dgm}(\mathbb{H}_n(X))$, $\mathrm{dgm}\left(\mathbb{H}_n^{\mathrm{emb},X}(\mathscr{H})\right)$ *and* $\mathrm{dgm}\left(\mathbb{H}_n^{\mathrm{emb},X}(X,\mathscr{H})\right)$ *for* $n \geq 0$. □

The exact triangle (17) yields matrix data on the correlations of the multi-layer persistence diagrams as follows. Let

$$\boldsymbol{\Phi}: \mathbb{V} \to \mathbb{W}$$

be a persistence morphism between ungraded persistence modules. Suppose that both $\mathbb{V}$ and $\mathbb{W}$ satisfy the unique factorization property with respect to decompositions in terms of interval persistence modules and let

$$\mathbb{V} = \bigoplus_{\alpha \in I_V} \mathbb{F}^{J_\alpha} \text{ and } \mathbb{W} = \bigoplus_{\beta \in I_W} \mathbb{F}^{J_\beta}.$$

Let $\Phi_{\alpha,\beta}$ be the composite

$$\Phi_{\alpha,\beta}: \mathbb{F}^{J_\alpha} \xrightarrow{\Phi|\mathbb{F}^{J_\alpha}} \mathbb{W} \xrightarrow{\text{proj}} \mathbb{F}^{J_\beta}.$$

According to [26, Proposition 16, Lemma 22], the set of persistence morphisms between two interval persistence modules is either a 1-dimensional vector space or 0. Thus $\Phi_{\alpha,\beta}: \mathbb{F}^{J_\alpha} \to \mathbb{F}^{J_\beta}$ is either a generator for $\text{Hom}\left(\mathbb{F}^{J_\alpha}, \mathbb{F}^{J_\beta}\right)$, or zero. Let the index sets I_V and I_W be totally ordered and define the *correlation matrix*

$$M(\Phi) = \left(m_{\alpha,\beta}\right)_{\alpha \in I_V, \beta \in I_W}$$

of Φ by setting

$$m_{\alpha,\beta} = \begin{cases} 1 & \text{if } \Phi_{\alpha,\beta} \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

The correlation matrix is an analogue of adjacency matrix of graphs, which gives correlations between $\text{dgm}(\mathbb{V})$ and $\text{dgm}(\mathbb{W})$ by adding directed edges. In summary, we have the following information data from super-persistent homology.

Proposition 3.9. *Let* $(\mathscr{H}, X)$ *be a super-hypergraph dominated by a directed/undirected (multi-)graph* G *with a scoring scheme. Suppose that* X *is of finite type. Then there are three multi-layer persistence diagrams* $\text{dgm}(\mathbb{H}_*(X))$, $\text{dgm}\left(\mathbb{H}_*^{\text{emb},X}(\mathscr{H})\right)$ *and* $\text{dgm}\left(\mathbb{H}_*^{\text{emb},X}(X,\mathscr{H})\right)$ *together with three correlation matrices* $M(\mathbb{J})$, $M(\mathbb{P})$ *and* $M(\partial)$ *between them.* □

An important point is that we allow $\mathscr{H}$ to be an arbitrary graded subset of X. If we fix X to be a Δ-set dominated by a graph G with a scoring scheme and allow $\mathscr{H}$ to be random, then $\text{dgm}(\mathbb{H}_*(X))$ is a deterministic barcode, while $\text{dgm}\left(\mathbb{H}_*^{\text{emb},X}(\mathscr{H})\right)$ and $\text{dgm}\left(\mathbb{H}_*^{\text{emb},X}(X,\mathscr{H})\right)$ are random. The correlation matrices may help further analyse the data.

Let X be a fixed Δ-set dominated by a directed/undirected (multi-)graph. It would be also interesting to consider the set $\mathbb{P}(X)$ of the isomorphic classes of persistence modules (16) for all graded subsets $\mathscr{H}$ of X. The inclusions $\mathscr{H}' \subseteq \mathscr{H} \subseteq X$ induce a morphism of super-hypergraphs $(\mathscr{H}', X) \to (\mathscr{H}, X)$. By taking super-persistent homology, one would get a quiver structure on the set $\mathbb{P}(X)$. Moreover, there is an *interleaving distance* between persistence modules introduced in [38] corresponding to *bottleneck distance* [39], which

gives the structure of a metric space on the quiver $\mathbb{P}(X)$. The following example illustrates that $\mathbb{P}(X)$ may give more robust information.

Example 3.10. Let X be a Δ-set. Let $\mathscr{H}$ be a graded subset of X consisting of *non-face* elements $\sigma \in X$. Here a non-face element σ means that there does not exist any element $\tau \in X$ such that $d_i\tau = \sigma$ for some i. In other words, $\mathscr{H}$ is given by removing all face elements in X. It is straightforward to see that

$$H_*^{\mathrm{emb},X}(\mathscr{H}) = \mathbb{F}(\mathscr{H}) \cap \mathbb{Z}(C_*(X)).$$

Therefore, the embedded homology can detect the cycles contributed from non-face elements. From the exact triangle (17), a part of the boundaries in $C_*(X)$ can also be detected. These detected elements would contribute to bar-codes through persistence. Hence, in addition to the persistence on the homology $H_*(X)$, $\mathbb{P}(X)$ can decode more topological features.

3.2. The ordinary persistent homology.

The classical persistent homology refers to the persistent homology of point cloud data using the Vietoris-Rips complex, the Čech complex or the witness complex. Persistent homology has been used as an important topological tool in data science. In this subsection, we rewrite on the classical persistent homology from the viewpoint of graphs with scoring schemes, and then give a natural generalization to persistent homology for graphs with reference maps to metric spaces.

3.2.1. The classical persistent homology.—

A *point cloud* dataset is a finite set $\mathscr{L}$ with a reference map that embeds $\mathscr{L}$ into a finite dimensional Euclidean space $\mathbb{R}^m$, thus we can consider $\mathscr{L}$ as a finite subset in $\mathbb{R}^m$. We now use a graph with a scoring scheme to describe the (persistent) Vietoris-Rips complex, Čech complex and witness complex in a unified way. The graph $G(\mathscr{L})$ is the complete graph having $\mathscr{L}$ as its vertex set. Intuitively, we assign one and only one edge to any two distinct points in $\mathscr{L}$. The main point is to show that different scoring schemes on $G(\mathscr{L})$ can obtain different persistent complexes that are currently widely used in TDA [36]. Below we give the scoring schemes for the Vietoris-Rips complex, the Čech complex, the strong witness complex and the weak witness complex. Let $\Lambda = \{l_0, l_1, \ldots, l_n\} \subseteq \mathscr{L}$ be a subset of $\mathscr{L}$.

For $x \in \mathbb{R}^m$, denote by $B(x, r)$ the closed ball of radius r centered at x. Define the following scoreing schemes on Λ:

1. The *Vietoris-Rips scoring* is given by

$$\mathfrak{M}^{VR}(\Lambda) = \frac{1}{2}\sup\{d(l_i, l_j) \mid l_i, l_j \in \Lambda \subseteq \mathbb{R}^m\}. \tag{18}$$

The balls $B(l_0, r), \ldots, B(l_n, r)$ pairwise intersect if and only if the score $\mathfrak{M}^{VR}(\Lambda) \leq r$.

2. The *Čech scoring* is defined by

$$\mathfrak{M}^{\check{C}}(\Lambda) = \inf_{x \in \mathbb{R}^m} \max\{d(x,l) \mid l \in \Lambda \subseteq \mathbb{R}^m\}. \tag{19}$$

Note that

$$\bigcap_{i=0}^{n} B(l_i, r) \neq \emptyset$$

if and only if the score $\mathfrak{m}^{\check{C}}(\Lambda) < r$.

3. The *Strong witness scoring* is defined by

$$\mathfrak{M}^{W^s}(\Lambda) = \inf_{x \in \mathbb{R}^m} \{\sup_{y \in \Lambda} d(x,y) - \inf_{z \in \mathscr{L}} d(x,z)\}. \tag{20}$$

4. The *Vietoris-Rips Strong witness scoring* is given by

$$\mathfrak{M}^{W^s_{VR}}(\Lambda) = \sup_{0 \leq i < j \leq n} \{\inf_{x \in \mathbb{R}^m} \{\max\{d(x,l_i), d(x,l_j)\} - \inf_{z \in \mathscr{L}} d(x,z)\}\}. \tag{21}$$

5. The *Weak witness scoring* is given by

$$\mathfrak{M}^{W^w}(\Lambda) = \inf_{x \in \mathbb{R}^m} \{\sup_{y \in \Lambda} d(x,y) - \inf_{z \in \mathscr{L} \setminus \Lambda} d(x,z)\}. \tag{22}$$

6. The *Vietoris-Rips weak witness scoring* is given by

$$\mathfrak{M}^{W^w_{VR}}(\Lambda) = \sup_{0 \leq i < j \leq n} \{\inf_{x \in \mathbb{R}^m} \{\max\{d(x,l_i), d(x,l_j)\} - \inf_{z \in \mathscr{L} \setminus \Lambda} d(x,z)\}\}. \tag{23}$$

Let X be the clique complex of $G(\mathscr{L})$ considered as a Δ-set. In other words, because G is complete, X is the set of all nonempty full subgraphs of $G(\mathscr{L})$. Here a *full subgraph* H of G is a subgraph H such that the edge set between any two vertices v and w in H is equal to the edge set between v and w in G. Choose a linear order on the vertex set $V(G(\mathscr{L}))$. Then (X, X) is a super-hypergraph. It is straightforward to check that the persistent super-hypergraph filtrations on (X, X) induced by the above scoring schemes coincide with the classical persistent filtrations in [36]. Here the witness scoring schemes defined in (3)–(6) are reformulations from [36, Definition 2.7, Definition 2.8].

3.2.2. Clique persistent homology of graphs with reference maps.—

Next, we will consider a canonical extension of ordinary persistent homology to the case of graphs with reference maps on vertices. Let G be a finite undirected (multi-)graph with a reference map that embeds the vertex set $V(G)$ into a finite dimensional Euclidean space $\mathbb{R}^m$ and let H be any subgraph. Then, any one of the six scoring schemes in (18)–(23) induces a persistent filtration on the clique complex clique(G). This gives the *clique persistent homology* on G.

The clique persistent homology on G could, in general, be quite different from the ordinary persistent homology of the vertex set of G under the reference map. For instance, if the clique complex clique(G) has non-trivial reduced homology, the resulting clique persistent homology converges to $H_*(\text{clique}(G))$ as $t \to \infty$, but the ordinary persistent homology converges to trivial homology as $t \to \infty$. The ordinary persistent homology of $V(G)$ under the reference map is obtained by rebuilding a new graph given by the complete graph on $V(G)$, that is, all edges in G are forgotten and the new edges are added in depending on the scoring schemes that are obtained through the reference map. On the other hand, when we consider the clique complex of G itself, the edges in G are accounted for.

The following example illustrates how clique persistent homology can describe shapes and therefore could be useful for data analysis on protein structures or image processing on 3D objects with complicated internal structures such as the heart.

Example 3.11. Let X be a polyhedron in $\mathbb{R}^m$. Let K be a simplicial complex that is a triangulation of X and let G be the graph given by the 1-skeleton of the barycentric subdivision of K. Let the reference map on G be given by the inclusion of G in $\mathbb{R}^m$. Then the geometric realization of the clique complex clique(G) is homeomorphic to X, and the persistent homology of clique(G) converges to $H_*(X)$. In particular, the number of infinite persistence modules in the n^{th} persistent homology of clique(G) is equal to the n^{th} Betti number of X. Thus the clique persistent homology detects the topological shape of X.

More generally we can remove the embedding hypothesis of reference maps. Let G be a finite undirected (multi-)graph and let

$$f : V(G) \to \mathbb{R}^m$$

be a function (without assuming injectivity). We can use *pull-back scoring* in the following sense. Let H be any subgraph of G. Then the image $f(V(H))$ is a finite subset located in $\mathbb{R}^m$. Let $\mathfrak{M}$ be a scoring scheme on point cloud data such as one of the six aforementioned scoring schemes. Define

$$\mathfrak{M}^f(H) = \mathfrak{M}(f(V(H))) \quad \text{(Pull-back Scoring)} \tag{24}$$

which induces a persistent filtration on the clique complex clique(G) depending on the reference map f. Different choices of f would result in different persistence diagrams. For instance, a constant function does induce a trivial persistence on $H_*(\text{clique}(G))$. The flexibility of f could be useful. For example, if f is randomly given, it induces corresponding random persistence diagram.

The following example illustrates that a pull-back scoring on clique persistent homology may be useful for detecting higher dimensional geometric shapes.

Example 3.12. Let $p: E \to B$ be a continuous map between polyhedra E and B. Assume that B is a subspace of $\mathbb{R}^m$. By triangulating B, we pull it back along p do define K^E and there is a simplicial map $p' : K^E \to K^B$ such that

1. K^E and K^B are simplicial complexes of triangulations of E and B, respectively.
2. Let $G(K^E)$ and $G(K^B)$ be the graphs given by the 1-skeleton of K^E and K^B, respectively. Then clique($G(K^E)$) = K^E and clique($G(K^B)$) = K^B.
3. p' is a simplicial approximation to p.

Let the reference map R on $V(G(K^B))$ be given by the inclusion $V\left(G\left(K^B\right)\right) \subseteq B \subseteq \mathbb{R}^m$. Let us take the pull-back scoring $\mathfrak{M}$ on $G(K^E)$). Then the persistent filtration on clique($G(K^E)$) induced by $\mathfrak{M}$ is the pull-back of the persistent filtration on Clique($G(K^B)$) induced by R. In particular, if $p : E \to B$ is a fibre bundle or, more generally, a fibration with fibre F, then we have a persistent Leray-Serre spectral sequence convergent to $H_*(E)$. It is well-known in algebraic topology that Leray-Serre spectral sequences are an important tool for computing $H_*(E)$ starting with $H_*(B)$ and $H_*(F)$.

3.3. Partition homology and persistent partition homology.

The methods of data science are typically aimed at finding structures and patterns within large datasets. Being able to glean information about the internal structures of graphical data would be useful in solving the typical problems given to machine learning algorithms. For example, classification problems, prediction and, in particular, partitioning data into clusters [114, Section 1.1.3].

If a collection of subgraphs forms a Δ-set structure, then we can calculate homology. A natural question is how to introduce a Δ-set structure on a collection of subgraphs in some natural way. More precisely, how to define face operations on subgraphs. We are going to show that any clustering on the vertex set can induce canonical face operations on subgraphs. For a dataset given by a graph, the topological features on collections of subgraphs under the face operations induced by a clustering may help for detecting correlations between the clusters.

