A Novel Approach to Topological Graph Theory

With R-K Tophedrons & LIGO Data Modelling

By: Animikh Roy[1] and Andor Kesselman[2]

February 22, 2021

Graph Theory and Topological Data Analytics are emerging as new and independent sets of highly effective tools for the multi-dimensional analysis of large datasets with intrinsic geometric properties. They are mathematically grounded methods that extract information from the inherent geometric nature of data by revealing concrete and useful analytical insights from multi-dimensional hidden layers of information. Graph Analytics has already proven to bring about a paradigm shift beyond the established capabilities of RDBMs through consistent and interdependent network hierarchy models. Whereas, Topological Data Analysis has been a more recently popularized approach for data compression and analysis of complex high dimensional data sets in the form of Phase Space projected Simplicial Complexes. The most common algorithm used in Topological Data Analysis is called Mapper, which uses partial clustering and persistent homology for Topological shape rendering and Isometric Data Compression. Although very powerful as a tool for scientific data analysis, the existing Mappers and Topological models have many drawbacks related to their sensitivity and consistency with Graph Network Analytics. In this paper, we aim to propose a novel approach for encoding vectorized associations between data points for the purpose of enabling smooth transitions between Graph and Topological Data Analytics. We also conclusively reveal effective ways of converting such vectorized associations to simplicial complexes in Topological Phase Space, resulting in filter specific, Homotopic Self-Expressive, 'Invariant-Topohedrons' with persistent Homology. Using filter based, self-expressed ML driven representations for data compression, we finally demonstrate the theory behind a shape invariant encoding in strong correspondence with Topological Network Entropy. The novel formulation of this work could lay the foundation for many future scientific and engineering applications for stable, high-dimensional data analysis purposes, especially with the advent of Cloud Distributed Massive Parallel Processing in the fields of Big-data Astronomy and beyond. It is focused to bring about the combined effectiveness of Topological Graph Theory formulations under a singularly effective framework.

1 Introduction

Since the advent of the "next-generation" high-throughput big-data revolution over the last decade, there has been an explosion of available scientific, business and social-media data, accelerating research, unprecedentedly in most domains of human society. These exponential advances in processing power coupled with distributed cloud-based file storage systems have revealed paradigm shifting implications in science and technology when aided with the correct, adaptable and scalable methods of analytics. Unfortunately figuring out the correct methodology with compatible robustness and scale has been a persistent challenge for researchers and analysts alike, due to the ever-growing size and complexity of high dimensional data sets. In fact, the very perplexing nature of scientific big-data coupled with their astronomical sizes are posing constant challenges to traditional computational methods, which are largely based on combinatorial mathematics and clustering techniques. In some cases, the nature of existing data is not suited to current approaches (for instance, the continuous nature of real-time data differentiation is not suited to conventional clustering methods); in others, the enormous size makes the analysis infeasible with current computing resources. Therefore, it is evident that new computational approaches

are required to boost existing Machine Learning (ML) driven analytical tools, systems and products to address these pertinent challenges.

Graph Theory (GT) and Topological Data Analysis (TDA) have recently emerged as independent novel frameworks for extracting hidden meaning and underlying insights from the study of geometric structure, shape and connections of such vast and complex datasets. However modern computational tools lack the technology, efficiency and flexibility to consistently carry out Graph Theory Network Analysis with hierarchical connections with localised clustering due to the inherent variability that could encode directed relationships in the affine connexions of the datasets to build homotopic manifolds and simplicial complexes. They also lack a consistent framework to mathematically define and classify the global properties of the same network through an effective means to smoothly transit between Graph and Topological structures without having to regenerate the entire data geometry from scratch due to lack of persistent homology between the two models.

Being able to showcase TDA and GT capabilities via smooth mathematical transformations on the same data network without the necessity to recreate its underlying geometric structure encompasses an enormous field of untapped potential in modern scientific big-data analytics. This academic paper explores that very possibility of consistently improving existing GT and TDA technologies with enhanced geometric compatibility while preserving their respective mathematical properties through simple Vectorized Associations in Topological Phase Space. This research also aims to facilitate a smooth transition between these two advanced analytical methodologies through a novel ML driven computational framework by building Self-Expressive Homotopic Topohedrons. These are shown as a special category of 3D Polyhedrons that maintain persistent Homology when projected onto a Topological Phase Space. These special Topohedrons are generated via select, filter-based ML driven optimizations on the underlying n-dimensional data set, preserved within the supressed Topological Space. This work also discusses the conditions involved with the preservation of Homotopy of such Topohedrons under continuous deformations brought about by any changes in Topological Network Entropy and formulates those implications through mathematical and computational models.

