Title
Topological tools for understanding complex systems

Author
Feng, Michelle H

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Topological tools for understanding complex systems

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Mathematics

by

Michelle Hsiao-Chin Feng

2020

ABSTRACT OF THE DISSERTATION

Topological tools for understanding complex systems

by

Michelle Hsiao-Chin Feng
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2020
Professor Mason A. Porter, Chair

The behavior of complex systems is often influenced by their structure. In mathematics, the field of algebraic topology has been especially useful for characterizing mathematical structures. Topological data analysis (TDA) is a growing field in which methods from algebraic topology are applied to studying the structure of data. TDA has been used in a variety of applications, including biological data, granular materials, and demography. Social interactions are heavily informed by space and have complex structure due to patterns in the way humans arrange themselves geographically. Consequently, social applications can benefit from the application of TDA.

In this dissertation, I develop topological methods for studying spatial networks and apply them to a wide variety of data sets. In particular, I study methods for building topological spaces (specifically, simplicial complexes) based on data. I present two novel simplicial-complex constructions, the adjacency complex and the level-set complex, for spatial data. I apply both constructions to random networks, cities, voting, and scientific images, gaining insights into the structure of these systems. I also propose a novel simplicial complex construction for studying patterns of neighborhood formation based on combining demographic

and spatial data. I present case studies in neighborhood segregation for two U.S. cities.

In addition to my topological research, I discuss two projects in the study of social systems using methods from network analysis. I present an extension to multilayer networks of the Hegselmann–Krause model for opinion dynamics and discuss preliminary findings on its convergence properties. I also present a framework for estimating homelessness underreporting in California Local Education agencies (LEAs).

The dissertation of Michelle Hsiao-Chin Feng is approved.

Michael Hill

Deanna Hunter

Stanley Osher

Mason A. Porter, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

List of Symbols

N	number of nodes
p	probability of an edge in Erdős-Rényi $G(N, p)$ random graph
K	mean degree in a Watts–Strogatz model
β	rewiring probability in a Watts–Strogatz model
H	homology group
k	dimension of a simplex
G	graph
V	set of nodes
E	edge set
S	simplicial complex
σ	simplex
F	field
C	group of k-chains
δ	boundary map
$\{X_i\}$	filtered simplicial complex
m	maximum dimension of a simplicial complex
b	birth time
d	death time
VR	Vietoris–Rips complex
$\vec{x}$	opinion vector
A	adjacency matrix
c	confidence boound
M	multilayer network
$\mathbb{L}$	layer set

$\mathcal{A}$	adjacency tensor
X	variable representing true homelessness
$\tilde{X}$	variable representing observed homelessness
M_X	variable representing misclassification

ACKNOWLEDGMENTS

This journey would not have been possible without the kind support of my family, collaborators and mentors, and friends.

To my supervisor, Mason Porter, you have always been an amazing and steadfast mentor to me. I am extremely grateful for the vast amounts of time and effort that you put into helping me through every step of this process. You have been incredibly supportive of my ideas and suggestions, patient and understanding when I encountered difficulties, and above all, eternally kind and caring. I cannot thank you enough for all of the help you have given me over the past few years.

Additionally, I want to thank all of my co-authors and collaborators for their contributions. Chapter 4 includes material from submitted papers [FP20a, FP20c], co-authored with Mason, who advised the projects and contributed substantially to the text.

Chapter 5 describes a project that involved several collaborators, begun at the 2018 Voting Rights Data Institute. Eion Blanchard provided expertise on GIS mapping tools and conceptualized and generated many of the visualizations. Austin Eide and Patrick Girardet contributed to model formulation and code. Moon Duchin provided ideas, helpful comments, and advised the project. I also thank Emilia Alvarez, Ruth S. Buck, Daryl DeFord, Max Hully, Everett Meike, Zach Schutzman, and Justin Solomon for helpful comments and conversations.

Chapter 6 contains a current project in collaboration with Heather Z. Brooks, Yacoub H. Kureh, and the PI Mason A. Porter. Heather Z. Brooks and Yacoub H. Kureh have contributed substantially to model formulation, code writing, running simulations, and analysis of simulation results. Chapter 6 also contains a current project in collaboration with Yacoub H. Kureh, Victor Leung, and Alexis Piazza. Victor Leung and Alexis Piazza contributed a wealth of legal knowledge and expertise on homelessness in Los Angeles schools. Yacoub H. Kureh contributed to data exploration efforts and model formulation.

Next, I thank my thesis committee members, Michael Hill, Deanna Needell, and Stanley Osher for their valuable comments, insights, and conversations on my thesis work. I also thank Marc Barthelemy, Abigail Hickock, Nina Otter, Giovanni Petri, and Bernadette Stolz for providing helpful feedback on projects and papers.

I am grateful for funding support from the Girsky Student Award and the NSF Division of Mathematical Sciences through award DMS-1922952.

To my friends and family, thank you for their unwavering support. In particular, I wish to acknowledge Katina Vradelis, whose steadfast friendship helped me through the most difficult year of my graduate career. Without your support, I could not have continued my studies uninterrupted. Also, thanks for the grammar advice and proofreading during this arduous thesis writing. Special thanks to my parents Daan and Chin Feng, for always prioritizing my educational development, and to all of my Los Angeles friends for late night dance practices and long conversations that helped keep me sane throughout this long process.

Finally, I cannot begin to express my appreciation for my loving and encouraging partner, Joshua Gensler. You were my first mathematical friend, many years ago, and you have aided me in every step of my mathematical career. From midnight conversations about mathematical philosophy to helping me study for my quals to in-depth code reviews, you have contributed a frankly incalculable amount of time and energy to helping me succeed. Your unwavering support for me, both scientifically and personally, cannot be understated, and I am so, so lucky to have someone who is truly my partner in every sense of the word.

VITA

2014 B.S. (Mathematics) with honors, University of Chicago.

PUBLICATIONS AND PREPRINTS

Michelle Feng and Mason A. Porter. "Persistent Homology of Geospatial Data: A Case Study With Voting." *arXiv:1902.05911*, 2019. (accepted to *SIAM Review*)

Michelle Feng and Mason A. Porter. "Quantifying 'Political Islands' with Persistent Homology." *SIAM News*, January–February 2020.

Michelle Feng and Mason A. Porter. "Spatial Applications of Topological Data Analysis: Cities, Snowflakes, Random Structures, and Spiders Spinning Under the Influence." *arXiv:2001.01872*, 2020. (preprint)

Michelle Feng, Abigail Hickok, Yacoub H. Kureh, Mason A. Porter, Chad M. Topaz. "Connecting the Dots: Discovering the 'Shape' of Data." *doi:10.31235/osf.io/7qd4t*, 2020. (preprint)

CHAPTER 1

Introduction

Studying structural properties can help us gain valuable insights into data, especially "big" and/or high-dimensional data. A variety of mathematical [New18], statistical [Kol09], and computational [CCK17] approaches to studying the structure of data have emerged as the field of data science continues to rapidly expand. In this dissertation, I discuss a variety of approaches that borrow ideas from algebraic topology, which is concerned with the study of topological shape via algebraic methods. Using these approaches, which extend the state-of-the-art, I study a variety of social and spatial applications.

In the remainder of this chapter, I provide brief intuitive introductions to some of the major mathematical concepts that will appear in my work. In Chapter 2, I discuss the empirical data sets that I study. In Chapter 3, I review in more detail the existing mathematical tools that form a basis for my original work. In Chapters 4–6, I discuss a variety of applications for which I helped develop novel topological methodologies. Specifically, Chapter 4 contains work done with Mason A. Porter in [FP20a] and [FP20c]. Chapter 5 contains work done in collaboration with the Voting Rights Data Institute. Chapter 6 contains a collaboration with Heather Z. Brooks (UCLA), Yacoub H. Kureh (UCLA), and Mason A. Porter on adaptations of bounded-confidence models. Chapter 5 also contains a project on homelessness in California schools in collaboration with Yacoub H. Kureh and the ACLU. Finally, in Chapter 7, I discuss conclusions and plans for continuing and expanding the aforementioned projects to future work.

1.1 The "shape" of social data

Recent advances in storage and computing technology have allowed people to collect and analyze massive amounts of new social data. Automation has allowed larger volumes of data to be processed than ever before. As residents of a three-dimensional (3D) world, individuals have a natural embedding into space. Interactions between individuals are influenced by a wide variety of factors, and because individuals do not interact randomly, the structure of social behaviors is often influenced by their embedding in space [Bar18]. For example, human migration is heavily informed by the existence of geographical landmarks and barriers [BBG18].

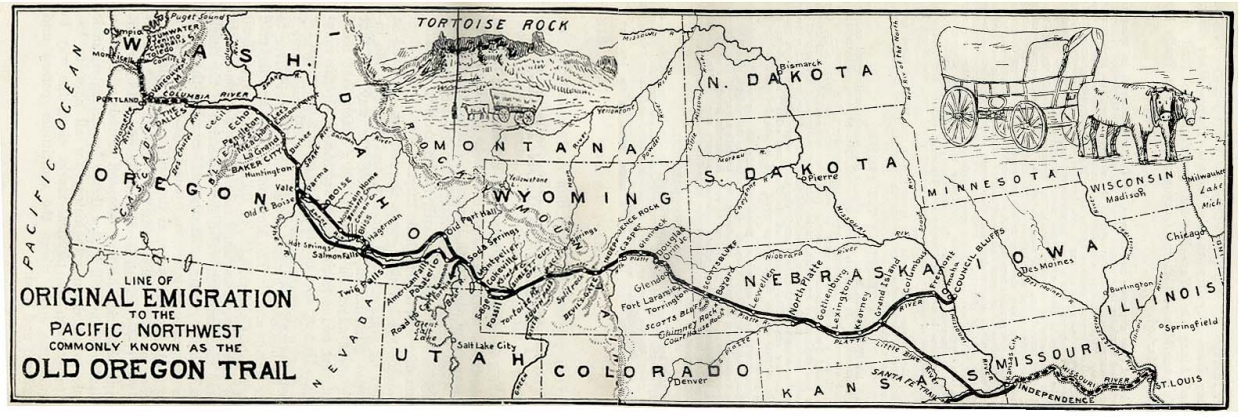


Figure 1.1: Geographical barriers affects human mobility. This image [from [Mee07], now in the public domain] shows the path of the famous Oregon trail. Note how the trail routes around certain mountains.

Additionally, social interactions can have structural properties that are not tied to three-dimensional space [WF94,Mar09]. For example, socioeconomic class has an enormous impact on how individuals interact with each other in the United States (including, but not limited to, the way that people form friendships [Mal17], marry [Jal03], or choose where they live [IW06]). Considering abstract demographic spaces such as socioeconomic class makes it possible to probe the structure of various social behaviors. One can also try to understand

other demographic spaces, like voting preferences or ideologies, using mathematical tools.

Achieving a thorough understanding social data (and social dynamics) requires achieving an understanding of structural properties in a variety of spaces. In this thesis, I focus on data sets that have social components (e.g., cities and social-media networks), though I have also studied some applications in the biological and physical sciences. In particular, many of my projects are geared towards the study of social structures and trying to identify and interpret these structures.

1.2 Topological methods for understanding global structure

While structure can be studied in many contexts, I focus mainly on the study of global structure in the topological sense [Hat02]. This is to be distinguished from global structure in a networks sense, which generally means very large-scale structure [New18]. Methods for understanding the global structure of a topological space rely intrinsically on information about the entire space. To illustrate this concept, consider a sphere. If I sample a neighborhood of any point on a sphere, I obtain a surface with the same properties as a plane. If I take a collection of a sphere's neighborhoods (which each resemble a plane) and stitch them together, I am able to obtain a lot of information about the sphere, but I am unable to describe the void in the center of the sphere. For example, a stereographic projection of a sphere covers the sphere's entire surface, but it fails to capture the void. To fully understand the structure of a sphere, I must consider the entire sphere at once. In this sense, the sphere has an inherent global structure that is not linked to any embedding of the sphere into a larger space.

This topological idea of global structure can be especially useful in instances where one does not want concepts of structure to rely on an embedding. For example, when considering social-science applications, it can be difficult to determine how to weight different factors relative to each other. If one embeds demographic data into Euclidean space, there is au-

tomatically a choice of embedding. Therefore, if one attempts to, for example, compute distances between data points in a demographic data set, those distances rely on the particular choice of embedding into Euclidean space (or another space with a well-defined metric). Because various demographic factors can have heterogeneous importance in individual experiences and decisions [PK93, Rot17, Sei98, RM10, KPM12, PKC10, Man18], one may wish to avoid characterizing individual behaviors based on a one-size-fits-all embedding. Topological methods can offer one way of answering questions about social structures that do not rely on these types of arbitrary choices, because topological properties are invariant to embedding.

One of the core tools of algebraic topology is the concept of *homology* [Hat02]. I leave a rigorous definition of homology to Section 3.3, but it's helpful to now give an intuition for homology. Generally, homology measures the "voids" in a topological space. For example, the void in the center of a sphere (or the one enclosed by a torus) is a 2D void. Understanding the homology groups (these are algebraic groups generated by the various voids in a space) gives a great deal of information about the space. Homology groups provide coarse information related to the homotopy class of a space. They have thus proven very useful in algebraic topology for helping to characterize spaces up to homotopy [Hat02]. One can extend this idea to using homological tools to provide a topological summary of a given data set [ZC04]. This topological summary can then be used to characterize the structure of that data set.

1.3 Networks

Many social applications are concerned with the interactions between individuals. This makes networks a very natural setting for studying social phenomena, as networks are used to study connections and relationships. Many social structures can be encoded readily as networks [WF94, New18], where nodes are individuals and edges represent some sort of relationship between individuals. In spatial applications, one can also encode spatial relationships as edges of a network. The level of abstraction that networks provide can be

leveraged on a wide variety of data sets, and because networks (as traditionally studied) are discrete mathematical objects, computations that involve networks are often tractable.

Network analysis has been applied broadly to numerous social systems [WF94], including migration networks [LBK89], prestige hierarchies in academia [CAL15], and analysis of social media data [GW16]. Biological applications like fungal networks [LFP17, HOG12], leaf venation networks [RLD15, RK16], and networks of neural activity [KBC14, FCP09, GGB16] have also been studied using network analysis. Applications to physics include granular materials [BOP15, KGK13, PPD18] and flows [RLK19]. Additionally, researchers have examined the behavior of dynamical processes like epidemics and other spreading processes [IPB19, TKH15, PCV15], opinion dynamics [BP20, MVP18], and coupled oscillators [ADK08, RPJ16] on networks.

1.4 Spatial systems

Many complex systems have a natural embedding in a low-dimensional space or are otherwise influenced by space, and it is often insightful to study such spatial complex systems using the formalism of networks [Bar18, New18]. In a spatial network, the location of nodes and edges in space can heavily inform both the structure of the network and the behavior of dynamical processes on it. Obtaining a meaningful understanding of power grids [SRC08, KOA18, AAN04], granular systems [PPD18], rabbit warrens [LCP14], and many other systems is impossible without considering the physical relationships between nodes in a network. For example, to examine traffic patterns on a transportation network in a meaningful way, it is important to include information about the physical distances between points and about the locations and directions of paths between heavily-trafficked areas [Bat17].

There are a variety of existing perspectives for studying spatial networks [Bar18, Bar11]. Many of them hail from quantitative geography [HC69, Pum20]. In the 1970s, geographers were already studying the role of space in the formation of networks and in the activi-

ties of individuals and goods over geographical networks. As data have become richer and more readily available, it has become possible to take increasingly intricate computational approaches to the study of spatial networks, and a variety of complex-systems approaches have contributed greatly to the literature on spatial networks [Bar18]. Researchers have also proposed various random models for spatial networks, and studying them yields baseline examples to compare to empirical networks [LP19, NBP19, SLC16, EEB11]. There have also been investigations of the effects of certain spatial network properties on the behaviors of several well-known dynamical processes, including the Ising model [BK87], coupled oscillators [ADK08], and random walks [YWB19].

CHAPTER 2

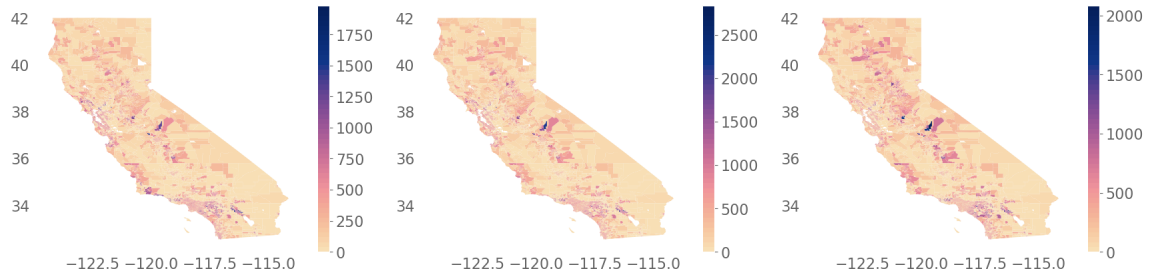
Data

The following sections describe various data sets that I study in this thesis. All of these are freely available online.

2.1 Voting data

For my work on a case study on voting [FP20a], discussed in Section 4.3, I use data from the *Los Angeles Times* California 2016 Election Precinct Maps [SFK16]. Compiled by the *LA Times* Data Visualization Team after the November 2016 elections, this data set has precinct-level results for all of California for every statewide race. Specifically, this encompasses results for the presidential race, California's senatorial race, and 17 statewide propositions. The data covers all of California's 24626 precincts (which are organized into 58 counties); for each one, it includes the number of votes for each choice in each race, along with an associated SHAPEFILE and other metadata.

This data can be used to generate a variety of precinct maps, illustrating voting preferences of various kinds. In Figure 2.1, I show choropleths that I generated using the results of various elections across the entirety of California.



(a) (b) (c)

Figure 2.1: California precincts colored by the numbers of votes for (a) Proposition 51, (b) Proposition 52, and (c) Proposition 53 in 2016. The horizontal axis is longitude, and the vertical axis is latitude.

2.2 Census and American Community Survey data

In my work on residential segregation (in collaboration with participants in the 2018 Voting Rights Data Institute; see Chapter 5) and in work on Los Angeles School District (LASD) homelessness (in collaboration with Yacoub H. Kureh, Victor Leung, and Alexis Piazza; see Section 6.2), I use data from census.gov. This data is drawn from the American Community Survey (ACS) [Bur16b,Bur16a], which occurs on a yearly basis and includes information on housing, education, employment, income, demographics, and much more. In particular, my segregation project uses information on race, ethnicity, and income at the census tract level (taken directly from a yearly ACS table). My homelessness project examines demographic and economic data at the school district level.

In Figure 2.2, I show various city choropleths that illustrate economic data from the ACS.

2.3 City street data

In a case study on city street networks [FP20c] (see Section 4.4), I use OSMnX [Boe17] (software written by Boeing et al.) to generate images of city street networks using lati-

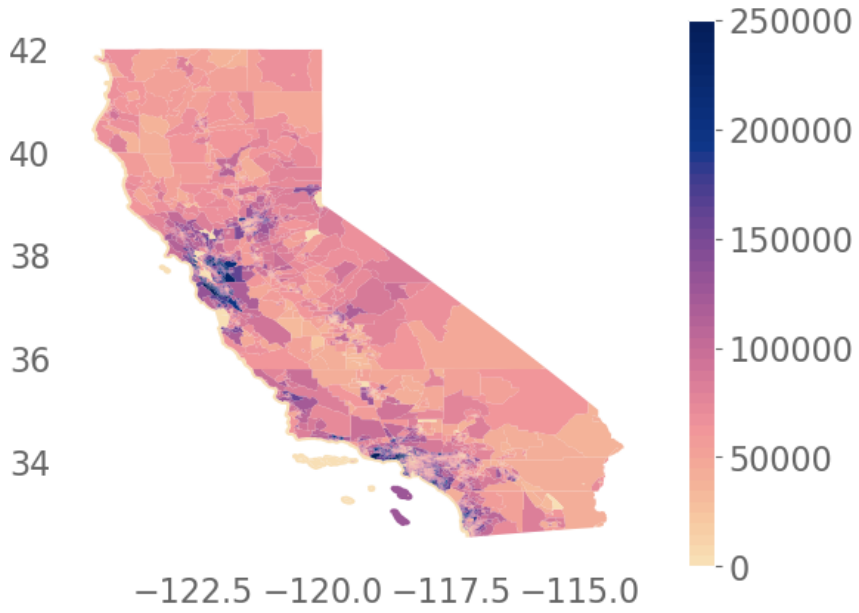


Figure 2.2: Census tracts of California, colored by median income in dollars.

tude/longitude coordinates. I obtained coordinates from SHAPEFILEs provided by ArcGIS [Son17, Bel15] and from a city coordinate data set downloaded from SimpleMaps [Sim19].

