Title

Applying Topological Data Analysis to Logistics Systems Analysis

Author

Carlsson, John G

Publication Date

2024-07-01

DOI

10.7922/G20C4T4B

Applying Topological Data Analysis to Logistics Systems Analysis

July 2024

A Research Report from the National Center for Sustainable Transportation

John Gunnar Carlsson, University of Southern California

National Center for Sustainable Transportation
A USDOT University Transportation Center

METRANS Transportation Consortium
USC CSULB

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. NCST-USC-RR-24-12	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A	
4. Title and Subtitle Applying Topological Data Analysis to Logistics Systems Analysis		**5. Report Date** July 2024	
		6. Performing Organization Code N/A	
7. Author(s) John Gunnar Carlsson, Ph.D., https://orcid.org/0000-0001-5346-8529		**8. Performing Organization Report No.** N/A	
9. Performing Organization Name and Address University of Southern California METRANS Transportation Consortium University Park Campus, VKC 367 MC:0626 Los Angeles, California 90089-0626		**10. Work Unit No.** N/A	
		11. Contract or Grant No. USDOT Grant 69A3551747114	
12. Sponsoring Agency Name and Address U.S. Department of Transportation Office of the Assistant Secretary for Research and Technology 1200 New Jersey Avenue, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Final Research Report (October 2022 – September 2023)	
		14. Sponsoring Agency Code USDOT OST-R	
15. Supplementary Notes DOI: https://doi.org/10.7922/G20C4T4B			
16. Abstract The purpose of this project is to apply computational tools from topological data analysis (TDA) to study logistical systems such as freight networks. TDA is a relatively nascent research area that allows one to describe geometric properties of a data set, such as connectivity, existence of holes, or clustering, in a way that imposes minimal assumptions on parametric structures like coordinate systems or forms of probability distributions. In recent years, TDA has been successfully applied to many different scientific domains, such as aviation, path planning, and time series analysis. To the best of our knowledge, this project will be the first to apply TDA to the logistics domain.			
17. Key Words Topological data analysis, freight transport		**18. Distribution Statement** No restrictions.	
19. Security Classif. (of this report) Unclassified	**20. Security Classif. (of this page)** Unclassified	**21. No. of Pages** 25	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

NCST

About the National Center for Sustainable Transportation

The National Center for Sustainable Transportation is a consortium of leading universities committed to advancing an environmentally sustainable transportation system through cutting-edge research, direct policy engagement, and education of our future leaders. Consortium members include: the University of California, Davis; California State University, Long Beach; Georgia Institute of Technology; Texas Southern University; the University of California, Riverside; the University of Southern California; and the University of Vermont. More information can be found at: ncst.ucdavis.edu.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

The U.S. Department of Transportation requires that all University Transportation Center reports be published publicly. To fulfill this requirement, the National Center for Sustainable Transportation publishes reports on the University of California open access publication repository, eScholarship. The authors may copyright any books, publications, or other copyrightable materials developed in the course of, or under, or as a result of the funding grant; however, the U.S. Department of Transportation reserves a royalty-free, nonexclusive and irrevocable license to reproduce, publish, or otherwise use and to authorize others to use the work for government purposes.

Acknowledgments

This study was funded, partially or entirely, by a grant from the National Center for Sustainable Transportation (NCST), supported by the U.S. Department of Transportation (USDOT) through the University Transportation Centers program. The authors would like to thank the NCST and the USDOT for their support of university-based research in transportation, and especially for the funding provided in support of this project.

NCST

Applying Topological Data Analysis to Logistics Systems Analysis

A National Center for Sustainable Transportation Research Report

July 2024

John Gunnar Carlsson, University of Southern California

NCST

[page intentionally left blank]

TABLE OF CONTENTS

List of Figures

Applying Topological Data Analysis to Logistics Systems Analysis

EXECUTIVE SUMMARY

The purpose of this project has been to apply computational tools from topological data analysis (TDA) to study logistical systems, with an emphasis on freight networks. TDA is a relatively nascent research area that allows one to describe geometric properties of a data set, such as connectivity, existence of holes, or clustering, in a way that imposes minimal assumptions on parametric structures like coordinate systems or forms of probability distributions. In recent years, TDA has been successfully applied to many different scientific domains, such as aviation, path planning, and time series analysis. To the best of our knowledge, this project has been the first to apply TDA to the logistics domain.

TDA is particularly useful for identifying coarse features in a datset, such as clusters, connected components, cycles, or holes. These all have natural interpretations within the context of freight network analysis; for example, a cluster of points likely corresponds to a large metropolitan region with major activity, and a cycle may correspond to desirable sequences of loads that shippers undertake so as to start and end at their home destination. The ability to identify these features represents a powerful tool in the analysis of a wide range of problems in freight and network analysis, such as identifying bottlenecks, cyclic behavior, or clustering.

The basic principle that we have exploited is that TDA excels at identifying coarse features in datasets using a technique called *persistence*, and is not sensitive to more localized phenomena. The fundamental data structure in TDA is called a *simplicial complex*, which is a generalization of a network structure that allows one to identify not only pairwise relationships (i.e. arcs or links in a network"), but also relationships between three or more entities (e.g., "these four cities are all part of the same metropolitan region). We have used these tools to make descriptive insights about the interconnectedness of freight networks.

Introduction

The purpose of this report is to apply techniques from topological data analysis (TDA) to problems in logistics systems analysis. TDA is a relatively nascent research area that allows one to describe geometric properties of a data set, such as connectivity, existence of holes, or clustering, in a way that imposes minimal assumptions on parametric structures like coordinate systems or forms of probability distributions. In recent years, TDA has been successfully applied to many different scientific domains, such as time series analysis, text mining, cancer biology, and materials science. To the best of our knowledge, this project has been the first to use TDA in the area of transportation, or operations research in general. The basic principle that we exploit is that TDA excels at identifying coarse features in datasets using a technique called *persistence*, and is not sensitive to more localized phenomena. This enables us to use its strengths in unique ways, such as identifying coarse features in network flows and improving the performance of local search methods for logistical optimization problems.

The structure of this report is as follows: Chapter 1 introduces the foundational techniques of TDA and *persistent homology* and illustrates its use via several examples. In Chapter 2, we define a new simplicial complex construction associated to a stochastic matrix and steady state vector. The persistence function is defined axiomatically in such a way that leverages the available structure of a random process. Chapter 3 investigates the complex's efficacy in recovering coarse features of Markov chains on data taken from origin-destination pairs in the US freight network.

1. Topological Data Analysis: A Primer

Topology is the branch of mathematics that studies shapes and spatial relations. Its application to the analysis of high-dimensional data sets is called *Topological Data Analysis (TDA)*. Topology has several features that make it ideal for applications to data [1]. First, topological techniques are "coordinate free," meaning the geometric properties being studied are intrinsic and do not depend on the choice of coordinates. Second, topology studies properties that are invariant under small deformations, making it possible to pick out the "shape" of objects despite variation and deformation. Thirdly, extensions of topological methods like homology allow one to construct summaries of the invariants of a space over a range of parameter values. This is useful because the results of point cloud techniques often depend on the choice of parameter, so a summary over a changing parameter value is oftentimes more valuable. These properties, which are illustrated in Figure 1, make topology an effective lens through which to view data.