The formulation of this work could have seminal implications on high-dimensional, complex scientific data sets especially in the fields of Astronomy and Particle Physics without the necessity of conventional, cumbersome clustering and binning techniques. It also replaces the existing Mapper algorithms with a holistic analytical framework that goes well beyond partial clustering and persistent homology for shape rendering and Isometric Data Compression native to conventional Topological Analytics.

1.1 Background and Motivations

The field of Topological Data Analysis is an emergent field with promising techniques for data compression and discovery. Topological Data Analysis is intended to help provide a toolset capable of exposing relevant features from high dimensional data by using geometric concepts. While geometric interpretation is a century old problem, modern conceptions of topological data analysis originated in 2002 with the work of Edelsbrunner et. al and have been considerably contributed to since then. Most notably, Professor Gunner Carlsson pioneered critical work in topological data computation in 2009. Since then it has been a growing field of research vaired set of research focuses.

Most of topological data analysis is centered around proving analysts with a *toolset* to understand fundamental geometric properties of high dimensional data. There are many challenges with current topological data analysis methods and in this paper we tackle the issue of persistence and dynamic systems. By combining concepts of phase space, topology, and graph together we propose a unified framework for analyzing dynamic systems. Our hope is that with future research using this paper as a launching point, researchers will have access to unparrelled flexibility in data analysis, by allowing them to have stable geometric analysis under both relational graph structures and geometric tensors. To implement such a radically new way of approaching data analysis, we have had to introduce a number of concepts such as "Association Vectors", "Roy Simplicies + Complexes", "Roy Topohedrons". Using the flow of information and entropic measurements, we are able to monitor and evaluate the development of our system. We describe the full system as the "Roy-Kesselman Model".

Details will be elaborated in the following sections however here is the coarse outline of our paper:

1. Topological Data Analysis - We describe current techniques as well as give a brief literature review.
2. Phase Space - We discuss phase space and it's implication on dynamic systems.
3. Pipeline - The main section of our paper with our pipeline and methodology for implementing our algorithm over dynamic systems.
4. Discussion - We will discuss implications for future work and improvements.

2 Topological Data Analysis: Literature Review

The field of Topological Data Analysis is an emergent field with promising techniques for data compression and discovery. Generally the pipeline for topological data analysis techniques follow the following pipeline[1]:

[1] **Michel2017**.

1. Input is a finite set of points with a notion of similarity or distance between them. It is generalized metric space of distance, and can be either induced on inherent.
2. A "countinous" shape is built on the data to highlight topology. They are built by covers over the input matrix and generate a group of simplicial complexes (called a filtration) that reflects the structure of the data at seperate scales.
3. Toplogical information is built from the data.
4. Analysis is applied over the topological set formed by the compression techniques.

There are a variety of challenges in topological data analysis, such as efficient convergence toward reebs graphs, sensitivity to resolutions and filtration, probabilistic interpretations, robust methods that are insentive to input type, flexibility for switching between relational and topological frameworks, persistence problems, and many more. We attempt to immediately address a few of these problems, mainly fusion of graph space and topological space, stability under perterbation, and dynamic systems.

3 Topological Phase Space

In the last century, the development of modern physics has been partially driven by the incorporation of a few key concepts. A phase space is the spatial representation of all possible states of a dynamical system, where each point uniquely identifies a state. A Topological Phase Space can be defined as the n-dimensional spatial representation of the same using a generalized Curvilinear Coordinate system allowing for all possible coordinate transformations and perfect isometric compressions while preserving geometric invariance.