Given a set of coordinates, OSMnX generates a city street network for a one-mile square block centered at that coordinate in NETWORKX [HSS08]. OSMnX also includes visualization capabilities that I used to save images of these street networks. These images are the ones that I analyzed in my investigation. In Figure 2.3, I show city street networks for several examples.

2.4 Scientific images

Networks constructed from real-world experiments in physics, chemistry, and biology are often informed by space and structure [Bar18]. In an application with Mason A. Porter [FP20c], I examined two different types of networks: spiderwebs (dicussed in Section 4.5.1) and snowflakes (discussed in Section 4.5.2). Both of these types of networks have planar

(a) (b) (c)

Figure 2.3: City street networks of (a) Los Angeles, (b) Barcelona, and (c) Madrid.

embeddings into $\mathbb{R}^2$, and I use images of these networks to examine those embeddings.

All of the images discussed in this section are grayscale images. To apply my methods (described in Section 4.1.2) to images, they need to be converted to black and white images. As a preprocessing step, I threshold the grayscale images and convert them to black and white images using GIMP [Tea20].

2.4.1 Spiderwebs

In 1948, Peter Witt began research on the effects of drugs on spiders to test whether garden spiders would shift their web-building hours if drugs were administered. Witt found that drugs affect the size and shape of the webs that are produced by spiders. He also found that higher doses of most drugs (100 μg per spider, as opposed to 10 μg per spider) tend to lead to larger changes in the shapes of webs, including yielding more irregular webs. Witt eventually published more than 100 papers and several books on the behavior of spiders and on their spider webs. For more information on his experiments with psychotropic substances and spiders, see his 1971 review article [Wit71].

In a 1995 technical briefing [NCR95], NASA (which was inspired by Witt's research) proposed that spiders who were administered more toxic substances produce webs that are

more deformed than the webs of spiders that were administered less toxic substances. Additionally, using techniques from statistical crystallography, they concluded that spiders fail to complete more sides of their webs when they are under the influence of more toxic substances.

In a case study of PH in spiderwebs [FP20c], I use five images from the NASA technical briefing [NCR95] and two images from Witt [Wit71] of various webs that were spun by spiders under the influence of a variety of psychotropic substances. I show these images in Figure 2.4

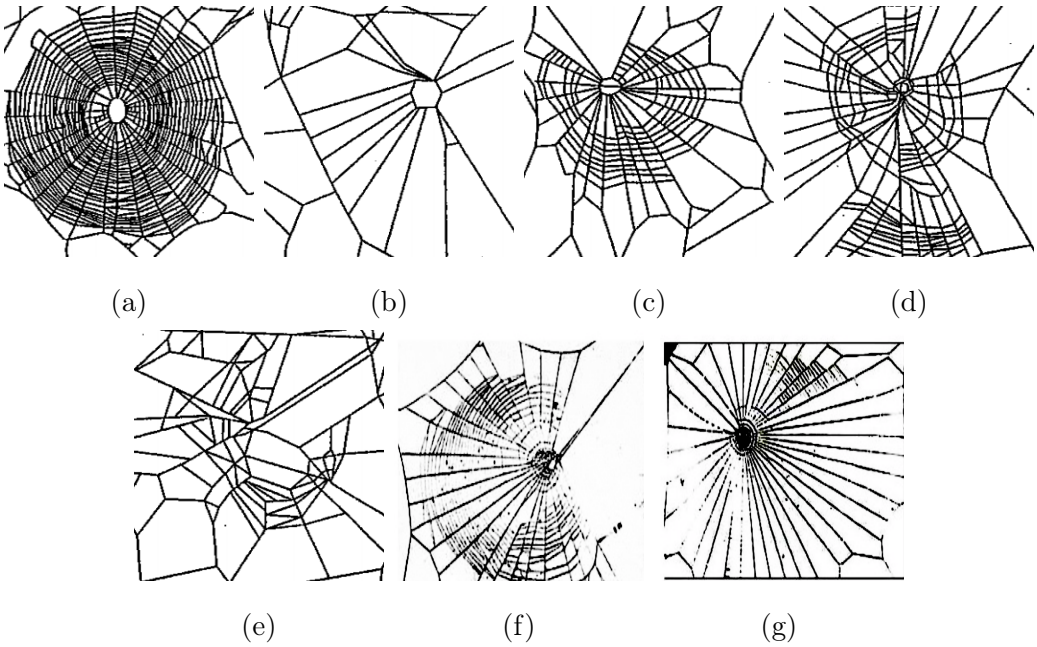


Figure 2.4: Webs spun by (a) drug-Free spiders, compared with webs spun by spiders that were under the influence of (b) chloral hydrate (sleeping pills), (c) marijuana, (d) speed, (e) caffeine, (f) peyote, and (g) LSD. [The images for panels (a)–(e) are from [NCR95], and the images for panels (f) and (g) are from [Wit71].]

2.4.2 Snowflakes

Kenneth G. Libbrecht of Caltech has studied crystal growth in snowflakes for many years, and in the process he has photographed a massive number of snowflakes. Some of his images

were put on a series of U.S. Postal Service stamps in 2006, and he has written several books on snowflake crystallography [Lib07, Lib16, Lib19]. In my work [FP20c], I take 12 snowflake images from his 2019 book [Lib19] and use them to analyze the spatial properties of different types of snowflake crystals.

In Figure 2.5, I show the images of all twelve snowflakes that I examine in Section 4.5.2.

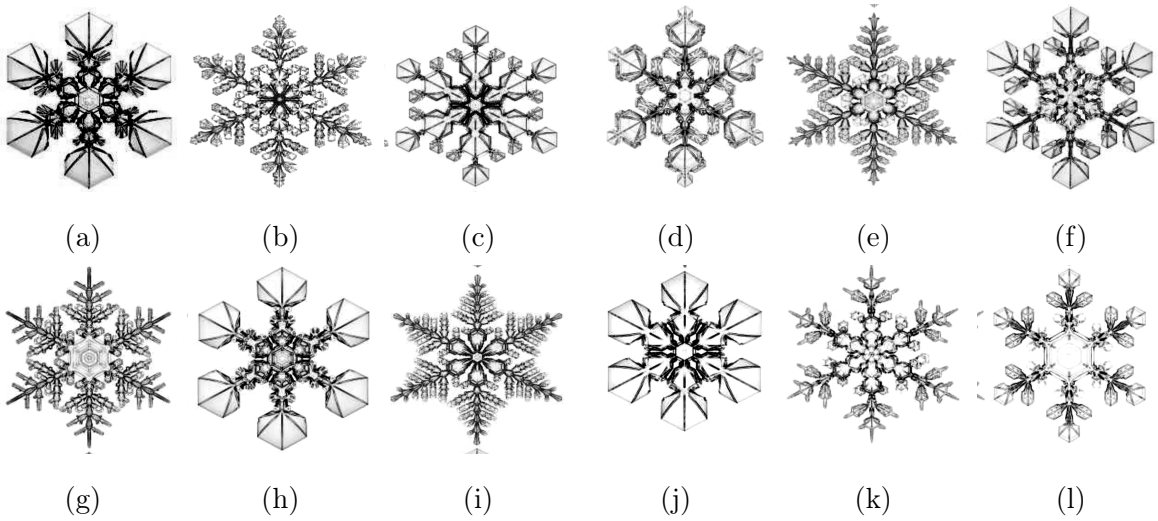


Figure 2.5: The full set of twelve snowflake images that I examine in Section 4.5.2. I label these snowflakes using the panel labels from this figure. I show Snowflake A in panel (a), Snowflake B in panel (b), and so on. [These images are from [Lib19].]

CHAPTER 3

Tools, techniques, and models

In this section, I give rigorous mathematical definitions for a variety of tools that I use in my research. In Sections 3.2–3.4, I discuss tools from algebraic topology. In Section 3.5, I describe a bounded-confidence model for studying continuous opinion dynamics on a network. In Section 3.6, I define two generalizations of networks that have been used in the literature to study higher-order network interactions.

3.1 Networks

I begin by defining some basic network terms. See [New18] for an introductory textbook about networks

Definition 1. *A **graph** (sometimes called undirected) is an ordered pair $G = (V, E)$, where V is a set whose elements are called **vertices** (also called **nodes**) and E is a subset of $V \times V$ whose elements are called **edges**.*

*The vertices u and v of an edge (u, v) are called the **endpoints** of the edge, and the edge (x, y) is said to be **incident** to u and v.*

*A **connected component** of an undirected graph G is a subgraph $H \subseteq G$ in which any two vertices are connected to each other by paths, and which is connected to no additional vertices in G.*

Definition 2. *A **directed graph** is an ordered pair $G = (V, E)$, where V is a set whose elements are called **vertices** and E is a set of ordered pairs whose elements are called **edges**.*

*The edge (x, y) is said to be **directed** from u to v.*

In this thesis, most of the networks I discuss take the form of undirected graphs, and I will refer interchangeably to "nodes" and "vertices". I discuss network representations that are not graphs in Section 3.6.

3.1.1 Graph models

Various models for generating graphs are studied in network science [New18]. In this section, I give definitions for a few graph models that appear in this thesis.

3.1.1.1 Cycle graphs

A *cycle graph* on N nodes is a graph in which the N nodes are connected in a closed chain. Note that N must be greater than or equal to 3. See Figure 3.1 for a visualization of a cycle graph.

3.1.1.2 Lattice graphs

A *lattice graph* (also known as a *mesh graph* or *grid graph*) is a graph whose drawing embedded in $\mathbb{R}^n$ forms a regular tiling. In this thesis, I use 2D rectangular lattice graphs on (M, N) where M is the number of vertices in one direction, and N is the number of vertices in the other. See Figure 3.2 for a visualization of an (M, N) lattice graph.

3.1.1.3 Erdős–Rényi random-graph model

An Erdős–Rényi (ER) random-graph model [ER59, SR51] generates random graphs in which edges occur between node pairs with equal, independent probabilities. In this thesis, I use the $G(N, p)$ ER model, where N is the total number of nodes and p is the probability of any potential edge (u, v) occurring in the graph. See Figure 3.3 for a visualization of an ER

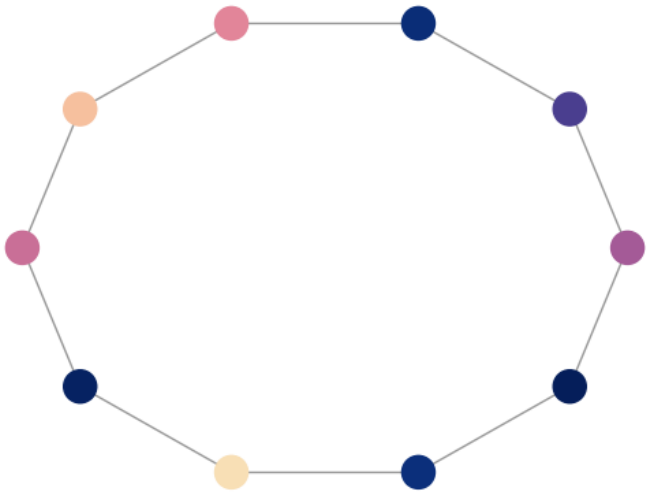


Figure 3.1: An example of a cycle graph with $N = 10$ nodes. For visual clarity, I use different colors for the nodes.

graph.

3.1.1.4 Watts–Strogatz small-world model

The Watts–Strogatz (WS) small-world model [WS98, New18] generates graphs with the "small-world" property. Given N, K, and β, there are N nodes in such a network. Each of these nodes is initially linked to its K closest neighbors ($K - 1$ neighbors if K is odd), as implemented in the NetworkX Python package. Some edges in the graph are replaced as follows: for each edge (u, v), with probability β, replace the edge with a new edge (u, w), where w is chosen uniformly at random. See Figure 3.4 for a visualization of a WS graph.

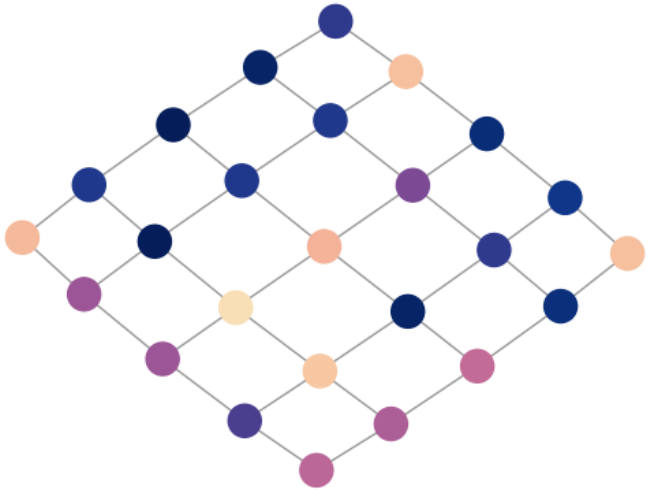


Figure 3.2: An example of a lattice graph with $M = 5$ by $N = 5$ nodes. For visual clarity, I use different colors for the nodes.

3.2 Simplicial complexes

In this section, I define some of the basic building blocks of simplicial homology. For more information, see [Hat02].

Definition 3. *A k-**simplex** is a k-dimensional polytope that is the convex hull of its $k+1$ vertices.*

Definition 4. *An **orientation** of a k-simplex is an ordering of its vertices, written as $(v_0, \ldots, v_k)$, with the rule that two orderings define the same orientation if and only if they differ by an even permutation.*

Definition 5. *An **m-face** is the convex hull of a subset of cardinality $m+1$ of a k-simplex, with $m < k$ and the orientation preserved. A **face** refers to an m-face of any dimension m.*

Definition 6. *A simplex A is a **coface** of a simplex B if B is a face of A.*

Figure 3.3: An example of a $G(N, p)$ graph with $N = 50$ and $p = 0.1$. For visual clarity, I use different colors for the nodes.

Definition 7. *A **simplicial complex** S is a set of simplices that satisfies the following conditions:*

1. *every face of a simplex from S is also in S;*
2. *the intersection of any two simplices $\sigma_1, \sigma_2 \in S$ is a face of both σ_1 and σ_2.*

Definition 8. *Given a simplicial complex S, we call the collection of all k-faces and their faces the k-**skeleton** of S.*

This definition of simplicial complex makes no use of orientation. However, in my later discussion of simplicial homology (see Section 3.3), orientation of simplices is very important.

3.3 Simplicial homology

In this section, I give the basic definitions of simplicial homology.

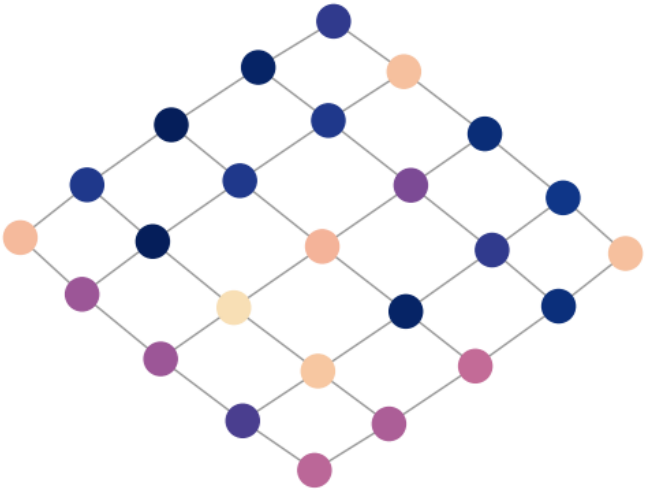


Figure 3.4: An example of a WS network with $N = 50$ nodes, $K = 5$ nearest neighbors, and rewiring probability $\beta = 0.1$. For visual clarity, I use different colors for the nodes.

Definition 9. *Let S be a simplicial complex. A **simplicial k-chain** is a finite formal sum*

$$\sum_{i=1}^{N} c_i \sigma_i \,,$$

where σ_i is an oriented k-simplex and each $c_i \in F$ for some field F.

I denote the group of k-chains on S by C_k. (With a consistent choice of orientation, this group can be considered as the free Abelian group on the basis of k-simplices in S.)

Definition 10. *Let $\sigma = (v_0, \ldots, v_k)$ be an oriented k-simplex. The **boundary operator***

$$\delta_k : C_k \to C_{k-1}$$

is the homomorphism defined by

$$\delta_k(\sigma) = \sum_{i=0}^{k} (-1)^i (v_0, \ldots, \hat{v_i}, \ldots, v_k) \,,$$

where $(v_0, \ldots, \hat{v_i}, \ldots, v_k)$ *is the oriented* $(k-1)$*-simplex that obtained by deleting the* i*-th vertex of* σ*.*

Elements of $Z_k = \ker \delta_k$ *are called* ***cycles****, and elements of* $B_k = \operatorname{im} \delta_{k+1}$ *are called* ***boundaries****.*

One can show by direct computation that $\delta^2 = 0$, so the groups (C_k, δ_k) form a chain complex.

Definition 11. *The* k***-th homology group*** H_k *of* S *over* F *is the quotient group*

$$H_k(S; F) = Z_k / B_k \,.$$

Note that $H_k(S; F)$ is nontrivial precisely when there are k-cycles on S that are not boundaries; this occurs when there are k-dimensional holes. For example, a cycle between three points gives a 2-cycle[1], and it is also a boundary precisely when the triangle with vertices at those three points is in the simplicial complex. Therefore, homological groups track features that one can construe intuitively as holes. In this thesis, I compute only H_0 and H_1, which track connected components (H_0) and loops in a space (H_1).

In most existing software packages for TDA [OPT17], one computes homology groups over the field $\mathbb{F}_2$ [OPT17]. Because $1 = -1$ in $\mathbb{F}_2$, orientation is not considered.

I now introduce the definition of a simplicial map, a fundamental building block in persistent homology (see Section 3.4).

Definition 12. *Let* S *and* T *be simplicial complexes. A* ***simplicial map*** $f : S \to T$ *is a function from the vertex set of* S *to the vertex set of* T *that preserves simplices.*

A simplicial map $f : S \to T$ also induces a homomorphism $f_* : H_k(S) \to H_k(T)$ for each integer k. The homomorphism f_* is associated with a chain map from the k-chain complex

[1] I distinguish homological cycles from cycles in a graph-theoretic sense

of S to the k-chain complex of T. This chain map is

$$(v_0, \ldots, v_k) \mapsto (f(v_0), \ldots, f(v_k)) \,,$$

where $(f(v_0), \ldots, f(v_k)) = 0$ if any pair of $f(v_0), \ldots, f(v_k)$ are not distinct.

This construction gives a functor from simplicial complexes to Abelian groups. This is essential to the theory of persistent homology that I discuss in Section 3.4.

3.4 Persistent homology

I now discuss persistent homology (PH), one of the primary tools of topological data analysis (TDA) [OPT17, Ghr08, ZC04].

Definition 13. *A **filtered simplicial complex** X is a sequence of simplicial complexes*

$$X_1 \subseteq X_2 \subseteq \cdots \subseteq X_l \,.$$

Suppose that we have experimental data X_{observed}, from which we have constructed a filtered simplicial complex $X = \{X_i\}$. There are many ways to construct such a sequence, including the commonly implemented Vietoris–Rips (VR) complex [Vie27], several fast approximations of VR complexes [EKS83, AEM07], and others [BEM10]. I discuss the existing constructions I use in this thesis in Section 3.4.1. I require that the sequence $\{X_i\}$ is increasing, as in the definition of a filtered simplicial complex; and I call each X_i a subcomplex. This filtered simplicial complex, along with inclusion maps between subcomplexes and chain and boundary maps of each of its subcomplexes, is called a *persistence complex*. I examine the homology of each subcomplex, noting that the inclusion map $X_i \hookrightarrow X_j$ induces a map $f_{i,j} : H_m(X_i) \to H_m(X_j)$, and that, by functoriality,

$$f_{k,j} \circ f_{i,k} = f_{i,j} \,. \tag{3.1}$$

Definition 14. *Let $X = \{X_i\}$, where $X_0 \subseteq X_1 \subseteq \cdots \subseteq X_l$, be a filtered simplicial complex.*

*The m-**th persistent homology** of X is the pair*

$$\left(\{H_m(X_i)\}_{1\leq i\leq l}, \{f_{i,j}\}_{1\leq i\leq j\leq l}\right),$$

*where $f_{i,j} : H_m(X_i) \to H_m(X_j)$, for all $i \leq j$ and m no larger than some maximum dimension (often chosen to be the dimension of X_{observed}), are the maps that are induced by the action of the homology functor on the inclusion maps $X_i \hookrightarrow X_j$. I refer to the collection of all m-th persistent homologies as the **persistent homology (PH)** of X.*

Most notably, the PH of a filtered simplicial complex encodes information about the maps between each subcomplex, thereby giving more information than the homologies of the individual subcomplexes. Each homology group with field coefficients $H_m(X_i)$ is a vector space whose generators correspond to holes in X_i, and the maps $f_{i,j}$ allow us to track these generators from $H_m(X_i)$ to $H_m(X_j)$. By choosing a convenient basis for $H_m(X_i)$, which one can do by the Fundamental Theorem of Persistent Homology [ZC04], we can construct a well-defined and unique collection of disjoint half-open intervals, where each generator $x \in H_m(K_i)$ corresponds to an interval $[b_x, d_x)$, with X_{b_x} denoting the subcomplex in which the generator (and its associated hole) first appears and X_{d_x} denoting the subcomplex in which the generator dies. More precisely, I say that $x \neq 0$ is born in $H_m(X_{b_x})$ if it is not in the image of f_{b_x-1,b_x}; it dies in $H_m(X_{d_x})$ if $d_x > b_x$ is the smallest index for which $f_{b_x,d_x}(x) = 0$. If $f_{b_x,j}(x) \neq 0$ for all $b_x < j \leq l$, then x lives forever and I associate the interval $[b_x, \infty)$ to it. For further details, see [OPT17, Ghr08, ZC04].

