A unified topological approach to data science—Joint with Prof Jelena Grbić from Southampton

Research Seminar on Geometry, Topology and Applications
Geometry and Topology of Novosibirsk State University
27 April 2020

Jie Wu

Hebei Normal University

April 27, 2020

A unified topological approach to data science

Motivations of a unified topological approach in data science

Topology of subgraphs

Super Persistent Homology

Pipeline of Current Topological Data Analysis (TDA)

1. input data consisting on a finite set of points coming with a notion of distance;
2. a “continuous shape” is built on top of the data: this results into an structure over the data;
3. topological and geometric information is extracted from the structures;
4. the topological and geometric information are the output of the approach and correspond to the new features of the data.

Such an approach can be naturally applied to point cloud data with a shortage that it **can not** be immediately or directly applied to **other data** such as graphs.

Our Goal/Hope—provide an uniform approach suitable for both point cloud data and graphic data

- In our setting, we explore topological structures on graphic data with scoring schemes.
- The current 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.
- This is a **theoretical research** on topological approaches in data science for hoping to make a tunnel between topology and data science.

Our Approaches

A. Homology Theory on **any collection** of subgraphs of a working graph. In theory, you choose whatever collection of subgraphs, you get homology on this collection of subgraphs.

B. Assign **any scoring scheme** on the working graph so that there is a **score** for any subgraph in the collection of subgraphs on your hand. Then it creates **persistent homology** as **new feature** for you.

C. Of course the **current persistent homology on point cloud data** should be answered from **A and B**.

Answer to C for Vietoris-Rips persistent homology

Let X be a point cloud data in $\mathbb{R}^N$. Mathematically, X is a finite set located in $\mathbb{R}^N$.

- **Step 1.** The **working graph** G is a **complete graph** by joining one edge for each pair of points in X.—**simple!**
- **Step 2.** The **collection of subgraphs**: any clique (complete subgraph) of G.—**simple!**
- **Step 3.** The **scoring scheme**: Let G' be a subgraph. Define its score

$$\mathfrak{M}^{VR}(G') = \frac{1}{2}\max\{d(v,w) \mid v,w \in V(G')\},$$

the half of the maximal distance between pairwise vertices.—**natural!**

Answer to C for Čech persistent homology

Let X be a point cloud data in $\mathbb{R}^N$. Mathematically, X is a finite set located in $\mathbb{R}^N$.

- **Step 1.** The **working graph** G is a **complete graph** by joining one edge for each pair of points in X.—**simple!**
- **Step 2.** The **collection of subgraphs**: any clique (complete subgraph) of G.—**simple!**
- **Step 3.** The **scoring scheme**: Let G' be a subgraph. Define its score

$$\mathfrak{M}^C(G') = \inf_{x \in \mathbb{R}^N} \max\{d(x, v) \mid v \in V(G')\},$$

the half of the maximal distance between pairwise vertices.—**also natural!**

Can we get anything new by looking scoring? Quick example 1

Let G be a graph located in $\mathbb{R}^m$. (e.g. graph data on $3D$ objects, data on protein structure.) Take VR-scoring on G. In stead of **complete graph on vertices of** G, we take **clique complex** $\mathrm{Clique}(G)$.

- $\Longrightarrow$ persistent homology converges to $H_*(\mathrm{Clqiue}(G))$.
- **Comparison.** *VR* persistent homology on point cloud data $V(G)$ converges to trivial homology.

- **Why is** $\mathrm{Clique}(G)$ **good?** Let $X = |K|$ be a polyhedron with K simplicial complex. Take G=1-skeleton of bary-centric subdivision of K. Then $|\mathrm{Clique}(G)| \cong |X|$.

Anything new? Quick example 2

Let us consider **pull-back scoring from a non-injective function from the vertex set to a Euclidean space**.

Let $p\colon E \to B$ be a fibration or fibre bundle with E, B polyhedra. Take triangulations on E and B to make p simplicial up to homotopy. Take graphs $G(E)$ and $G(B)$ as 1-skeletons of the bary-centric subdivisions of simplicial models for E and B.

Take scoring scheme on $G(E)$ as the **pull-back** of

$$V(G(E)) \xrightarrow{\text{proj}} V(G(B)) \xhookrightarrow{\text{embedding}} \mathbb{R}^m$$

Consider clique complexes $\mathrm{Clique}(G(E))$ and $\mathrm{Clique}(G(B)) \Longrightarrow$ **persistent Leray-Serre spectral sequence**.