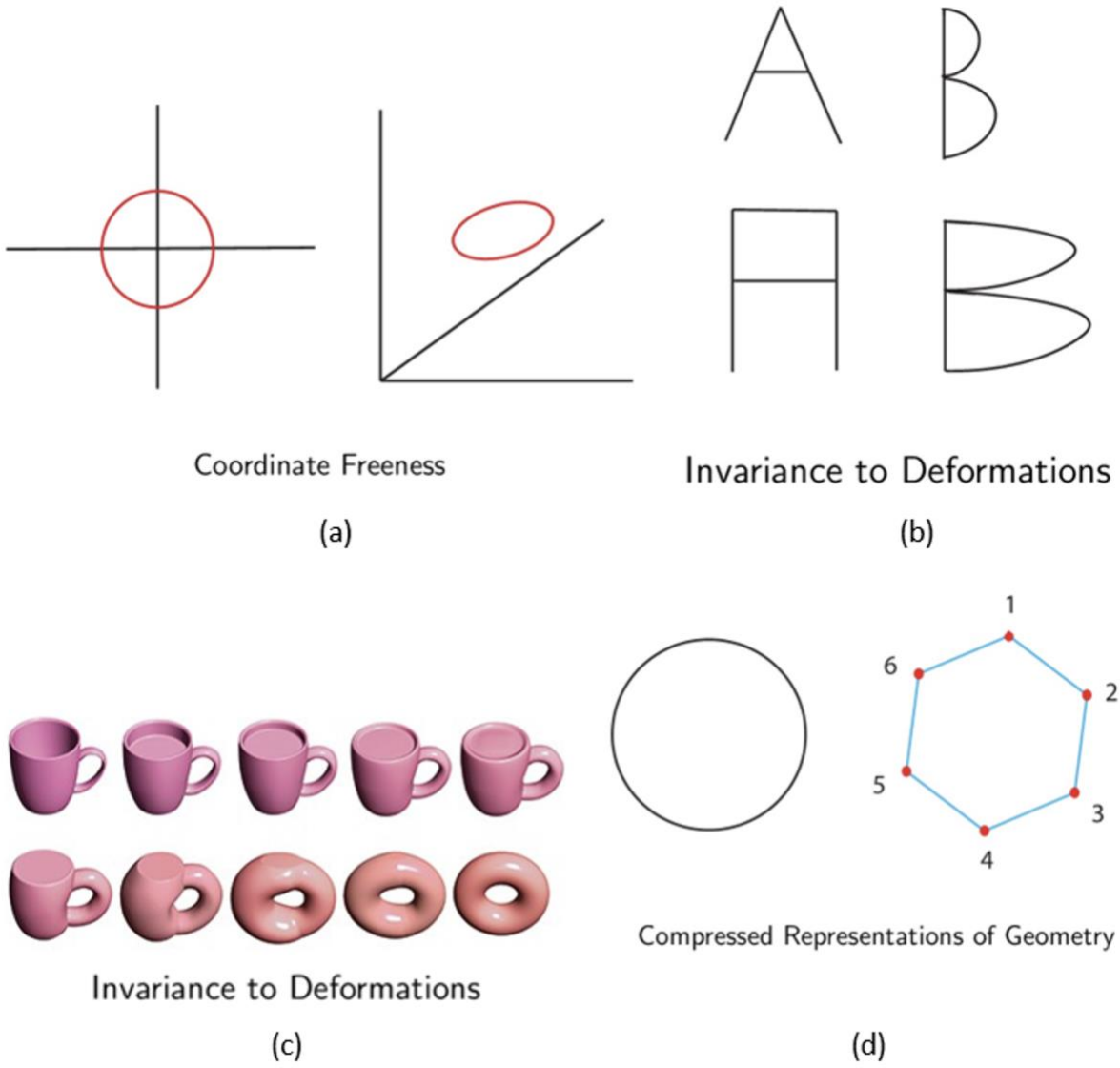

Figure 1. Essential features that make topology ideal for data applications.

In recent years, TDA has been successfully applied to many different scientific domains such as time series analysis [2], image processing [3, 4, 5], text mining [6], and materials science [7]. Figure 2 shows an example of a TDA pipeline applied to point cloud data from a human hand, in which coarse features - the finger and thumbs - are recognized as distinct entities. In the following sections of this report, we will utilize TDA methodology in a novel application to logistics systems analysis, exploiting the ability of TDA to identify coarse features in datasets via a technique called *persistence*.

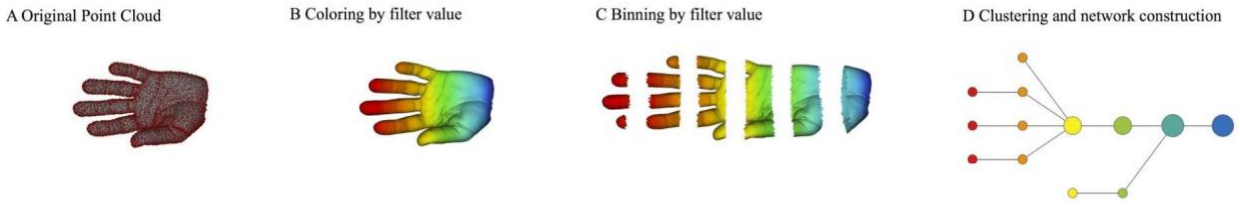

Figure 2. A computational topology analysis of point cloud data from the human hand from [1].

1.1 Homology and persistence

When considering a topological space, we often wish to characterize its intrinsic properties. Specifically, we want information regarding its connected components, loops, and higher dimensional analogues. Algebraic topology formalizes this notion of connectivity information via the *homotopy group*, the set of equivalence classes of loops under an equivalence relation which encodes the "sameness" or essential difference of loops. Unfortunately, the homotopy group of a space is typically difficult to compute. However, a more computable extension called the *homology group* exists. From the homology group of a space we can derive a vector of integers called *Betti numbers*, where the *k*-th Betti number counts the number of equivalence classes of *k*-dimensional surfaces in the space under the extended equivalence relation. Informally speaking, the lower dimensional Betti numbers have natural, visual interpretations: the zeroth Betti number β_0 counts the number of connected components, the first Betti number β_1 counts the number of 1-dimensional holes, or loops, and the second Betti number β_2 counts the number of voids of a shape. Figure 3 shows a few examples of some shapes and their corresponding Betti numbers.

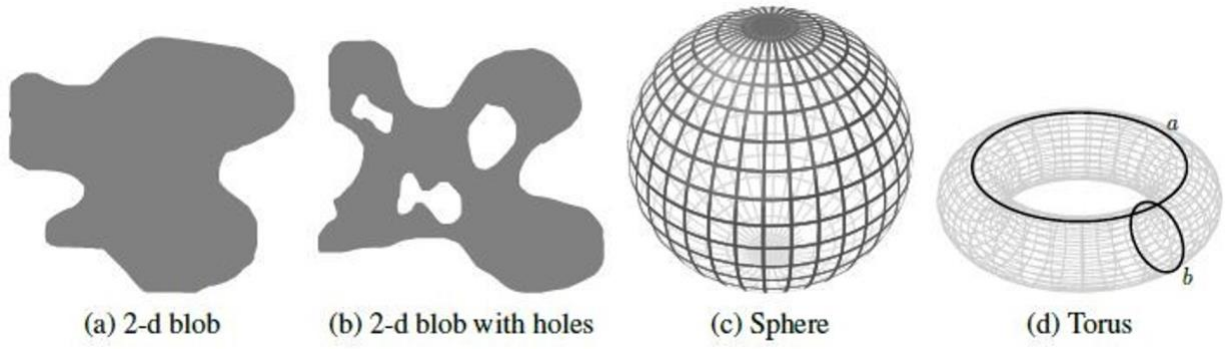

Figure 3. A 2-dimensional "blob" has β0 = 1 because it consists of one component, and βk = 0 for all other k (3a). The 2-dimensional blob with holes in (3b) still has $\beta_0 = 1$, but $\beta_1 = 3$ because there are three holes present. The (hollow) sphere in (3c) has $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$, and $\beta_k = 0$ for all other *k*; the $\beta_2 = 1$ is due to the interior of the sphere. Finally, the torus in (3d) has $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$, and $\beta_k = 0$ for all other *k*; the $\beta_1 = 2$ is due to the two holes indicated with thickened lines (one for the "donut" hole and one that wraps around the thickness of the torus). The $\beta_2 = 1$ is due to the interior of the torus.