In my work, I focus on H_0 and H_1 in particular, due to the increased computational complexity associated with building higher-dimensional simplicial complexes and the 2D nature of many of the data sets explored in this thesis.

Conventional wisdom suggests that the features with the longest persistence are the most important ones. Intuitively, features that have longer persistence occur across a larger variety of scales. True topological features are scale-invariant, so it makes sense that one may conclude that these long-persistence features are the most meaningful features. However, in

practice, the PH depends heavily on the filtered simplicial complex that one constructs. In many data sets, one can choose a filtered simplicial complex construction where short bars are meaningful [BHP20, SHP17].

Choosing a filtered simplicial complex is a very important step in the computation of PH. In my research, I seek to construct a filtered simplicial complex whose PH is readily interpretable. This entails trying to construct filtered simplicial complexes whose most persistent features correspond to a meaningful feature in the data set (for example, strength of voting preference, in a case study on voting [FP20a]).

3.4.1 Methods for constructing a filtered simplicial complex

In this subsection, I review several common methods for constructing filtered simplicial complexes from point clouds.

3.4.1.1 Vietoris–Rips complex

One of the most prevalent constructions is the Vietoris–Rips (VR) complex, which one constructs using the pairwise distances between points in a point cloud [Vie27, EH10].

Let X be a data set in the form of a point cloud. Given a real number $\epsilon > 0$, define the VR complex $\mathrm{VR}_\epsilon(X)$ as follows:

$$\mathrm{VR}_\epsilon(X) := \{\sigma \subseteq X : \ \forall\, x, y \in \sigma\,,\ d(x,y) \leq \epsilon\}\,.$$

If there are n points in X, the maximal possible VR complex is the $(n-1)$-simplex that consists of all points in X and all of its subsimplices. By taking a collection $\{\epsilon_i\}$, with $0 = \epsilon_0 < \epsilon_1 < \epsilon_2 < \cdots < \epsilon_k$, and considering

$$X = \mathrm{VR}_{\epsilon_0}(X) \subseteq \mathrm{VR}_{\epsilon_1}(X) \subseteq \cdots \subseteq \mathrm{VR}_{\epsilon_k}(X)\,,$$

we obtain a filtered simplicial complex whose PH we can compute. It is straightforward to construct a VR complex, because we only need to compute pairwise distances. Additionally,

there are various fast algorithms for constructing a VR complex [Zom10]. Unfortunately, for large point clouds, the worst-case VR complex has $2^{|X|} - 1$ simplices and dimension $|X| - 1$.

Vietoris–Rips complexes are employed commonly, because it is relatively easy to construct them, they are intuitively appealing, they have important theoretical guarantees from the Nerve Theorem [KS13], and (perhaps most importantly) they have been implemented widely in existing PH packages.

3.4.1.2 Alpha complex

The alpha complex [EKS83], which we denote by $\mathrm{A}_\epsilon(X)$, also relies on a distance parameter and is defined as follows. Let $\epsilon > 0$, and let $X_\epsilon := \bigcup_{x \in X} B(x, \epsilon)$. Additionally, let $(V_x)_{x \in X}$ be the Voronoi diagram of X. Consider the intersection $V_x \cap B(x, \epsilon)$ for each $x \in X$, and note that the collection of these sets covers X_ϵ. We then have

$$\mathrm{A}_\epsilon(X) = \{\sigma \subseteq X : \ \forall\, x_i \in \sigma\,,\ \bigcap_i (V_x \cap B(x, \epsilon)) \neq \emptyset\}\,.$$

Because of the restriction of the ϵ-balls to the Voronoi diagram, the alpha complex restricts the dimension of the space in which X is embedded. In our case, because our data is embedded in $\mathbb{R}^2$, the alpha complex has 2D simplices (e.g., faces) as its highest-dimensional simplices.

3.4.2 Visualizations

Given $\{X_i\}$, the collection of half-open intervals is known as the *barcode* [Ghr08] of $\{X_i\}$, and one uses it to visualize the m-th persistent homology. Generators with longer associated half-open intervals are more persistent. In general, one uses the persistence of features to distinguish signal from noise, but recent work (including our own [FP20a], which I discuss in this thesis) indicates that persistence is not always readily interpretable in a meaningful way [SHP17, HMM19, BHP20]. There are a variety of ways to visualize barcodes; in the following paragraphs, I describe several of the most common visualizations.

3.4.2.1 Barcodes

Persistence barcodes [Ghr08] are a method for visualizing PH that represent each homological feature as a line whose left endpoint is the birth time and whose right endpoint is the death time. We represent infinite-persistence features with an arrow at the right endpoint. Persistence barcodes allow one to quickly see each homological feature separately. Additionally, because more persistent features correspond to longer lines, the most persistent features are the most visually prominent ones.

In Figure 3.5, I show an example of a computed persistence barcode.

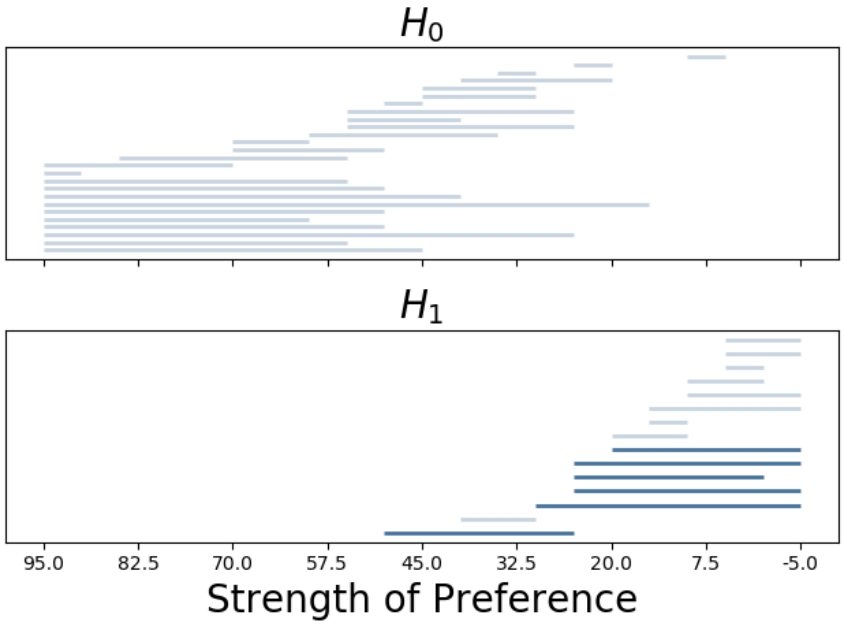


Figure 3.5: I show an example of barcodes for H_0 and H_1. The filtration parameter is on the horizontal axis, and each bar is given its own vertical coordinate that separates the bars from each other. Note that each bar is visually distinct, and that the longer bars dominate the image.

In barcodes in this thesis, I display each dimension of the PH as a separate barcode.

3.4.2.2 Persistence Diagrams

Persistence diagrams (PDs) [OPT17] associate to each homological feature an ordered pair (b, d), where b is the birth time and d is the death time. We plot infinite-persistence features (b, ∞) on a line at infinity. More persistent features are located farther from the diagonal, and less persistent features appear close to the diagonal. We can plot different dimensions of the PH on the same diagram by plotting each dimension in a different color or shape.

In Figure 3.6, I show an example of a computed PD.

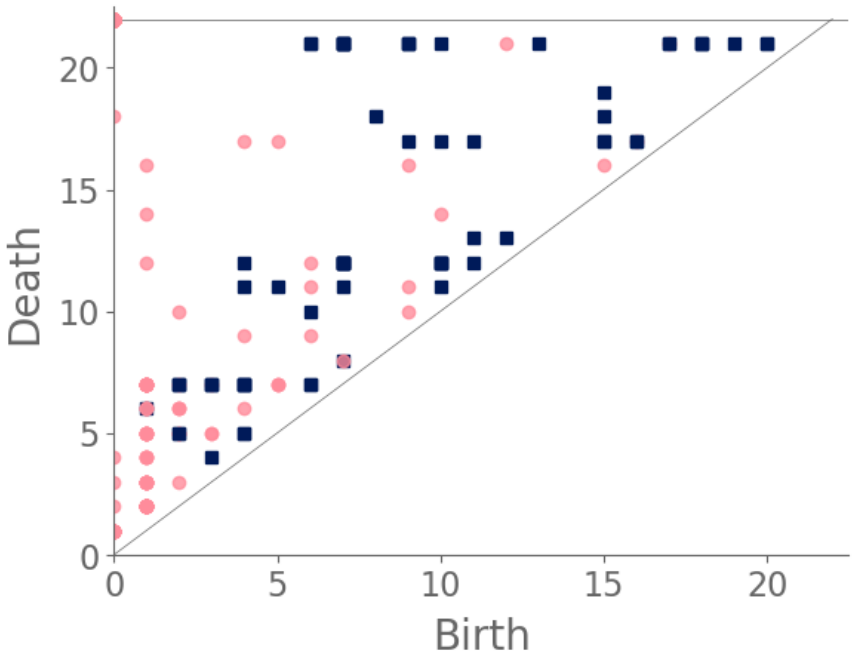


Figure 3.6: I show a PD. Birth is on the horizontal axis, and death is on the vertical axis. I plot features in H_0 as pink circles and features in H_1 as blue squares.

In this thesis, I plot both H_0 and H_1 on the same PD; computations of higher-dimension PHs do not occur in my thesis.

One advantage of PDs is the ability to equip the space of PDs with an easily computed metric called the bottleneck distance.

Definition 15. *The **bottleneck distance** between two PDs is the shortest distance b for which there exists a perfect matching between the points of the two diagrams (completed by adding points on the diagonal) such that any pair of matched points are at distance at most b, where the distance between points is the sup norm in $\mathbb{R}^2$.*

Bottleneck distance is included in many PH software packages.

3.4.2.3 Other visualizations

In this section, I discuss two other visualizations used in PH. These visualizations do not appear in my work, but they are useful for applying statistical methods to the results of PH computations. Persistence landscapes [Bub15] are a method of summarizing PH that provides several statistical benefits. Persistence landscapes allow a well-defined notion of a mean of a collection of persistence landscapes, and they also allow one to compute confidence intervals.

Definition 16. *The **persistence landscape** is a function $\lambda : N \times \mathbb{R} \to \overline{\mathbb{R}}$, where $\overline{\mathbb{R}}$ denotes the extended reals. We define*

$$\lambda(k,t) = \sup\left(m \geq 0 : \beta^{t-m,t+m} \geq k\right) ,$$

*where $\beta^{i,j}$ is the **Betti number** of the persistence module from X_i to X_j. I define*

$$\beta^{a,b} = \dim(\mathrm{Im}(\iota : X_i \to X_j)) .$$

Persistence images [AEK17] provide a method to represent a PD as a finite-dimensional vector. This allows persistence images to be used with vector-based machine-learning tools like support vector machines [VC64], linear models [Hoc83], and many more.

Given a probability distribution $g_u(x,y)$ (which one often chooses to be the normalized symmetric Gaussian with mean u and variance σ^2) and a nonnegative weighting function $f : \mathbb{R}^2 \to \mathbb{R}$ that is 0 along the horizontal axis, continuous, and piecewise differentiable, we define a transformation from a PD to a persistence surface.

Definition 17. *Given a PD B in birth–death coordinates, the corresponding* ***persistence surface*** $\rho_B : \mathbb{R}^2 \to \mathbb{R}$ *is the function*

$$\rho_B(z) = \sum_{u \in B} f(u) g_u(z) \,.$$

Intuitively, one can imagine placing a Gaussian over each point of a PD, thereby "blurring" the diagram.

Given a persistence surface, we discretize the surface into a persistence image as follows.

Definition 18. *Given a PD B, the* ***persistence image*** *of B is the collection $\{I(\rho_B)_P = \int\int_P \rho_B \, dy \, dx\}$ of pixels, where P is the area of the pixel.*

Persistence surfaces and images can be computed for any choice of weighting function and distribution, but the standard choices are to give the largest weight to the most persistent features and to use a Gaussian distribution. The persistence image of a single dimension of the PH is a vector, and vectors for PHs of different dimensions can be concatenated together to provide a single persistence image that represents all dimensions.

3.4.3 Previous and current research in TDA

PH is a quickly evolving field, with contributions from a diverse set of communities. PH methods have been applied to a wide range of fields, including neuroscience [LEF16, YKA16, BMM16], time-series analysis [SHP17], dynamical systems [YB20], the social sciences [BZ18, ID19, WSK17], biology [XW14, KBN16], hurricanes [TMD20], granular materials [KGK13, BHO18], and many others.

Researchers are also developing new methods for use in PH. This includes developing new types of filtered simplicial complexes [BHM19,KDS16], making PH more tenable to statistical analysis [Bub15, MTB15], developing theoretical guarantees for PH methods [AA17], and more. Other research in TDA includes development of clustering methods like Mapper

[SMC07], computations in knot theory [LHS19] and homotopy [BM13], and the development of persistent homotopy [MZ19].

3.5 Bounded-confidence models

Bounded-confidence models, introduced by Hegselmann and Krause [HK02] and Deffuant et al. [DNA00], model continuous opinion dynamics on a network. They capture the idea that individuals adjust their opinions based on a tolerance threshold [MVP18]. At each time, individuals' opinions take on a value (generally in $\mathbb{R}$, but theoretically in any metric space) based on some function of their neighbors' opinions, as long as their neighbors' opinions are within a certain distance of their own. In the following subsections, I give more precise specifications of various versions of bounded-confidence models.

3.5.1 Hegselmann–Krause model

In this subsection, I discuss the Hegselmann–Krause model for continuous opinion dynamics [HK02]. Given a group of N agents, denote each agent's opinion at time t by $x_i(t) \in [-1, 1]$. That is, $\vec{x} : \mathbb{R}^n \to [-1, 1]^n$ describes agent i's opinion at time t. Let A be the adjacency matrix, with A_{ij} denoting the edge weight of (i, j). Given a confidence bound c, at each discrete time step, update all agents' opinions at time $t + 1$ according to the following rule:

$$x_i(t+1) = \frac{x_i(t) + \sum_{j=1}^{N} A_{ij} x_j(t) \mathbb{1}_{|x_i(t) - x_j(t)| < c}}{1 + \sum_{j=1}^{N} A_{ij} \mathbb{1}_{|x_i(t) - x_j(t)| < c}}, \tag{3.2}$$

where $\mathbb{1}$ is the indicator function.

3.5.2 Deffuant model

In the Deffuant model [DNA00], only randomly-selected agents change their opinions in a given time step. Specifically, we choose a pair of neighboring agents, x_i and x_j uniformly at

random, and their opinions are updated according to the rule

$$x_i(t+1) = x_i(t) + m(x_j(t) - x_i(t))\mathbb{1}_{|x_i(t)-x_j(t)|<c}\,,$$
$$x_j(t+1) = x_j(t) + m(x_i(t) - x_j(t))\mathbb{1}_{|x_i(t)-x_j(t)|<c}\,.$$

Note that the selection of two neighboring agents (or equivalently, an edge) introduces stochasticity.

3.5.3 Previous research on bounded-confidence models

Bounded-confidence models have been applied to study various applications, and they have also been generalized in numerous ways. Researchers have studied consensus formation [Lor07, Dit01], polarization [SPG19, DSC17], and various adaptations of the model (including rejection [HDJ14], incorporation of media [BP20], and others [BCV14, Var14]). Applications include the spread of opinions on social networks [BP20] and random networks [MVP18], decision-making processes [ZDZ19], and formation of linguistic opinions [DCL16].

3.6 Generalizations of graphs

In network science, it is most traditional to study networks in the form of graphs [New18]. However, in many applications, it is not always suitable to use a framework in which interactions are always pairwise or in which there is only one type of interaction [Por20, LRS19]. In this section, I discuss generalizations of graphs that are designed to study networks. In particular, I focus on "multilayer networks" [KAB14] in Subsection 3.6.1 and on "simplicial networks" in Subsection 3.6.2. There are also other generalizations of graphs, but I do not discuss them in this thesis. They include hypergraphs [New18], time-dependent networks [HS12, Hol15], and many more.

3.6.1 Multilayer networks

Multilayer networks [KAB14, Por18, AM19] allow nodes to interact with other nodes across different layers. This allows different types connections between nodes, depending on the layer (or layers) on which nodes and edges occur. Imagine, for example, a transportation network in a city. There may be multiple modes of transportation that connect point A and point B. A multilayer network can model the two points as nodes and the various forms of transportation as edges on different layers. While various types of multilayer networks, such as multiplex networks, have been studied for decades [Mit69, Ebe97, WF94], I present the mathematical formulation given in Kivelä et al. [KAB14], as this notation is convenient for the multilayer network analysis.

Definition 19. *A **multilayer network** is a tuple $M = (V_M, E_M, V, \mathbb{L})$, where*

- *V is the set of nodes. A node can exist on any layer to belong to V. This set of nodes represents the entities in the network.*
- *$\mathbb{L} = \{L_a\}_{a=1}^{d}$ is a sequence of sets of elementary layers, where each L_a is a set of layers for a fixed aspect a and there are d aspects. An "aspect" is a type of layer. The set of all possible layers is given by the Cartesian product $L_1 \times L_2 \times \cdots \times L_d$.*
- *$V_M \subseteq V \times L_1 \times \cdots \times L_d$ is the set of node-layer tuples. This encodes the occurence of a node v on a specific layer $(l_1, \ldots, l_d)$.*
- *$E_M \subseteq V_M \times V_M$ is the set of edges between node-layer tuples.*

Note that I have put no requirements on E_M, so edges can occur between nodes on the same layer (an intralayer edge) or between nodes on different layers (an interlayer edge).

Definition 20. *The **adjacency tensor** $\mathcal{A} : V \times V \times L_1 \times L_1 \times \cdots \times L_d \times L_d \to \mathbb{R}$ is the tensor in which $\mathcal{A}(u, v, \alpha, \beta) = w$ indicates that there is an edge of weight w between node-layer pairs (u, α) and (v, β).*

Multilayer networks have been studied in the context of biological systems [PPP17, FSP19, GMD18, SJ15], transportation [GB15], and more.

3.6.2 Simplicial networks

In some networks, higher-order (i.e. non-pairwise interactions) can occur; consider, for example, team sports, or peer pressure. These non-pairwise interactions cannot be modeled using a graph, unless one assumes that all interactions with three or more agents are equivalent to the collection of pairwise interactions. Simplicial complexes, however, present a natural framework for modeling interactions of k individuals via k-simplices. A simplicial network is simply a simplicial complex S. I use the terminology "simplicial network" to reframe the simplices of the complex in terms of network relationships. The 0-simplices of a network are nodes, 1-simplices are edges, and higher-order simplices are faces.

There has been some work on generalizing network models and concepts to simplicial networks, but the difficulty of dealing with orientation makes the study of simplicial networks a largely open area. For example, some work on infectious disease modeling has been done by Iacopini et al [IPB19], and a simplicial extension of voter models was examined in [HK20].