Let G be a directed/undirected (multi-)graph. Assume that there is a disjoint clustering $\mathbf{p}$ on the vertex set $V(G)$. In other words, there is a disjoint union

$$V(G) = \coprod_{i=0}^{m} V_i(G)$$

under the clustering $\mathbf{p}$, where each $V_i(G)$ is a cluster. Let H be a subgraph of G. Then there exists a unique sequence $(k_0, k_1, \ldots, k_n)$ with $0 \le k_0 < k_1 < \ldots < k_n \le m$ such that $V(H) \cap V_i(G) \neq \varnothing$ for $i \in \{k_0, k_1, \ldots, k_n\}$ and $V(H) \cap V_i(G) = \varnothing$ if $i \notin \{k_0, k_1, \ldots, k_n\}$. We call H a subgraph of G *linked to* $(n+1)$ *clusters*. Viewing H as an abstract n-simplex, we define the j^{th}-face map, $d_j^{\mathbf{p}}(H)$, as the full subgraph of H formed after removing all of those vertices $v \in V(H) \cap V_{k_j}(G)$ together with the edges incident to (or from) such v. The resulting subgraph $d_j^{\mathbf{p}}(H)$ is linked to n clusters with $V\left(d_j^{\mathbf{p}}(H)\right) \cap V_{k_j}(G) = \varnothing$. It is straightforward to show that the Δ-identity

$$d_i^{\mathbf{p}} d_j^{\mathbf{p}} = d_j^{\mathbf{p}} d_{i+1}^{\mathbf{p}}$$

for $i \geq j$ holds. The face operation $d_j^{\mathbf{p}}$ is *induced by the disjoint clustering* $\mathbf{p}$.

Now let $\mathscr{H}$ be any collection of subgraphs of G. Define $\mathscr{H}_n$ to be the subset of $\mathscr{H}$ consisting of those subgraphs in $\mathscr{H}$ linked to $(n+1)$ clusters. This gives a graded structure on $\mathscr{H} = \{\mathscr{H}_n\}_{n \geq 0}$. Let $X(\mathscr{H})$ be the collection of subgraphs of G given by all of the elements in $\mathscr{H}$ together with all iterated faces. Then $X(\mathscr{H})$ is a Δ-set and $(\mathscr{H}, X(\mathscr{H}))$ is a super-hypergraph. The resulting homology groups

$$H_*(X), H_*^{\text{emb}, X}(\mathscr{H}), H_*^{\text{emb}, X}(X, \mathscr{H})$$

are called the *partition homology*.

In general, $X(\mathscr{H})$ may not be a simplicial complex. In simplicial complexes, the assumption that simplices are determined by their vertices is too strict with applications in mind. For exploring topological structures arising from disjoint clusterings, a Δ-set is a more suitable notion.

In practice, for studying possible correlations between clusters, one could start with a collection of one or more subgraphs, $\mathscr{H}$, linked with some or all clusters, and then produce the Δ-set $X(\mathscr{H})$. Due to the nature of simplicial homology, higher dimensional homology of X, and higher dimensional embedded homology of $(\mathscr{H}, X)$ would give topological features measuring the group correlations between more clusters. In particular, we have set up the Δ-set $X(\mathscr{H})$ such that the homological dimension of a given subgraph H in X is $n-1$, where n is the number of clusters linked with H.

In theory, we can start with any collection of subgraphs as the initial data $\mathscr{H}$. For instance, we can start with $\mathscr{H}$ given by all or some of the k-regular subgraphs of G, Eulerian subgraphs, traceable subgraphs or Hamitonian subgraphs as initial data for constructing the Δ-set $X(\mathscr{H})$. This would give different topological approaches to understanding the internal structures of $\mathscr{H}$ in addition to what we will survey in Appendix A.

If there is a scoring scheme on G, then we can calculate the super-persistent homology of $(\mathscr{H}, X)$ called *persistent partition homology*. A scoring scheme on G can be deterministically or randomly given. If a scoring scheme is randomly given, it may not be regular. This means that in the induced persistent filtration on X, the graded subset $X(t)$ may not be a Δ-subset of X for all t. In this case, the persistence system can be modified replacing chains related to the terms $X(t)$ by infimum or supremum chains on $X(t)$. The resulting persistence modules and persistence diagrams can be modified accordingly.

From the perspective of data processing, a clustering may be compared against certain optimization properties. Currently, we use discrete Morse theory which works well on

simplicial complexes and cell complexes [63], and chain complexes [88] with applications in data analysis [115]. Moreover, the combinatorial Laplacian operator works well on simplicial complexes and chain complexes, where cohomology with coefficients in real numbers can be expressed as the null space of the Laplacian operator on cochains [82].

3.4. Other face operations.

We have shown how disjoint clusterings can induce face operations on subgraphs. This construction works well as a theory which unifies many constructions such as the clique complex and the neighborhood complex. However, if we are interested in subgraphs having some special properties, this construction has some disadvantages. For instance, if we are interested in collections of finite connected subgraphs H, then the subgraph of H given by removing some of its vertices may not be connected. In the following subsection, we will discuss some alternative ways of getting natural constructions of face operations.

3.4.1. Link-blowup face operations.—The idea of link-blowup face operations comes from geometric constructions on tubular neighborhoods of submanifolds or regular neighborhoods of subcomplexes. Let H be a subgraph of some graph G and let S be a subset of the vertex set $V(H)$. If we remove S from H, then we could add some edges from the working graph G to make a blowup for the subgraph $H \setminus S$. A natural way to add these edges is to consider the neighbors of vertices of S in H and to add the edges between these neighbors from the working graph G.

Let G be a directed/undirected (multi-)graph and let $E^G(v, w)$ denote the edge set between v and w for vertices $v, w \in V(G)$. Let $S \subseteq V(G)$ be a subset and define the *induced subgraph of S in G*, $G[S]$, to be the full subgraph of G having S as its vertex set. The *closed neighborhood set* $N[S]$ is defined by

$$N[S] = S \cup \{u \mid u \in V(G) \text{ adjacent to a vertex } v \in S\}$$

namely, $N[S]$ is the union of the neighborhoods of vertices $v \in S$. The *link set* of S in G is

$$\mathrm{Lk}(S) = N[S] \setminus S .$$

Let H be a subgraph of G and let $S \subseteq V(H)$ be a subset of the vertex set of H. The *link-blowup*[2] of H along S is defined as

$$H[V(H) \setminus S] \cup G[\mathrm{Lk}(S) \cap V(H)] .$$

Note that the graph $H[V(H) \setminus S] \cup G[\mathrm{Lk}(S) \cap V(H)]$ has the same vertex set of $H[V(H) \setminus S]$.

[2]This definition is taken from a geometric setting. We may consider the subgraph $G[N[S]]$ as a regular neighborhood of S. Then $\mathrm{Lk}(S)$ are the vertices located in the "*boundary*" of the regular neighborhood. Geometrically, we add all of edges in G joining vertices in $\mathrm{Lk}(S) \cap V(H)$ to form a blowup on $H[V(H) \setminus S]$.

Now suppose that there is a disjoint clustering **p** on the vertex set $V(G)$ so that there is a disjoint union

$$V(G) = \coprod_{i=0}^{m} V_i(G)$$

under **p** with each $V_i(G)$ a cluster. Let H be a subgraph. Then there exists a unique sequence $(k_0, k_1, \ldots, k_n)$ with $0 \le k_0 < k_1 < \ldots < k_n \le m$ such that $V(H) \cap V_i(G) \neq \varnothing$ for $i \in \{k_0, k_1, \ldots, k_n\}$ and $V(H) \cap V_i(G) = \varnothing$ if $i \notin \{k_0, k_1, \ldots, k_n\}$. Let $V_j(H) = V(H) \cap V_{k_j}(G)$. Define the *link-blowup face operation* as

$$d_j^{\mathrm{lk}}(H) = H\left[V(H)\backslash V_j(H)\right] \cup G\left[\mathrm{Lk}\left(V_j(H)\right) \cap V(H)\right] \tag{25}$$

that is, the link-blowup of H along $V_j(H) = V(H) \cap V_{k_j}(G)$ for $0 \le j \le n$.

Remark 3.13. For helping to understand the link-blowup face operation, one can give a coloring on the vertices of G so that the vertices in each cluster has the same color under the disjoint clustering **p**, and perform the link-blowup on the vertices having the same color.

Proposition 3.14. *Given a disjoint clustering* **p**, *let* d_j^{lk} *be defined as above. Then*

$$d_i^{\mathrm{lk}} d_j^{\mathrm{lk}} = d_j^{\mathrm{lk}} d_{i+1}^{\mathrm{lk}} \tag{26}$$

for $i \ge j$.

Proof. Let H be a subgraph of G. Let V_j denote $V_j(H)$ defined as above and let d_j denote d_j^{P}, the induced face operations from the disjoint clustering **p**. We will use the fact that

$$V\left(d_l^{\mathrm{lk}}(H)\right) = V(d_l(H)) = V(H)\backslash V_l\,.$$

Now

$$\begin{aligned} d_i^{\mathrm{lk}}\left(d_j^{\mathrm{lk}}(H)\right) &= d_j^{\mathrm{lk}}(H)\left[V\left(d_j^{\mathrm{lk}}(H)\right)\backslash V_{i+1}\right] \cup G\left[\mathrm{Lk}(V_{i+1}) \cap V\left(d_j^{\mathrm{lk}}(H)\right)\right] \\ &= d_j^{\mathrm{lk}}(H)\left[V\left(d_j^{\mathrm{lk}}(H)\right)\backslash V_{i+1}\right] \cup G\left[\mathrm{Lk}(V_{i+1}) \cap \left(V(H)\backslash V_j\right)\right] \\ &= d_j^{\mathrm{lk}}(H)\left[V\left(d_j^{\mathrm{lk}}(H)\right)\backslash V_{i+1}\right] \cup G\left[\mathrm{Lk}(V_{i+1}) \cap \left(V(H)\backslash V_j \backslash V_{i+1}\right)\right] \end{aligned}$$

because $\mathrm{Lk}(V_{i+1}) \cap V_{i+1} = \varnothing$.

The term $d_j^{\mathrm{lk}}(H)\left[V\left(d_j^{\mathrm{lk}}(H)\right)\backslash V_{i+1}\right]$ is the induced subgraph of $d_j^{\mathrm{lk}}(H)$ on its vertex subset

$$V\left(d_j^{\mathrm{lk}}(H)\right)\backslash V_{i+1} = V(H)\backslash V_j \backslash V_{i+1}\,.$$

By definition, $d_j^{\mathrm{lk}}(H) = H\left[V(H)\backslash V_j(H)\right] \cup G\left[\mathrm{Lk}\left(V_j(H)\right) \cap V(H)\right]$. Restricting to the vertex subset

$$W = V(H)\backslash V_j \backslash V_{i+1}$$

we have

$$d_j^{\mathrm{lk}}(H)\left[V\left(d_j^{\mathrm{lk}}(H)\right)\backslash V_{i+1}\right] = H[W] \cup G\left[\mathrm{Lk}\left(V_j\right) \cap W\right]$$

and so

$$d_i^{\mathrm{lk}}\left(d_j^{\mathrm{lk}}(H)\right) = H[W] \cup G\left[\mathrm{Lk}\left(V_j\right) \cap W\right] \cup G[\mathrm{Lk}(V_{i+1}) \cap W].$$

By the same arguments,

$$d_j^{\mathrm{lk}}\left(d_{i+1}^{\mathrm{k}}(H)\right) = H[W] \cup G\left[\mathrm{Lk}\left(V_j\right) \cap W\right] \cup G[\mathrm{Lk}(V_{i+1}) \cap W].$$

□

3.4.2. Face operations on subgraphs with marked starting-vertices.—Let H be a subgraph of G and let S be a subset of the vertex set $V(H)$. When we remove S from H, we wish to add as few edges as possible to make a blowup for the subgraph $H \backslash S$ with the aim of preserving particular properties of H. For this construction, consider an extension of the working graph G by adding an extra edge between any two distinct vertices v, w in G labeled as ∞_{vw}. This is an analogue to the idea of compactification in geometry. To showcase this we consider a special family of subgraphs.

Let G be a directed/undirected (multi-)graph.

Definition 3.15. A *subgraph with marked starting-vertices* of G is a pair (H, SV(H)) satisfying the following conditions:

1. H is a subgraph of G,
2. SV(H) is a subset of $V(H)$,
3. Every vertex v of H is reachable by a directed/undirected path out from a vertex in SV(H).

In the case of digraphs, one may require further that there are no directed edges in H incident into any vertex in SV(H). In this case, SV(H) acts as a source set for H. We want to give a description of face operations that is consistent across graphs and digraphs, therefore we do not require such an extra condition. There could be some redundant vertices contained in SV(H). A trivial example is to choose SV(H) = $V(H)$, in which case there are no face operations. If SV(H) is a proper subset of $V(H)$, there will be nontrivial face operations.

Let $(H, \mathrm{SV}(H))$ be a subgraph with marked starting-vertices of G. We will now recursively construct a partition on the vertex set $V(H)$, called the *neighborhood-extension partition*, in the following way. Let $V_0(H) = \mathrm{SV}(H)$ and suppose that $V_j(H)$ is constructed with $j \geq 0$. Define $V_{j+1}(H)$ as follows:

1. In the undirected case, let V_{j+1} be the link set of the subgraph $H[V_j]$ in the graph H.
2. In the directed case, let V_{j+1} be the *out-link set* of the subgraph $H[V_j]$ in the graph H. Here, for a (multi-)digraph Γ and a subgraph Γ', the *closed out-neighborhood set* of Γ' in Γ is the union of Γ' and the out-neighbors of Γ' in Γ, denoted by $N^{\mathrm{out}}(\Gamma')$. The *out-link set* of Γ' is defined as

$$\mathrm{Lk}^{\mathrm{out}}(\Gamma') = N^{\mathrm{out}}(\Gamma') \backslash V(\Gamma').$$

   For a subset S of $V(\Gamma)$, let $N^{\mathrm{out}}(S) = N^{\mathrm{out}}(\Gamma[S])$ and $\mathrm{Lk}^{\mathrm{out}}(S) = \mathrm{Lk}^{\mathrm{out}}(\Gamma[S])$.

If $(H, \mathrm{SV}(H))$ is a finite subgraph with marked starting-vertices of G, then above recursive construction will stop after finitely many steps, hence this gives a finite partition on $V(H)$.