1.2 Filtered simplicial complexes using persistent homology

Rather than computing Betti numbers at a single filtration value of the complex, we compute a summary of the Betti numbers over a range of filtration values to examine features in the point cloud which "persist." An example is explained in detail below.

We begin with a point cloud $X = \{(-1, 0), (1, 0), (-1, 2), (1, 2), (0, 3)\}$ (Figure 4). These points form a house-like shape in R^2. We can construct a *Vietoris-Rips* complex $VR(X, t)$ with filtration value t in the following way: take the set X as the vertex set for the complex, and include the k-simplex $\{x_0, x_1, \ldots, x_k\}$ if and only if $d(x_i, x_j) \leq t$ for all $0 \leq i, j \leq k$. We then allow the filtration value t to vary from 0 to a maximum value $t_{max} = 4$ and observe how the complex changes in terms of its Betti numbers. As long as the maximum filtration value is larger than the diameter of X, all edges will eventually be included. We will illustrate this process below, using balls of radius t centered at each vertex point. An edge $[a, b]$ is included when $b \in B_t(a)$ and $a \in B_t(b)$, where $B_t(x)$ is the ball of radius t centered at point x.

First, analyzing β_0, we begin with five connected components at $t = 0$ (each point is its own component). For filtration values $0 \leq t < \sqrt{2}$, none of the balls intersect, and we maintain five connected com- ponents. At filtration value $t = \sqrt{2}$, edges $[(-1, 2), (0, 3)]$ and $[(0, 3), (1, 2)]$ are included in the simplex, reducing the number of connected components to three. This structure persists until $t = 2$, where edges $[(-1, 2), (-1, 0)]$, $[(-1, 0), (1, 0)]$ and $[(1, 0), (1, 2)]$ are included. This reduces the number of connected components to one. Similarly for β_1, there are no 1-dimensional holes until filtration value $t = 2$, at which point the hollow square appears. This hole is filled at $t = \sqrt{8}$, at which point the edges $[(-1, 2), (1, 0)]$ and $[(-1, 0), (1, 2)]$ are added to the complex and the 2-dimensional simplices that make up the square are included.

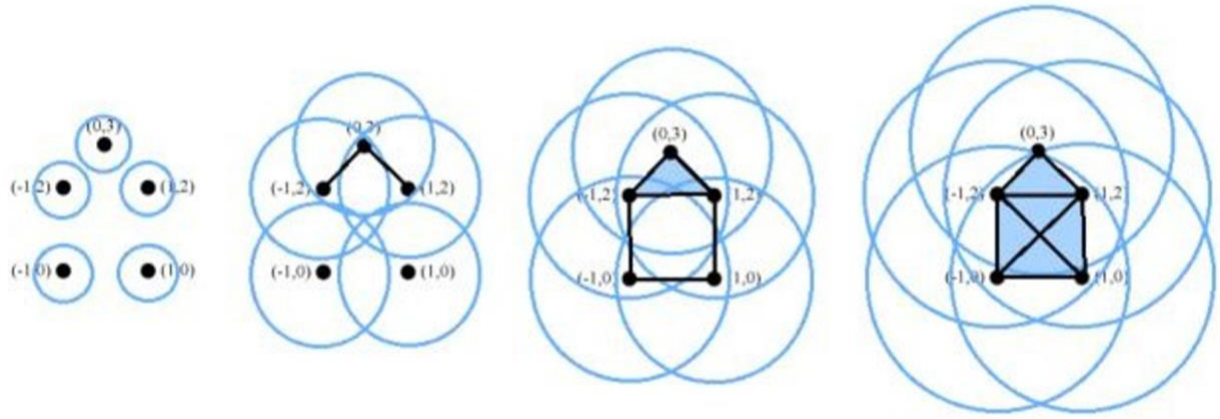

Figure 4. Evolution of $VR(X)$ for increasing values of t. Intervals shown: $0 \leq t < \sqrt{2}$, $\sqrt{2} \leq t < 2$, $2 \leq t < \sqrt{8}$, $\sqrt{8} \leq t$.

The relationship between β_0, β_1 and t can be summarized in *Betti barcodes* (Figure 5, which show the intervals of t over which each connected component (first-dimensional hole) exists. These barcodes provide a concise summary of the set's key geometric features, including those

that persist over large intervals of filtration values and those that appear and "die" over comparatively shorter intervals.

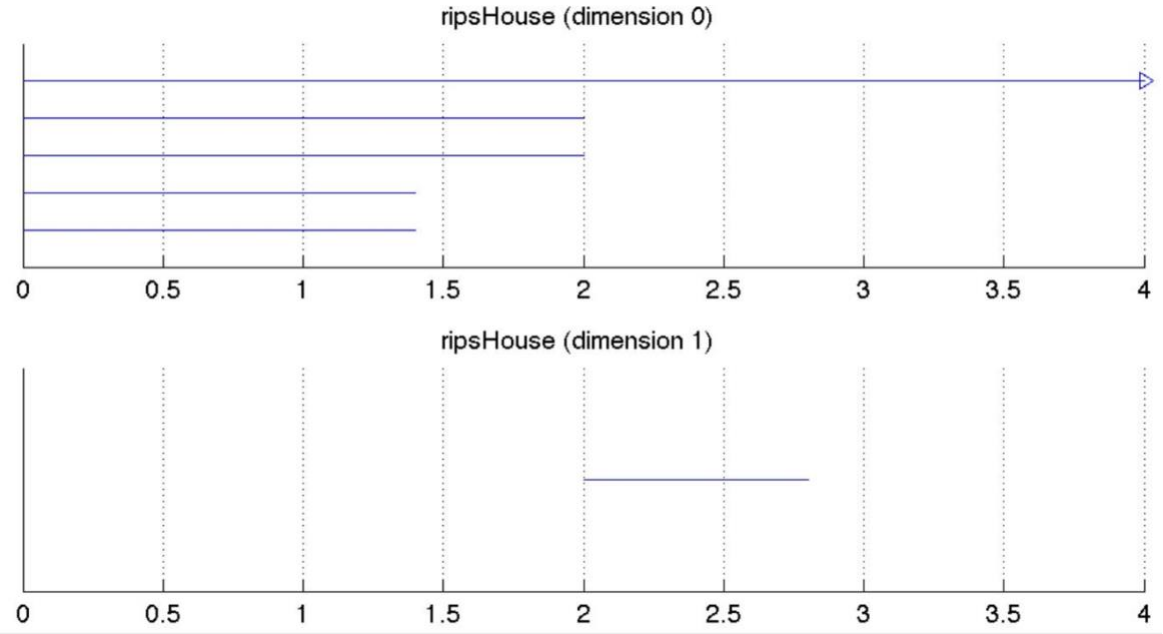

Figure 5. Zeroth and first Betti barcodes for the house point cloud under Vietoris-Rips complex.

2. The Persistent Homology of a Transportation Flow Network

In this section, we introduce a new filtered persistence complex construction associated to a transportation flow network. So as to side-step issues in non-conservation of flows, we find it most helpful to frame our construction in the language of Markov chains by normalizing the rows of the flow matrix, thus resulting in a stochastic matrix together with a steady state vector. We then consider several examples of varying complexity. To the best of our knowledge, no other Markov chain-based filtered complex constructions exist. First, we begin with some intuition behind the complex.

Markov chains are used to model many different things; one advantage they possess is that one can use (for example) first transition analysis to derive coarse algebraic features of a model. The best example of such a feature is the stationary distribution, which roughly speaking provides information about the importance of the various states, but other features include transient versus recurring states, degree sequences, hitting times, and information from first transition analysis. These "features" can aid in making observations and conclusions about the process. The purpose of this chapter is to use topological data analysis to identify coarse features in Markov chains that are not accessible via traditional methods.