CHAPTER 4

Persistent homology of spatial data

The following sections are adapted from two original papers [FP20a, FP20c] that I co-authored with my advisor Mason A. Porter. I begin in Section 4.1 by defining and describing new methods in PH that we developed in those papers. Section 4.3 describes a case study in map-based voting data [FP20a]. The data set is described in Section 2.1. I adapt Sections 4.2, 4.4, and 4.5 from a recent paper on applications of PH to spatial networks [FP20c].

4.1 Novel methods for PH

In this section, I describe methods for generating simplicial complexes based on data. We introduced these methods in [FP20a]. The "adjacency construction" in Section 4.1.1 is designed to work on network-based data, and the "level-set construction" in Section 4.1.2 is designed to work with image-based or manifold-based data. We wrote a popular exposition of these methods for *SIAM News* in [FP20b].

4.1.1 Adjacency construction of PH

I now describe a way to construct a filtered simplicial complex based on network adjacencies. Consider a network in the form of a graph (V, E), with numerical data $f(v)$ associated with each node v. For a given filtration step X_i, let the 0-simplices of X_i be given by $v \in V$ such that $f(v) \leq \epsilon$ for some value ϵ. For any edge $(u, v) \in E$, if $u \in X_i$ and $v \in X_i$, add (u, v) to fX_i. Finally, to X_i, add all triangles (u, v, w) such that (u, v), (v, w), and (u, w) are in X_i.

Repeat this process for X_{i+1} using a larger value of ϵ. By construction, each $X_i \subseteq X_{i+1}$, and so the collection $\{X_i\}$ is a valid filtered simplicial complex. See Figure 4.1 for an illustration of such a filtered simplicial complex.

In some of our applications, we use an alternate adjacency construction in which we associate data $g(u, v)$ to the edges, instead of to the nodes. This construction differs from the one above only in that we can define the function $f'(v) = \min_{u:(u,v)\in E}\{g(u, v)\}$. We then proceed with the above adjacency construction, but we substitute f' for f. We recently introduced our main adjacency construction in [FP20a], and we introduced our adaptation of it to edge-based data in [FP20c].

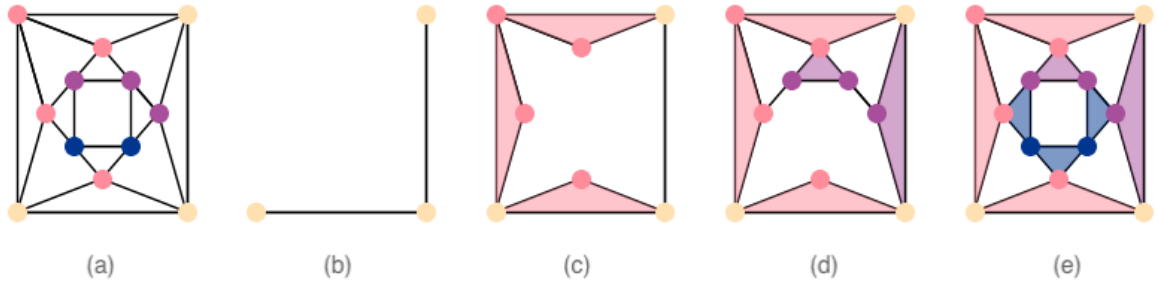


Figure 4.1: We illustrate an adjacency construction of PH on (a) a planar graph, whose nodes we color according to a function value from yellow to dark blue. At each filtration step (see panels (b)–(e)), we add all nodes with a given range of function values. We also add any edges between these new nodes, as well as any edges between these new nodes and existing nodes, and we fill in any triangles that form. Only cycles of length three form triangles, so the graph in panel (a) yields five infinite-length features in H_1 (as one can see from the five holes that remain in panel (e)).

4.1.2 Level-set construction of PH

The other PH construction (also introduced in [FP20a]) that I use in this thesis involves describing data as a manifold, rather than as a graph. Let M denote a 2D manifold, such

as data in an image format. We consider the boundary Γ of M and construct a sequence

$$M = M_0 \subseteq M_1 \subseteq \cdots \subseteq M_n$$

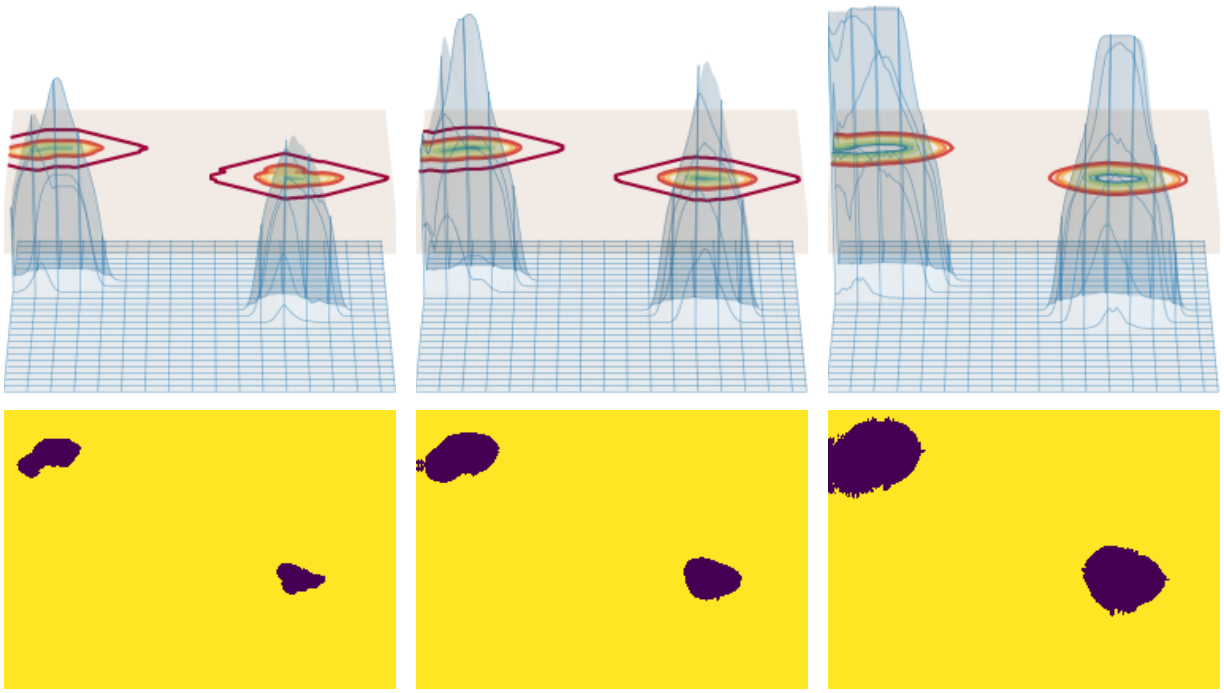
of manifolds, where at each time step, we evolve the boundary Γ_t of M_t outward according to the level-set equation. (See [OF03] for a thorough exposition of the level-set equation and level-set dynamics.) That is, if the manifold M is embedded in $\mathbb{R}^2$, we define a function $\phi(\vec{x}, t) \colon \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}$, where $\phi(\vec{x}, t)$ is the signed distance function from $\vec{x}$ to Γ_t at time t. We propagate Γ_t outward at velocity v (in our case, v is a constant 1) using the equation

$$\frac{\partial \phi}{\partial t} = v|\nabla \phi| \,. \tag{4.1}$$

Because this evolution gives a signed distance function at every time step t, we take M_t to be the set of points $\vec{x}$ such that $\phi(\vec{x}, t) > 0$. This corresponds to points inside the boundary Γ_t. We stop at time T large enough that M_T covers the entire image (alternatively, when $\phi(\vec{x}, t) > 0$ for all x in the domain of the base image).

Intuitively, in terms of a geographical map, we can visualize the graph of $\phi(\vec{x}, 0)$ as a mountain (or multiple mountains, if there is more than one connected component), with the boundary of the map at sea level, the interior of the map above water, and the complement of the map below water. The set M_0 is the set of points $\vec{x}$ that are at or above sea level. As we evolve ϕ, we move the entire mountain upward, increasing the amount of land above water. The new region that is at or above water is our expanded manifold M_t. In Figure 4.2, we show the evolution of the 0-superlevel set (i.e., all points $\vec{x}$ such that $\phi(\vec{x}, t) \geq 0$) as t increases, along with the graph of ϕ to help visualize the corresponding evolution of the level-set equation (4.1).

We then impose $\{M_t\}$ over a triangular grid of points to construct a filtered simplicial complex $\{X_t\}$ in the following manner. For every grid point $\vec{x}$, if $\phi(\vec{x}, t) > 0$, add $\vec{x}$ to X_i as a 0-simplex. For two points $\vec{x}, \vec{y}$ that are neighbors in the grid, if $\phi(\vec{x}, t) > 0$ and $\phi(\vec{y}, t) > 0$, add $(\vec{x}, \vec{y})$ to X_i as a 1-simplex. Finally, if $\vec{x}, \vec{y}, \vec{z}$ are the vertices of a triangle in the grid,

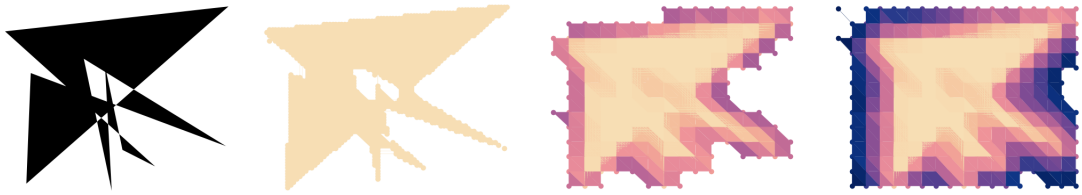


(a) $T = 0$ (b) $T = 4$ (c) $T = 10$

Figure 4.2: Evolution of (top row) a level set, with corresponding (bottom row) contour plots of ϕ. As t increases, the graph of ϕ translates upward, so the 0-superlevel set expands. (Clipping of minimum and maximum values, which we do for computational efficiency, leads to flat areas at the minimum and maximum values of ϕ.)

add $(\vec{x}, \vec{y}, \vec{z})$ to X_i as a 2-simplex. In [FP20a, FP20c], we used a triangular grid by first constructing a regular rectangular grid whose grid cells have a width and height of 5 pixels. This results in grid points every five pixels. We connect these points to their neighbors in the north, south, east, and west directions. We then connect each grid point to the diagonal adjacent grid points to the northwest and southeast, resulting in a triangular grid.

This procedure constructs a corresponding simplicial complex X_t for each M_t. In Figure 4.3, I show a visualization of this simplicial complex. Because the level-set equation (4.1) evolves continually outward, this automatically satisfies that condition that $X_t \subseteq X_{t+1}$, so $\{X_t\}$ is a filtered simplicial complex.



(a) (b) (c) (d)

Figure 4.3: Illustration of a level-set adjacency construction of PH. In (a), I show a synthetic image that I used as an initial manifold for level-set evolution. In (b)–(d), I show various filtration steps of the filtered simplicial complex that I generate by performing a level-set evolution on the image in panel (a). Panel (b) shows the simplicial complex that I obtain by overlaying the image in panel (a) on a triangular grid. In panels (c) and (d), I add new vertices, edges, and triangles to the image as it evolves outward. Darker colors indicate simplices that enter the filtration at a later time step.

In the examples in this thesis, I construct level-set complexes for 2D images. However, the level-set method generalizes to higher dimensions, so one can also construct level-set complexes for higher-dimensional manifold data.

4.1.3 Sizes and computation times

In general, construction of simplicial complexes can be very slow, as one must check all possible simplices. The number of simplices grows as n^m, where n is the number of vertices and m is the maximum simplex dimension that one is considering. Consequently, methods that build smaller simplicial complexes tend to be faster. In Table 4.1, we compare the number of simplices in the simplicial complexes that we construct using the various methods.

In this section, I discuss comparisons of computation time and simplicial complex size to state-of-the-art methods. These computation times (and Table 4.1) appear in our work in [FP20a], and all of the tables in this section are based on the California precinct-level election data from [SFK16]. However, the size scaling and computation time properties displayed in these tables is generalizable to many 2D spatial networks, for reasons I discuss in the following paragraph. See Section 4.3 for more information on the problem statement and particular choices used in the construction of simplicial complexes for the voting data. In these tables, we only construct one of the VR or alpha complexes. Because of long computation times, for counties with at least 151 precincts, we computed alpha complexes (approximations to VR complexes) instead of VR complexes.

From Table 4.1, we see that the adjacency and level-set complexes do not scale in size as rapidly as the VR complexes. This arises from how we construct these complexes. In adjacency complexes based on 2D spatial data, the number of neighbors can be almost constant for any number of nodes, as there are sometimes practical bounds on the number of nodes that can border another nodes. For example, any district on a map is likely to have a constant number of neighboring districts. The beneficent scaling of the level-set complexes with respect to the size of image data arises from our specific choices for how we construct them. Because we take each vertex of a simplicial complex to be a point on a triangular grid, it has at most six neighboring vertices (one for each of its cardinal directions, as well as one to its upper left and one to its lower right), and it can thus be a member of at most

six 2-simplices. One can make different choices of triangular grids—in our case, we simply added a northwest/southeast diagonal to each square in a square grid—and the number of neighbors is $O(1)$, as long as the grid is composed of triangles that are roughly the same size and shape (as is true for many grids). In practice, the size of a level-set complex is depends mostly on how much of an image is covered by the initial manifold M. That is, if M has large contiguous regions that span most of an image, then most of the possible grid points, edges, and faces possible will be added to the filtered simplicial complex, leading to a larger simplicial complex.

In Table 4.2, I compare the computation times for the construction and computation of PH for several of the larger (and therefore more computationally intensive) complexes that we studied in [FP20a]. From Table 4.2, it is significantly faster to construct the adjacency and level-set complexes than it is to construct the VR complexes. This is especially striking in light of the fact that we did not optimize our implementations of the new methods to make them as fast as possible. (For the level-set complexes, it is possible to make the computations much faster using existing implementations of level-set dynamics [GFO18].)

Our constructions are only slightly slower than or of similar computation time to the construction of alpha complexes. These speed gains are due largely to the significantly smaller number of simplices that we need for our new types of simplicial complexes. In 2D geospatial applications, the number of simplices is smaller than for other applications because of constraints from our starting geographical maps. In other applications, one does not typically benefit from such a built-in limitation in numbers. (For example, networks in general do not satisfy the property that the degrees of the nodes are roughly constant for any total number of nodes [New18].) However, other spatial applications (e.g., granular materials, transportation networks, and various examples in biology) will likely also benefit from these ideas. We illustrate that with a few examples in Sections 4.2, 4.4, and 4.5.

Table 4.1: Sizes (i.e., number of simplices) of simplicial complexes constructed based on voting data [FP20a]. We first partitioned each county into precincts that voted for Clinton (C) and precincts that voted for Trump (T). We did not consider precincts that did not favor one of the two candidates. We then computed VR (or alpha), adjacency, and level-set complexes for each of these sets of precincts. (We computed VR complexes for counties with at most 150 precincts and alpha complexes for counties with 151 or more precincts.)

County	# Precincts	VR		Alpha		Adjacency		Level-set	
		C	T	C	T	C	T	C	T
Alameda	1156	–	1967	5843	–	5755	70	3327	3578
Alpine	5	2	1	–	–	11	1	11962	1505
Amador	30	3	884	–	–	2	168	46	3979
Calaveras	29	8	641	–	–	6	92	1897	5195
Colusa	17	19	74	–	–	10	46	1665	5329
Contra Costa	711	–	3551	3561	–	3240	126	4135	3215
Del Norte	18	5	204	–	–	4	61	3584	6385
El Dorado	196	2397	89301	–	–	136	1123	782	4965
Fresno	592	–	–	1825	1431	1540	1192	2031	4788
Glenn	34	8	1152	–	–	4	156	329	5247
Humboldt	127	45998	680	–	–	504	119	15211	7323
Imperial	179	32496	6320	–	–	313	129	4375	6223
Inyo	25	33	216	–	–	14	51	4169	2242
Kern	642	–	–	1125	2119	928	2083	1429	5033
Kings	183	6305	69786	–	–	155	599	4849	7338
Lake	70	2279	779	–	–	99	73	4468	11275
Lassen	51	1	5920	–	–	1	250	193	11439
Los Angeles	4988	–	–	26551	1747	27705	1067	8587	6686
Madera	67	927	1947	–	–	103	132	925	5139
Marin	182	–	3	1037	–	1074	3	7893	621
Mariposa	25	5	401	–	–	7	91	2241	4485
Mendocino	250	–	692	1115	–	946	51	11901	1400
Merced	268	139832	54664	–	–	546	435	2213	6999
Modoc	21	0	399	–	–	0	94	0	7995
Mono	12	41	5	–	–	35	4	2499	3452
Monterey	467	–	13887	2297	–	1059	135	3597	4370
Napa	170	170093	56	–	–	858	15	10414	4968
Nevada	82	2569	2242	–	–	230	201	2946	2495
Orange	1668	–	–	5391	3811	4373	2632	5719	6513
Placer	363	5085	–	–	1685	141	1902	1210	3354
Plumas	30	8	618	–	–	6	102	723	6609
Riverside	1126	–	–	2291	2833	1602	2081	2231	2617
Sacramento	1267	–	–	2935	1275	15893	3459	4263	6748
San Benito	54	1804	276	–	–	152	67	699	6357
San Bernardino	2654	–	–	6206	4953	3658	2465	1700	6487
San Diego	2111	–	–	8007	3329	7480	2977	4680	7447
San Francisco	599	–	0	3499	–	3728	0	6826	0
San Joaquin	500	–	–	1659	1091	1490	902	7115	13419
San Luis Obispo	161	24600	14301	–	–	307	351	1319	4321
San Mateo	467	–	8	2573	–	2457	4	13865	782
Santa Barbara	250	–	11950	971	–	835	287	3488	6542
Santa Cruz	267	–	28	1307	–	1301	7	4737	295
Shasta	121	3	75177	–	–	2	745	941	5973
Sierra	22	3	233	–	–	2	57	417	3677
Solano	258	125438	13096	–	–	727	338	4589	5891
Sonoma	491	–	886	2355	–	2204	32	6031	899
Stanislaus	218	45984	51289	–	–	420	493	2536	6219
Sutter	52	62	3558	–	–	23	266	588	10689
Tehama	46	0	4261	–	–	0	241	0	5007
Trinity	25	25	243	–	–	12	60	5485	10344
Tulare	250	13096	–	–	921	235	1032	2242	7763
Tuolomne	68	18	10605	–	–	6	334	3380	3997
Yolo	129	49597	486	–	–	559	70	5089	4597
Yuba	46	5	3422	–	–	3	199	1909	8521

Table 4.2: Computation times of selected county–candidate pairs from California precinct-level voting data, where I show the fastest method for each example in bold. I present several larger counties to show that our methods are substantially faster than computing VR complexes. For small counties, such as Imperial and Tulare, the improvement in computation time is less noticeable. Computing level-set complexes is not substantially faster for small counties than for large counties, as the number of simplices in a level-set complex is based on the image resolution of a geographical map, rather than on its number of precincts. [Table appears in [FP20a].]

County	VR		Alpha		Adjacency		Level-set	
	Complex	PH	Complex	PH	Complex	PH	Complex	PH
El Dorado (T)	180.426 s	0.783 s	–	–	**0.090 s**	**0.008 s**	2.580 s	0.011 s
Imperial (C)	0.739 s	0.154 s	–	–	**0.0137 s**	0.009 s	9.29 s	**0.007 s**
Los Angeles (C)	–	–	15.479 s	0.065 s	39.264 s	0.069 s	**8.842 s**	**0.045 s**
Merced (C)	488.823 s	0.669 s	–	–	**0.0217 s**	**0.009 s**	6.677 s	0.025 s
Napa (C)	654.803 s	0.980 s	–	–	**0.048 s**	**0.010 s**	9.161 s	0.042 s
San Bernardino (C)	–	–	1.765 s	0.032 s	**0.691 s**	0.030 s	4.385 s	**0.019 s**
Tulare (T)	–	–	**0.0515 s**	0.016 s	0.129 s	0.015 s	5.180 s	**0.006 s**

4.2 Random spatial networks

In this section, I discuss applications of our adjacency PH construction to a dynamical process on synthetic networks in which space plays an important role. The text and figures are adapted from [FP20c].