To define face operations taking marked starting-vertices in to account, we embed G into a larger graph $\hat{G}$, where $V(\hat{G}) = V(G)$ and $E(\hat{G})$ is the extension of $E(G)$ by adding one edge ∞_{vw} for $v \neq w \in V(G)$ in the undirected case, and by adding two directed edges ∞_{vw} from v to w and ∞_{wv} from w to v for vertices $v \neq w \in V(G)$ in the directed case.

Now let $(H, \mathrm{SV}(H))$ be a finite subgraph with marked starting-vertices of $\hat{G}$ with the neighborhood extension partition

$$V(H) = \coprod_{i=0}^{n} V_i.$$

We assume that $H \neq \emptyset$ and so $V_0 = \mathrm{SV}(H) \neq \emptyset$. From the recursive definition, $V_{i+1} \neq \emptyset$ implies that $V_i \neq \emptyset$. Therefore, V_n is the last nonempty set in the recursive procedure.

We define the face operation d_j^{SV} on H, $0 \leq j \leq n$, with $n > 0$ as follows:

1. $d_0^{\mathrm{SV}}(H) = H[V(H) \backslash V_0]$ with $SV\left(d_0^{\mathrm{SV}}(H)\right) = V_1$.
2. $d_n^{\mathrm{SV}}(H) = H[V(H) \backslash V_n]$ with $SV\left(d_n^{\mathrm{SV}}(H)\right) = V_0$.
3. For $0 \leq j < n$,

$$d_j^{\mathrm{SV}}(H) = H\left[V(H) \backslash V_j\right] \cup \mathscr{E}^H\left(V_{j-1}, V_{j+1}\right)$$

   with $\mathrm{SV}\left(d_n^{\mathrm{SV}}(H)\right) = V_0$, where $\mathscr{E}^H\left(V_{j-1}, V_{j+1}\right)$ is a subset of the edge set $E(\hat{G})$ consisting of ∞_{vw} for $v \in V_{j-1}$ and $w \in V_{j+1}$ satisfying the property that there does not exist an edge from v to w in H[3].

Let $V_{-1} = V_{n+1} = \varnothing$. Then we can write in a unified way that

$$d_j^{\mathrm{SV}}(H) = H\left[V(H)\backslash V_j\right] \cup \mathscr{E}^H\left(V_{j-1}, V_{j+1}\right) \text{ with } \mathrm{SV}\left(d_j^{\mathrm{SV}}(H)\right) = V_{\delta_{j,0}} \tag{27}$$

for $0 \leq j \leq n$, where $\delta_{a,b}$ is the Kronecker δ symbol.

We need to show that $\left(d_j^{\mathrm{SV}}(H), \mathrm{SV}\left(d_j^{\mathrm{SV}}(H)\right)\right)$ is also a finite subgraph with marked starting-vertices of $\hat{G}$. This is straightforward because we add in ∞_{vw} for possible missing edges in H from V_{j-1} to V_{j+1}. We also need to show that the Δ-identity holds for these face operations. Let $i \geq j$. For $i > j$, we get

$$d_i^{\mathrm{SV}}\left(d_j^{\mathrm{SV}}(H)\right) = d_j^{\mathrm{SV}}\left(d_{i+1}^{\mathrm{SV}}(H)\right) = H\left[V(H)\backslash V_j \backslash V_{i+1}\right] \cup \mathscr{E}^H\left(V_{j-1}, V_{j+1}\right) \cup \mathscr{E}^H\left(V_i, V_{i+2}\right).$$

For $i = j$, we have

$$d_j^{\mathrm{SV}}\left(d_j^{\mathrm{SV}}(H)\right) = d_j^{\mathrm{SV}}\left(d_{j+1}^{\mathrm{SV}}(H)\right) = H\left[V(\mathscr{H})\backslash V_j \backslash V_{j+1}\right] \cup \mathscr{E}^H\left(V_{j-1}, V_{j+2}\right).$$

This gives the following proposition.

Proposition 3.16. *In the set of finite subgraphs with marked starting-vertices of $\hat{G}$, the operations d_j^{SV} are well-defined and satisfy the Δ-identity for face operations.* □

Now let $\mathscr{H}$ be a collection of finite subgraphs with marked starting-vertices of G. Under the extension $G \preceq \hat{G}$, $\mathscr{H}$ is also a collection of finite subgraphs with marked starting-vertices of $\hat{G}$. Let X be the collection of finite subgraphs with marked starting-vertices of $\hat{G}$ given by all elements in $\mathscr{H}$ together with all of their iterated faces under face operations d_j^{SV}. Then X is Δ-set dominated by $\hat{G}$. We call $(\mathscr{P}(G) \cap X, X)$ the *super-hypergraph generated* by $\mathscr{H}$.

3.4.3. Revisiting path complexes.—We now come back to the work of Yau's school on path (co-)homology of graphs [76]. Following their terminology, a simple digraph G is a pair (V, E), where V is any set and $E \subseteq \{V \times V \setminus \mathrm{diag}\}$. We will show that d_j^{SV} can be used to describe the face operations given in [76]. To do this, we consider the "*largest quotient simple digraph*" $\tilde{G}$ on $\hat{G}$. Let $V(\tilde{G}) = V(\hat{G}) = V$, and for any ordered pair $(v, w) \in V \times V \setminus \mathrm{diag}$, identify all the directed edges from v to w to give one such directed edge. Since G is a simple graph, $\tilde{G}$ can be chosen to be a quotient of $\hat{G}$. Then G is a subgraph of the simple digraph $\tilde{G}$. Let $v, w \in V(G) = V(\tilde{G})$ be two distinct vertices. If there exists a directed edge e_{vw} from v to w, then ∞_{vw} is identified with e_{vw}. Otherwise ∞_{vw} is isolated. The graph $\tilde{G}$ can be considered as the completion of G in the sense that it is the smallest complete simple digraph containing G. Here a complete simple digraph means a simple digraph with the property that for any two distinct vertices v and w, there are exactly two directed edges with one from v to w and another from w to v.

[3] In the directed case, we only consider directed edges from a vertex in V_{j-1} to a vertex in V_{j+1} in H. If there are no directed edges in H from $v \in V_{j-1}$ to $w \in V_{j+1}$, we add ∞_{vw} to join them with direction from v to w.

Replacing $\hat{G}$ by $\tilde{G}$, our face operation d_j^{SV} on paths in $\tilde{G}$ coincides with the face operation in [76, Section 4] in the sense that it describes the j-th term in the boundary operators for chains and cochains on the path complex. Here, to match the definitions, a regular elementary path in [76, Section 4] is a path in $\tilde{G}$ and an allowed regular elementary path in [76, Section 4] is a path in the subgraph G. Then one can obtain the same objects by going via the definition of path homology in [76] or the definition of embedded homology of hypergraphs given above. In the undirected case, the operations d_j^{SV} on paths in $\tilde{G}$ coincide with the face operations in [76, Section 5].

For directed multi-graphs (quivers), the operations d_j^{SV} describe the face operations in [77]. Similarly to the undirected case, we need to do a certain identification on $\hat{G}$. Following the argument in [77, Section 3], for a complete quiver G, ∞_{vw} is identified with the 1-chain given by the sum of all directed edges from v to w in G. This would define a chain complex on the path complex of a complete quiver. For an arbitrary quiver G, one can embed G into its completion $\bar{G}$, and take the infimum chain complex (in Proposition 2.2) of the path complex of G in the chain complex of the path complex of $\bar{G}$ to define path homology for the quiver G, see [77] for details.

As notions of paths and walks are commonly used in data analytics, generalising them to higher dimensional combinatorial objects, such as path complexes, could provide new tools for various applications. From a data science point of view, a graph G is assumed as a working data. Then the path homology gives some topological information on G. Using a scoring scheme of G, one could get persistent path homology of G, this gives a persistence diagram/barcode of G as a topological feature. However, if the graphic data-set G is large, the computational complexity may be an issue. From such a perspective, it would be reasonable to consider a selected sub complex, or more generally a selected graded subset of the path complex. The general theory developed in this article gives a framework that makes it possible to explore topological features from subcomplexes or graded subsets of path complexes.

3.5. Descriptions of simplicial homology via Δ-sets.

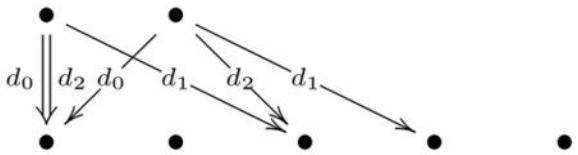
In this subsetion we only consider mod 2 homology, hence the coefficients are taken in $\mathbb{F} = \mathbb{Z}/2$.

3.5.1. Δ-neural network.—

We can interpret a Δ-set as a quiver or network. A similar object is a feed-forward neural network, as defined in [141, Section 3.6.1].

Feed-forward neural networks, are the simplest form of artificial neural networks. The feed forward neural network was the first and arguably simplest type of artificial network devised. In this network, the information moves in only one direction, forward, from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. In a feed-forward system, processing elements are arranged into distinct layers with each layer receiving input from the previous layer and outputting to the next layer.

Let $X = \{X_n\}_{n \geq 0}$ be a Δ-set. Each element in X is considered as a node (vertex), that is, the node set X is partitioned by layers labeled by X_0, X_1,…. For each $x \in X_n$, assign one and only one arrow (directed edge) $d^n_{i,x}: x \to d_i(x)$ for $0 \leq i \leq n$. This forms a set of arrows whose tails lie in X_n and whose heads lie in X_{n-1}. So it forms a quiver. The following picture illustrates the arrows from X_2 to X_1.



Rephrasing the definition of Δ-set into the terminology of network, a Δ-*neural network* is a quiver with distinct layers of nodes labeled by X_0, X_1,… such that for each node $x \in X_n$ with $n > 0$, there are arrows $d^n_{i,x}$, $0 \leq i \leq n$, tailed at x and headed at some node $d^n_i(x) \in X_{n-1}$ such that

$$d^{n-1}_i\left(d^n_j(x)\right) = d^{n-1}_j\left(d^n_{i+1}(x)\right) \tag{28}$$

for $0 \leq j \leq i \leq n-1$. In a Δ-neural network, the information only flows in one direction, from input nodes that could be located in different layers to the output nodes.

The adjacency relationship from the n^{th} layer X_n to the $n-1^{\text{th}}$ layer X_{n-1} can be described by $(n+1)$ matrices as follows. For $x \in X_n$ and $y \in X_{n-1}$, let

$$w^i_{x,y} = \begin{cases} 1 & \text{if } y = d_i(x) \\ 0 & \text{otherwise} \end{cases}$$

for $0 \leq i \leq n$. Let

$$W_n(i) = \left(w^i_{x,y}\right) x \in X_n, y \in X_{n-1}$$

be a $|X_n| \times |X_{n-1}|$ matrix, which is the matrix for the face operation $d_i: X_n \to X_{n-1}$. Equation (28) can be rewritten as the following formula

$$W_n(j) W^t_{n-1}(i) = W_n(i+1) W^t_{n-1}(j) \tag{29}$$

for $0 \leq j \leq i \leq n-1$, where A^t is the transpose of a matrix A.

Let $\mathscr{H}$ be a graded subset of the Δ-set X. In our theory of super-hypergraphs, $\mathscr{H}$ carries part of the Δ-set structure of X. More precisely, the face operation $d_i: X_n \to X_{n-1}$ induces a partially defined face operation $d_i: \mathscr{H}_n \rightsquigarrow \mathscr{H}_{n-1}$. By considering this as a network, $\mathscr{H}$ is full subnetwork of the Δ-neural network induced by X, where a *full subnetwork* is the induced network of the nodes of $\mathscr{H}$ in the neural network induced by X. Therefore, one can

get homology on any full subnetwork of a Δ-neural network using embedded homology of super-hypergraphs.

Remark 3.17. The product rule (28) is important to define the boundary operator on the chains. But variations are possible. For example, one could vary the product rule for weighted simplicial complexes [116], or the boundary operators on cochains could be varied to account for twisted de Rham cohomology [56].

3.5.2. Descriptions of mod 2 homology.—We proceed by giving the ideas behind the intuition of mod 2 homology of Δ-sets and super-hypergraphs. Let $(\mathscr{H}, X)$ be a super-hypergraph. Then the n-chains on X are linear combinations of the elements in X_n with coefficients in $\mathbb{Z}/2$. So each n-chain α corresponds to a subset $\{x_1, \ldots, x_k\} \subseteq X_n$ given by $\alpha = x_1 + x_2 + \cdots + x_k$. Since the coefficients are in $\mathbb{Z}/2$,

$$\partial(\alpha) = \sum_{i=1}^{k} \partial(x_i) = \sum_{i=1}^{k} \Sigma_{j=0}^{n} d_j(x_i)$$

which is the *trace* of the multi-subset $\{d_j(x_i) \mid 0 \le j \le n, 1 \le i \le k\}$ of X_{n-1}. Here the multiplicity of $y = d_j(x_i) \in X_{n-1}$ is the number of pairs (j', i') such that $d_{j'}(x_{i'}) = d_j(x_i) = y$, that is, the in-degree of the node $y \in X_{n-1}$ in the Δ-neural network. Hence we have the following proposition.

Proposition 3.18. *An n-chain $\alpha = x_1 + x_2 + \cdots + x_k$ is a mod 2 cycle (that is, $\partial(\alpha) = 0$) if and only if any node in the subset*

$$\{d_j(x_i) \mid 0 \le j \le n, 1 \le i \le k\}$$

of X_{n-1} has even in-degree. □

This proposition indicates that one can consider the nodes in X_{n-1} with even in-degrees in search for possible mod 2 cycles in n-chains.

The following proposition follows from the fact that $Z_n\left(\inf_*^{C_*(X)}(\mathscr{H})\right) = \mathbb{Z}/2(\mathscr{H}_n) \cap Z_n(C_*(X))$.

Proposition 3.19. *A mod 2 cycle $\alpha = x_1 + x_2 + \cdots + x_k$ in the chains $C_n(X)$, with all x_i distinct represents a cycle for the mod 2 embedded homology $H_n^{\mathrm{emb}, X}(\mathscr{H})$ if and only if $\{x_1, \ldots, x_k\} \subseteq \mathscr{H}_n$.* □

An n-chain $\alpha = x_1 + \cdots + x_k$ with all x_i distinct is a boundary in the chain complex $C_*(X)$ if and only if the equation

$$\alpha = \partial(\beta)$$

where $\beta = y_1 + \cdots + y_m$ with all x_i distinct in X_{n+1} has a solution. If there is a solution $\alpha = \partial(\beta)$, then $\{x_1, \ldots, x_k\}$ is the set of nodes in the multi-set $\{d_s(y_t) \mid 0 \le s \le n+1, 1 \le t \le m\}$ which have odd in-degrees. This proves the following statement.