The idea is a math formation of a practical method

Guo-Wei Wei, *Persistent homology analysis of biomolecular data*, **SIAM News 50 (10)**, December 1, 2017:

> However, persistent homology neglects chemical and biological information ... and is thus **not as competitive as** geometry or physics-based representation in quantitative predictions. **Element-specific persistent homology**, or multi-component persistent homology built on colored biomolecular network, has been introduced... This approach enciphers biological properties—such as hydrogen bonds, van der Waals interactions, hydrophilicity, and hydrophobicity—into topological invariants, rendering a **potentially revolutionary representation** for biomolecules.

Element-specific=subnetworks only having C or O or "C and O"...

Remarks

Dynamics is also important in biology. They concerns network motif.

- Ron Milo, Shai S Shenorr, Shalev Itzkovitz, Nadav Kashtan, Dmitri B Chklovskii, Uri Alon, *Network motifs: simple building blocks of complex networks*, **Science**, 298 (5594) (2002), 824-827.

Discrete Morse theory, and combinatorial Hodge-Laplacian are very important in applied topology. Some of my students work on these topics.

In this talk, we only discuss generalizations of persistent homology by notions.

The mathematical question

Let G be a working graph. Let $\mathcal{H}$ be a family of finite subgraphs.

Question. What is a natural way to define homology of $\mathcal{H}$?

Rationality of Question: Abstract simplicial complex is a family of (finite) subsets that is closed under subset-operation. There is a well-established **simplicial homology theory**.

New Situation:

1) $\mathcal{H}$ is a family of finite subgraphs, rather than a family of finite sets; and
2) **no hypothesis** that $\mathcal{H}$ is closed under subgraph-operation.

Topology of subgraphs has been extensively studied

A list of complexes from subgraphs:

- **clique (flag) complex:** A simple graph determines a simplicial complex whose k-simplices are the $(k+1)$-cliques of the graph. —A. V. Ivashchenko, 1994.
- **neighborhood complex** of a graph G is a simplicial complex whose vertices are the vertices of G and whose simplices are those subsets of the vertex set $V(G)$ which have a common neighbor.— L. **Lovász**, 1978.
- **graph complex:** abstract simplicial complex on the edge set.— Jacob Jonsson, book in 2008; Victor A. **Vassiliev**, 2018.
  - **Kontsevich**'s graph complex is defined differently.
- **path complex:** pathes in a digraph.— A. Grigor'yan, Y. Lin, Y. Muranov and **S.T. Yau**.
- Also **Hom complex**, Eric Babson and Dmitry N. Kozlov; **Independence Complex**...

Two types of topology on $\mathcal{H}$

For creating topology, we regard a subgraph in $\mathcal{H}$ as a **simplex** in certain dimension. It requires a **face-operation** so that we can "**glue**" together.

Face-operation 1. Vertex-deletion: clique complex, neighborhood complex, path complexes.

Face-operation 2. Edge-deletion: graph complex.

Edge-deletion Topology—Need homology of hypergraphs

Let G be a working graph. Let $\mathcal{H}$ be a family of finite subgraphs.

Consider $\mathcal{H}$ as a family of finite subsets of the edge set $E(G)$.

Each subgraph is determined by its edge set.

$\mathcal{H}$ is a **hypergraph** under **edge-deletion operation**.

There is a homology theory (as extension of simplicial homology theory) on hypergraphs: [1]

[1] Stephane Bressan, Jingyan Li, Shiquan Ren, Jie Wu, *The Embedded Homology of Hypergraphs and Applications*, Asia J. Math. **23** (2019), no. 3, 479-500.

Vertex-deletion Topology—Need homology of super-hypergraphs

Let G be a working graph. Let $\mathcal{H}$ be a family of finite subgraphs.

Consider $\mathcal{H}$ as a family of finite subsets of the vertex set $V(G)$.

Each subgraph **may not be** determined by its vertex set.

Example. Let G be a multi-graph with vertices a and b and two edges f_1, f_2 between a and b. Then af_1b and af_2b are two subgraphs having the same vertices.

If we want to explore topology of subgraphs, the notion of hypergraph is insufficient.

We need a new notion. We call it **super-hypergraph**.