For each network (V, E), we run the Watts threshold model (WTM) [Wat02] on it. Given a graph, we select a fraction $\rho_0 = 0.05$ of its nodes uniformly at random to be "infected" at time 0. At each time step, we then compute the fraction of each node's neighbors that are infected. (That is, we synchronously update the states of the nodes [PG16].) If the fraction of a node's neighbors that are infected meets or exceeds a threshold (in our case, the threshold is $\upsilon = 0.18$ for all nodes), the node becomes infected. We take this implementation of the

WTM to be the generator of a function $f : V \to \mathbb{N}$, where $f(v)$ is the time at which node v becomes infected. We say that infected nodes are in the set I. If v never becomes infected, we set $f(V) = \max_{v \in I} f(v) + 1$, so we eventually add all nodes to a filtered simplicial complex. The resulting filtered simplicial complex consists of the subgraphs that are generated by I at each time step. See [TKH15, MTP18, Yin13] for some research on the properties of the Watts threshold model on spatial networks.

At each time step, the set of nodes which are "infected" and the edges between them make up the "infected subgraph" of the network. We examine topological changes in the infected subgraph of three different types of synthetic networks. We first examine random geometric graphics (RGGs) [Pen07]. For each instance of an RGG, we pick 100 nodes uniformly at random from the unit square. If the Euclidean distance between two nodes is less than or equal to 0.1, we add an edge between them. Our second type of synthetic network is a square lattice network with 100 nodes. We arrange the 100 nodes in a 10×10 grid on the unit square, and we then connect the nodes along the grid lines (as in Figure 4.4). Our third type of synthetic network is a Watts–Strogatz (WS) small-world network [WS98, Por12] (see Section 3.1.1.4). We begin with a ring of 100 nodes. We then connect each node to its $K = 4$ nearest neighbors. We then rewire each edge uniformly at random with a probability of $\beta = 0.1$ using the implementation of the WS model in NetworkX [HSS08].

For each type of synthetic network, we consider 100 instances, which we generate using NetworkX. For the RGG and WS networks, each instance is a different graph; the square lattice network is deterministic in nature. For all three types of networks, each instance has a different initial set of infected nodes. We show visualizations of each of these types of networks (with WTM dynamics on it) in Figure 4.4.

Our adjacency construction begins by selecting the initially infected nodes and the edges between them of a network to create an infected subgraph that we call an "infection network". As the infection spreads, we add more nodes and edges to the infection network until eventually we have added all nodes and edges to it.

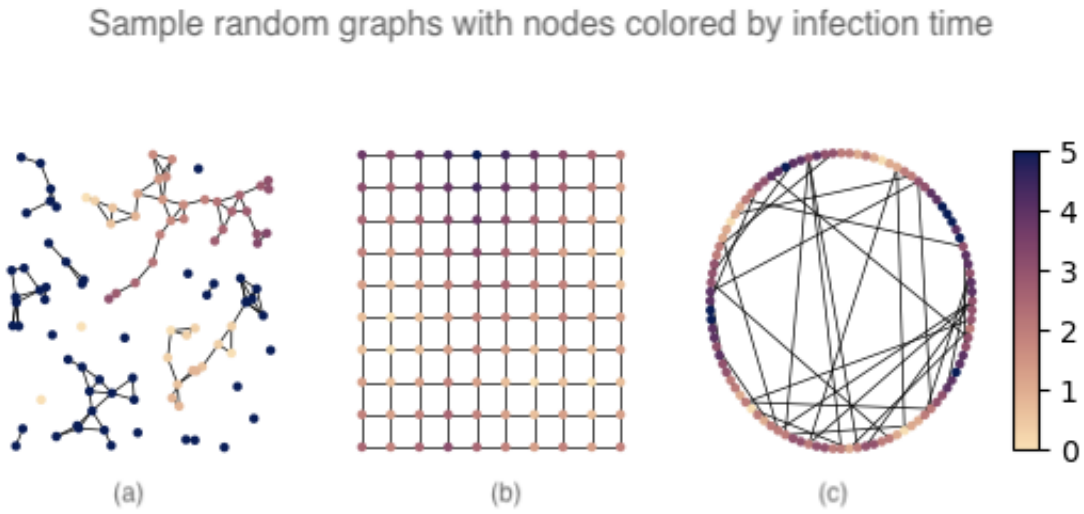


Figure 4.4: An instance of each of our synthetic networks with Watts threshold model (WTM) dynamics on it. The corresponding PDs are in Figures 4.5–4.7. We color the nodes based on the time that they become infected. The three types of synthetic networks are (a) a random geometric graph, (b) a square lattice network, and (c) a Watts–Strogatz small-world network.

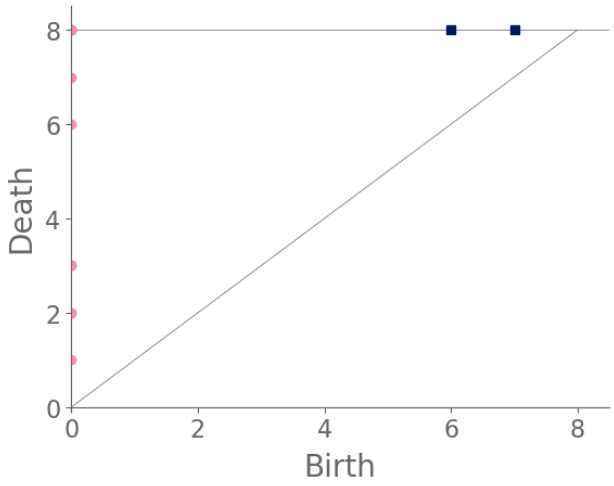


Figure 4.5: PD for an instance of the WTM on an RGG. We plot each feature as a point on the PD, for which the horizontal coordinate represents the birth time and the vertical coordinate represents the death time. We plot features with infinite persistence (i.e., features that do not die within the range of filtration parameters that we use for a PH computation) on a horizontal line at the top of the PD. We plot features in H_0 (which indicates the connected components) as pink circles, and we plot features in H_1 (which indicates the 1D holes) as dark-blue squares.

Examining the PHs of the RGGs (see Figure 4.5), we see that for our parameter values, the infection network tends to have several connected components, resulting in a large number of features in H_0. However, because of the spreading behavior of the WTM, new nodes can become infected only via their infected neighbors. Because features in H_0 record connected components of a graph, new infected nodes join existing connected components. Therefore, features can only be born at time 0 or in the last step, which is when we add all remaining uninfected nodes to our filtered simplicial complex. By contrast, features in H_1 are relatively rare, as most cycles that occur in an RGG are filled because of the uniform probability distribution of the node locations.

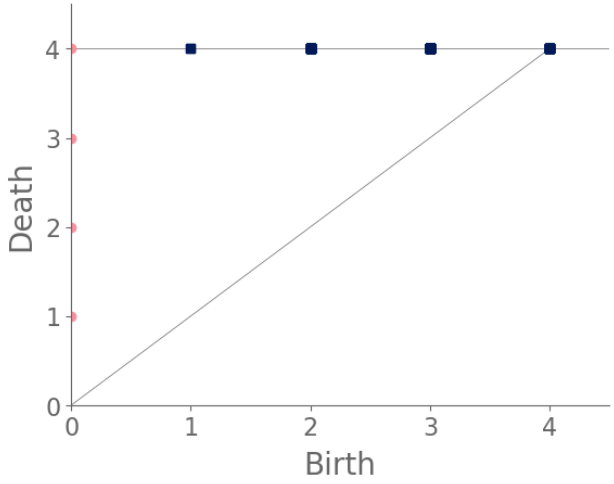


Figure 4.6: PD for an instance of the WTM on a square lattice network.

For a square lattice network (see Figure 4.6 for a PD of the WTM on such a network), we first note that there is only a single infinite-length feature in H_0, as the final infection network necessarily consists of a single connected component. Consequently, H_0 consists of a set of features that are born at time 0 and eventually merge (and therefore die), resulting in a single infinite-length feature. Additionally, there are a constant number (81, to be precise) of features in H_1, because when we construct a simplicial complex, every grid cell of the lattice is a feature in H_1 at the final filtration. However, these features can be born at a variety of times, as all nodes must enter the filtration for the filtration to incorporate every

lattice cell.

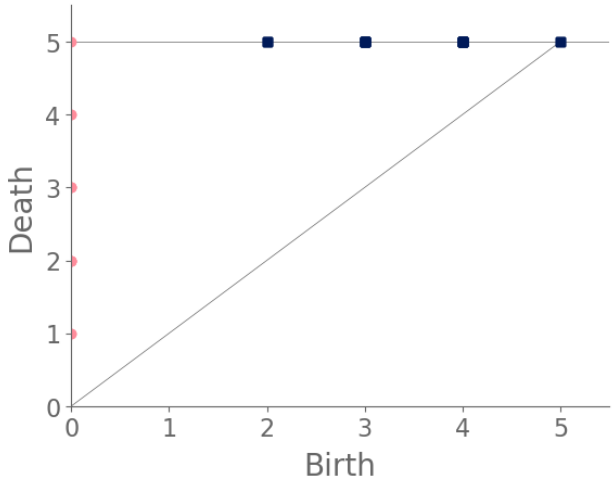


Figure 4.7: PD for an instance of the WTM on a WS network.

From Figure 4.7, we see a WS small-world network also eventually has an infection network that consists of a single connected component. However, the WS networks consistently have more features in H_1 than the RGG networks, because the former's (non-geometric) shortcut edges usually result in splitting an existing cycle (and hence a feature in H_1) into two cycles.

We summarize our observations about the various synthetic networks in Table 4.3, in which we give the means and standard deviations of the number of features during the temporal evolution of the WTM for each type of synthetic graph. Our counts include features that appear at any time during the WTM dynamics.

4.3 The "shape" of voting

In this section, I discuss an application of adjacency and level-set complexes to California voting maps. The text and figures in this section are adapted from [FP20a].

For the adjacency complexes, we build networks based on the voting maps as follows.

Table 4.3: Means and standard deviations of the numbers of features in H_0 and H_1 during the temporal evolution of the WTM across all instantiations of each type of synthetic graph. (Our counts include features that appear at any time during the WTM dynamics.)

	Mean (H_0)	STD (H_0)	Mean (H_1)	STD (H_1)
RGG	23.16	3.1897	1.2	1.0
Square lattice	4.56	0.5886	81	0
WS	8.29	2.0214	26.95	5.2314

Each precinct is a node in the network, and there is an edge between two precincts if they are queen adjacent. Precincts are queen adjacent if they touch at any point, including corners. This can be contrasted with rook adjacency, which occurs when any two regions of a map share a border.

For the level-set complexes, we use images of the voting maps as input data. For any county, we take the map of all precincts that voted for the same candidate to be the base manifold M. In visualizations, we color precincts according to voting preference. Precincts that voted for Clinton are colored in blue, while precincts that voted for Trump are colored in red. The darker the color, the stronger the preference for a candidate. Precincts with equal numbers of votes for each candidate are colored in white.

We generate two types of visualizations for our PH results. The first takes the form of barcodes, where we display each feature as a bar whose length corresponds to its persistence. The second is a map visualization, where we mark the locations of the features that we find by computing PH by drawing a cycle that passes through all of the generators of a feature. (We call this a "feature map".) These generators are not necessarily unique, and we select our generators by using a standard PH algorithm (specifically, by using the row-reduced boundary matrix) [ZC04]. Although the non-uniqueness of generators is a potential concern, in our study, any set of generators results in some group of precincts that surround a voting

island. We color the cycle according to the political party of the candidate. For example, if we find a blue hole in a sea of red, we draw a red cycle. To help illustrate the various interpretations of persistence, we highlight "long-persistence" features in H_1. Specifically, if an element $[x] \in H_1$ has persistence interval $[\text{birth}([x]), \text{death}([x])]$, we compute

$$l = \frac{\text{death}([x]) - \text{birth}([x])}{\max_{[y]\in H_1} [\text{death}([y]) - \text{birth}([y])]} \,. \tag{4.2}$$

If $l \geq 0.75$, we consider $[x]$ to be a long-persistence feature. We color long-persistence features in dark red or dark blue, depending on the political party of the candidate, and we color other features in lighter shades of red or blue. We also color long-persistence features with darker bars in the barcodes. We discuss results for two counties in this section.

For our first example, we compare the barcodes and feature maps that we obtain by computing PH of the alpha, adjacency, and level-set complexes that we generate from red precincts (i.e., those with a majority who voted for Donald Trump) in Tulare County (see Figure 4.8). Tulare County is relatively small, with only 250 precincts. The county is predominantly rural, although it has a few small urban areas towards its western side. Politically, Tulare is a strongly Republican county, and only a very small proportion of its precincts voted blue (i.e., for Hillary Clinton) in the 2016 election. Looking at a voting map of Tulare, we observe several pockets of blue voters that we hope to be able to detect using PH. To detect these blue pockets, we consider the topological structure of simplicial complexes that we construct using only the part of the map with red precincts, and we seek to find holes in these complexes. In Figure 4.9, we show the results of the three different constructions.

For the alpha complex, we observe that the dimension-1 barcodes indicate that most features do not have long persistences. The loops that surround the blue holes are light red, indicating that they are not long-persistence features. Additionally, the single long-persistence feature corresponds to a loop in the northwest part of the voting map (see Figure 4.9); it connects three precincts whose union is disconnected, and it does not surround any blue areas. It thus exhibits two problems, which we call "scaling" and "contiguity" problems. "Scaling" refers to the phenomenon in which precincts occur at wildly different size scales.

When we use Euclidean distance between precinct centroids to build a simplicial complex, it is influenced by the size of the precincts. For example, the spacing of the three precincts forming this loop is such that the pairwise distances between them are similar, but this spacing is at a distance that is larger than the precincts themselves, causing them to form a loop even though none of them are adjacent to each other on the map. Because this loop corresponds to the only long-persistence bar in the barcode, it is difficult to use persistence to distinguish fake loops like this one from real loops in the western region of the map. This particular feature also demonstrates a "contiguity" problem; that is, the alpha complex does not recognize that the three precincts do not form a contiguous region. As a result, the loop that the alpha complex detects does not correspond to a contiguous loop of precincts. Overall, the alpha complex does detect some pockets (of voting results that are surrounded by different voting results), but it misses a few of them just southeast of the central area. It also detects many features that are not real.

In contrast to our observations when using the alpha complex, generator precincts in the adjacency complex mostly form contiguous loops. By virtue of our construction, edges cannot occur between the centroids of precincts that are not adjacent to each other. A few features that are disconnected from their (graph-theoretic) neighbors do still appear on the resulting feature map, largely because the precincts themselves have complicated shapes. For example, some of them are not simply connected and others have multiple connected components. Some work in mathematical gerrymandering has focused on tackling some of these issues by quantifying the idea that electoral districts ought to be "compact" [BS18, DT18]. However, for the most part, the generator precincts surround blue and light-red holes in the voting map. Additionally, there are fewer bars in the dimension-1 barcode in the adjacency complex than in the alpha complex, and more of the bars in the adjacency complex correspond to long-persistence features. The longest bar corresponds to the large hole in the middle that includes both blue and light-red precincts. Although these light-red precincts do eventually join the filtered simplicial complex, the blue precincts in the middle ensure that this hole

never closes. Keeping in mind that the generators of a feature are not necessarily unique, the particular algorithm that we use to compute PH selects the group of darker red precincts that surround that area. We also observe several small light-red holes (which correspond to early-birth bars) and several blue holes (which correspond predominantly to the bars in the barcode that are born late). The adjacency complex is able to locate most of the blue areas of the voting map—the exceptions are a few areas near the edges (and there is no hope of detecting several of these as holes, because they lie on the county's borders and thus cannot be surrounded)—and it has little noise. All of the long-persistence feature are true features, and we can therefore do a better job of distinguishing signal from noise for Tulare County with an adjacency complex than with an alpha complex.

Finally, we examine the level-set barcode and feature map for Tulare County. We only consider features that start at time 0, because features that start at later times do not exist in the initial image and are instead created by the level-set evolution. Observe that the dimension-1 barcode has several features—some with long persistence and others without long persistence—that start at time 0; there is one feature that starts at a much later step of the filtration. These bars that start at time 0 correspond to several of the holes in the western area of the voting map. We detect only six of these holes, as some of them occur on size scales that are too small for us to capture in our level-set complex because of our choice of grid when constructing the coomplex. We also observe that the persistence of a bar is correlated positively with the size of its associated hole. The single long-persistence feature corresponds to the largest blue hole. Overall, the level-set complex captures most of the blue areas on the map and avoids most of the noise, although it does fail to detect some of the smaller regions.

For our second example, we consider Imperial County's blue precincts, which we show in the map in Figure 4.10. In contrast to Tulare County, it is not immediately evident for Imperial County where there may be holes. There do seem to be a few very small red precincts that are surrounded by blue precincts, so we hope to be able to capture some of

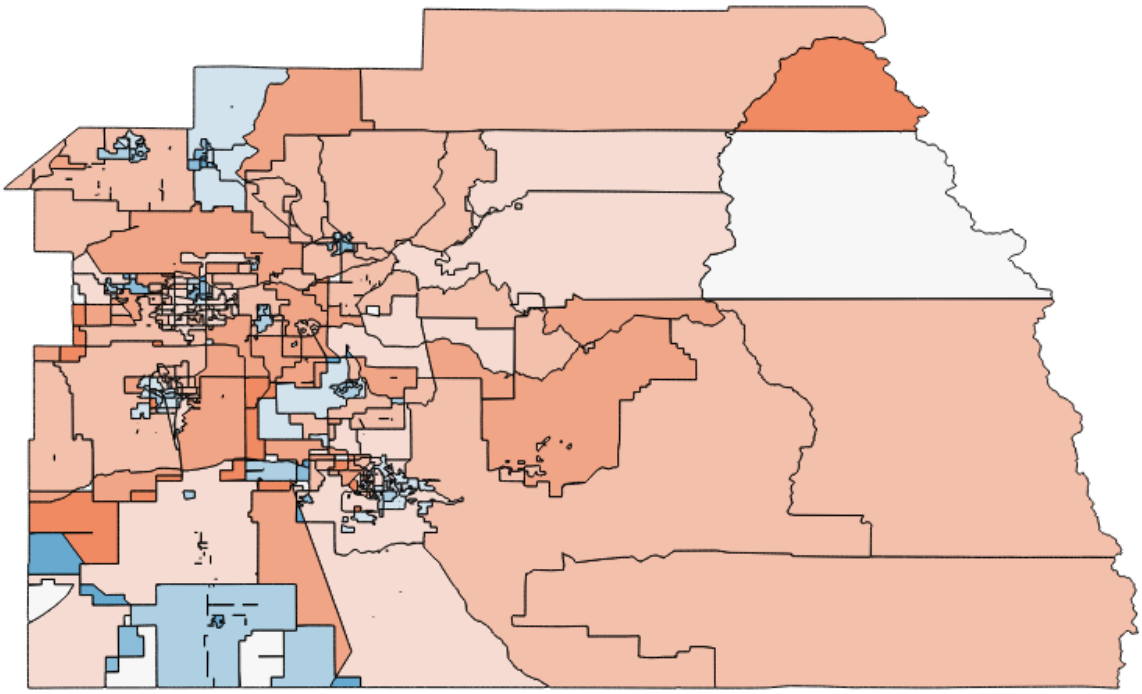
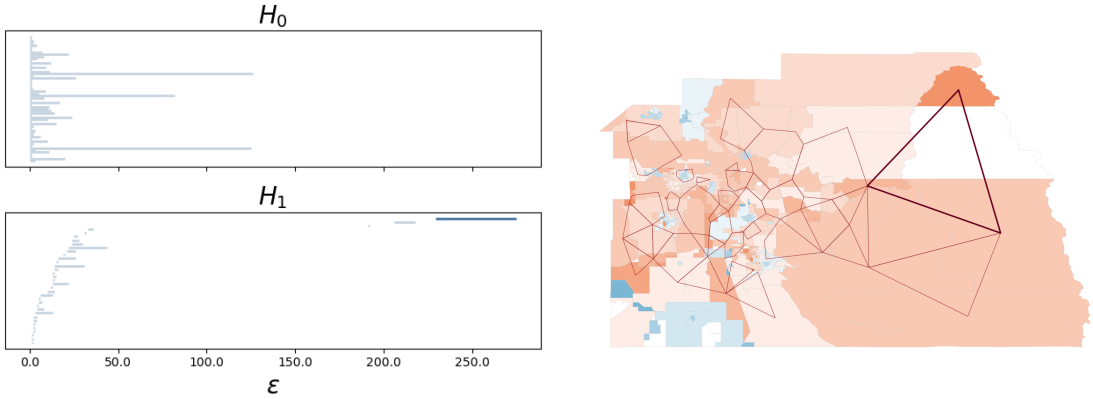
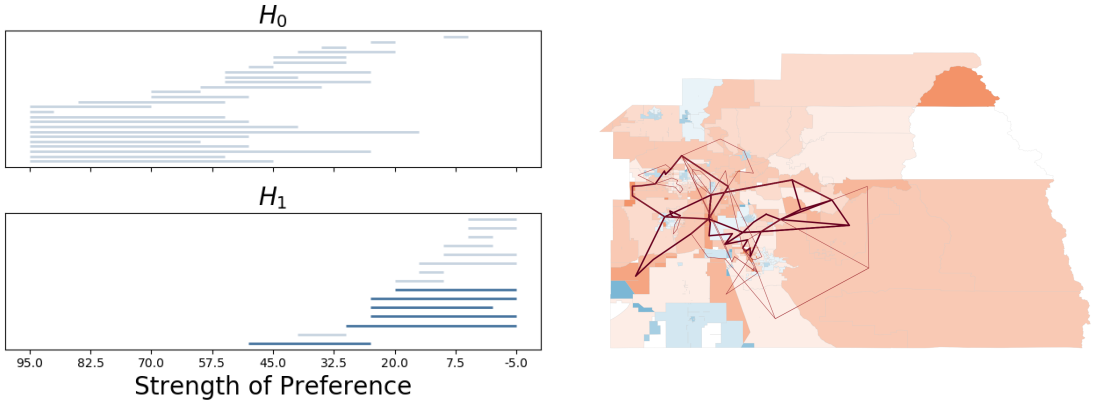