Proposition 3.20. *An n-chain $\alpha = x_1 + \cdots + x_k$ with all x_i distinct is a boundary in the mod 2 chain complex $C_*(X)$ if and only if there exists a subset $\{y_1, \ldots, y_m\} \subseteq X_{n+1}$ with $y_1, \ldots, y_m$ distinct such that $\{x_1, \ldots, x_k\}$ is the set of nodes in the multi-set $\{d_s(y_t) \mid 0 \le s \le n+1, 1 \le t \le m\}$ which have odd in-degrees.* □

Note that $B_n\left(\inf_*^{C_*(X)}(\mathscr{H})\right) = \mathbb{Z}/2(\mathscr{H}_n) \cap \partial(\mathbb{Z}/2(\mathscr{H}_{n+1}))$.

Proposition 3.21. *Let $\alpha = x_1 + \cdots + x_k \in \mathbb{Z}/2(\mathscr{H}_n)$ with $x_1, \ldots, x_k$ distinct. Then α is a boundary in $\inf_*^{C_*(X)}(\mathscr{H})$ if and only if there exists a subset $\{y_1, \ldots, y_m\} \subseteq \mathscr{H}_{n+1}$ with $y_1, \ldots, y_m$ distinct such that $\{x_1, \ldots, x_k\}$ is the set of nodes in the multi-set $\{d_s(y_t) \mid 0 \le s \le n+1, 1 \le t \le m\}$ which have odd in-degrees.* □

4. Potential applications.

4.1. Potential applications in bio-molecular structures and drug design.

Applications of persistent homology to molecular biology has achieved great success in computer aided drug design [106, 131, 137]. According to [137, Paragraph 0005], theoretical models for the study of the structure-function relationships of biomolecules are conventionally based on purely geometric techniques. Mathematically, these approaches make use of local geometric information such as: coordinates, distances, angles, areas and curvatures for the physical modeling of biomolecular systems. However, conventional purely geometry based models tend to be overwhelmed by too much structural detail and are frequently computationally intractable. Topological approaches to determining the nature of structure-function relationships of biomolecules provide a dramatic simplification compared to conventional geometry based approaches [137, Paragraph 0053].

However, persistent homology neglects chemical and biological information during topological simplification and is thus not as competitive as geometry or physics-based alternatives in quantitative predictions [136]. Element-specific persistent homology, or multi-component persistent homology built on colored biomolecular networks, has been introduced to retain chemical and biological information in topological abstractions [30]. This approach encodes biological properties—such as hydrogen bonds, van der Waals interactions, hydrophilicity, and hydrophobicity—into topological invariants, rendering a potentially revolutionary representation for biomolecules, according to the SIAM news [136].

Recently, we have proposed hypergraph based persistent cohomology (HPC) for molecular representations in drug design [93]. In our HPC model, the protein-ligand interactions at the molecular level are represented as a series of element-specific hypergraphs. Figure 5 illustrates our hypergraph model for a protein-ligand complex with ID 3PB3. Its binding core region is divided into a series of element-specific atom-sets. From these

atom sets, element-specific hypergraphs can be constructed to characterize the interactions between protein atom-sets and ligand atom-sets at the level of atoms. Further, we have proposed a distance-related filtration process as illustrated in Figure 6. With the embedded homology model for hypergraphs, we have developed the hypergraph persistent homology and cohomology for molecular characterization. Molecular features and descriptors can be obtained from hypergraph persistent barcodes and hypergraph enriched barcodes, and this information can be further combined with machine learning models, in particular, the gradient boosting tree (GBT). Our HPC-GBT model has performed well for protein-ligand binding affinity predictions. Its Pearson correlation coefficients (PCCs) for the three PDBbind datasets, including PDBbind-v2007, PDBbind-v2013 and PDBbind-v2016, are consistently better than traditional machine learning models with molecular descriptors.

A molecular representation based on super-hypergraphs could give more flexibility in molecular structure and interaction characterization. Unlike simplicies and hyperedges, super-hyperedges can incorporate local topological structures, that is, subgraphs. This provides a unique way to identify and describe molecular motifs, function groups, and domains. Further, boundary operators can be defined through vertex-deletion and edge-deletion, which provide ways to define different types of homology groups and thus characterize different types of inner topological connections. Moreover, different filtration processes can be defined by considering different scoring functions, which in turn will induce different super-hypergraph based persistent homology/cohomology. Finally, molecular descriptors/fingerprints can be generated from super-hypergraph models and further combined with machine learning models for molecular data analysis in materials, chemistry and biology.

4.2. Potential applications in networks with group interactions.

The abstract of a recent review article [12], citing more than 800 references, reads,

> The complexity of many biological, social and technological systems stems from the richness of the interactions among their units. Over the past decades, a variety of complex systems has been successfully described as networks whose interacting pairs of nodes are connected by links. Yet, from human communications to chemical reactions and ecological systems, interactions can often occur in groups of three or more nodes and cannot be described simply in terms of dyads…We review the measures designed to characterize the structure of these systems and the models proposed to generate synthetic structures, such as random and growing bipartite graphs, hypergraphs and simplicial complexes. We introduce the rapidly growing research on higher-order dynamical systems and dynamical topology, discussing the relations between higher-order interactions and collective behavior…

Here we can see that simplicial complexes, a fundamental notion in algebraic topology, has been extensively used for providing representations of higher-order interactions [12, First paragraph of Section 2.1.3]. In some practical problems, the limitations of simplicial complexes due to completeness and vertax-determination present a problem. Hypergraphs provide a more general and unconstrained description of higher-order interactions [12, Paragraphs 2–4, page 7].

Recent progress shows that simplicial homology can be naturally extended as a homology theory on hypergraphs [23]. This provides topological invariants for geometric models using hypergraphs which have had successful applications in biomolecular structures and drug design, described in the previous subsection.

As an extension of hypergraphs, super-hypergraphs would provide more a general and unconstrained description of higher-order interations. If we assume that the higher-order interactions take place among the nodes in a working graph, which indicates the pre-existence of the pairwise bonds or the primary pairwise links between the nodes, then the most general and unconstrained description of higher-order interations would be a collection of finite subgraphs of the working graph, which is exactly the topic explored in this article.

5. Conclusion.

In this paper, we introduced a new mathematical theory which allows for topological invariants to be applied to broader range of problems, in particular enriching the methods of TDA. This new theory is suitable for both graph data and point cloud data analysis, while overcoming various limitations of the standard persistent homology theory such as the topological noise and the constraining requirements to use data with metric. Using this new theory, the upgraded pipeline of TDA becomes indetermiinistic in nature allowing for flexibility and adjustments. Moreover, various new topological invariants can be constructed in our flexible setting. As highlighted in Subsections 3.2–3.5, based on this topological approach, more computational tools of algebraic topology will find applications in data science. For example, in algebraic topology the computation of simplicial homology of a space can be largely simplified by homotopically deforming it into a simpler shape, see [80].

As each simplex of a simplicial complex is uniquely determined by its vertieces, simplicial complexes cannot model collections of subgraphs. To explore topological structures on space of subgraphs, in this paper we use Δ-sets. Furthermore, we introduce the notion of sup-hypergraph, as a generalization of hypergraphs, which sets a stage for the exploration of topological structures on subgraphs. The homology theory of super-hypergraph, established in Section 2, endows any collection of subgraphs with topological features.

In this work we also use the notion of scoring scheme. As highlighted in Section 3, scoring schemes are used to introduce persistence in an abstract setting without the use of any notion of metric. The classical constructions in persistent simplicial homology theory can be recovered using various scoring schemes.

We should point out that this work presents a theoretic research resulting in a framework that provides an upgraded topological approach to data science, with the aim to foster further interactions between topology and data science.

Further research on super-hypergraphs is needed as this is a new and challenging mathematical concept. From a topological perspective there are many interesting questions to consider such as, the algebraic structure of the homology of these objects as well as the homotopy aspects of super-hypergraphs that are far less-understood. Furthermore, developments in the topological study of super-hypergraphs will feed into new innovative

methods in TDA and other wide-ranging applications. Additionally, the computational complexity of the homology theory of super-hypergraphs is comparable to that of simplicial homology, therefore algorithms methods stemming from our approach will be similarly feasible as classical computation.

Acknowledgments

The work of JW was supported in part by Natural Science Foundation of China (NSFC grant no. 11971144) and High-level Scientific Research Foundation of Hebei Province. The third author was supported by Nanyang Technological University Startup Grant M4081842 and Singapore Ministry of Education Academic Research fund Tier 1 RG109/19, Tier 2 MOE-T2EP20220-0010, and Tier 2 MOE-T2EP20120-0013. The work of GWW was supported by NIH grant GM126189, NSF grants DMS-1761320, IIS-1900473, and DMS-2052983, and NASA grant 80NSSC21M0023. We wish to thank the referees most warmly for important suggestions that have improved the exposition of this paper.

Appendix A.: Topological structures associated to graphs.

Throughout mathematics numerous topological and categorical structures on graphs have been explored. In this appendix, we will survey various simplicial complexes associated to graphs that allow us to consider the space of subgraphs of a given graph from a topological perspective.

A *directed (multi-)graph* (or *multi-digraph* or *quiver*) is a pair $G = (V(G), E(G))$ together with a function end: $E(G) \longrightarrow V(G) \times V(G)$ given by

$$\mathrm{end}_G(e) \mapsto (i(e), t(e))$$

where $V(G)$ is the *vertex* set, $E(G)$ is the *edge* set, $i(e)$ is the *initial* vertex of the edge e, and $t(e)$ is the *terminal* vertex of e.

An *undirected (multi-)graph*[4] is a pair $G = (V(G), E(G))$ together with a function $\mathrm{end}_G : E(G) \longrightarrow (V(G) \times V(G))/\Sigma_2$ given by

$$\mathrm{end}_G(e) \mapsto \{i(e), t(e)\}$$

where $(V(G) \times V(G))/\Sigma_2$ is the orbit set of $(V(G) \times V(G))$ modulo the Σ_2-action given by permuting the coordinates, $V(G)$ is the *vertex* set, $E(G)$ is the *edge* set, end_G is an *incidence relation* that associates with each edge of G an unordered pair of, possibly equal, elements of $V(G)$. In this definition of a directed/undirected (multi-)graph, the empty graph is allowed.

A *subgraph* H of a directed/undirected (multi-)graph G is a graph $H = (V(H), E(H))$ with $V(H) \subseteq V(G)$, $E(H) \subseteq E(G)$ and $\mathrm{end}_H = \mathrm{end}_G|_{E(H)}$.

[4]We follow the definition of a multi-graph in [51, 10]. In some literature such as [113], a multi-graph is defined by requiring the edge set to be a multi-set. The difference is that the edges between two vertices are labeled by $E(G)$ together with the incidence map end_G. Such a definition coincides with the definition on quiver (as directed multi-graph) [121].

A directed/undirected graph G is *simple* if end_G is injective and the image $\mathrm{end}_G(E(G))$ is disjoint from the diagonal $\Delta(V(G))$ in $V(G) \times V(G)$ or $(V(G) \times V(G))/\Sigma_2$. This means that there are no loops or multi-edges between two vertices.

From the perspective of applications, the initial data is represented by a given graph G and let $\mathscr{K}$ be a collection of subgraphs of G. Our goal is to investigate the possible topological structures on $\mathscr{K}$. However, before we address this general question, we review some classical constructions of simplicial complexes associated to graphs.

A.1. Clique complexes.

Typically, the study of collections of subgraphs has focused on measuring how strongly connected different parts of a graph are. A clique (or flag) complex and an independence complex (the clique complex of the complementary graph) are topological spaces that contain information about the connectivity of a graph. These are widely used objects in mathematics and its applications, see [4, 11, 58, 60, 87] for some recent works.

A *complete graph* is a simple graph $G = (V(G), E(G))$ with the property that every pair of distinct vertices of G are adjacent in G.

A *clique* of a graph G is a complete subgraph of G.

The *clique complex* of a simple graph G is the abstract simplicial complex $\mathrm{Clique}(G)$ whose simplices consist of all cliques of G. An n-simplex σ in $\mathrm{Clique}(G)$ is a clique of G with $(n + 1)$ vertices, and a face of a simplex $\sigma \in \mathrm{Clique}(G)$ is a complete subgraph obtained by deleting some vertices of σ.

When working with a multi-graph $G = (V, E)$, the set of cliques $\mathrm{Clique}(G)$ is generally not a simplicial complex as this requires that all simplices are uniquely determined by their vertex set. For example, let G be a multi-graph with two vertices v and w and two edges e_1 and e_2 joining them. Then

$$\mathrm{Clique}(G) = \{\bar{e}_1, \bar{e}_2, v, w\}$$

has two 1-simplices $\bar{e}_1$ and $\bar{e}_2$ sharing the same vertices v and w, see Figure 7. Therefore, a more suitable object for describing the topological structure of $\mathrm{Clique}(G)$ is a Δ-set.

For an undirected multi-graph G, the Δ-set structure on $\mathrm{Clique}(G)$ is given in the following way. Assign a total ordering to $V(G)$ and define $\mathrm{Clique}_n(G)$ to be the set of cliques of G that have exactly $n+1$ vertices. For $\sigma \in \mathrm{Clique}_n(G)$ with vertices $v_0 < v_1 < \cdots < v_n$, define $d_i\sigma = \sigma - v_i$, the subclique of σ obtained by deleting the vertex v_i and the edges incident to v_i for $0 \leq i \leq n$. It is straightforward to check that $\mathrm{Clique}_*(G)$ forms a Δ-set[5].

[5]The definition of the Δ-set $\mathrm{Clique}_*(G)$ depends on the given order on vertices of G, but the homology of $\mathrm{Clique}_*(G)$ is independent on this choice because the geometric realization of a Δ-set is a Δ-complex [139, Proposition 1.39, p. 51] in the sense of Hatcher [80].

A.2. Neighborhood complexes and Jonsson's graph complexes.