The concept of Phase Space in itself is a simple but powerful idea that emerged in the second half of the 19th century, during the golden era of differential geometry, and it is at the core of modern classical, quantum, and statistical mechanics. The trajectory that a dynamical system describes in the phase space as it evolves with time contains rich information about the system. For instance, by looking at the shape of the trajectories that a pendulum describes in its phase space, we can infer the existence of different dynamical regimes, or the ratio between the length of the pendulum and the acceleration of gravity.

The phase space of a simple pendulum is a two-dimensional cylinder, where the periodic coordinate corresponds to the angle (q) of the pendulum with respect to the vertical, and the longitudinal coordinate to its angular velocity (v). Each point in this space specifies a unique combination of the position and velocity and uniquely determines the subsequent evolution. For small angular velocities, the pendulum oscillates back and forth around the equilibrium point. For large velocities, the pendulum describes a circular motion.

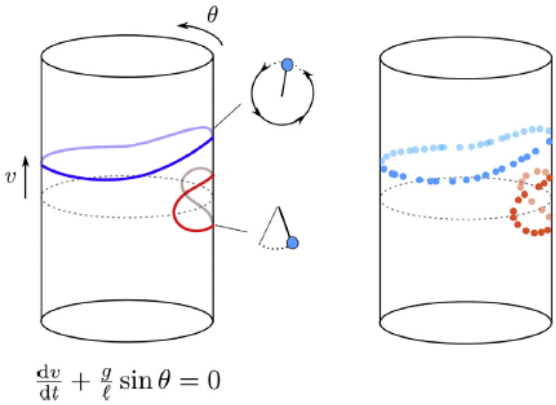

Figure 1: *Pendulum in Phase Space*

These two regimes are represented by qualitatively different trajectories in the phase space which cannot be continuously deformed into each other (in mathematical terms, they are homotopically inequivalent). By just looking at the shape of the trajectories in the phase space, we can extract information about a dynamical system. The dynamics of the simple pendulum is fully described by a differential equation depending on the length of the pendulum (l) and the acceleration of gravity (g). In more complex biological systems for example, such mathematical equations describing trajectories in the phase space are usually unknown, but current technologies allow to reconstruct trajectories from high-throughput measurements.

4 Pipeline

In this section we elaborate on the full algorithmic pipeline in respect to the sections as described in the previous sections. By combining isometric compression, association vectors, machine learning, phase space projections, and filters we are able to accomplish a novel approach toward data analysis which allows for two important properties: Classification and identification of stable shapes in phase space and a smooth transition between topological and graph structures. This section will demonstrate the component flow in specificsand the full pipeline. Some of the specific techniques used such as clustering, cuts, entropy measures, etc will be proposed in the following sections, however the best implementation of each pipeline component needs to be tested rigorsly and often may be application based.

See Figure 10.2 for the pipeline.

4.1 Step 0: Input Space

Let $\mathbb{X}$ or the *Input Space* be superset that consists of a set of $n \times m$ input matrices X where $X_i = [0, N] \subseteq \mathbb{R}^d$ whose union into $\mathbb{X}$ form the toplogical set . The set of X form a topological set τ, due to the following properties:

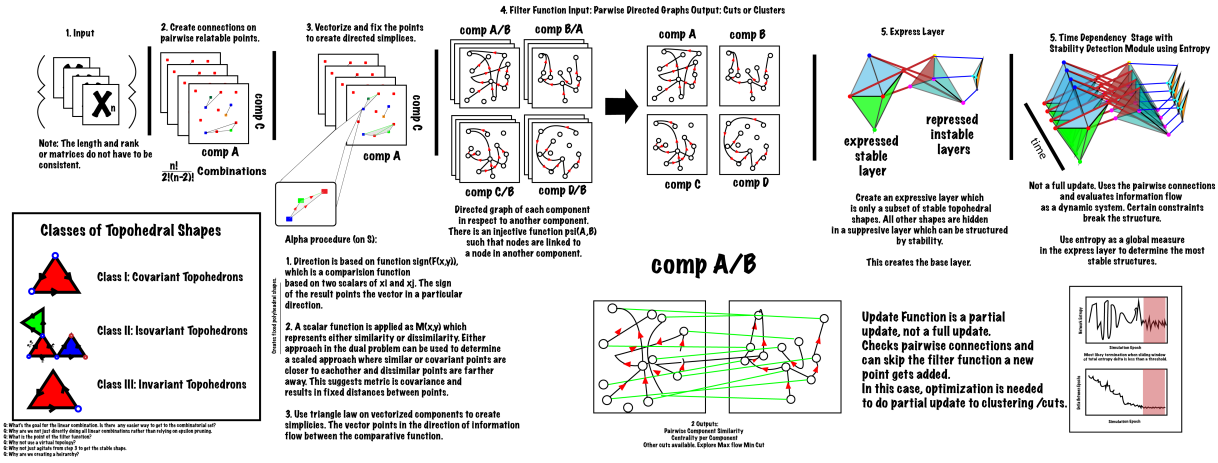

Figure 2: *Pipeline*

- $X \in \tau, \emptyset \in \tau$
- $\{O_i\}_{i \in I} \subseteq \tau \Rightarrow \bigcup_{i \in I} O_i \in \tau$ (the union of τ is in τ)
- $\{O_i\}_{i \in I}^{n} \subseteq \tau \Rightarrow \bigcap_{i \in I} O_{i=1}^{n} \in \tau$ (intersections in τ are in τ)

The topological space forms a generalization of the metric space and so any metric space as an input would be sufficient to continue the problem. Due to some of our later constructions in the pipeline it is necessary that there is a bijective mapping ψ for each component within τ such that $\forall \psi : x \in X \rightarrow y \in X_{y \neq x}$. Moreover, $\forall \cup_{i=0}^{n} x \in X$ it is one-to-one and forms X.

4.2 Step 1: Association Vectors

From point $p \in X$ and set $S \subseteq X$ we denote the minimum distance ϵ and $\delta(p, S)$ to be the minimum distance function from $p \rightarrow S$. Similar to Vietoris–Rips complex, we evaluate all points within $\delta(p, S) \leq \epsilon$, by drawing them as connected components.

Computationally, creating the connected components is the quadratic of the points in each layer however[2] there has been a variety of work done to make Vietoris–Rips approximations more computationally tractable such as Donald Sheely who achieves simplicial complexity of $O(n \log n)$ time, however we will leave future papers to discuss optimization strategies necessary for fast computation as this will require modification of traditional Vietoris–Rips constructions.

Let us call the set connected chain C whose properties include:

- $C \in X$ and $C \leq X$
- $\mathbb{C} = \bigcup_{i \in I} C$ such that the union of all connected chains are the set of all connected components.
- $\|\vec{c}\|$ denotes the length of the connected component, which is the number of edges.
- $\|\vec{c}\| = n - 1$ where n is the connected vertices.
- $\| \cap C_{i \in I} / \mathbb{C} \| \geq 0$ such that the intersection of all connectd chains is greater than or equal to 0.
- Each component in C can be represented by their pairwise components such that $C_{i,j}$ represent the connection between $C_{i,j}$
- Order invariant such that $C_{j,k} = C_{k,j}$
- $\mathbb{C} = \{C_{j,k} \rightarrow C_{k,i} \rightarrow C_{i,l} \rightarrow ...\}$ such that $\mathbb{C}$ represents a connected chain which contains no disjoint components. In other words, one could trace the chain without ever having to lift his/her pencil.

To encode direction and magnitude into our pairwise connected components, we define the **Alpha Function** as stated below.

Definition 4.1 (Alpha function)**.**

Let the α function be an encoding function on a connected set C such that $\forall (c \in C)$, there is a direction and magnitudal component such that $\forall C, (m, d \in c)$. We call each c an association vector which after encoding is denoted as $\vec{c}$. Let the set $\vec{A}$ be bijective and $\vec{A} = C$ however let C be a more generalized form. We call the encoding function α, such that $\alpha(C)$ provides the necessary encodings for each vector. There are a number of consequences with this encoding, such as the fact that $\|\vec{c}\|$ is fixed and contingent to the α function.

In dynamic systems, these values change as the data underneath changes, thereby these values are only snapshot invariant. With perterbation to the data and time our underlying encoding undergoes state changes. As will be clear later, this is important to consider because we will use techniques to determine stable structures under pertebation and time.