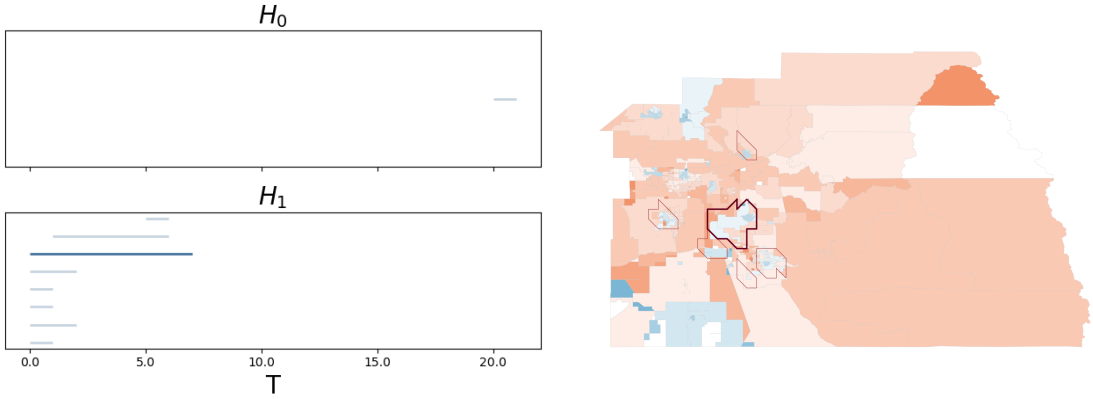
Figure 4.8: Tulare County, which we color based on the voting for president in the 2016 election. Red precincts have a majority who voted for Trump, and blue precincts have a majority who voted for Clinton. Darker colors indicate stronger majorities. For the level-set complex, we plot only features which start at time 0, as features starting at later times are noise created by the level-set evolution.



(a) Alpha complex



(b) Adjacency complex



(c) Level-set complex

Figure 4.9: Barcodes and generated loops for red precincts in Tulare County. We mark long-persistence features using darker loops with thicker line widths. For the level-set filtration, we show only features that start at time 0.

those. Overall, however, we expect to observe relatively few features.

Examining the results from the various constructions, we observe that the VR complex picks up some noise, and only one of the features appears to surround a hole. Instead, it finds several areas where the blue precincts are tightly clustered, but they do not seem to surround any red precincts. Furthermore, all of the features have similar persistences, and they are all categorized as long-persistence features. Unfortunately, because so many of the precincts in Imperial County are small, it is unsurprising that all of the features have similar persistences, so it is difficult to distinguish signal from noise. Moreover, as we will see, our findings from the adjacency complex and level-set complex imply that the VR complex is not picking up any real holes.

The adjacency complex picks up one long-persistence feature and two other features. On inspection, these appear to be small white or light-blue holes that are surrounded by darker blue districts. All three of the holes appear to be around either white precincts or red precincts, and the single long-persistence feature is composed of relatively dark-blue precincts. The long-persistence feature also seems to be the only feature that corresponds to a feature from the VR construction.

In contrast to the adjacency and VR complexes, which include very few features, the level-set complex picks up a large number of dimension-1 features, but none of them start at time 0. This occurs because, as the level set evolves, the separate connected components eventually combine, creating a larger number of holes than the ones that actually exist in the original voting map. This illustrates one of the problems with the level-set complex: as time passes, the simplicial complex tends toward becoming progressively more connected, which can create some false features when the simplicial complex starts with many connected components. However, if one considers only those features that exist at time 0, one can distinguish between genuine and false features. Most of the counties have relatively homogeneous voting patterns, with small pockets of dissimilarity, so few of the California counties exhibit this behavior in practice. Additionally, including only features that begin

at time 0 results in reasonable feature maps.

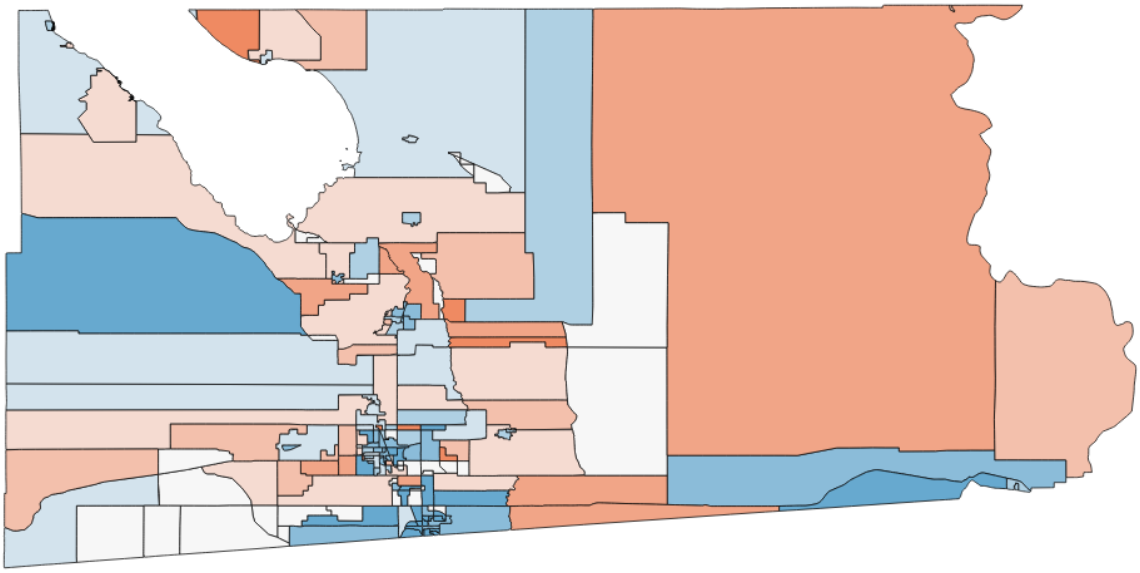
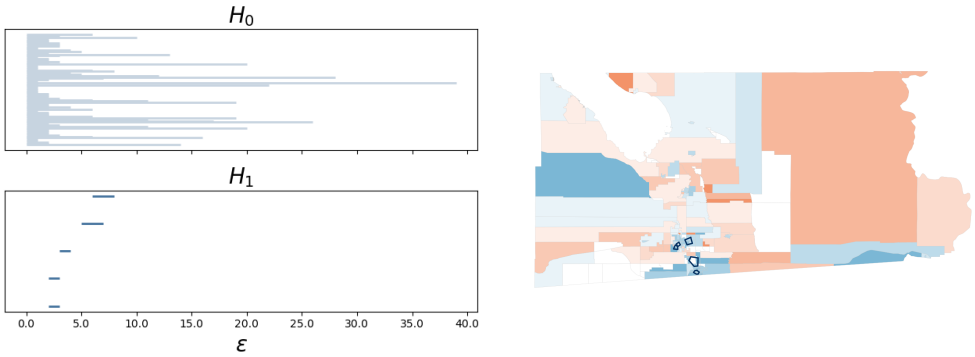


Figure 4.10: Imperial County, which we color based on presidential voting. Red precincts have a majority who voted for Trump, and blue precincts have a majority who voted for Clinton. Darker colors indicate stronger majorities.

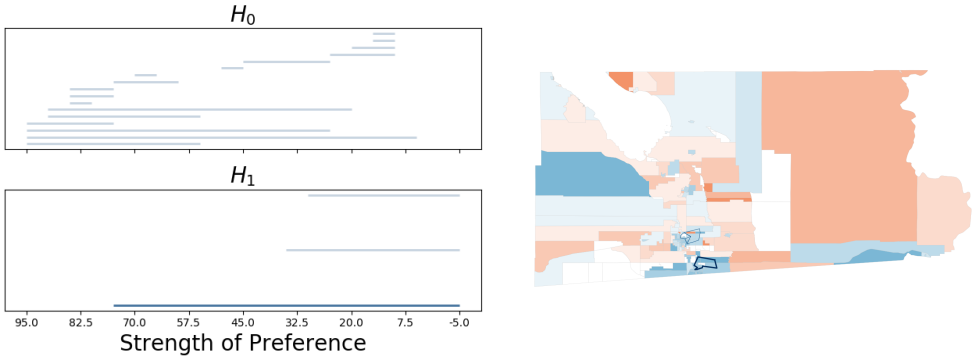
4.3.1 Comparison of our results to "ground truth"

We conclude our case study of voting data with some discussion of the accuracy with which we are able to use long-persistence features to find true features in the *LA Times* voting data. In Table 4.4, we show the proportion of long-persistence features that indicate an actual hole, as determined by the human eye. We highlight the most successful method for each county in bold. We see that our adjacency and level-set approaches outperform the VR and Alpha constructions. This indicates that our methods are less likely than the traditional distance-based approaches to detect noise as significant features in these examples.

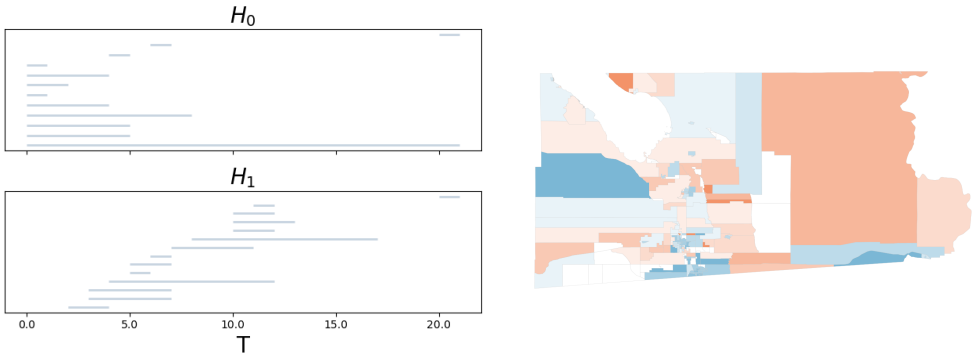
For source code and more computations, see https://github.com/mhcfeng/precinct.



(a) VR complex



(b) Adjacency complex



(c) Level-set complex

Figure 4.11: Barcodes and generated loops for blue precincts in Imperial County. The VR complex results in several false "features". The adjacency complex detects two white holes and one red hole. The level-set complex is unable to detect any holes that start at time 0, because there do not exist sufficiently large white or red holes in the first step of the filtration.

Table 4.4: Proportion of long-persistence features that identify a real feature in simplicial complexes based on voting data. Bold text shows the method with the highest proportion of long-persistence features that correspond to a voting island. The adjacency and level-set methods perform very well, with most counties detecting only "true" voting islands.

County	VR		Alpha		Adjacency		Level-set	
	C	T	C	T	C	T	C	T
Alameda	–	0.00	**1.00**	–	**1.00**	–	**1.00**	–
Alpine	–	–	–	–	–	–	–	–
Amador	–	**1.00**	–	–	–	–	–	–
Calaveras	–	**1.00**	–	–	–	–	–	**1.00**
Colusa	–	**1.00**	–	–	–	–	–	**1.00**
Contra Costa	–	0.00	0.00	–	**1.00**	–	**1.00**	**1.00**
Del Norte	–	0.00	–	–	–	**1.00**	–	0.00
El Dorado	0.00	**1.00**	–	–	**1.00**	**1.00**	–	**1.00**
Fresno	–	–	0.00	0.00	0.66	0.00	–	**1.00**
Glenn	–	0.00	–	–	–	0.00	–	**1.00**
Humboldt	0.00	0.00	–	–	0.50	–	**1.00**	**1.00**
Imperial	0.20	**1.00**	–	–	**1.00**	**1.00**	–	**1.00**
Inyo	–	0.00	–	–	–	**1.00**	–	–
Kern	–	–	0.00	**1.00**	**1.00**	**1.00**	–	**1.00**
Kings	0.00	0.00	–	–	**1.00**	0.67	–	0.87
Lake	**1.00**	0.00	–	–	–	–	**1.00**	–
Lassen	–	**1.00**	–	–	–	–	**1.00**	**1.00**
Los Angeles	–	–	0.00	0.00	–	–	**1.00**	–
Madera	**1.00**	**1.00**	–	–	**1.00**	**1.00**	–	**1.00**
Marin	–	–	**1.00**	–	**1.00**	–	**1.00**	–
Mariposa	–	**1.00**	–	–	–	–	–	–
Mendocino	–	0.00	**1.00**	–	**1.00**	–	**1.00**	–
Merced	0.11	**1.00**	–	–	0.5	**1.00**	–	**1.00**
Modoc	–	0.00	–	–	–	–	–	–
Mono	0.00	–	–	–	–	–	–	–
Monterey	–	0.00	0.00	–	**1.00**	0.00	**1.00**	**1.00**
Napa	0.25	0.00	–	–	**1.00**	–	.75	–
Nevada	0.00	**1.00**	–	–	**1.00**	**1.00**	**1.00**	**1.00**
Orange	–	–	0.00	0.00	0.00	0.50	**1.00**	**1.00**
Placer	0.50	–	–	0.00	–	**1.00**	**1.00**	**1.00**
Plumas	–	**1.00**	–	–	–	**1.00**	–	**1.00**
Riverside	–	–	0.00	.33	**1.00**	**1.00**	**1.00**	**1.00**
Sacramento	–	–	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**
San Benito	**1.00**	0.00	–	–	**1.00**	–	–	**1.00**
San Bernardino	–	–	0.00	0.00	–	0.75	–	**1.00**
San Diego	–	–	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
San Francisco	–	–	0.00	–	**1.00**	–	**1.00**	–
San Joaquin	–	–	0.00	0.00	0.75	**1.00**	**1.00**	**1.00**
San Luis Obispo	0.00	0.14	–	–	**1.00**	**1.00**	–	**1.00**
San Mateo	–	–	**1.00**	–	**1.00**	–	**1.00**	–
Santa Barbara	–	**1.00**	0.00	–	.67	**1.00**	–	**1.00**
Santa Cruz	–	–	**1.00**	–	0.00	–	**1.00**	–
Shasta	–	0.00	–	–	–	**1.00**	–	–
Sierra	–	–	–	–	–	–	–	–
Solano	0.00	0.00	–	–	**1.00**	**1.00**	**1.00**	**1.00**
Sonoma	–	0.00	0.00	–	**1.00**	–	**1.00**	–
Stanislaus	0.00	0.00	–	–	**1.00**	**1.00**	–	**1.00**
Sutter	–	0.00	–	–	–	**1.00**	–	**1.00**
Tehama	–	0.00	–	–	–	**1.00**	–	–
Trinity	–	0.00	–	–	–	0.00	**1.00**	**1.00**
Tulare	0.00	–	–	0.00	–	**1.00**	–	**1.00**
Tuolomne	–	0.00	–	0.00	–	**1.00**	–	**1.00**
Yolo	0.00	–	–	–	**1.00**	**1.00**	0.00	0.00
Yuba	–	0.00	–	–	–	**1.00**	–	**1.00**

4.4 Classification of cities based on their street networks

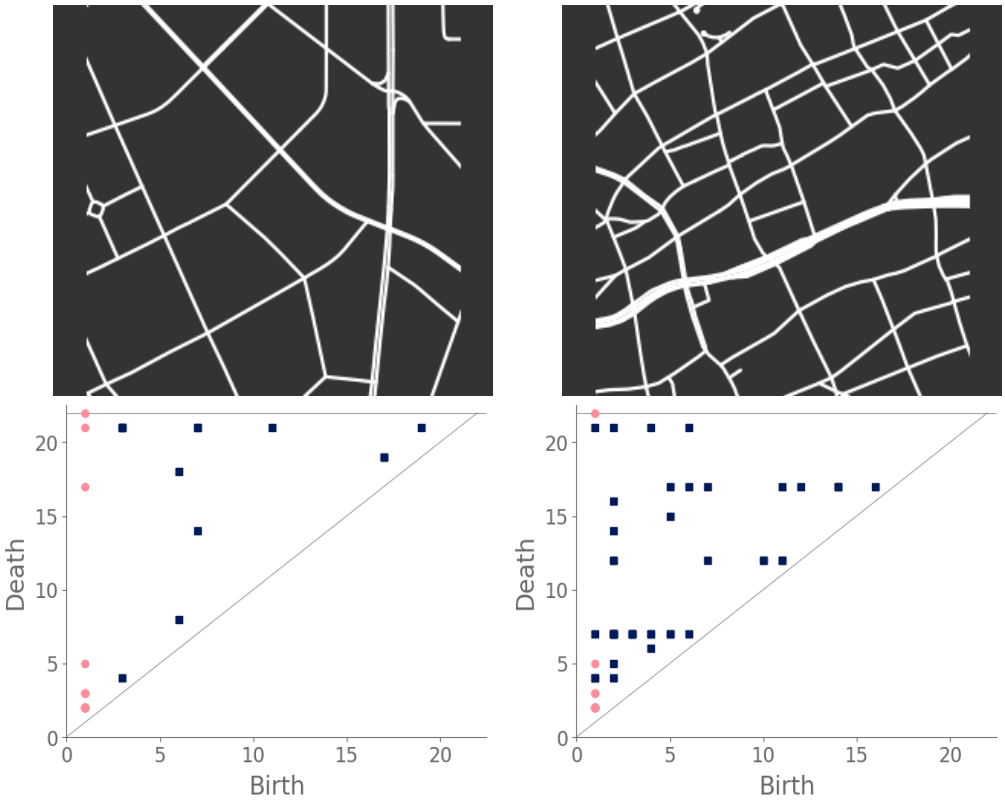
The field of urban analytics has grown rapidly in the last several years [Bar18,Bar19,Pum20]. To give one example that is of central interest to us, increasingly powerful computational tools have allowed researchers to characterize cities in terms of their street networks [Boe19a]. Indeed, a variety of tools from network analysis have been applied to the study of urban street networks [Boe18,Bar17,CSL06,Bat17]. In the present section, we use city street networks as base manifolds for constructing level-set complexes, and we thereby characterize cities based on their PHs. We use these PHs to compare city morphologies both within a single city and across a variety of cities.

We use our level-set construction to obtain topological descriptors in the form of persistence for city street networks. We then use these city persistences to compare (1) different regions of the same city and (2) different cities to each other. We obtain all of our city street networks with the software package OSMnx [Boe17] using latitude–longitude coordinates and taking a 1 km block that is centered at specified coordinates. In each example below, we will indicate how we choose these coordinates.

The initial filtration of the filtered simplicial complex that results from our level-set PH construction consists only of the streets in a network. As we increase the filtration time, we slowly add city blocks to the complex, and the topology changes as those blocks are filled in. More regular city blocks are more likely to be filled in without creating any new homological features, and larger blocks take longer to be filled in. Our construction is thereby able to capture information about the size and regularity of city blocks. The existence of dead ends tends to lead to the "pinching" of blocks into multiple homological features — as dead ends expand, they lengthen and eventually meet with nearby streets, cutting through blocks in the process — so our approach also yields information about dead ends.

4.4.1 Comparing different regions of the same city

We sample 169 points from the city of Shanghai using a SHAPEFILE of Shanghai's neighborhood boundaries that we downloaded from ARCGIS [Son17]. From the SHAPEFILE, we obtain a bounding box for each neighborhood. We sample uniformly within this bounding box, discarding points that do not lie within the polygon neighborhood geometry that is defined in the SHAPEFILE. The sampling ends when we reach the desired number of points. In total, we sample ten points from each administrative district, and we also include nine historical landmarks with coordinates from Google Maps [Goo19]. In Figure 4.12, we show maps and their associated PDs for two examples.