2.1 Conceptual example: the predator-prey model

The *Lotka-Volterra model* studies the interactions of a population of X prey particles (animals, agents) and Y predator particles. We assume the following three possible reaction events:

1. Prey reproduction, $X \rightarrow 2X$
2. Prey consumption generates a predator, $X + Y \rightarrow 2Y$
3. Predator death, $Y \rightarrow \phi$

Furthermore, each prey reproduces at rate α, prey and predator encounters occur at rate β, and predators die off at rate γ. Letting X(t) and Y(t) be the predator and prey populations as functions of time, the Lotka-Volterra equations are given by

$$\frac{dX}{dt} = \alpha X(t) - \beta X(t)Y(t)$$
$$\frac{dY}{dt} = \beta X(t)Y(t) - \gamma Y(t)$$

Solutions to this set of equations are characterized by boom and bust cycles within the predator and prey populations, pictured in Figure 6. Intuitively, these capture the phenomenon that when predator populations are high, prey are consumed at a higher rate and experience population decline. Similarly, when prey populations are low because of over-hunting, predators no longer have a robust food source and begin experiencing population decline until prey populations recover. Examining the state space diagram of the Lotka-Volterra equations reveals an orbital structure constrained by the initial state ($X(0)$, $Y(0)$), as shown in Figure 7.

A continuous-time Markov chain (CTMC) version of this model is given in [8]. Let (X, Y), the number of prey and predator particles, be the states of the CTMC and define the following state-dependent transition rates:

- $(X, Y) \rightarrow (X + 1, Y)$ with rate c_1X (prey reproduction)
- $(X, Y) \rightarrow (X - 1, Y + 1)$ with rate c_2XY (predator-prey encounter)
- $(X, Y) \rightarrow (X, Y - 1)$ with rate c_3Y (predator death)

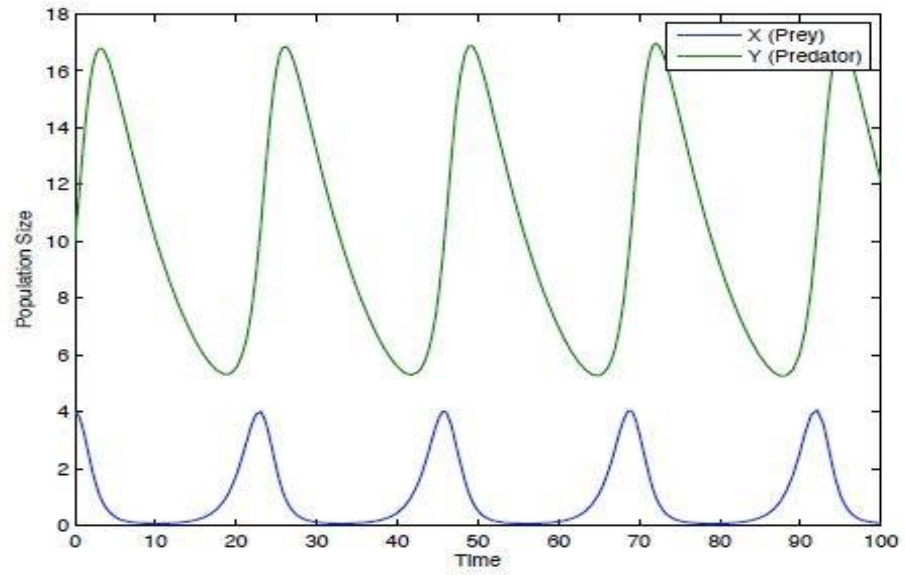

Figure 6. Boom and bust cycles in the predator-prey model.

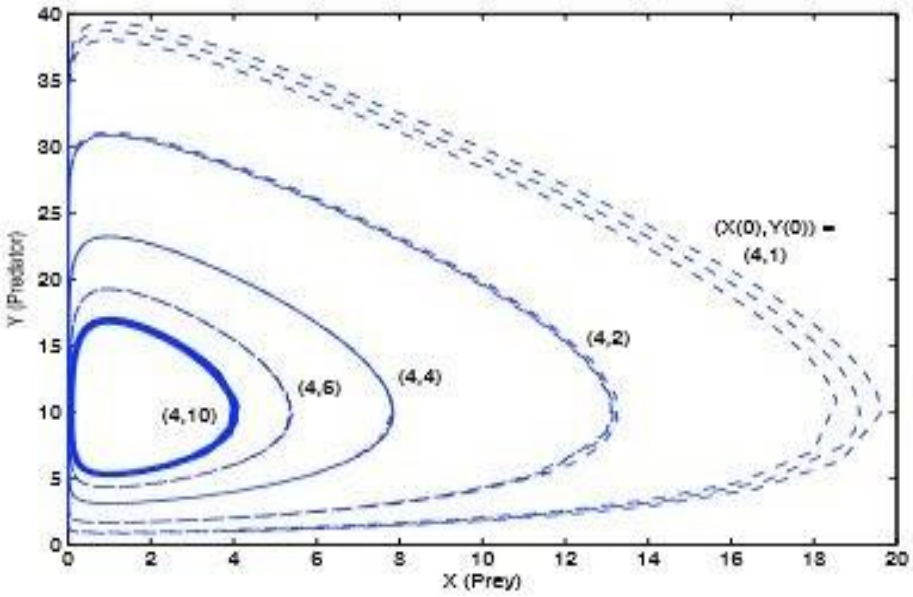

Figure 7. State space diagram for the Lotka-Volterra equations showing orbits for several initial conditions.

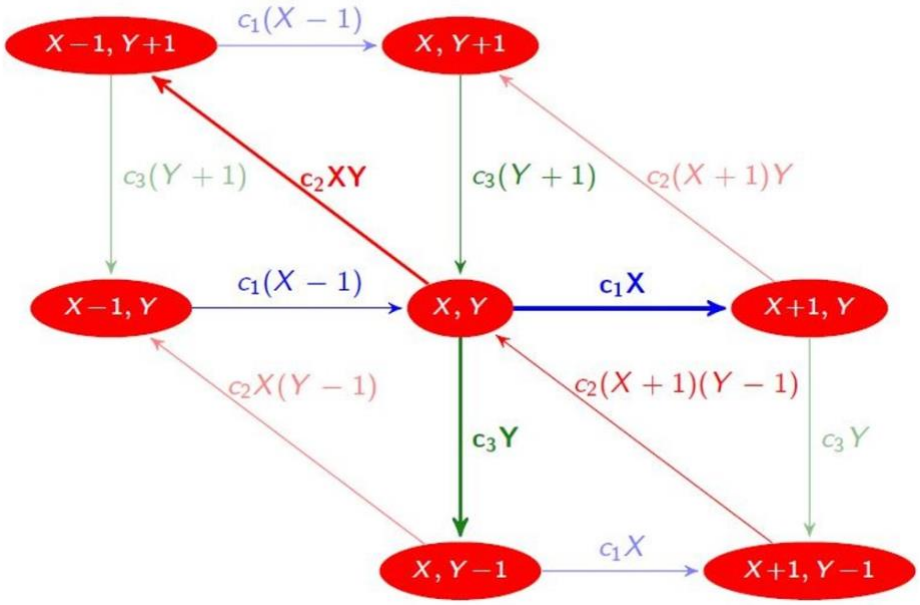

Figure 8. Transition structure of the CTMC model for predator-prey interactions as defined by [21].