α_M: Magnitude Function of Alpha

Let the magnitude function be consistent with the definition of a distance function, who's length is inversely porportional to the distance encoding. Such is that larger distances are localized with further seperation and closer distances are more proximal with ea-

[2] **Piegl**.

chother. Let further formulization be described as below:

1. $D(A_{a,b}, B_{c,d}) > 0$
2. $D(A_{a,b}, B_{c,d})$ is finite.
3. $D(A_{a,b}, B_{c,d})$ is ordinal with higher similarity being inversely porportional to the value.
4. The relative encodings are scale invariant.

α_D: Directional Function of Alpha

Let the directional function be a singleton choice in the set $B = \{Source, Target, \emptyset\}$ where the directional component represents flow of information between source and target. For the encoding, we declare a comparitive interface $\alpha_D(source, target)$ which chooses an element in B which provides the directional component of the association vector. To measure information flow, we will typically use the function: $sign(f(source, target))$ where f is a comparative function that measures information flow between the source and the target.

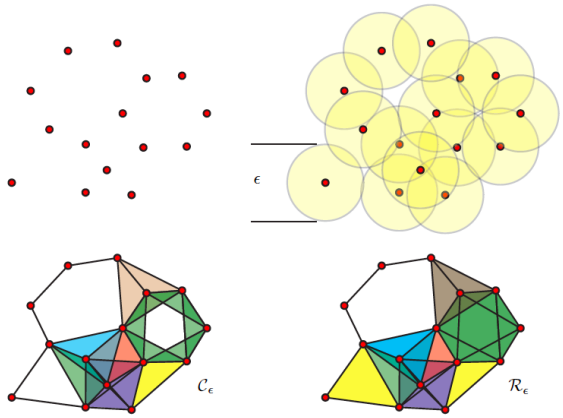

Figure 3: *A fixed set of points. Top right: Closed balls of radius $\epsilon/2$ centered at the points. Bottom left: Cech complex has the homotopy type of the $\epsilon/2$ cover. $(S^1 \vee S^1 \vee S^1)$ Bottom right: Vietoris-Rips complex has a different homotopy type $(S^1 \vee S^2)$. Image from R.*

4.3 Step 2: Directed Simplicial Complex

Using the concept of collinearity, we can create a constrainted and directed simplex by use of the triangle law. We will term each 2-simplex as a "Roy Simplex" and the directed simplicial complex that will eventually be formed after the pipeline is finished as a "Roy Topoheadron".

Definition 4.2 (Roy Simplex)**.** *Let a roy simplex be a directed simplicial complex where one side is directed vector made from the collinear combination of the two other vectors that compose the 2-simplex.*

Let a triangle be completable if it can be represented by a linear combination between two vectors in vectorspace. As a basis for this completion, we will use the simple vector triangle law to create collinear connected components. We represent this by equation 10 below. Using another comparative function, we look at the information flow between collinear points to provide a direction for the connecting vector. We will call this this collinear association vector, and this vector is critical for later on in the pipeline for creating stable structures.

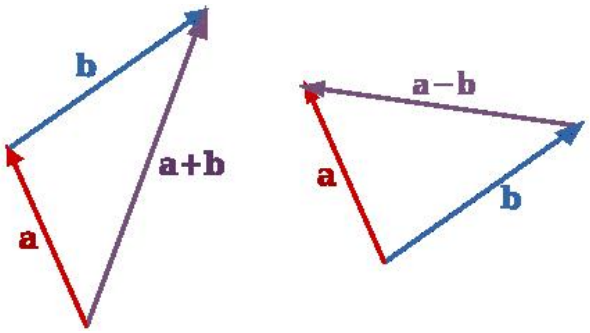

Figure 4: *Vector Addition. The left is tail-head oriented vectors. To the right is coinitial vectors*

Triangle of Vector Law Addition

If CoInitial Vectors:

$$\vec{c} = \vec{a} - \vec{b} \tag{1}$$

If Tail-Tail Vectors:

$$\vec{c} = \vec{a} + \vec{b} \tag{2}$$

where:
$\vec{c}$ is the connecting vector between vectors a and b, originating at a singleton if it is coinitial. If the vectors are tail-head oriented then $\vec{c}$ originiate from different points.