(a) (b)

Figure 4.12: Two sampled street networks from (a) PuDong New District and (b) ZhaBei district. [We generated both maps using OSMNX [Boe17].]

After computing PH (in the form of a PD) for each map, we compute the bottleneck distance between each pair of maps. Bottleneck distance (see Definition 15) is a metric that is defined on the space of PDs. It gives the shortest distance d for which there exists a perfect matching between the points of the two PDs and all diagonal points, such that any pair of matched points are at most a distance d from each other, where we use the sup norm in $\mathbb{R}^2$ to compute the distance between points. Once we have pairwise bottleneck distances between PDs, we perform average-linkage hierarchical clustering into three clusters. (We chose to have three clusters based on looking at the dendrogram.) In Figure 4.13, we show the sampled points (which we color according to their cluster). We observe that the three clusters consist largely of historical neighborhoods ("City center"), concession-era neighborhoods ("Transition"), and modern neighborhoods ("New construction"). In Figure 4.14, we show administrative districts along with the year that they were constructed. We break them down by the percentage of the sample points that are in each cluster.

4.4.2 Comparing street networks from different cities

We continue our analysis of cities by characterizing and comparing the structures of street networks of 306 cities across the globe. We downloaded latitude and longitude coordinates from SimpleMaps [Sim19] and selected all cities with a population at least 1.5 million people. Given these latitude and longitude coordinates, we use OSMnx [Boe17] to obtain street networks. We then compute PH for each city and cluster their PDs using average-linkage hierarchical clustering with three clusters. We sometimes refer to a city in a given cluster as a city of a certain "type". Our results depend on the specific latitude and longitude coordinates in our downloaded data set. Accordingly, our results are influenced by the particular location of a city's coordinates, which are the standard ones in SimpleMaps.

In the following paragraphs, we describe our clusters of cities. We define "blocks" to be the cells of a planar street graph. Although our level-set construction for computing PH is not designed explicitly to characterize blocks, we take advantage of the fact that our level-set

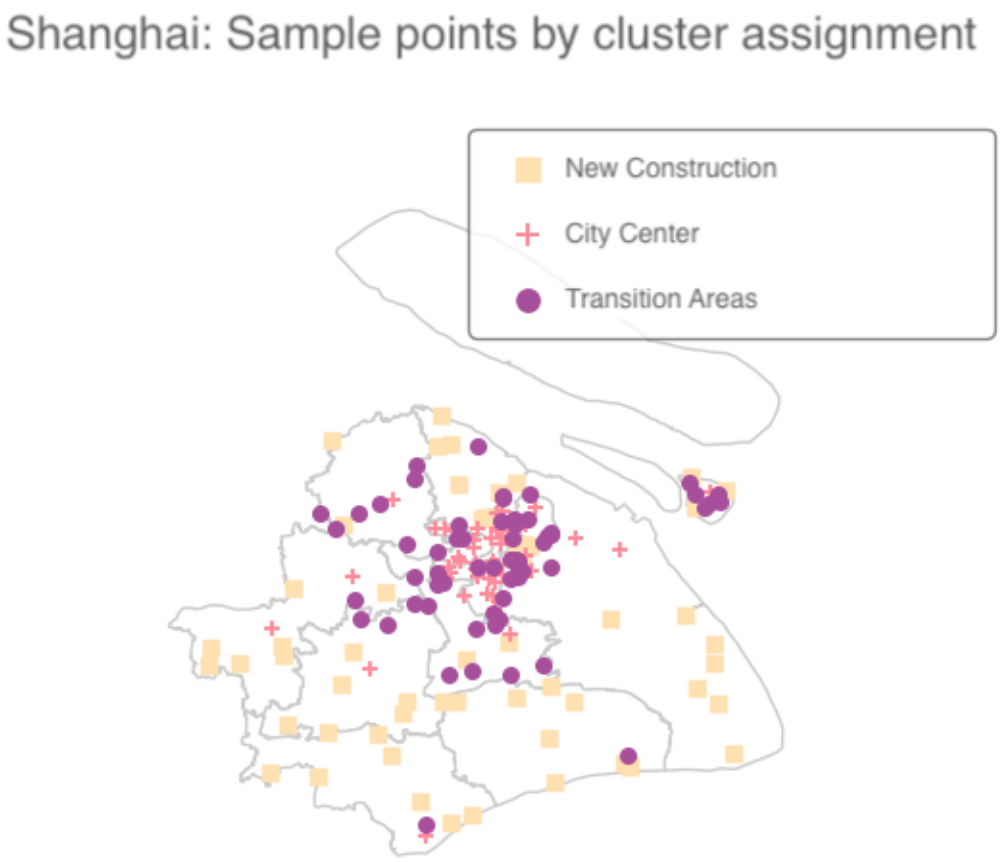


Figure 4.13: Sampled points in Shanghai. We color these points according to their cluster assignment from average-linking hierarchical clustering of neighborhoods of Shanghai into three clusters.

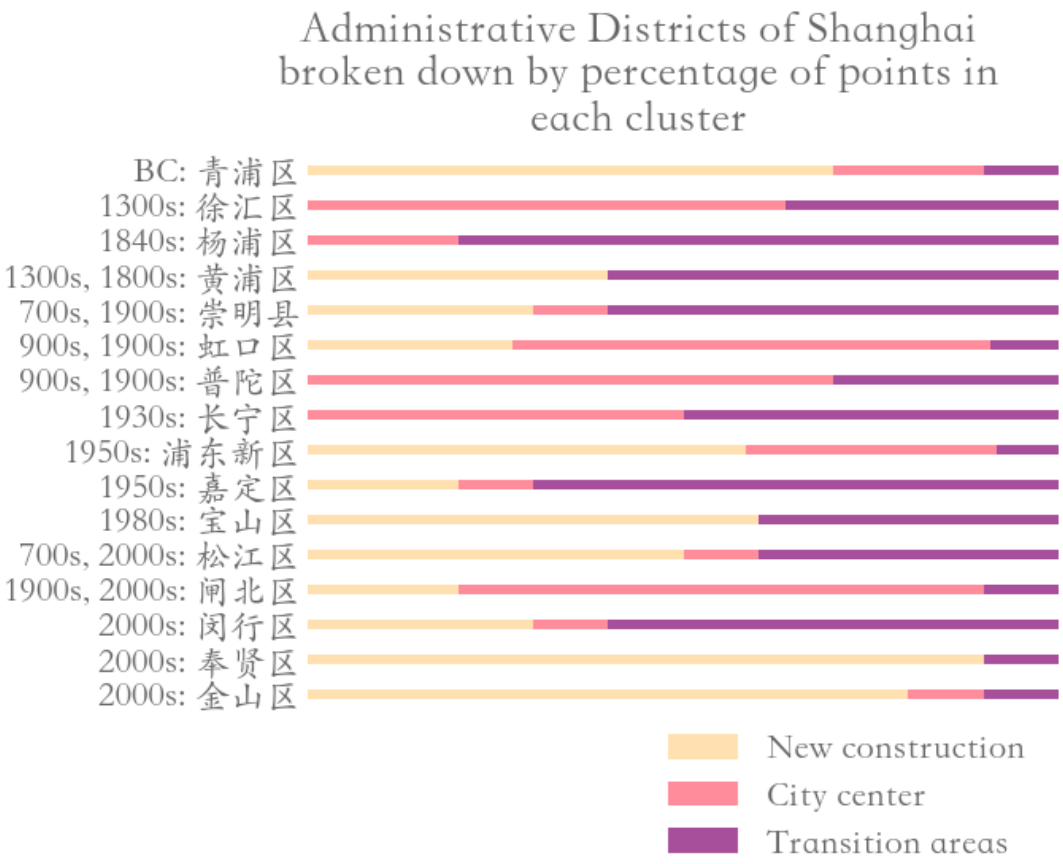


Figure 4.14: Breakdown of administrative districts in Shanghai into our three clusters. (We order the districts roughly by their year of development.) Most of the older districts have a larger percentage of points that are assigned to the “City center" cluster, whereas the points in the “Transition” cluster tend to occur in districts that included development in the 19th and early 20th centuries. The “New construction" cluster is the most common assignment for neighborhoods that were built in the 1950s or later.

construction takes the set of streets as its initial manifold. As the streets expand outward according to Equation (4.1), they fill in the blocks. Larger blocks take longer to fill in, and blocks fill in more evenly when they are closer to circular in shape. Roughly, we characterize block sizes based on the death times of H_1 features: small sizes correspond to early death times (specifically, less than 10), medium sizes correspond to death times between 10 and 15, and large sizes correspond to late death times (specifically, more than 15). We also designate blocks as "regular" (when they are close to a regular polygon) or "irregular" (for blocks that do not resemble a rectangle or some other convex polygon). If a block is very irregular, then as its streets expand, it is possible that narrow parts of the block shrink and close off, such that the block segments into smaller blocks. We refer to this phenomenon as "pinching". Our three main clusters are dominated by (1) gridlike cities, (2) cities with gridlike patches that are interspersed with larger, non-gridlike blocks, and (3) cities that have a large number of non-gridlike structures (specifically, dead ends or large holes) that interrupt other structures. We use the term "interrupted grid" for cities that are either mostly gridlike with some patches that are not gridlike or that consist of patches of disparate grids that are stitched together (with other features between them).

Our first major cluster has 99 cities and is dominated by cities with small, gridlike blocks. All regions of the world have some cities of this type, but North America has the largest percentage (relative to all of the cities that we sample from that continent) of these gridlike cities and Europe has the smallest percentage of them. The block sizes in this cluster tend to be small or medium, resulting in filtrations whose maximum filtration value tends to be small in comparison to cities in the other two clusters. In the PDs, we also observe that the distributions of death times of features in H_1 tends to be close to uniform and over a small range. Such distributions occur because these gridlike cities tend to have even distributions of block sizes, even though they include some areas with slightly smaller and/or slightly larger grid sizes. They do not have large blocks, so they do not have features in H_1 with late death times.

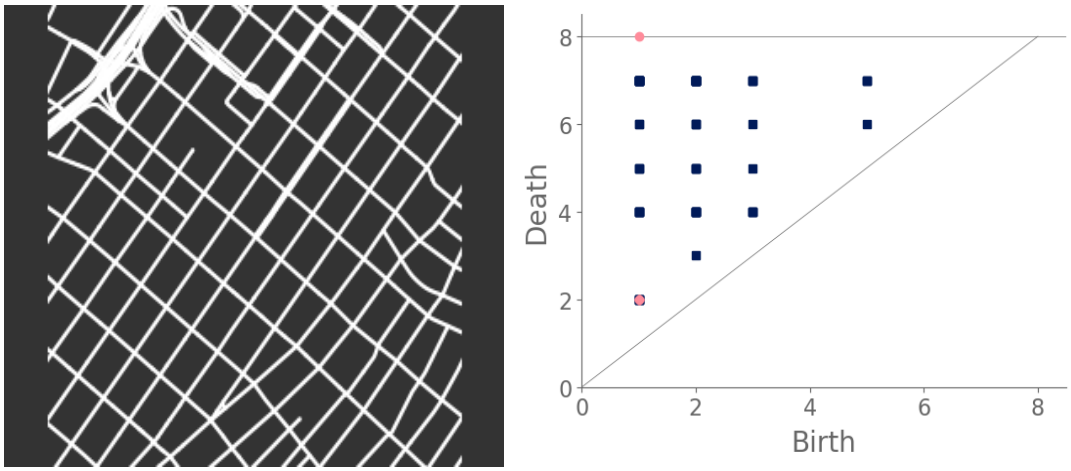


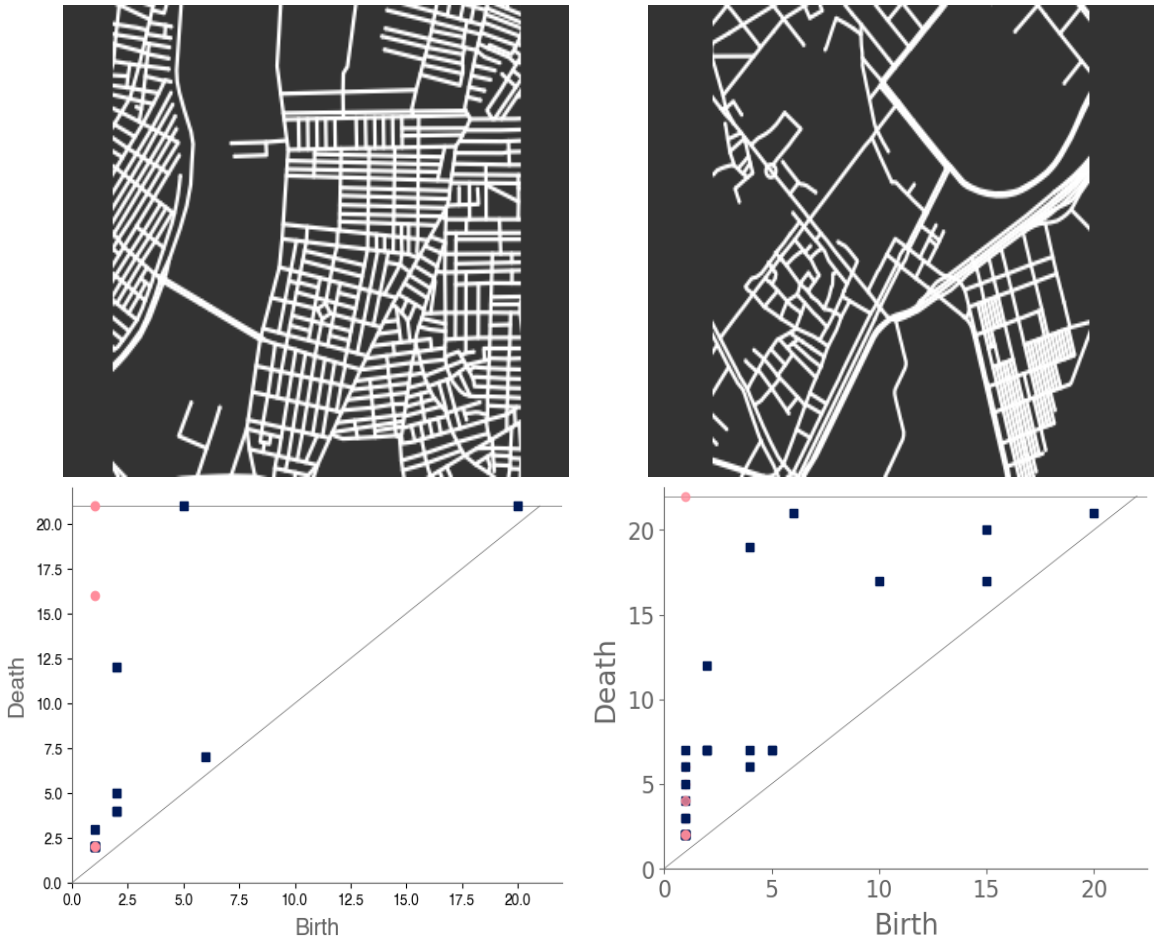
Figure 4.15: Cities in our first major cluster have gridlike street layouts. One example of such a city is Los Angeles, which we show in this figure. We show its street network on the left and its associated PD on the right.

Our second major cluster has cities with patches of grids that are interspersed with structures that are not gridlike. This cluster, with 149 cities, is the largest of our three clusters. The PDs in this cluster tend to have larger maximum death times than the PDs for the cities in our first cluster. In the PDs, gridlike blocks yield collections of features in H_1 with early death times, and the larger, non-gridlike structures yield features in H_1 with late death times. The non-gridlike areas in these cities do tend to have fairly regular shapes, resulting in a relatively small number of features in H_1 with late birth times. Such late-birth-time features usually correspond to the pinching of blocks, which can occur either via dead ends or via shape irregularities. By examining the dendrogram from our hierarchical clustering, we can further separate the second cluster into two subclusters, which we show in Figure 4.16. The first of these subclusters consists mostly of cities that have large patches of gridlike structure, with a small number of large blocks that interrupt the grids. The PDs for cities in this subcluster tend to have a large number of features in H_1 with early death times, and they tend to have only a small number of isolated features in H_1 with late death times. The second subcluster of our second major cluster consists mostly of cities with

small patches of grids that are mixed with large irregular blocks. The PDs for cities in this subcluster tend to have a larger number of features in H_1 with late death times than is the case for the cities in the other subcluster of cluster two.

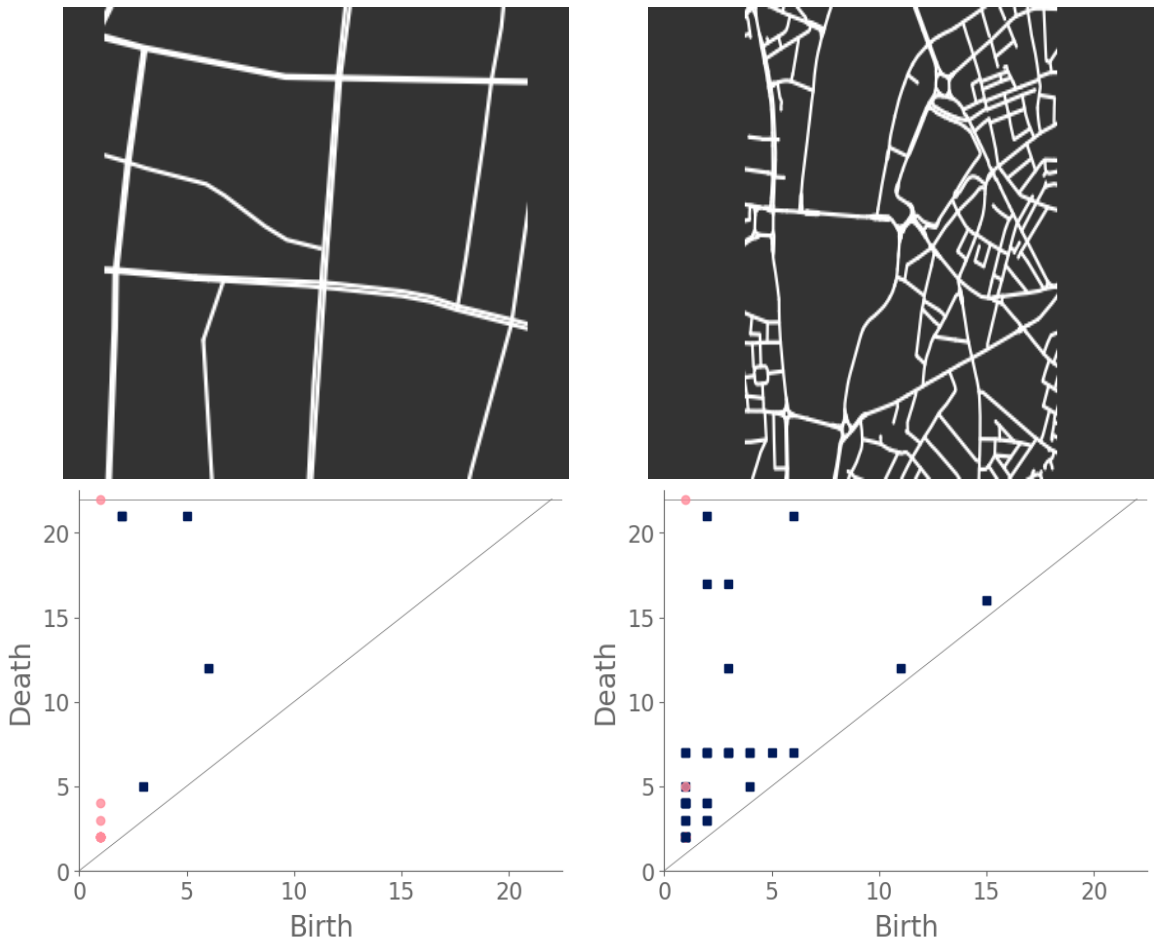
Our third major cluster, with 58 cities, consists of cities with a large number of non-gridlike structures. In particular, many of these cities have a large number of dead ends, rectangular blocks that are not arranged in a grid, or both. Some of our observations include streets that do not continue through particular blocks (e.g., there is a street, it is obstructed, but then it continues after the obstruction), which leads to a mixture of block sizes even in areas of a city that tend to have regular blocks. We refer to these situations as "obstructions". The PDs of the cities in this cluster have a larger number of features in H_1 with medium death times (specifically, in the range 10–15), and many of these features are close to the diagonal. This is common when large blocks are pinched into several regions, as the smaller regions are born at the pinching time, rather than near the beginning of the filtration. Therefore, they do not survive long enough to have a late death time. By examining the dendrogram from our hierarchical clustering, we see two clear subclusters. However, one of these subclusters consists of only two cities (Beirut and Nanyang). The PDs of both of these cities are dominated by two features in H_1 with late death times, and they also have several features in H_1 with medium death times. In Figure 4.17, we show examples of cities in cluster three.