We proceed by considering a collection of simplicial complexes associated to graphs which will naturally lead to new constructions suitable for studying spaces of subgraphs. We start with a famous construction of the neighborhood complex of a graph. This was introduced by Lovász [94] in 1978 in his work on Kneser's conjecture which laid the foundations of topological combinatorial by introducing homotopy theoretical methods to combinatorics. Nowadays, the research area of topological combinatorics is very active and fruitful. The generalization by Lovász of the neighborhood complexes to the Hom complex [7, 89], which has the same homotopy type as the clique complex of an exponential graph [54, Remark 3.6], was used in a breakthrough work of Babson and Kozlov [8] to solve the Lovász conjecture which relates the chromatic number of a graph with the homology of its Hom complex. Our theory is based on the exploration of the interplay between topology and combinatorics.

The *neighborhood complex* $\mathcal{N}(G)$ of a graph G is a simplicial complex on vertex set $V(G)$ in which an n-simplex is a subset of $V(G)$ with $n+1$ vertices such that all vertices are adjacent to an other vertex in G.

As we discussed in the previous subsection, Clique(G) may not be a simplicial complex for a multi-graph G. However, for any graph G the neighborhood complex $\mathcal{N}(G)$ is a simplicial complex.

The topology on the geometric realization of $\mathcal{N}(G)$ can be quite different from that of Clique(G) in general. For example, let G be a graph with three vertices a, b, c and two edges given by ab and bc. Then $\mathcal{N}(G) = \{\{a, c\}, \{a\}, \{b\}, \{c\}\}$, which is not connected, see Figure 8b, and Clique(G) = {{a, b}, {b, c}, {a}, {b}, {c}} which is connected, see Figure 8a. This indicates that there are various topological structures one could construct for a given working graph G.

In [84, p.26], Jonsson defines a graph complex in the following way. A *graph complex*[6] on a finite vertex set V is a family $\mathscr{C}$ of simple graphs on the vertex set V such that $\mathscr{C}$ is closed under deletion of edges; if $H \in \mathscr{C}$ and $e \in H$, then $H - e \in \mathscr{C}$. Identifying $H = (V, E) \in \mathscr{C}$ with the edge set E, we may interpret $\mathscr{C}$ as a simplicial complex. There are potentially different graph complexes on a given vertex set V because the collection of simple graphs can be chosen in a different way.

With a slight modification to Jonsson's definition, namely adding a hypothesis that the simple graphs in $\mathscr{C}$ are subgraphs of G, we retain the central ideas of Jonsson's construction but also gain control over the space of subgraphs. In contrast to clique complexes and neighborhood complexes, the face operations in Jonsson's graph complex are given by deleting edges. Also, the construction of a graph complex is not fully determined by G as there are various choices for families $\mathscr{C}$ of simple subgraphs of G that can form graph complexes. A non-deterministic characteristic of these complexes might be useful in data science as the family $\mathscr{C}$ can be adjusted for each iteration of the analysis.

[6]Kontsevich also introduced graph complexes with a different defintion [86].

A.3. Path complexes.

Considering hypergraphs as a combinatorial generalization of simplicial complexes allows the construction of the path complex of a given digraph.

The topological exploration of path complexes was first introduced by Shing-Tung Yau and his collaborators in a series of papers [73, 74, 75, 76, 68, 72, 69, 70, 71, 77]. Motivated by ideas from physical applications, A. Dimakis and F. Müller-Hoissen attempted to construct the cohomology of digraphs [53, 52]. They considered path complexes on an intuitive level without a precise definition of the corresponding cochain complex.

In this subsection, we survey the main ideas of path complexes of simple digraphs.

Let G be a simple digraph. A *directed path* in G is an alternating sequence $\lambda = v_0a_1v_1a_2v_2\cdots a_kv_k$, with all vertices v_i distinct for $0 \le i \le k$ and the edges, a_i, are incident out of v_{i-1} and incident into v_i for $1 \le i \le k$.

Let $\mathscr{P}$ be the set of directed paths in G. We want to associate a combinatorial object to G built out of directed paths. Since G is simple, there is at most one edge joining two distinct vertices. So a directed path $\lambda = v_0a_1v_1a_2v_2\cdots a_kv_k$ is determined by its vertices $v_0, v_1, \ldots, v_k$. Thus we consider $\lambda = v_0a_1v_1a_2v_2\cdots a_kv_k$ as an abstract k-simplex $\{v_0, v_1, \ldots, v_k\}$. For $\mathscr{P}$ to be a simplicial complex, any nonempty subset of $\{v_0, v_1, \ldots, v_k\}$ must be a simplex. In other words, any subsequence $\left(v_{i_0}, v_{i_1}, v_{i_2}, \ldots, v_{i_t}\right)$, $0 \le i_0 < i_2 < \cdots < i_t \le k$ of λ must forms a directed path in G. This is not true in general. For example, if $v_0a_1v_1a_2v_2$ is a directed path in G, then there may not exist an edge incident out of v_0 and incident into v_2 in G, that is, (v_0, v_2) may not form a directed path. Therefore, a structure to consider on the set $\mathscr{P}$ is that of a hypergraph.

The set $\mathscr{P}$ becomes a hypergraph with its vertex $V(G)$ and the hyperedge set given by directed paths in G. By definition, an abstract simplicial complex is a hypergraph with the additional condition that any nonempty subset of a hyperedge is a hyperedge. Therefore, a hypergraph can be viewed as a simplicial complex with some faces missing, where a hyperedge of cardinal $k+1$ is a k-simplex in the terminology of simplicial complexes. The approaches in [53, 52] and Yau's school lead to the embedded homology of hypergraphs as an extension of simplicial homology theory as introduced in [23].

By allowing vertex repetition in directed paths, we get directed walks. The walk complex $\mathscr{W}(G)$ for a digraph or quiver (i.e. directed multi-graph) G, is similar to the path complex but with we replace directed paths with directed walks. Therefore, $\mathscr{W}(G)$ is an extension of the notion of the nerve of a category in the following sense. Consider a category $\mathscr{C}$ as a quiver with the composition operation on head-to-tail arrows. Then the nerve of category $\mathscr{C}$ is the walk complex of quiver $\mathscr{C}$.

A.4. Vertex-deletion topology.

Let G be a directed/undirected (multi-)graph. Let $\mathscr{H}$ be a collection of finite subgraphs of G. Assign to $\mathscr{H}$ the grading function $f_v: \mathscr{H} \to \mathbb{N} = \{0, 1, 2, \ldots\}$ given by the size, namely, for

$H \in \mathscr{H}$, let $f_V(H) = |V(H)| - 1$. Let $\mathscr{H}_n = f_V^{-1}(n)$. Next step is to define face operations to obtain a topological structure. There are several natural approaches available.

A.4.1. Primary vertex-deletion topology.

Assume that the vertex set $V(G)$ is totally ordered. A geometric way to define face operations is to delete a vertex together with all edges incident to this vertex. More precisely, let $H \in \mathscr{H}_n$ with vertices $v_0, v_1, \ldots, v_n$. Define $d_i(H)$ for $0 \le i \le n$ to be the subgraph of H by deleting v_i together with any edges joining with v_i. This vertex deletion does not ensure that $d_i(H)$ lies in $\mathscr{H}_{n-1}$. Let

$$\Delta(\mathscr{H}) = \left\{ d_{i_1} d_{i_2} \cdots d_{i_t}(H) \middle| H \in \mathscr{H}, \quad 0 \le i_1 < i_2 < \cdots < i_t \le |V(H)| - 1 \right\} \tag{30}$$

be the family of subgraphs of G obtained from $\mathscr{H}$ together with iterated faces on the subgraphs in $\mathscr{H}$. It is straightforward to check that $\Delta(\mathscr{H})$ is a Δ-set, and $\mathscr{H} \subseteq \Delta(\mathscr{H})$ is a graded subset[7]. Hence $(\mathscr{H}, \Delta(\mathscr{H}))$ is a super-hypergraph.

Definition A.1. Let G be a directed/undirected (multi-)graph. Let $\mathscr{H}$ be a collection of finite subgraphs of G. The *primary vertex-deletion topological structure* on $\mathscr{H}$ is the super-hypergraph structure defined as above.

Similarly to clique complexes on multi-graphs, $\Delta(\mathscr{H})$ may not be a simplicial complex in general. Therefore, the notion of a super-hypergraph is the most natural and suitable topological description for $\mathscr{H}$.

The super-hypergraph $(\mathscr{H}, \Delta(\mathscr{H}))$ has a structure of fibrewise topology as follows.

Let

$$V(\mathscr{H}) = \{V(H) \mid H \in \mathscr{H}\} \text{ and } V(\Delta(\mathscr{H})) = \{V(H) \mid H \in \Delta(\mathscr{H})\}$$

be a family of finite subsets of $V(G)$. Then $V(\Delta(\mathscr{H}))$ is a simplicial complex, and $V(\mathscr{H})$ is a hypergraph whose simplicial closure is $V(\Delta(\mathscr{H}))$. Moreover we have a Δ-map

$$V : \Delta(\mathscr{H}) \to V(\Delta(\mathscr{H}))$$

and a morphism of super-hypergraphs

$$V : \mathscr{H} \to V(\mathscr{H}).$$

By taking geometric realization, we have a continuous map

$$|V| : |\Delta(\mathscr{H})| \to |V(\Delta(\mathscr{H}))|$$

[7]From the Δ-identity (5), $\Delta(H)$ contains all iterated faces on the subgraphs in $\mathscr{H}$, which is the smallest family of subgraphs of G containing $\mathscr{H}$ that is closed under the face operation.

which is a fibrewise topology in the sense of James [83].

Clique complexes are typical examples of primary vertex-deletion topology, where $\mathscr{K}$ is given by cliques in a grpah G. In this case, $\mathscr{K}$ itself is already a Δ-set so $\mathscr{K} = \Delta(\mathscr{K})$ and the map $V : \mathscr{K} \to V(\mathscr{K})$ is an isomorphism.

The neighborhood complex is another good example that admits a fibrewise topological structure as follows. Let

$$\widetilde{\mathcal{N}(G)} = \{H \mid H \text{ is a subgraph of } G \text{ and } V(H) \in \mathcal{N}(G)\}. \tag{31}$$

Then it is straightforward to check that

$$\widetilde{\mathcal{N}(G)} = \Delta(\widetilde{\mathcal{N}(G)}) \text{ and } V(\widetilde{\mathcal{N}(G)}) = \mathcal{N}(G) \tag{32}$$

with a continuous map

$$\left|V\right| : \left|\widetilde{\mathcal{N}(G)}\right| \to \left|\mathcal{N}(G)\right| \tag{33}$$

which is called a *fibrewise neighborhood topology* of G.

A.4.2. Secondary vertex-deletion topology.

Consider the path complex of a simple digraph G and its face operation d_i. Let $\lambda = v_0 a_1 v_1 a_2 v_2 \cdots a_n v_n$ be a directed path. Then $d_i(\lambda)$ is given by deleting the vertex v_i. However, we have to add back the directed edge from v_{i-1} to v_{i+1} provided that it exists to ensure that $d_i(\lambda) \in \mathscr{P}_n$. This gives a different type of topological structure, in which we need to redefine the edges to match the vertex removal of the face operation. This can be generalized in the following way.

Let G be a directed/undirected simple graph and let $\mathscr{K}$ be a family of finite subgraphs of G. Let the vertex set $V(G)$ be totally ordered. For $H \in \mathscr{K}$, as a finite subgraph of G with vertices $v_0 < v_1 < \cdots < v_n$, define $d_i H$ to be the subgraph of G by removing v_i from H and adding the edge between v_{i-1} and v_{i+1} if it exists. Then $\mathscr{K}$ forms a super-hypergraph in a similar way as in the case of primary vertex-deletion topology. Here the notion of a super-hypergraph is necessary because there could be two subgraphs in $\mathscr{K}$ sharing same vertices. For instance, if there is an edge joining two distinct vertices v and w in G, then the subgraphs consist of two vertices v and w with the edge joining them and without the edge joining them, respectively, are different.

Definition A.2. Let G be a directed/undirected simple graph. Let $\mathscr{K}$ be a collection of finite subgraphs of G. The *secondary vertex-deletion topological structure* on $\mathscr{K}$ is the super-hypergraph structure defined as above.

The secondary vertex-deletion topology naturally applies to subgraphs of a simple graph. However, to construct a topological structure on a space of subgraphs of a multi-graph in this a way would be more complicated.

There are other possible topological structures on special families of subgraphs. Analogously to various techniques developed in simplicial homotopy theory, for special families of subgraphs having good patterns, one could delete more than one vertex under each elementary face operation d_i.

A.5. Edge-deletion topology.

Let G be a directed/undirected (multi-)graph and $\mathscr{H}$ be a collection of finite subgraphs of G. Another reasonable way to assign the grading function $f_e: \mathscr{H} \to \mathbb{N} = \{0, 1, 2, \ldots\}$ is by counting edges, that is, for $H \in \mathscr{H}$ let $f_e(H) = |E(H)| - 1$. Then $\mathscr{H}_n = f_e^{-1}(n)$. Note that a subgraph H of G is uniquely determined by its edge set $E(H)$. We do not need to use the notion of a Δ-set for describing topological structure on $\mathscr{H}$ from edge-deletion. If $\mathscr{H}$ is closed under edge-deletion operation, then it forms a simplicial complex, which is exactly a path complex in the sense of Jonsson. Otherwise, $\mathscr{H}$ is only a hypergraph.

For a fixed graph G, the edge-deletion topology could be quite different from the vertex-deletion topology because already the grading functions f_v and f_e could be quite different. The edge-deletion operation may not commute with the vertex-deletion operation, so the relationship between the edge-deletion topology and the vertex-deletion topology is not immediately clear. To better understand these structures, more exploration of the relationship between different topological structures on families of subgraphs is needed.

Finally, we should point out that there are many other ways to introduce topological structures on subgraphs, for example following ideas related to Hom complexes. The frontier of research in topological combinatorics has potential to provide new mathematical tools in data science.

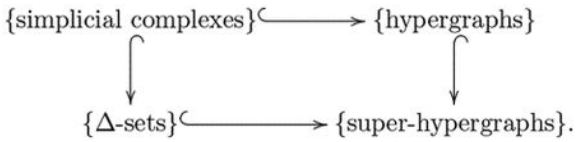
Appendix B.: The connections of the concepts in the article.

There are various concepts in the article, including new concepts such as super-hypergraph. We highlight the connections between them in this appendix.

The following statements are well known and important:

- A simple graph is a 1-dimensional simplicial complex.
- A multi-graph with its vertices totally ordered is a 1-dimensional Δ-set/Δ-complex.
- A quiver (multi-digraph) is a 1-dimensional Δ-set/Δ-complex.

In the following table, the arrow $\hookrightarrow$ means an inclusion of sets.



REFERENCES

[1]. Dionysus: The persistent homology software, Software available at http://www.mrzv.org/software/dionysus.