These transitions are summarized in Figure 8. Of course there are other ways of capturing the predator-prey relationship in a stochastic model. However, for the purpose of TDA applications, we choose this particular stochastic model to study. We consider the transition matrix *P* derived from the associated jump chain with modifications to ensure that there are no absorbing states (that is, no population extinctions or growth beyond maximum values M_{prey}, M_{pred}). The entries of *P* are defined as follows:

$$P\big((i,j),(i+1,j)\big) = \frac{c_1 i}{c_1 i + c_2 ij + c_3 j}, i \neq M_{prey}$$

$$P\big((i,j),(i-1,j+1)\big) = \frac{c_2 ij}{c_1 i + c_2 ij + c_3 j}, i \neq 1, j \neq M_{pred}$$

$$P\big((i,j),(i,j-1)\big) = \frac{c_3 j}{c_1 i + c_2 ij + c_3 j}, j \neq 1$$

$$P\Big(\big(i,M_{pred}\big),\big(i+1,M_{pred}\big)\Big) = \frac{c_1 i}{c_1 i + c_3 M_{pred}}, i \neq M_{prey}$$

$$P\big((i,M_{pred}),(i,M_{pred}-1)\big) = \frac{c_3 M_{pred}}{c_1 i + c_3 M_{pred}}$$

$$P\Big(\big(M_{prey},j\big),\big(M_{prey}-1,j+1\big)\Big) = \frac{c_2 M_{prey} j}{c_2 M_{prey} j + c_3 j}, j \neq M_{pred}$$

$$P\Big(\big(M_{prey},j\big),\big(M_{prey},j-1\big)\Big) = \frac{c_3 j}{c_2 M_{prey} j + c_3 j}, j \neq 1$$

$$P((M_{prey}, M_{pred}), (M_{prey}, M_{pred} - 1)) = 1$$

$$P\big((1, j), (1, j - 1)\big) = \frac{c_3 j}{c_1 = c_3 j}, j \neq 1$$

$$P\big((1, j), (2, j)\big) = \frac{c_1}{c_1 + c_3 j}$$

$$P\big((i, 1), (i + 1,1)\big) = \frac{c_1 i}{c_1 i + c_2 i}, i \neq M_{prey}$$

$$P\big((i, 1), (i - 1,2)\big) = \frac{c_2 i}{c_1 i + c_2 i}, i \neq 1$$

$$P\left(\big(M_{prey}, 1\big), \big(M_{prey} - 1,2\big)\right) = 1$$

$$P\big((1,1), (2,1)\big) = 1$$

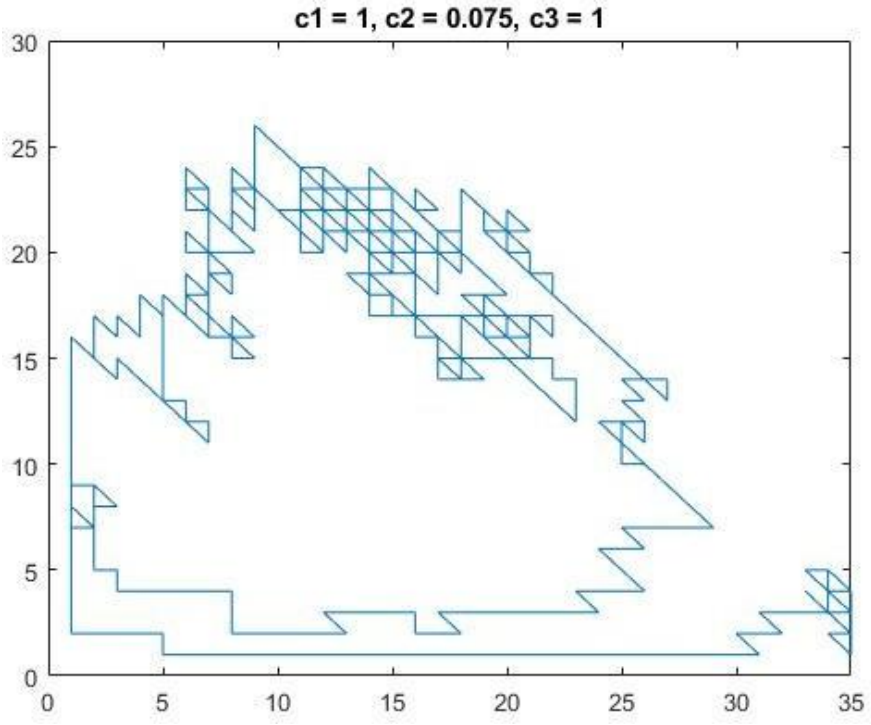

Figure 9. A simulation of 500 steps from the Markov chain defined by Equations (3)-(17).

Figure 9 shows a simulation of 500 steps from the resulting Markov chain starting from a randomly selected state. Parameter choices were $M_{prey} = M_{pred} = 35$, $c_1 = 1$, $c_2 = 0.075$, $c_3 = 1$ and $M_{prey} = M_{pred} = 20$, $c_1 = 1$, $c_2 = 0.2$, $c_3 = 1$. [8] states that there is no single limit cycle, but rather a family of perturbed cycles. However, common sense says that there is really just one cycle, but we lack the machinery to identify it. Our objective in the next section is to construct a complex that captures this cycle.

2.2 A filtered complex for network transportation flows

We now describe in detail the construction of the new filtered complex for modelling flows in a transportation network, as interpreted as movement on a Markov chain. Let P be a stochastic matrix for a Markov chain with n states, with stationary distribution π such that $\pi P = \pi$. Let $[P^m]_{ij}$ denote the ij-th entry of that matrix P^m. We propose the *Carlsson-Sweitzer-Siojo (CSS) function* which induces a matrix Q by

$$Q_{ij} = \max_{m \geq 1}([P^m]_{ij} - \pi_j)$$

Another possibility is to consider a matrix Q induced by the *discrete Green function* [9]

$$Q_{ij} = \sum_{m \geq 1}([P^m]_{ij} - \pi_j)$$

For each of these functions, the entry Q_{ij} quantifies how much node i contributes to node j over the long run. Next, since stationary distributions have the convention that high values of π_i correspond to significant states i, we will reverse the usual convention for persistence of simplices. That is, if two simplices satisfy the inclusion relation $\Delta' \subseteq \Delta$, then the persistence function should satisfy $f(\Delta') \geq f(\Delta)$. We now construct the persistence function for the new complex in an axiomatic fashion as interpreted via flows in a transportation network.

First, the persistence value of a simplex $\{j_0, \ldots, j_k\}$ should be an aggregate of all the contributions to it, from all the n nodes in the chain. Furthermore, nodes with low stationary distribution values should make proportionally small contributions. Thus, our persistence function should be of the form

$$f(\Delta) = \sum_i \pi_i h(Q_{ij_0}, Q_{ij_1}, \ldots, Q_{ij_k})$$

Next, the function h should only be large if *all* its arguments $Q_{ij_0}, Q_{ij_1}, \ldots, Q_{ij_k}$ are large. This is because $f(\Delta)$ should always be thought of as a property of the entire simplex, not merely its components. We therefore take a product of the entries $Q_{ij_0}, Q_{ij_1}, \ldots, Q_{ij_k}$. To summarize, the final form of the persistence function for our complex construction is

$$f(\{j_0, \ldots, j_k\}) = \min_{p \in \Delta^k} \sum_i \pi_i Q_{ij_0}^{p_0} \cdots Q_{ij_k}^{p_k}$$

Where $p \in \Delta^k$ denotes an entry of the probability simplex.

2.2.1 Lotka-Volterra revisited

We now return to the Lotka Volterra model from the previous section. The persistence barcodes obtained from applying the Markov chain complex to the transitions matrix P are given in Figure 10. The longest β_1 barcode indicates a cycle in the Markov chain, which is best interpreted as a cycle of increasing and decreasing predator and prey populations. While the state space diagram shows individual orbits tied to specific initial conditions, the TDA technique reveals the overall cyclic nature of the Markov chain. This coarse feature is not uncovered by

traditional methods. Furthermore, the β_0 barcodes indicate that there are many states (i, j) that are disconnected, i.e., not typically visited by the chain. This experiment not only demonstrates the use of the new filtered complex and TDA techniques applied to a theoretical model, but also the use of higher order Betti numbers to analyze the structure of Markov chains.