Δ-set

A Δ**-set** means a sequence of sets $X = \{X_n\}_{n\geq 0}$ with *faces* $d_i : X_n \to X_{n-1}$, $0 \leq i \leq n$, such that

$$d_i d_j = d_j d_{i+1}$$

for $i \geq j$, which is called the Δ*-identity*.

The notion of Δ-set is a generalization of (abstract) simplicial complex by ruling out **face-operation**.

Simplicial homology can be defined using the notion of Δ-set.

Super-hypergraph

A **super-hypergraph** is a pair $(\mathcal{H}, X)$, where X is a Δ-set and $\mathcal{H}$ is a graded subset of X.

We call $\mathcal{H}$ a **super-hypergraph born from** X, and X is called **a parental Δ-set** of $\mathcal{H}$.

Example. Let G be a multi-graph with vertices a and b and two edges f_1, f_2 between a and b. Let $\mathcal{H} = \{af_1b, af_2b\}$ be two simple subgraphs. Then $\mathcal{H}$ can be viewed as a super-hypergraph with two 1-simplex with sharing the same missing vertices.

Algebraic Lemmas

Let G_* be a chain complex of groups and let D_* be a graded subgroup of G_*. Here we do not assume that G_n is commutative. Define

- $\sup_*^{G_*}(D_*)$ is the intersection of subcomplexes C_* of G_* with property that $D_n \leq C_n$ for $n \in \mathbb{Z}$.
- $\inf_*^{G_*}(D_*)$ is the product of subcomplexes E_* of G_* with property that $E_n \leq D_n$ for $n \in \mathbb{Z}$.

We briefly denote $\sup_*(D_*)$ for $\sup_*^{G_*}(D_*)$ and $\inf_*(D_*)$ for $\inf_*^{G_*}(D_*)$ if the embedding of $D_* \subseteq G_*$ is clear.

Algebraic Lemmas

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

1. $\sup_*(D_*)$ is the least 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 $\partial_n^{-1}(D_{n-1})$.

Algebraic Lemmas

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

1. The inclusion
$$\inf_*(D_*) \longrightarrow \sup_*(D_*)$$
induces an injective mapping on homology.

2. Suppose that $\partial^{G_*}_{n+1}(D_{n+1})$ is contained in the normalizer of D_n for each n. Then the inclusion
$$\inf_*(D_*) \longrightarrow \sup_*(D_*)$$
induces an isomorphism on homology. In particular, if D_n is normal in G_n for $n \in \mathbb{Z}$, then the inclusion $\inf_*(D_*) \longrightarrow \sup_*(D_*)$ induces an isomorphism on homology.

Embedded Homology of Super-hypergraphs

Let $(\mathcal{H}, X)$ be a super-hypergraph. Let A be an abelian group. The **embedded homology** $H_*^{\mathrm{emb},X}(\mathcal{H}; A)$ **with coefficients in** A of $(\mathcal{H}, X)$ is defined by the homology of the chain complex of $\inf_*$ and $\sup_*$ of the graded subgroup $\mathbb{Z}(\mathcal{H}) \otimes A$ in the chain complex $C_*(X; A)$.

Note. The **gap complex** $\sup_*(\mathbb{Z}(\mathcal{H}) \otimes A)/\inf_*(\mathbb{Z}(\mathcal{H}) \otimes A)$ is contractible. If there are some additional information, one may get further information on the gap complex. For instance, if there is a group G-action, then one may look at homology $H_*((\sup_*/\inf_*) \otimes_{\mathbb{Z}(G)} M)$ for G-modules M.

Remark

Embedded Homology of a hypergraph/super-hypergraph **may not be equal to** homology of a simplicial complex.

Let $\mathcal{H}$ be the boundary of a 2-simplex with **removing all three vertices**. Let X be the boundary of the 2-simplex. Then $H_1(X) = \mathbb{Z}$, $H_0(X) = \mathbb{Z}$, and $H_1(\mathcal{H}) = \mathbb{Z}$, $H_0(\mathcal{H}) = 0$.

No nonempty space whose unreduced 0-th homology is 0.

hypergraphs/superhypergraphs seem **geometry-like** objects.

Geometric gap complex: Let $\Delta\mathcal{H}$ be the minimal Δ-subset of X containing $\mathcal{H}$, and let $\delta\mathcal{H}$ be the maximal Δ-subset of X contained in $\mathcal{H}$. The inclusion $\delta\mathcal{H} \to \Delta\mathcal{H}$ may not be homotopy equivalent.