[2]. Adams H, Emerson T, Kirby M, Neville R, Peterson C, Shipman P, Chepushtanova S, Hanson E, Motta F and Ziegelmeier L, Persistence images: A stable vector representation of persistent homology, J. Mach. Learn. Res, 18 (2017), Paper No. 8, 35 pp.

[3]. Adcock A, Carlsson E and Carlsson G, The ring of algebraic functions on persistence bar codes, Homology Homotopy Appl., 18 (2016), 381–402.

[4]. Aharoni R, Berger E and Ziv R, Independent systems of representatives in weighted graphs, Combinatorica, 27 (2007), 253–267.

[5]. Ahmed M, Fasy BT and Wenk C, Local persistent homology based distance between maps, In Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, (2014), 43–52.

[6]. Anirudh R, Thiagarajan JJ, Kim I and Polonik W, Autism spectrum disorder classification using graph kernels on multidimensional time series, preprint, arXiv:1611.09897.

[7]. Babson E and Kozlov DN, Complexes of graph homomorphisms, Israel J. Math, 152 (2006), 285–312.

[8]. Babson E and Kozlov DN, Proof of the Lovász conjecture, Ann. of Math, 165 (2007), 965–1007.

[9]. Bae W, Yoo JJ and Ye JC, Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification, In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), (2017), 1141–1149.

[10]. Balakrishnan R and Ranganathan K, A Textbook of Graph Theory, 2nd edition, Universitext, Springer, New York, 2012.

[11]. Barmak JA, Star clusters in independence complexes of graphs, Adv. Math, 241 (2013), 33–57.

[12]. Battiston F, Cencetti G, Iacopini I, Latora V, Lucas M, Patania A, Young J-G and Petri G, Networks beyond pairwise interactions: Structure and dynamics, Phys. Rep, 874 (2020), 1–92.

[13]. Bauer U, Ripser: A lean C++ code for the computation of Vietoris-Rips persistence barcodes, Software available at https://github.com/Ripser/ripser.

[14]. Bauer U, Kerber M and Reininghaus J, Distributed computation of persistent homology, In Meeting on Algorithm Engineering and Experiments (ALENEX), SIAM, (2014), 31–38.

[15]. Bauer U, Kerber M, Reininghaus J and Wagner H, PHAT–persistent homology algorithms toolbox, In Mathematical software¡ªICMS, 8592 (2014), 137–143.

[16]. Bendich P, Cohen-Steiner D, Edelsbrunner H, Harer J and Morozov D, Inferring local homology from sampled stratified spaces, In IEEE Symposium on Foundations of Computer Science (FOCS'07), (2007), 536–546.

[17]. Bendich P, Edelsbrunner H and Kerber M, Computing robustness and persistence for images, IEEE Transactions on Visualization and Computer Graphics, 16 (2010), 1251–1260. [PubMed: 20975165]

[18]. Bendich P, Gasparovic E, Harer J, Izmailov R and Ness L, Multi-scale local shape analysis and feature selection in machine learning applications, In International Joint Conference on Neural Networks (IJCNN), IEEE, (2015), 1–8.

[19]. Bendich P, Wang B and Mukherjee S, Local homology transfer and stratification learning, In Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms, (2012), 1355–1370.

[20]. Bergomi MG, Ferri M, Vertechi P and Zuffi L, Beyond topological persistence: Starting from networks, Mathematics, 9 (2021).

[21]. Binchi J, Merelli E, Rucco M, Petri G and Vaccarino F, jholes: A tool for understanding biological complex networks via clique weight rank persistent homology, Electron. Notes Theor. Comput. Sci, 306 (2014), 5–18.

[22]. Bonis T, Ovsjanikov M, Oudot S and Chazal F, Persistence-based pooling for shape pose recognition, In Computational Topology in Image Context, 9667 (2016), 19–29.

[23]. Bressan S, Li J, Ren S and Wu J, The embedded homology of hypergraphs and applications, Asian J. Math, 23 (2019), 479–500.

[24]. Bubenik P, Statistical topological data analysis using persistence landscapes, J. Mach. Learn. Res, 16 (2015), 77–102.

[25]. Bubenik P and Kim PT, A statistical approach to persistent homology, Homology Homotopy Appl, 9 (2007), 337–362.

[26]. Bubenik P and Vergili T, Topological spaces of persistence modules and their properties, J. Appl. Comput. Topol, 2 (2018), 233–269.

[27]. Cameron PJ, Automorphisms and cohomology of switching classes, J. Combinatorial Theory Ser. B, 22 (1977), 297–298.

[28]. Cameron PJ, Cohomological aspects of two-graphs, Math. Z, 157 (1977), 101–119.

[29]. Cang Z, Mu L and Wei G-W, Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening, PLOS Computational Biology, 14 (2018), 1–44.

[30]. Cang Z and Wei G, Integration of element specific persistent homology and machine learning for protein-ligand binding affinity prediction, International Journal for Numerical Methods in Biomedical Engineering, 34 (2018), e2914.

[31]. Cang Z and Wei G-W, Analysis and prediction of protein folding energy changes upon mutation by element specific persistent homology, Bioinformatics, 33 (2017), 3549–3557. [PubMed: 29036440]

[32]. Cang Z and Wei G-W, Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions, PLOS Computational Biology, 13 (2017), 1–27.

[33]. Carlsson G, Ishkhanov T, Silva V and Zomorodian A, On the local behavior of spaces of natural images, Int. J. Comput. Vis, 76 (2008), 1–12.

[34]. Carlsson G, Singh G and Zomorodian A, Computing multidimensional persistence, In Algorithms and Computation, 5878 (2009), 730–739.

[35]. Carlsson G and Zomorodian A, The theory of multidimensional persistence, Discrete Comput. Geom, 42 (2009), 71–93.

[36]. Carlsson G, Topology and data, Bull. Amer. Math. Soc. (N.S.), 46 (2009), 255–308.

[37]. Cerri A and Landi C, The persistence space in multidimensional persistent homology, In Discrete Geometry for Computer Imagery, 7749 (2013), 180–191.

[38]. Chazal F, Cohen-Steiner D, Glisse M, Guibas LJ and Oudot SY, Proximity of persistence modules and their diagrams, In SCG ‘09: Proceedings of the twenty-fifth annual symposium on Computational Geometry, (2009), 237–246.

[39]. Chazal F, de Silva V, Glisse M and Oudot S, The Structure and Stability of Persistence Modules, SpringerBriefs in Mathematics, Springer, [Cham], 2016.

[40]. Chazal F, Fasy B, Lecci F, Michel B, Rinaldo A and Wasserman L, Subsampling methods for persistent homology, In Proceedings of the 32nd International Conference on Machine Learning, (eds. Bach F and Blei D), PMLR, Lille, France, 37 (2015), 2143–2151.

[41]. Chazal F and Michel B, An introduction to topological data analysis: Fundamental and practical aspects for data scientists, Front. Artif. Intell, 2021.

[42]. Cheng Y and Wells AL Jr., Switching classes of directed graphs, J. Combin. Theory Ser. B, 40 (1986), 169–186.

[43]. Chevyrev I, Nanda V and Oberhauser H, Persistence paths and signature features in topological data analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 42 (2018), 192–202. [PubMed: 30530312]

[44]. Chung FRK and Graham RL, Cohomological aspects of hypergraphs, Trans. Amer. Math. Soc, 334 (1992), 365–388,

[45]. Cohen-Steiner D, Edelsbrunner H and Morozov D, Vines and vineyards by updating persistence in linear time, In Computational geometry (SCG’06), (2006), 119–126.

[46]. Crawley-Boevey W, Decomposition of pointwise finite-dimensional persistence modules, J. Algebra Appl, 14 (2015), 1550066, 8 pp.

[47]. Curtis EB, Simplicial homotopy theory, Advances in Math, 6 (1971), 107–209.

[48]. de Silva V and Ghrist R, Homological sensor networks, Notices Amer. Math. Soc, 54 (2007), 10–17.

[49]. De Silva V, Morozov D and Vejdemo-Johansson M, Persistent cohomology and circular coordinates, Discrete Comput. Geom, 45 (2011), 737–759.

[50]. Dey TK and Mandal S, Protein classification with improved topological data analysis, In LIPIcs. Leibniz Int. Proc. Inform, 113 (2018), 13 pp.

[51]. Diestel R, Graph Theory, vol. 173 of Graduate Texts in Mathematics, 5th edition, Graduate Texts in Mathematics, 173. Springer, Berlin, 2017.

[52]. Dimakis A and Mu F¨ller-Hoissen, Differential calculus and gauge theory on finite sets, J. Phys. A, 27 (1994), 3159–3178.

[53]. Dimakis A and Mu F¨ller-Hoissen, Discrete differential calculus: Graphs, topologies, and gauge theory, J. Math. Phys, 35 (1994), 6703–6735.

[54]. Dochtermann A, Hom complexes and homotopy theory in the category of graphs, European J. Combin, 30 (2009), 490–509.

[55]. Duval AM and Reiner V, Shifted simplicial complexes are Laplacian integral, Trans. Amer. Math. Soc, 354 (2002), 4313–4344.

[56]. Dwork B, On the zeta function of a hypersurface, Inst. Hautes Etudes Sci. Publ. Math, (1962), 5–68.

[57]. Edelsbrunner H, Letscher D and Zomorodian A, Topological persistence and simplification, Discrete Comput. Geom, 28 (2002), 511–533.

[58]. Ehrenborg R and Hetyei G, The topology of the independence complex, European J. Combin, 27 (2006), 906–923.

[59]. Emtander E, Betti numbers of hypergraphs, Comm. Algebra, 37 (2009), 1545–1571.

[60]. Engström A, Independence complexes of claw-free graphs, European J. Combin, 29 (2008), 234–241.

[61]. Fasy BT, Kim J, Lecci F and Maria C, Introduction to the r package tda, preprint, arXiv:1411.1830.

[62]. Fasy BT and Wang B, Exploring persistent local homology in topological data analysis, In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2016), 6430–6434.

[63]. Forman R, Morse theory for cell complexes, Adv. Math, 134 (1998), 90–145.

[64]. Frosini P and Landi C, Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval, Computer Analysis of Images and Patterns, 6854 (2011), 294–301.

[65]. Gabriel P, Unzerlegbare Darstellungen. I, Manuscripta Math., 6 (1972), 71–103.

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

[67]. Giansiracusa N, Giansiracusa R and Moon C, Persistent homology machine learning for fingerprint classification, IEEE International Conference On Machine Learning And Applications (ICMLA), 2019.

[68]. Grigorian A, Lin Y, Muranov Y and Yau S-T, Homologies of path complexes and digraphs, ArXiv.

[69]. Grigorian A, Muranov Y and Yau S-T, Homologies of digraphs and künneth formulas, Comm. Anal. Geom, 25 (2017), 969–1018.

[70]. Grigor'yan A, Muranov YV and Yau S-T, Graphs associated with simplicial complexes, Homology Homotopy Appl, 16 (2014), 295–311.

[71]. Grigor'yan A and Muranov YV, Cohomology theories of simplicial complexes , algebras , and digraphs.

[72]. Grigor'yan AA, Lin I, Muranov YV and Yau S, Path complexes and their homologies, Fundam. Prikl. Mat, 21 (2016), 79–128.

[73]. Grigor'yan A, Jimenez R, Muranov Y and Yau S-T, On the path homology theory of digraphs and Eilenberg-Steenrod axioms, Homology Homotopy Appl., 20 (2018), 179–205.

[74]. Grigor'yan A, Jimenez R, Muranov Y and Yau S-T, Homology of path complexes and hypergraphs, Topology Appl., 267 (2019), 106877, 25 pp.

[75]. Grigor'yan A, Lin Y, Muranov Y and Yau S-T, Homotopy theory for digraphs, Pure Appl. Math. Q, 10 (2014), 619–674.

[76]. Grigor'yan A, Lin Y, Muranov Y and Yau S-T, Cohomology of digraphs and (undirected) graphs, Asian J. Math, 19 (2015), 887–931.

[77]. Grigor'yan A, Muranov Y, Vershinin V and Yau S-T, Path homology theory of multi-graphs and quivers, Forum Math, 30 (2018), 1319–1337.

[78]. Guo W, Manohar K, Brunton SL and Banerjee AG, Sparse-tda: Sparse realization of topological data analysis for multi-way classification, IEEE Transactions on Knowledge and Data Engineering, 30 (2018), 1403–1408.

[79]. Han YS, Yoo J and Ye JC, Deep residual learning for compressed sensing ct reconstruction via persistent homology analysis, preprint, arXiv:1611.06391.

[80]. Hatcher A, Algebraic Topology, Cambridge University Press, Cambridge, 2002.

[81]. Hiraoka Y, Nakamura T, Hirata A, Escolar EG, Matsue K and Nishiura Y, Hierarchical structures of amorphous solids characterized by persistent homology, Proceedings of the National Academy of Sciences, 113 (2016), 7035–7040.

[82]. Horak D and Jost J, Spectra of combinatorial Laplace operators on simplicial complexes, Adv. Math, 244 (2013), 303–336.

[83]. James IM, Fibrewise Topology, Cambridge Tracts in Mathematics, 91. Cambridge University Press, Cambridge, 1989.

[84]. Jonsson J, Simplicial Complexes of Graphs, Lecture Notes in Mathematics, 1928. Springer-Verlag, Berlin, 2008.

[85]. Kališnik S, Tropical coordinates on the space of persistence barcodes, Foundations of Computational Mathematics, 1–29.

[86]. Kontsevich M, Derived Grothendieck-Teichmüller group and graph complexes [after T. Willwacher], Séminaire Bourbaki, 2016/2017 (2019), 1120–1135; Exposés, 1126 (2019), 183–211.

[87]. Kozlov DN, Complexes of directed trees, J. Combin. Theory Ser. A, 88 (1999), 112–122.

[88]. Kozlov DN, Discrete Morse theory for free chain complexes, C. R. Math, 340 (2005), 867–872.

[89]. Kozlov DN, Simple homotopy types of Hom-complexes, neighborhood complexes, Lovász complexes, and atom crosscut complexes, Topology Appl, 153 (2006), 2445–2454.

[90]. Kramár M, Levanger R, Tithof J, Suri B, Xu M, Paul M, Schatz MF and Mischaikow K, Analysis of Kolmogorov flow and Rayleigh-Bénard convection using persistent homology, Phys. D, 334 (2016), 82–98.