3. An Empirical Study

One “real-world” application is to a set of commercial flight departure and arrival data available through BTS [10]. The dataset includes details from commercial flights between October 1987 and April 2008 with nearly 120 million records. Taking just the "Origin" and "Destination" columns from the data set for a single year, we build a Markov chain that uses the frequency of origin-destination pairs to derive transition probabilities for each location pair. That is, the probability of a transition from location x to location y is simply the number of flights from x to y divided by the total number of flights departing from x. Additionally, in order to avoid absorbing states, we delete any locations that are either a destination but never an origin, or an origin but never a destination. We can now apply the filtered complex to the resulting transition probability matrix and observe the resulting persistence barcodes. The barcodes obtained from performing these steps for the 1987 data are shown in Figure 11. There is one β_1 barcode of significant length, and we can visualize the cities and edges that make up a representative cycle for this barcode, as shown in Figure 12.

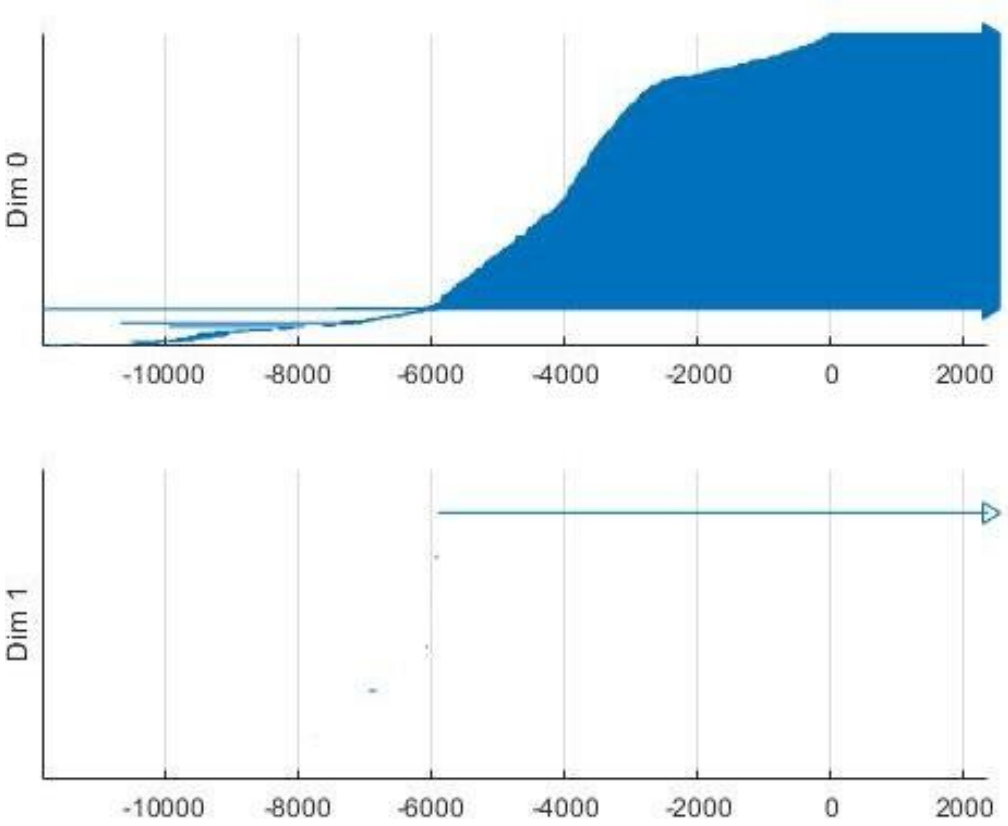

Figure 10. β0 and β1 barcodes for a stochastic model of predator-prey interactions obtained from the Markov chain complex.

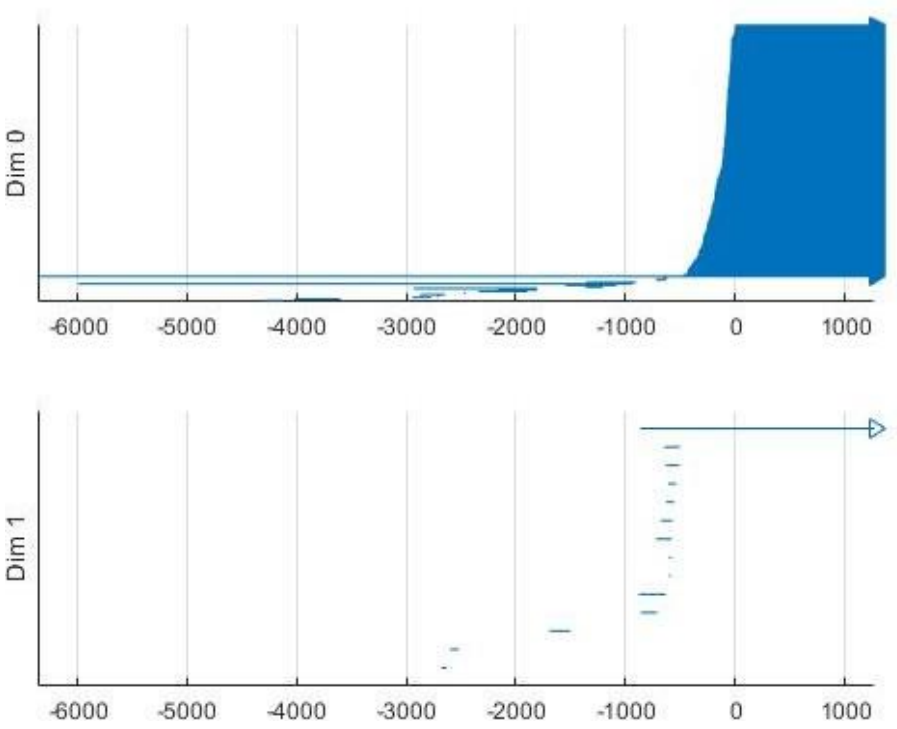

Figure 11. β0 and β1 persistence barcodes for airline data from 1987.

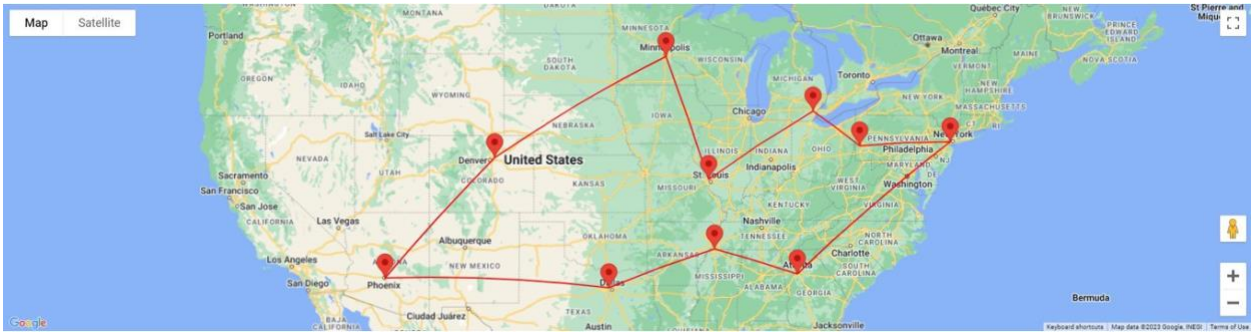

Figure 12. A representative cycle for the longest β1 barcode.