Mayer-Vietoris Sequence

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

Theorem. With the notation above, there is a commutative diagram

Mayer-Vietoris Sequence

$$
\begin{array}{ccccc}
 & & & & \sup_*^X(\mathcal{H}) \\
 & & & & \| \\
\sup_*^A(\mathcal{H}^A) \cap \sup_*^B(\mathcal{H}^B) & \hookrightarrow & \sup_*^A(\mathcal{H}^A) \oplus \sup_*^B(\mathcal{H}^B) & \xrightarrow{j_A - j_B} & \sup_*^A(\mathcal{H}^A) + \sup_*^B(\mathcal{H}^B) \\
\uparrow & & \uparrow \simeq & & \uparrow \\
\inf_*^A(\mathcal{H}^A) \cap \inf_*^B(\mathcal{H}^B) & \hookrightarrow & \inf_*^A(\mathcal{H}^A) \oplus \inf_*^B(\mathcal{H}^B) & \xrightarrow{j_A| - j_B|} & \inf_*^{A)}(\mathcal{H}^A) + \inf_*^B(\mathcal{H}^B), \\
\| & & & & \\
\inf_*^{A \cap B}(\mathcal{H}^{A \cap B}) & & & &
\end{array}
$$

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

Super Persistent Homology

Let G be a directed or undirected multi-graph with a **scoring scheme** $\mathfrak{M}$.

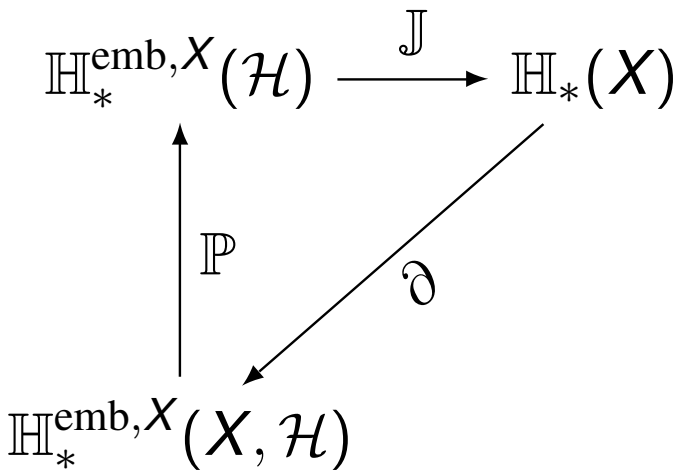
Let $(\mathcal{H}, X)$ be a super-hypergraph dominated by G, i.e X is a collection of finte subgraphs with face operation defined in certain way. Then there is a **filtration** $(\mathcal{H}^t, X^t)$, $t \in \mathbb{R}$, defined by

$$\mathcal{H}^t = \{G' \in \mathcal{H} \mid \mathfrak{M}(G') \leq t\} \text{ and } X^t = \{G' \in X \mid \mathfrak{M}(G') \leq t\},$$

The *associated persistent homology* of this filtration is called **super-persistent homology** of $(\mathcal{H}, X)$ under scoring scheme $\mathfrak{M}$.

Persistence modules of super-persistent homology

Let $(\mathcal{H}, X)$ be a super-hypergraph dominated by a directed/undirected (multi-)graph G with a scoring scheme. There is an exact triangle of graded persistence modules



with ∂ of degree -1.

Structure Theorem

Let $(\mathcal{H}, X)$ 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 $\mathbb{H}_*(X)$, $\mathbb{H}_*^{\mathrm{emb},X}(\mathcal{H})$ and $\mathbb{H}_*^{\mathrm{emb},X}(X,\mathcal{H})$ admit unique direct sum decompositions in terms of graded **interval persistence modules** up to the order of factors in the category of graded persistence modules.

Proof follows from a theorem of Crawley-Beovey derived from the classical Gabriel Theorem in representation theory.

Interval persistence module $\mathbb{F}^J$: J is an interval, $\dim(\mathbb{F}_t^J) = 1$ if $t \in J$, and 0 otherwise. The map $\mathbb{F}_t^J \to \mathbb{F}_s^J$, $t \leq s$, is defined in the canonical way.

What does Structure Theorem tell?