[91]. Kramár M, Goullet A, Kondic L and Mischaikow K, Persistence of force networks in compressed granular media, Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 87 (2013), 042207. [PubMed: 23679407]

[92]. Li C, Ovsjanikov M and Chazal F, Persistence-based structural recognition, In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2014), 1995–2002.

[93]. Liu X, Wang X, Wu J and Xia K, Hypergraph-based persistent cohomology (HPC) for molecular representations in drug design, Briefings in Bioinformatics, 22 (2021).

[94]. Lovász L, Kneser's conjecture, chromatic number, and homotopy, J. Combin. Theory Ser. A, 25 (1978), 319–324.

[95]. Makarenko N, Kalimoldayev M, Pak I and Yessenaliyeva A, Texture recognition by the methods of topological data analysis, Open Engineering, 6 (2016).

[96]. Mallows CL and Sloane NJA, Two-graphs, switching classes and Euler graphs are equal in number, SIAM J. Appl. Math, 28 (1975), 876–880.

[97]. Maria C, Filtered complexes, in GUDHI User and Reference Manual, GUDHI Editorial Board, 2015, https://gudhi.inria.fr/doc/3.4.1/group__simplex__tree.html.

[98]. Meng Z and Xia K, Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction, Science Advances, 7 (2021), eabc5329. [PubMed: 33962954]

[99]. Mielants W and Leemans H, Z_2-cohomology of projective spaces of odd order, In Combinatorics '81 (Rome, 1981), Ann. Discrete Math., North-Holland, Amsterdam-New York, 18 (1983), 635–651.

[100]. Mischaikow K, Mrozek M, Reiss J and Szymczak A, Construction of symbolic dynamics from experimental time series, Physical Review Letters, 82 (1999), 1144–1147.

[101]. Mischaikow K and Nanda V, Morse theory for filtrations and efficient computation of persistent homology, Discrete Comput. Geom, 50 (2013), 330–353.

[102]. Munkres JR, Elements of Algebraic Topology, Addison-Wesley Publishing Company, Menlo Park, CA, 1984.

[103]. Nakamura T, Hiraoka Y, Hirata A, Escolar EG and Nishiura Y, Persistent homology and many-body atomic structure for medium-range order in the glass, Nanotechnology, 26 (2015), 304001. [PubMed: 26150288]

[104]. Nanda V, Perseus: The persistent homology software, Software available at http://www.sas.upenn.edu/~vnanda/perseus.

[105]. Nguyen DD, Cang ZX and Wei GW, A review of mathematical representations of biomolecular data, Physical Chemistry Chemical Physics, 2020.

[106]. Nguyen D, Gao K, Wang M and Wei G-W, MathDL: Mathematical deep learning for D3R grand challenge 4, Journal of Computer-Aided Molecular Design, 34 (2020), 131–147. [PubMed: 31734815]

[107]. Niyogi P, Smale S and Weinberger S, A topological view of unsupervised learning from noisy data, SIAM J. Comput, 40 (2011), 646–663.

[108]. Obayashi I, Hiraoka Y and Kimura M, Persistence diagrams with linear machine learning models, J. Appl. Comput. Topol, 1 (2018), 421–449.

[109]. Pachauri D, Hinrichs C, Chung MK, Johnson SC and Singh V, Topology-based kernels with application to inference problems in Alzheimer's disease, IEEE Transactions on Medical Imaging, 30 (2011), 1760–1770. [PubMed: 21536520]

[110]. Parks A, Lipscomb S and VA NSWCD., Homology and Hypergraph Acyclicity: A Combinatorial Invariant For Hypergraphs, Defense Technical Information Center, 1991, https://apps.dtic.mil/sti/citations/ADA241584.

[111]. Pokorny FT, Ek CH, Kjellström H and Kragic D, Persistent homology for learning densities with bounded support, In Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS'12, Curran Associates Inc., Red Hook, NY, USA, 2 (2012), 1817–1825.

[112]. Qaiser T, Tsang YW, Taniyama D, Sakamoto N, Nakane K, Epstein D and Rajpoot N, Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features, Medical Image Analysis, 55 (2019), 1–14. [PubMed: 30991188]

[113]. Rahman MS, Basic Graph Theory, Undergraduate Topics in Computer Science, Springer, Cham, 2017.

[114]. Rebala G, Ravi A and Churiwala S, An Introduction to Machine Learning, 2019.

[115]. Reininghaus J, Günther D, Hotz I, Prohaska S and Hege H-C, TADD: A computational framework for data analysis using discrete Morse theory, In Mathematical Software—ICMS 2010, Lecture Notes in Comput. Sci., 6327 (2010), 198–208.

[116]. Ren S, Wu C and Wu J, Weighted persistent homology, Rocky Mountain J. Math, 48 (2018), 2661–2687.

[117]. Robins V and Turner K, Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids, Phys. D, 334 (2016), 99–117.

[118]. Robins V, Computational topology for point data: Betti numbers of α-shapes, In Morphology of Condensed Matter, 600 (2002), 261–274.

[119]. Saadatfar M, Takeuchi H, Robins V, Francois N and Hiraoka Y, Pore configuration landscape of granular crystallization, Nature communications, 8 (2017), 15082.

[120]. Said A and Torra V, Data Science in Practice, Studies in Big Data, Springer International Publishing, 2019.

[121]. Schiffler R, Quiver Representations, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, Springer, Cham, 2014.

[122]. Seidel JJ, A survey of two-graphs, In Colloquio Internazionale Sulle Teorie Combinatorie (Rome, 1973), Tomo I, Atti dei Convegni Lincei, (1976), 481–511.

[123]. Seidel JJ and Taylor DE, Two-graphs, a second survey, In Algebraic Methods in Graph Theory, Vol. I, II (Szeged, 1978), Colloq. Math. Soc. János Bolyai, 25 (1981), 689–711.

[124]. Seversky LM, Davis S and Berger M, On time-series topological data analysis: New data and opportunities, In IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), (2016), 1014–1022.

[125]. Singh G, Memoli F, Ishkhanov T, Sapiro G, Carlsson G and Ringach DL, Topological analysis of population activity in visual cortex, Journal of Vision, 8 (2008).

[126]. Skraba P, Ovsjanikov M, Chazal F and Guibas L, Persistence-based segmentation of deformable shapes, In IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, (2010), 45–52.

[127]. Tausz A, Vejdemo-Johansson M and Adams H, Javaplex: A research software package for persistent (co)homology, Software available at http://code.google.com/p/javaplex, 2011.

[128]. Tierny J, Topological Data Analysis for Scientific Visualization, Springer-Verlag, Berlin, 2017.

[129]. Turner K, Mukherjee S and Boyer DM, Persistent homology transform for modeling shapes and surfaces, Inf. Inference, 3 (2014), 310–344.

[130]. Umeda Y, Time series classification via topological data analysis, Transactions of the Japanese Society for Artificial Intelligence, 32 (2017), 1–12.

[131]. Wang M, Cang Z and Wei G-W, A topology-based network tree for the prediction of protein-protein binding affinity changes following mutation, Nature Machine Intelligence, 2 (2020), 116–123.

[132]. Wang R, Nguyen DD and Wei G-W, Persistent spectral graph, Int. J. Numer. Methods Biomed. Eng, 36 (2020), e3376, 27 pp.

[133]. Wang Y, Ombao H and Chung MK et al. , Persistence landscape of functional signal and its application to epileptic electroencaphalogram data, ENAR Distinguished Student Paper Award.

[134]. Wee J and Xia K, Forman persistent Ricci curvature (FPRC)-based machine learning models for protein–ligand binding affinity prediction, Briefings in Bioinformatics, 22 (2021).

[135]. Wee J and Xia K, Ollivier persistent ricci curvature-based machine learning for the protein–ligand binding affinity prediction, J. Chem. Inf. Model, 61 (2021), 1617–1626. [PubMed: 33724038]

[136]. Wei G-W, Persistent homology analysis of biomolecular data, SIAM NEWS, 2017, https://sinews.siam.org/Details-Page/persistent-homology-analysis-of-biomolecular-data.

[137]. Wei G, Nguyen D and Cang Z, System and methods for machine learning for drug design and discovery, US Patent App., 16 (2019), 239–327.

[138]. Wells AL Jr., Even signings, signed switching classes, and (−1, 1)-matrices, J. Combin. Theory Ser. B, 36 (1984), 194–212.

[139]. Wu J, Simplicial objects and homotopy groups, In Braids, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap, World Sci. Publ., Hackensack, NJ, 19 (2010), 31–181.

[140]. Xia K and Wei G-W, Persistent homology analysis of protein structure, flexibility, and folding, Int. J. Numer. Methods Biomed. Eng, 30 (2014), 814–844.

[141]. Yadav N, Yadav A and Kumar M et al., An Introduction to Neural Network Methods for Differential Equations, SpringerBriefs in Applied Sciences and Technology. Springer, Dordrecht, 2015.

[142]. Zaslavsky T, Characterizations of signed graphs, J. Graph Theory, 5 (1981), 401–406.

[143]. Zeppelzauer M, Zieliński B, Juda M and Seidl M, A study on topological descriptors for the analysis of 3d surface texture, Computer Vision and Image Understanding, 167 (2018), 74–88.

[144]. Zhang ZF, Song Y, Cui HC, Wu J, Schwartz F and Qi HR, Early mastitis diagnosis through topological analysis of biosignals from low-voltage alternate current electrokinetics, In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE, (2015), 542–545.

[145]. Zhou Z, Huang YZ, Wang L and Tan TN, Exploring generalized shape analysis by topological representations, Pattern Recognition Letters, 87 (2017), 177–185.

[146]. Zhu XJ, Persistent homology: An introduction and a new text representation for natural language processing., In IJCAI, (2013), 1953–1959.

[147]. Zielinski B, Juda M and Zeppelzauer M, Persistence codebooks for topological data analysis, Artificial Intelligence Review volume, 54 (2021), 1969–2009.

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

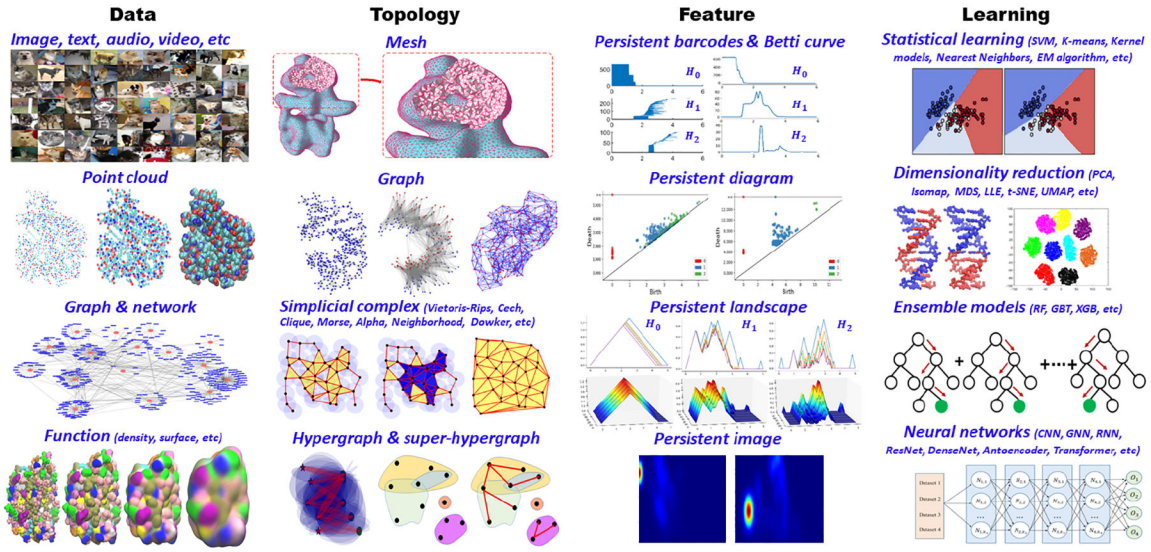


Figure 1.

Illustration of TDA and TDA-based learning models for data analysis. Generally speaking, all TDA-based learning models have four components, including data, topology, feature and learning. More specifically, data is collected and preprocessed firstly. Second, topological representations and models are constructed to describe the inner structural and interactional information of the data. Note that efficient representations are of key importance to machine learning. Third, a series of topological features are generated by using persistent homology models. Topological-invariant-based features provide a better characterization of the most fundamental and intrinsic properties of the data, thus they have a better generalizability and transferability for machine learning models. Finally, the topological features are combined with machine learning models for various classification and regression tasks.

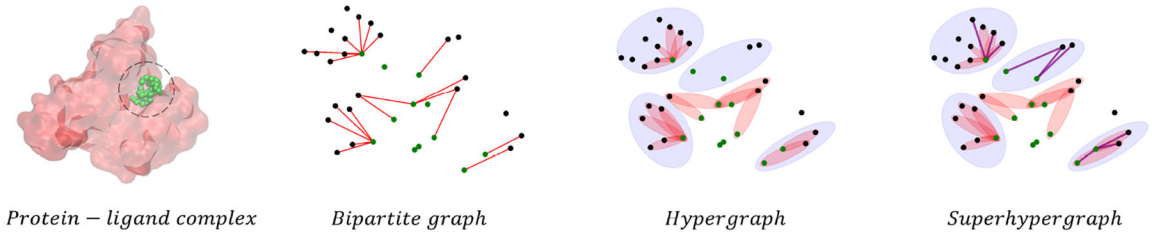


Figure 2.

Illustration of a super-hypergraph model constructed from the protein-ligand complex (ID: 3E6Y). The ligand (green color) is a drug that is used to cure the disease caused by the protein (red color). The potency and efficacy of the drug is directly determined by the atomic interactions between the ligand and the protein. Traditionally, atomic interactions are modeled by a graph (A). However, graphs can only characterize pair-wise interactions (by edges) and fall short for many-body interactions. Hypergraph models (C) use the hyperedge, i.e., a set of vertices, to represent many-body interactions and have demonstrated great power for biomolecular data analysis (See Section 4 for details). Mathematically, a n-hyperedge contains $n + 1$ vertices in it. Note that 1-hyperedges are denoted by red ellipses and n-hyperedges ($n > 1$) are represented by blue ellipses. The super-hypergraph (D) provides an even more flexible representation and incorporates detailed local topology within each hyperedge. Note that the hyperedge in super-hypergraph is a subgraph, i.e., a set of vertices together with edges. If we only consider vertex part of the subgraph, the super-hypergraph reduces to a hypergraph.