Following the same procedure with the 2008 data, we obtain the barcodes shown in Figure 13, which show three significant β_1 barcodes. Javaplex can identify a (nonunique) representative cycle for each of the three longest β_1 barcodes, which we display in Figure 14. Viewing the subcomplexes in this way reveals three empty 2-simplices composed of the following edges (1-simplices): ABQ-DFW, DFW-PHX, PHX-ABQ; ABQ-DFW, DFW-DEN, DEN-ABQ; DEN-DFW, DFW-PHX, PHX-DEN. Recall that, intuitively, a simplex is added to this Markov chain complex if there are sufficiently many nodes with large enough stationary distribution values that contribute to it over the long run. Furthermore, if $Qij' = 0$ for some j' in the simplex, then the term for node i in the persistence value is zero. Thus, the appearance of an empty 2-simplex implies that in the long run, there are nodes that contribute to both cities in each edge of the simplex, but not to all three. To confirm this, we can examine the matrix Q. Specifically, taking the empty 2 simplex

ABQ-DFW-DEN as an example, we consider the top 10% of values $Q(i, j)$ for $j \in \{$ABQ, DFW, DEN$\}$ and all possible nodes i and look for contributing nodes j that are in common to two but not all three cities in the 2-simplex. The Venn diagrams in Figure 15 and Figure 16 show the results of this process for both the ABQ-DFW-DEN and ABQ-DFW-PHX empty 2-simplices. Additionally, plotting these contributing cities on a map of the United States shows them as belonging to three geographically separated regions (see Figure 17 and Figure 18). The top contributors for an edge are also smaller local metropolitan regions, possibly implying that the two edge cities but not the third in the simplex act as a "hub" for the geographic region. Cycles such as these can be interpreted as potential opportunities for agglomeration, as their components have many individual commonalities in terms of flows.

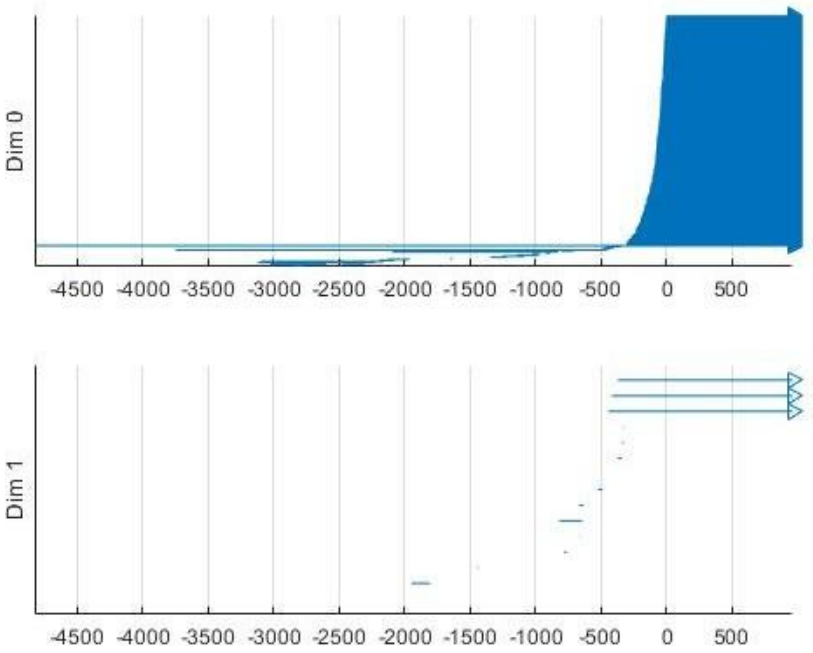

Figure 13. β0 and β1 persistence barcodes for airport flow data from 2008.

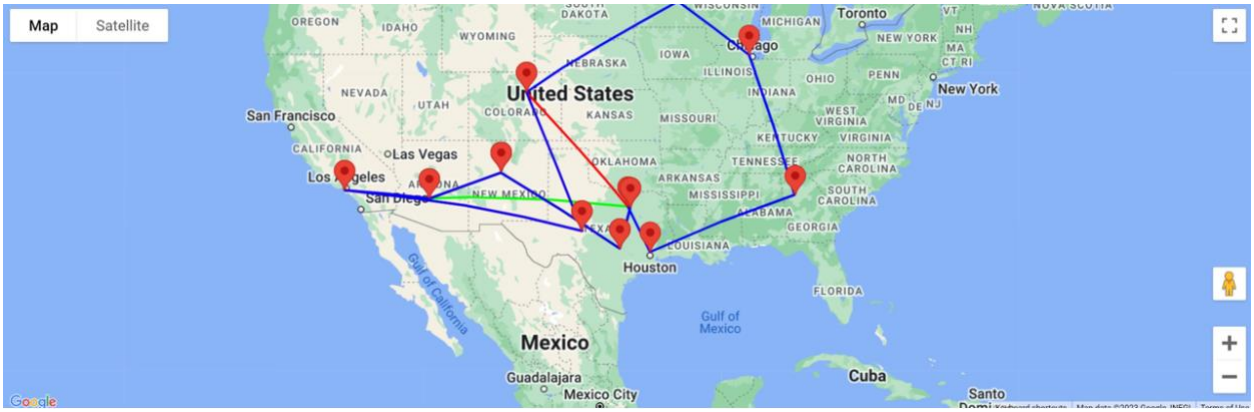

Figure 14. Representative cycles for the three longest β1 barcodes: ABQ-AUS-DAL-DFW-PHX-ABQ, ABQ-AUS-DAL-DFW-DEN-SJT-LAX-PHX-ABQ, and ABQ-AUS-DAL-DFW-IAH-ATL-ORD-MSP-DEN-SJT-LAX-PHX-ABQ.

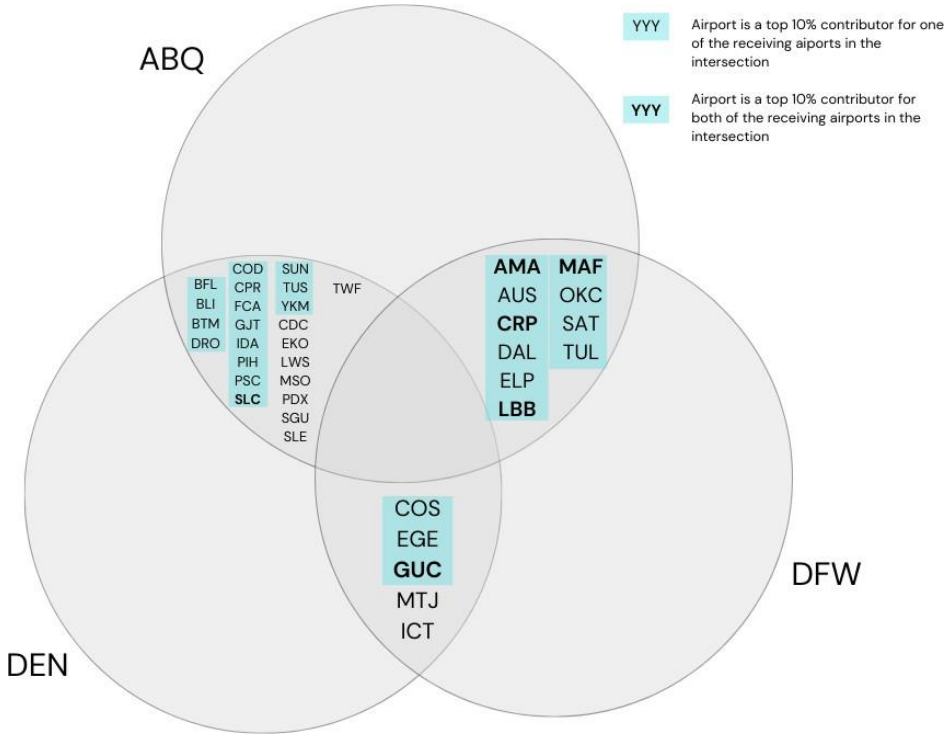

Figure 15. Cities from the top 50 contributors to ABQ, DEN, and DFW were considered.