Let $\mathbb{M}$ and $\mathbb{N}$ be persistence modules that admit unique (up to order of factors) decompositions in terms of interval persistence modules. Let $\phi\colon \mathbb{M} \to \mathbb{N}$ be a morphism of persistence modules. Let $\mathbb{F}^{I_\alpha}$ be a factor of $\mathbb{M}$, and let $\mathbb{F}^{J_\beta}$ be a factor of $\mathbb{N}$. Then the composite

$$\mathbb{F}^{I_\alpha} \hookrightarrow \mathbb{M} \xrightarrow{\phi} \mathbb{N} \xrightarrow{\text{proj.}} \mathbb{F}^{J_\beta}$$

is either isomorphic or zero. Hence we have a **correlation matrix** M for ϕ with $m_{\alpha\beta} = 1$ if ϕ is an isomorphism and 0 if ϕ is zero.

$\Longrightarrow$ super-persistent homology gives (1) **three multi-layered barcodes**, and (2) **three multi-layered correlation matrices**.

Partition Homology and Persistent Partition Homology

Assume that there is a disjoint clustering **p** on the vertex set $V(G)$. In other words, there is a disjoint union

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

where each $V_i(G)$ is a cluster.

Let H be a subgraph of G. There exists a unique sequence $(k_0, k_1, \ldots, k_n)$ with $0 \leq k_0 < k_1 < \ldots < k_n \leq m$ such that $V(H) \cap V_i(G) \neq \emptyset$ for $i \in \{k_0, k_1, \ldots, k_n\}$ and $V(H) \cap V_i(G) = \emptyset$ if $i \notin \{k_0, k_1, \ldots, k_n\}$.

Viewing H as certain abstract n-simplex, we define the j-face $d_j^{\mathbf{p}}(H)$ as a full subgraph of H by removing all of those vertices $v \in V(H) \cap V_{k_j}(G)$ together the edges joining with v.

Other Constructions of Face Operations

- **link-blowup Face Operations.** Graph interpretation of geometric construction of a kind of *blowup*: Let H be a subgraph of G. We can remove a subset V_i of $V(H)$ together with removing edges joining V_i and filling back edges from G to those vertices of H that are neighbors to V_i. (Think as simplicial compex. Consider regular neighborhood of V_i and the link of V_i.
- **Face Operations on Subgraphs with Marked Starting-Vertices.** Let H be a subgraph with a subset of $V(H)$ as marked starting vertices V_0. Then we can take V_1 as the set of neighbors of V_0, and so on.
- All of the constructions on simplicial/Δ-structures on a collection of subgraphs would contribute to persistent homology such as **persistent path homology, persistent random walk homology**.

Scoring Scheme and sequence of cochains

Any scoring scheme $\mathfrak{M}$ **restricted to** $\mathcal{H}$ gives a sequence of cochains $\mathfrak{M}^n = \mathfrak{M}| : \mathcal{H}_n \to \mathbb{R}$.

Conversely, any sequence of cochains extends to a scoring scheme on G.

The relationships and product structures on chains and cochains might help for studying the **adjustment or learning** on scoring schemes.

Also **geometry** might help for studying the **adjustment or learning** on scoring schemes.

Remarks on further research

Clustering is a fundamental problem in data science. Let G be a graph data. Given any clustering **p** on $V(G)$, it means that $V(G)$ is covered by a collection of subsets $\{V_1, \ldots, V_n\}$. We may introduce topological structures on **p** in several ways such as the poset structure by inclusions, Čech complex by considering intersections of V_i's, or ceratin new graph whose vertices are given by the clusters and whose edges are given by certain **correlations**.

An interesting point is that, whence we decide certain topological structure on the vertex set according to a given clustering, we can look at collections of subgraphs whose vertex sets are clusters. The forgetful mapping from subgraphs to their vertex sets would induce Δ-maps if we manage well on face operations. Then the ideas from **fibrewise topology, fibrations, fibre bundles** may help for analyzing the clustering with providing some topological features.

Reference of the talk, and some remarks

Jelena Grbić and Jie Wu, *A unified topological approach to data science*, preprint.

It will be uploaded to ArXiv soon. (I hope)

Question. Would topology and dynamics be good tools for studying virus?

When the earth becomes warmer, many unknown ancient virus may come out from frozen land. Need mathematical methods to predict drugs for treating not-well-understood virus.

Need to classify protein structures to predict unknown virus in future—it is related to knot theory, but people in biology commented that classical knot theory seems not useful to them.

Thank You for Your Attention!