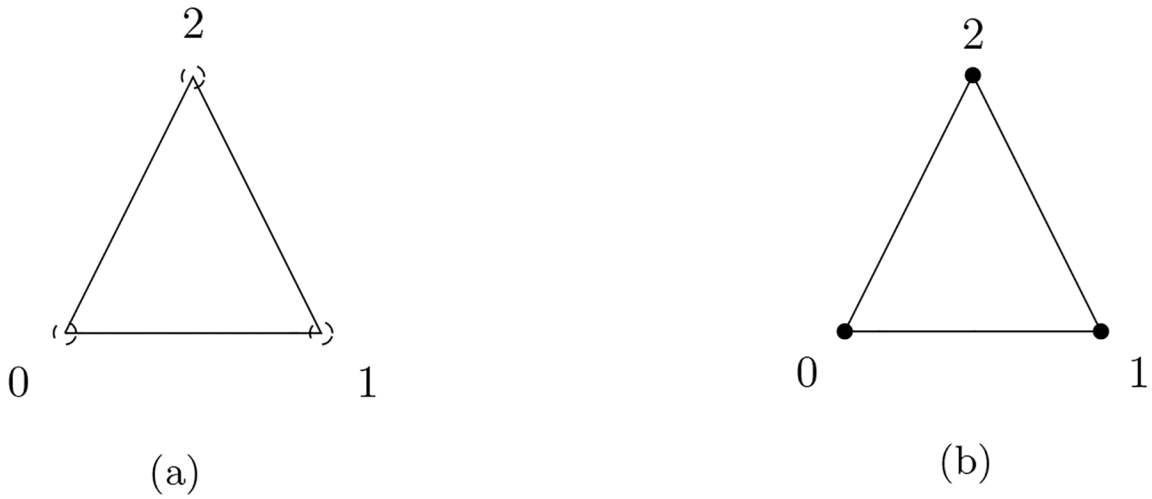


Figure 3.
(a) The hypergraph $\mathscr{K}$, where the cross indicates that a vertex is missing. (b) $\Delta(\mathscr{K})$, the smallest Δ-set that contains $\mathscr{K}$.

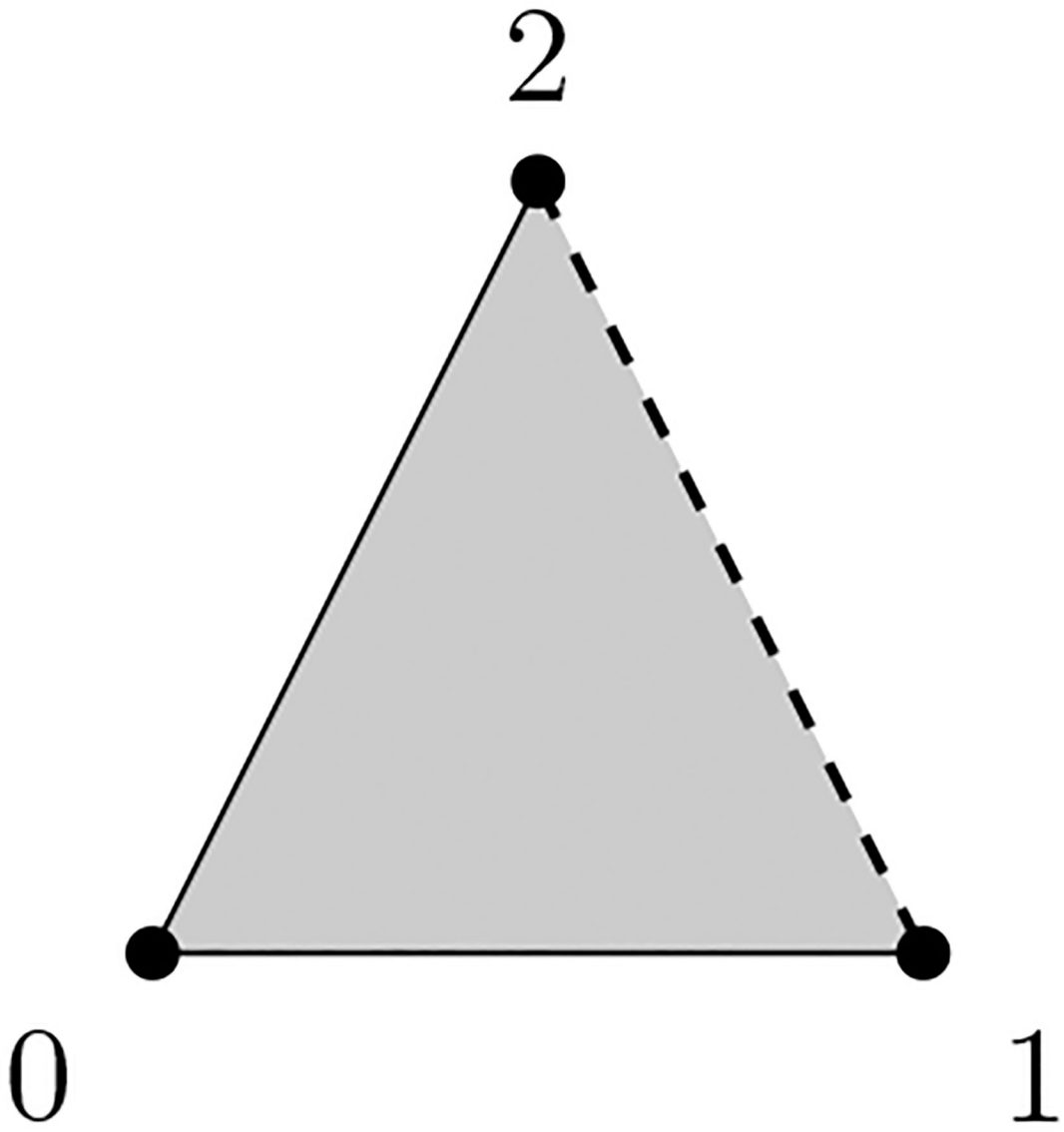


Figure 4.
The hypergraph $\mathscr{H}$ is a standard 2 simplex where the dotted edge is missing.

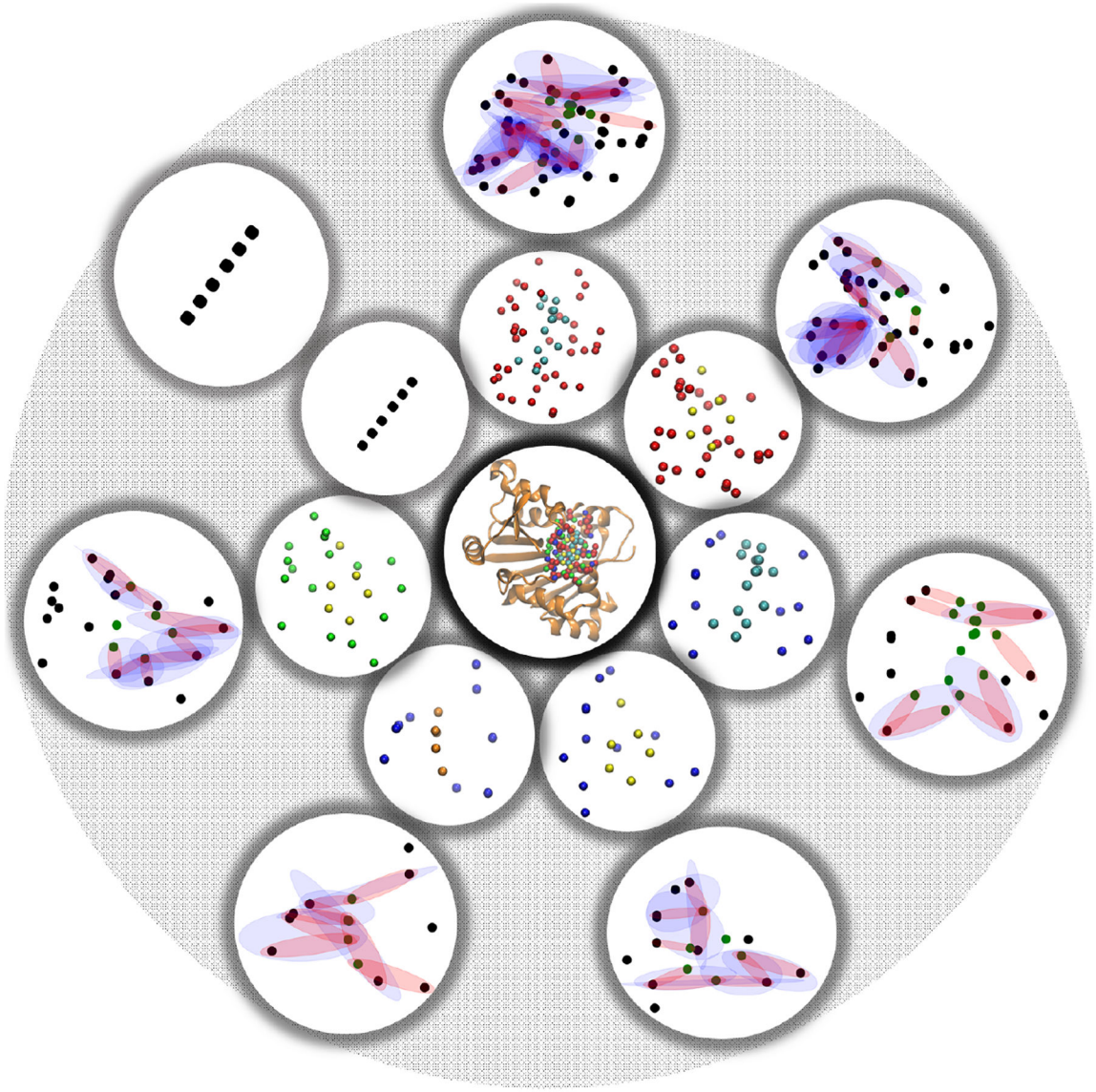


Figure 5.

Illustration of an element-specific hypergraph model for a protein-ligand complex (ID 3PB3). The binding core region of the complex is decomposed into a series of element-specific atom-sets. The interactions between protein atom-sets and ligand atom-sets are modeled as a series of hypergraphs.

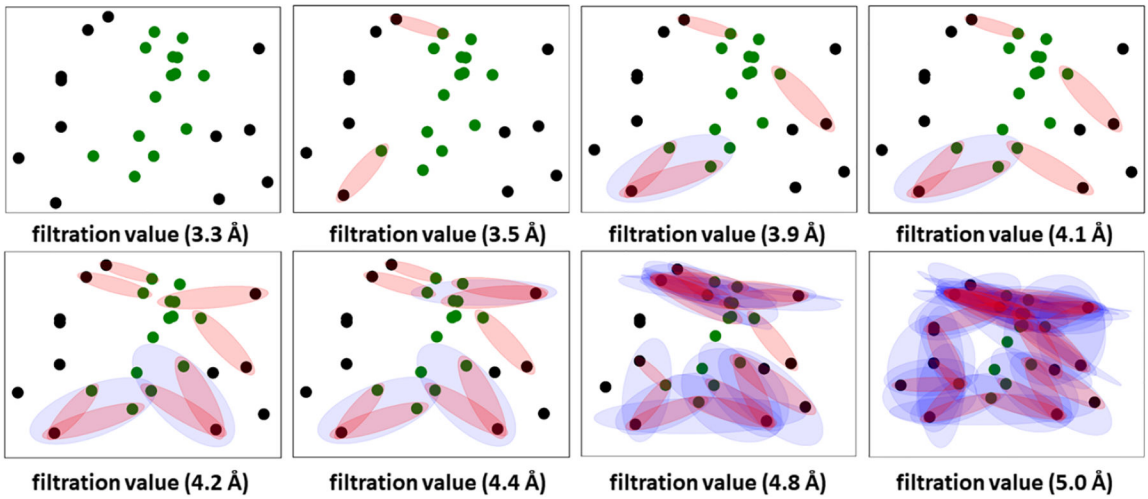


Figure 6.
Illustration of a hypergraph-based filtration process for the protein-ligand complex with ID 3PB3.

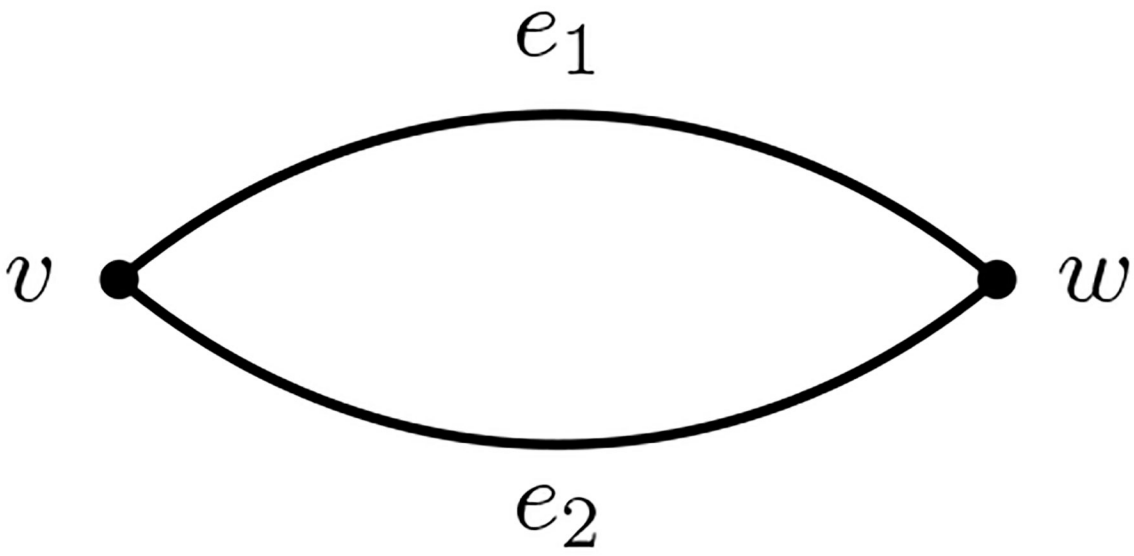


Figure 7.
The multi-graph G, which looks the same as the clique complex Clique(G).

Figure 8.
(a) The graph G, which is the same as Clique(G). (b) The neighborhood complex of G, $\mathcal{N}(G)$.

Table 1.

Topological structures associated to graphs

Constructions	Complex Type	Face Type
clique complex of a simple graph	simplicial complex	vertex-deletion
clique complex of a multi-graph	Δ-set	vertex-deletion
neighborhood complex	simplicial complex	vertex-deletion
Jonsson's graph complex	simplicial complex	edge-deletion
path complex of a simple graph	hypergraph	vertex-deletion
path complex of a multi-graph/quiver	super-hypergraph	vertex-deletion