We term the *Dominating Vector* to be larger magnitude of two vectors |A| and |B|. To find the dominate vector is very simple. The formulate $sign(\|\vec{a}\| - \|\vec{b}\|)$ gives the dominate vector with respect to the first component. When providing direction to our generated vector, we point the vector towards the dominate vector. This is intended to mimic the flow of information from one point to another.

The generated vector we term as the **dependent vector**.

4.3.1 Roy Simplex Classifications

Let us denote the information state of the network as W and the velocity of the information toward the beginning and end of each timestep as W_i and W_f respectively. Let us denote $\Delta(W)$ as the total change of information in the network through over timestep h and $\partial(W)$ to be related to the component differentials of W.

The formation of Roy Simplicial Complexes and ostensibly Roy Polyhedrons have geometric implications

that will be described below. We classify the roy simplex into three different classifications based upon the intrinsic nature of the simplex.

[Class I] - Invariant Topohedrons: are topohedrons that perserve symmetry under all cases. They are formed as a colinear combination of two vectors and are stable structures in the topohedron. As the flow of information changes between the dependent vector, the structure is stable and the toplogy does not change.

[Class II] - Covariant Topohedrons: are topohedrons that perserve symettry however are linked simplicials on a single vertex.

[Class III] - Isovariant Topohedrons: are formed by cycles in information flow and thus have no dominate vector. There are few properties unique to an isovariant topohedron.

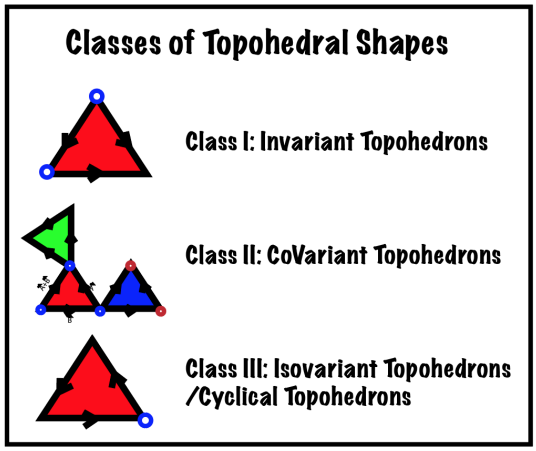

Figure 5: *Classes Explained*

4.4 Express Steps

Definition 4.3 (Component Map)**.** *Let $\psi : A \to B$ be the component map from component A to component B which is an injective map between pairwise components.*

Let us call the express step in the pipeline a lossy compression step where we isometrically compress the simplicial simplices formed in Step 3 into multi-level classifications. To do this, we employ partial clustering on filter mechanisms such as centrality and pairwise simlarity to form cardinality on the set. We use the highest rank of cardniality to form the set we call the *Express Layer* and follow the expressed layer by disjoint sets called *Repressed Layers*.

After filtration, three componential "Roy Simplexes" and connected to form a "Roy Topoheadron". The Roy Topoheadron has a number of special properties.

4.5 Entropy Measurements and Stability Detection

As a dynamic system, we can produce these "Roy Topoheadrons" over time, which allow us to monitor the development of the topology through phase space. We define the phase space of a roy topoheadron to be the shadow projection of the topological shape over time. By using a measurement of network entropy.

Gibbs Entropy

$$\sigma = 1/N log F \tag{3}$$

where: F denotes the cardinality of the ensemble. Other measurements include metrics such as:

Shannon Entropy:

$$S = (\mathcal{L}) = -\sum_{i<j} \sum_{\alpha} \pi_{ij}(\alpha) log \pi_{ij}(\alpha)$$

There is more work to be done to find the best entropy measurements for a dynamic system of directed topological objects and graphs.[3]

4.6 Updating Topology

Our update function does not need to recompute all steps from scratch. There are a variety of opimization steps that can be employed to increase the efficieny of update function. After the first run, we can optimize so each step updates minimalistic. The optimizations will be discussed in a seperate paper.

5 Discussion

Discussion points:

1. Optimization strategties
2. Entropy Measures
3. Pertebation

5.1 Application Discussion

5.2 Further Work

[3] **Bianconi2009**.