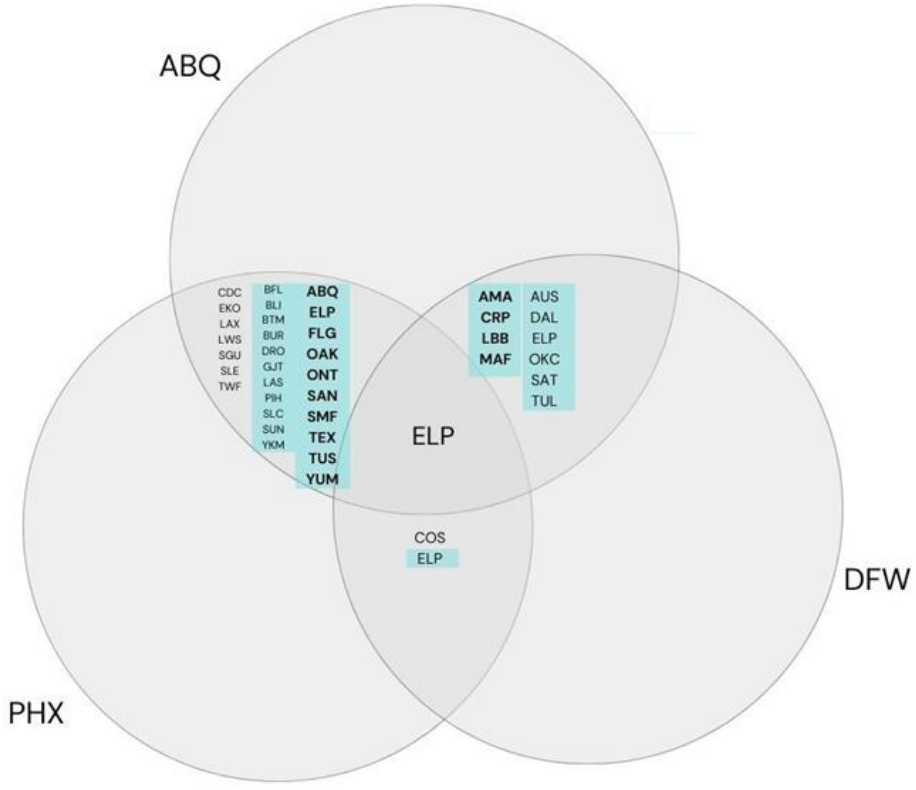

Figure 16. Cities from the top 50 contributors to ABQ, PHX, and DFW were considered.

Figure 17. A Google map showing the highlighted cities from Figure 15 and the three cities that compose the empty 2-simplex (in yellow).

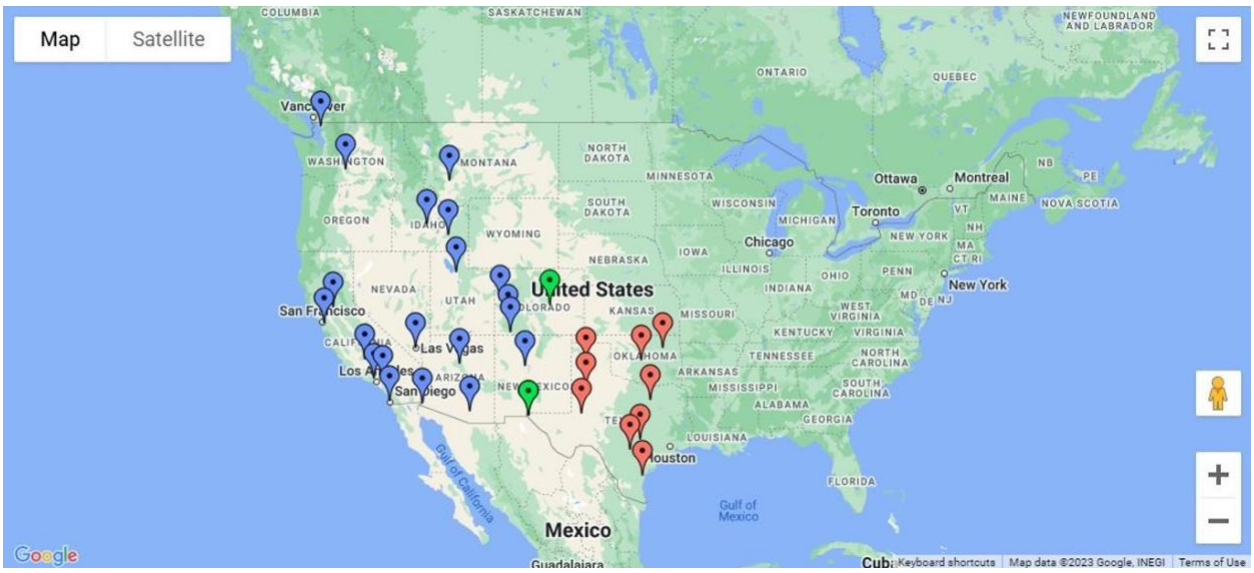

Figure 18. A Google map showing the highlighted cities from Figure 16.

4. Conclusions

This project has attempted to bridge the gap between topological data analysis and logistics systems analysis. We have introduced a novel simplicial complex construction applicable to transportation flow networks that capture similarity structures and cycles between cities. Its use in exploring the underlying structure and coarse features of a stochastic process identifies a new way to determine coarse features in Markov chains such as cycling or clustering. Our results broaden the relevance and suitability of topological data analysis as an interdisciplinary tool for many fields of study, and its future applications in transportation.

References

[1] P. Y. Lum, G. Singh, A. Lehman, T. Ishkanov, M. Alagappan, J. Carlsson, G. Carlsson, and Mikael Vilhelm Vejdemo Johansson. “Extracting insights from the shape of complex data using topology”. English. In: Scientific Reports 3 (Feb. 2013). Issn: 2045-2322. doi:10.1038/srep01236.

[2] Marian Gidea and Yuri Katz. Topological Data Analysis of Financial Time Series: Landscapes of Crashes. Papers 1703.04385. arXiv.org, Mar. 2017. URL: https://ideas.repec.org/p/arx/papers/1703.04385.html.

[3] Gunnar E. Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. “On the Local Behavior of Spaces of Natural Images”. In: International Journal of Computer Vision 76 (2007), pp. 1–12.

[4] Arian Maleki, Morteza Shahram, and Gunnar Carlsson. “A near optimal coder for image geometry with adaptive partitioning”. In: 2008 15th IEEE International Conference on Image Processing. 2008, pp. 1061–1064. doi:10.1109/ICIP.2008.4711941.

[5] Jose A. Perea and Gunnar E. Carlsson. “A Klein-Bottle-Based Dictionary for Texture Representation”. In: International Journal of Computer Vision 107 (2013), pp. 75–97.

[6] Hubert Wagner, Pawel Dlotko, and Marian Mrozek. “Computational Topology in Text Mining”. In: CTIC. 2012.

[7] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G. Escolar, Kaname Matsue, and Yasumasa Nishiura. “Hierarchical structures of amorphous solids characterized by persistent homology”. In: Proceedings of the National Academy of Sciences 113 (2016), pp. 7035–7040.

[8] Gonzalo Mateos. Predator-Prey Population Dynamics. Slide presentation. 2018.

[9] Robert B. Ellis. “Discrete Green’s functions for products of regular graphs”. In: arXiv: Combinatorics (2003).

[10] https://www.census.gov/programs-surveys/cfs.html

Data Summary

Products of Research

The data that were collected were obtained from the Bureau of Transportation Statistics and are free to use by the public.

Data Format and Content

The Commodity Flow Survey (CFS) is the primary source of national and state-level data on domestic freight shipments by American establishments. Data are provided on the types of commodities being moved, along with their origins and destinations, values, weights, modes of transportation, distance shipped, and ton-miles of commodities shipped. The CFS is a component of the Census Bureau's economic census and is conducted every five years.

Data Access and Sharing

The general public can access the data by visiting https://www.bts.gov/product/commodity-flow-survey.

Reuse and Redistribution

There are no restrictions on reuse and redistribution of the data used in this report.