We color our cities according to their major cluster and show them on a world map in Figure 4.18. In Figure 4.19, we show the breakdown of cities from each continent into the various clusters. We calculate PH for only four major cities in Oceania, so we cannot draw strong conclusions from the cluster assignments of those cities. Among the other regions, we observe that North America has the largest proportion of cities with gridlike street layouts and the smallest proportion of cities with non-gridlike layouts. By contrast, Europe has the smallest proportion of cities with gridlike street layouts. This is consistent with the common perception that North American cities are much more gridlike than European cities. In all



(a) (b)

Figure 4.16: Cities in our second major cluster have patches of gridlike structure that are mixed with large blocks. As examples of cities in this cluster, we show (a) Aleppo and (b) Barcelona. We show their street networks in the top row and their associated PDs in the bottom row. Aleppo illustrates the idea of having holes in a large grid and is an example of a city in the first subcluster of cluster two. Barcelona, which is in the second subcluster, is an example of a city with small patches of gridlike structure.



(a) (b)

Figure 4.17: Examples of cities in our third major cluster include (a) Nanyang and (b) London. We show their street networks in the top row and their associated PDs in the bottom row. Cities in our third major cluster include dead ends, irregular blocks, and obstructions. This leads to a large range of block sizes and hence to features in H_1 that have medium death times. Such features are rare in the other two major clusters. For example, Nanyang has several streets with obstructions, and London has dead ends and a broad distribution of block sizes.

regions, we also observe that a large fraction of the cities are interrupted grids. Additionally, we observe that South America, Africa, and Asia have similar distributions of city types.

Interestingly, from the map in Figure 4.18, South America, Asia, and Africa appear to have areas that are dominated by specific major clusters. We observe non-gridlike cities in the northern part of South America, whereas we see gridlike cities along its east coast. In Africa, most of the non-gridlike cities occur along the western coastline. In Asia, we see few non-gridlike cities throughout most of Southeast Asia. Across the map, there appears to be a potential equatorial band of non-gridlike cities. We do not have an explanation for these patterns, but they are fascinating and seem worthy of future research efforts.

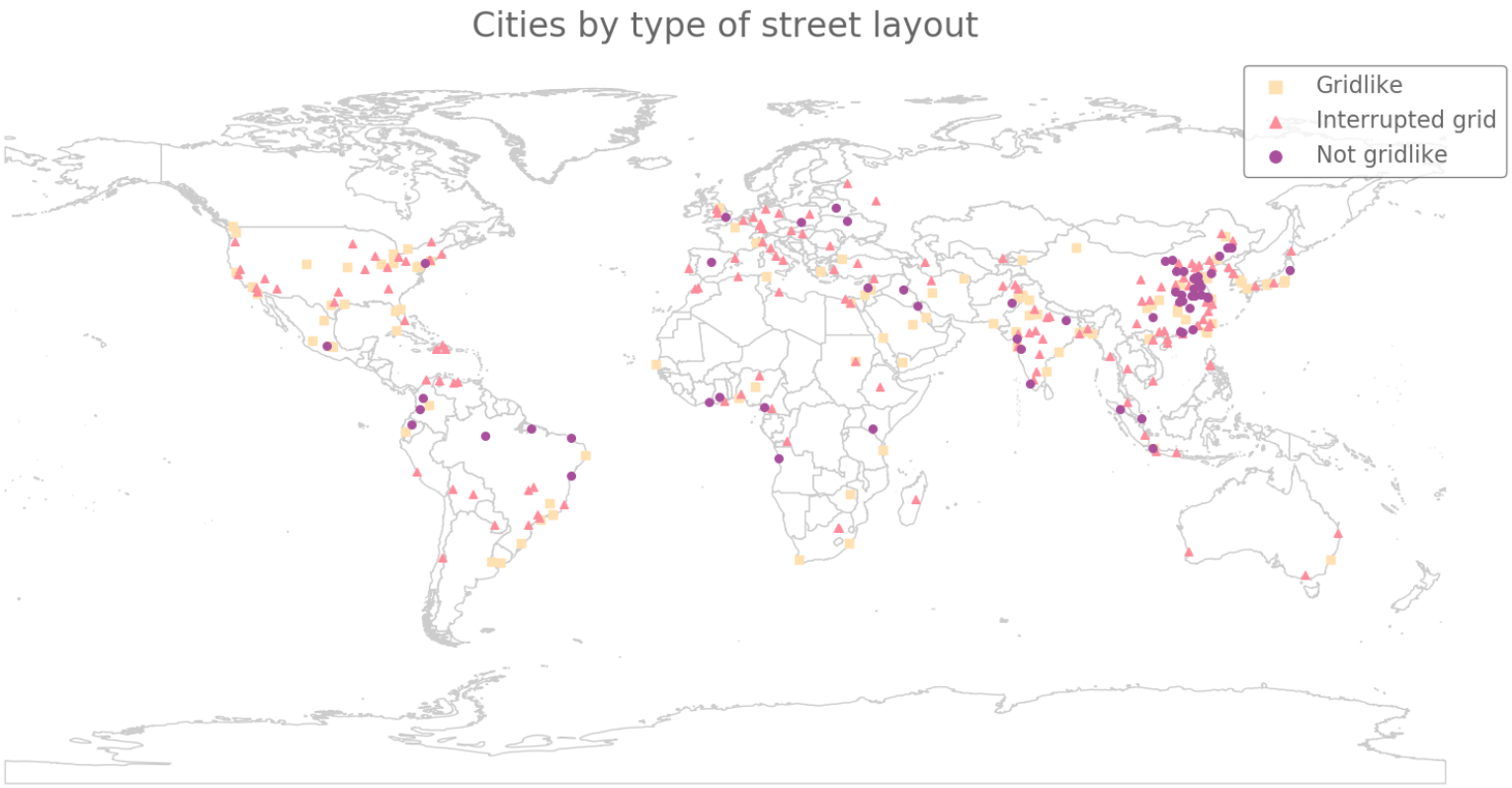


Figure 4.18: Cities colored by their cluster assignments from average-linkage hierarchical clustering of cities into three clusters. [The SHAPEFILE of the world map is from [Bel15].]

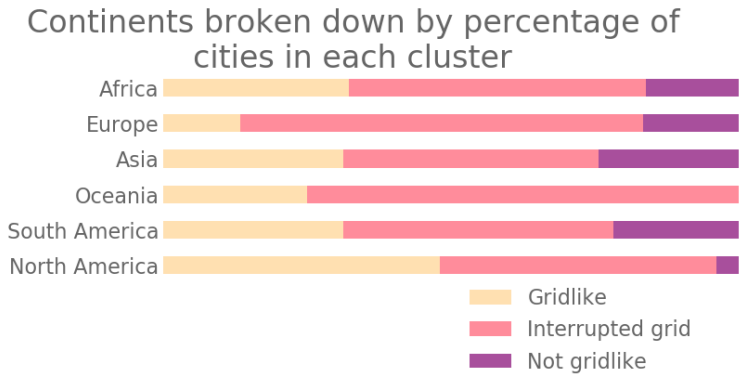


Figure 4.19: Continents broken down based on the distribution of cities into our three major clusters.

4.4.3 Comparison of our classification to that of Louf and Barthelemy [LB14b]

Classifying cities based on street networks is an area of interest in urban analytics. Street networks can affect congestion [LB14a], accidents [MG11,TSW20], and property crime [BBB94]. Prior papers have sought to characterize street networks based on network features [HSY20], machine-learning methods [TSW20], orientation [Boe19b], and other approaches.

We compare our results to the city classification of Louf and Barthelemy [LB14b], who associated each city with a conditional probability distribution that captures the areas and shapes of its blocks. We choose their method as a point of comparison because they studied a wide range of cities and (like us) codified cities from a block-based perspective. They used the word "fingerprint" as a monicker for their block-based representation of cities. In our method, we codify cities according to their PHs, which we generate using the level-set construction of Section 4.1.2. Both the approach of [LB14b] and our approach capture information that is based on city blocks, although our PH representation differs substantially from the fingerprints of [LB14b].

Louf and Barthelemy clustered cities into four groups, whereas we have chosen to cluster our cities into three groups. In [LB14b], European and North American cities largely

inhabit the same cluster (group three in [LB14b]), but they appear in distinct subclusters, demonstrating that there is a substantive difference between cities from the two regions. Our method finds that North America has largest proportion of cities with gridlike streets among all of the regions and that Europe has the smallest proportion of such cities.

In contrast to the above situation, Africa, Asia, and South America have a fairly balanced composition of city types, with a potential equatorial band of non-gridlike cities. Louf and Barthelemy observed several clusters (groups one, two, and four in [LB14b]) that occur predominantly in Africa, Asia and Oceania (which they combined into one entity), and South America. Notably, none of our clusters are as dominant as group three (which they described as having heterogeneous block sizes and shapes) in [LB14b], although we do observe that our cluster of cities with interrupted grids (such cities are characterized in part by their heterogeneous block sizes) is also our largest cluster.

Now that we have compared our results to those of [LB14b], we briefly compare and contrast the types of information that the two methods can capture. Recall that our level-set construction for PH generates filtered simplicial complexes that first consist of streets and then expand outward to absorb the blocks between them. The PH of these filtered simplicial complexes thereby gives a low-dimensional representation of the original image of a city street network. Because irregularly shaped blocks are absorbed into the surrounding streets at a different pace than regular blocks, we capture information about the regularity of each block. Louf and Barthelemy's method also uses information about the regularity of block shape. See Equation (3.2) in [LB14b] for a precise mathematical statement of how they measured the regularity of blocks. It is related to so-called "compactness measures" [Gil02] used in the study of gerrymandering [BS18, DT18], which compare the area of a shape to the area of a circle in which the shape is circumscribed.

Because the original image of a street network includes information about the spatial relationships between blocks, the PH that results from our approach also encodes some of this information. By contrast, Louf and Barthelemy's fingerprints do not encode information

about the spatial relationships of blocks to each other. Additionally, our method captures information from dead ends, which Louf and Barthelemy discarded.

Overall, although both our approach and that of [LB14b] use a block-based representation to characterize cities, there are subtle differences in the way that the two approaches encode block information. The commonality of a block-based perspective results in some similarities. For example, the clusters that result from the two approaches seem to be based heavily on block size and regularity. However, our approach appears to prioritize spatial relationships between different clusters of blocks (specifically, whether blocks are arranged in a grid); such information is not captured in the approach of [LB14b]. Consequently, the two approaches capture different city morphologies, and we expect them to be useful as complementary techniques for studying structures in spatial complex systems.

4.5 Scientific images

4.5.1 Spiderwebs

In this section, I discuss an application to the topology of spiderwebs from [FP20c]. See Section 2.4.1 for background on the data set that I used in this case study.

We use five images from the NASA technical briefing [NCR95] and two images from Witt [Wit71] of various webs that were spun by spiders under the influence of a variety of psychotropic substances, apply a level-set construction to compute PH, and perform average-linkage hierarchical clustering to yield the dendrogram in Figure 4.20. We show the images of the spiderwebs and their associated PDs in Figure 4.21.

Our classification places the drug-free spider into its own cluster. The drug-free spiderweb is characterized by a clear central hole, threads radiating outward at approximately even intervals, and completed rings of threads that surround the center. We place marijuana, peyote, and LSD in the same cluster. In these webs, there is a clearly identifiable center,

and most of the radial threads are evenly-spaced, straight, and radiate outward directly from the center. However, for the webs in this cluster, rings of threads are either difficult to see or are incomplete. The final cluster consists of chloral hydrate, caffeine, and speed. In the caffeinated spider's web, one cannot even clearly identify a center[1]. One can locate a center in the webs of the spiders that were under the influence of speed or chloral hydrate (a sedative that is used in sleeping pills), but many of the radial threads do not join the center and some of the radial threads are not straight. Almost no complete rings of thread are visible in any of the three webs in this cluster.

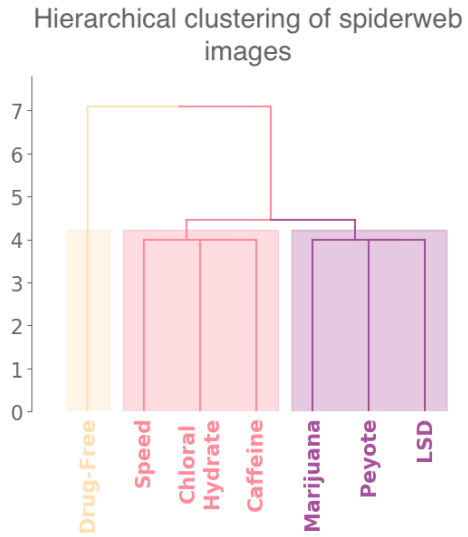


Figure 4.20: Classification of webs that were spun by spiders under the influence of various psychotropic substances.

4.5.2 Snowflakes

In this section, I discuss an application to snowflake crystals from [FP20c].

We start with twelve different images (from [Lib19]) of snowflakes with different crystalline structures. (See Figure 2.5 in Section 2.4.2.) From these images, we compute level-set

[1]The web that was produced by the caffeinated spider is always fun to point out when giving presentations.

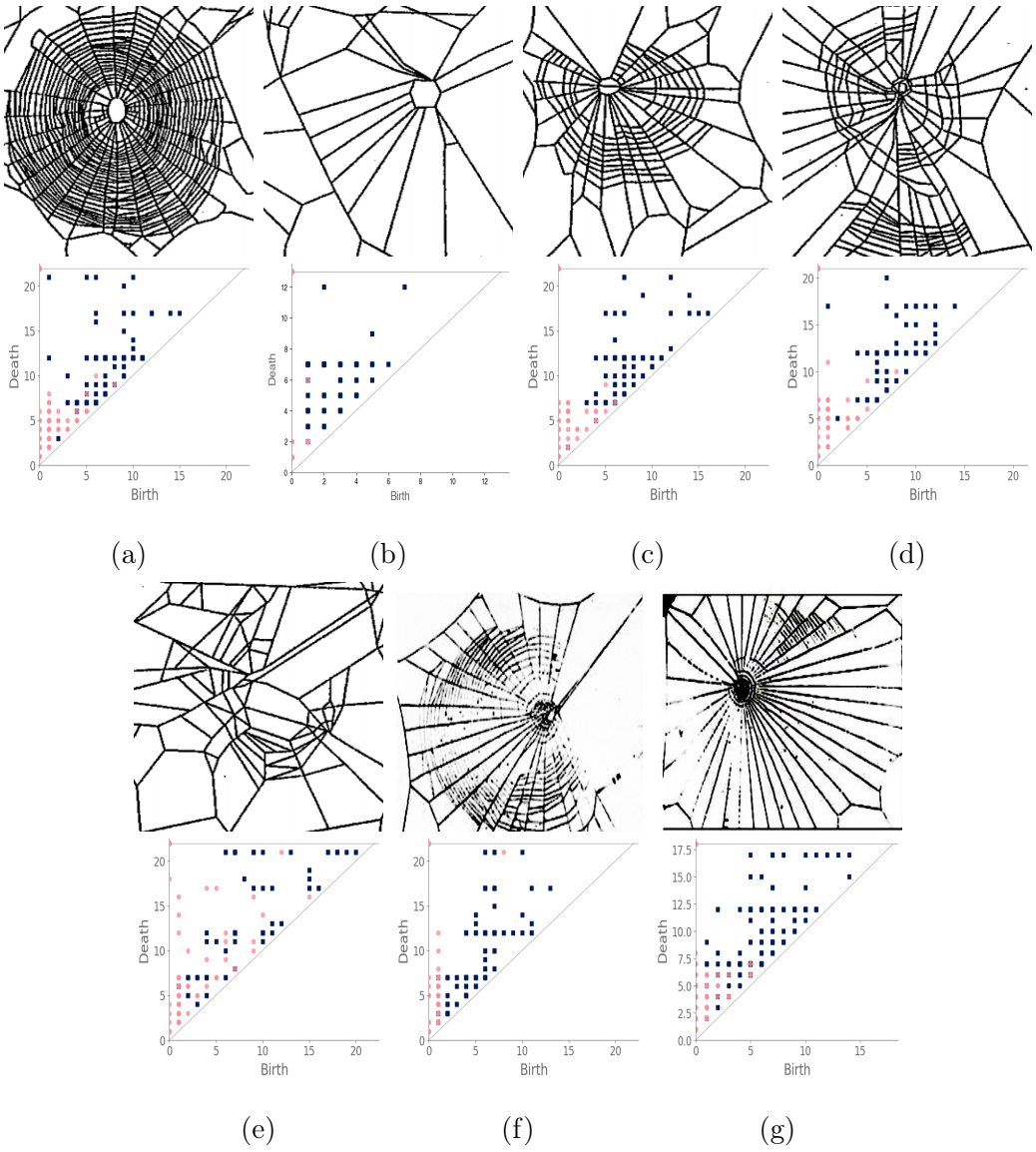
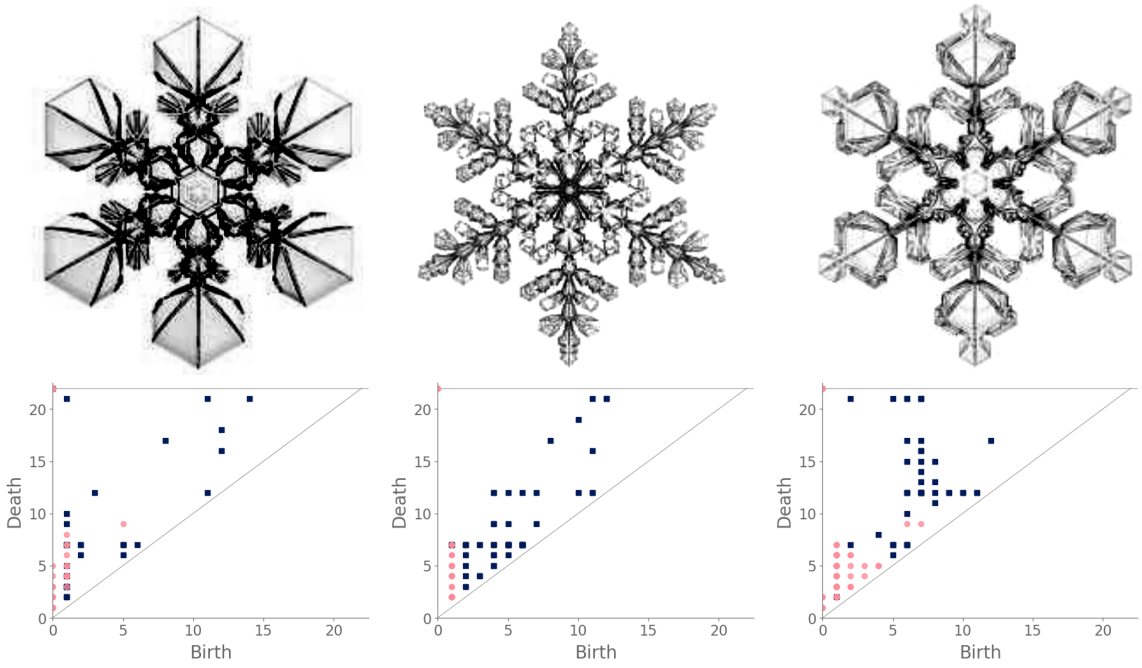


Figure 4.21: Webs spun by (a) drug-free spiders, compared with webs spun by spiders that were under the influence of (b) chloral hydrate (sleeping pills), (c) marijuana, (d) speed, (e) caffeine, (f) peyote, and (g) LSD. [The images for panels (a)–(e) are from [NCR95], and the images for panels (f) and (g) are from [Wit71].]



(a) (b) (c)

Figure 4.22: Snowflakes can have a variety of crystalline structures, as we illustrate with (a) Snowflake A, (b) Snowflake B, and (c) Snowflake D. We show the snowflake structures in the top row and their associated PDs in the bottom row. We show the structures of our full set of snowflakes in Figure 2.5. [The images in the top row are from [Lib19].]

complexes and PHs. We then perform average-linkage hierarchical clustering on the PDs to produce the dendrogram in Figure 4.23.

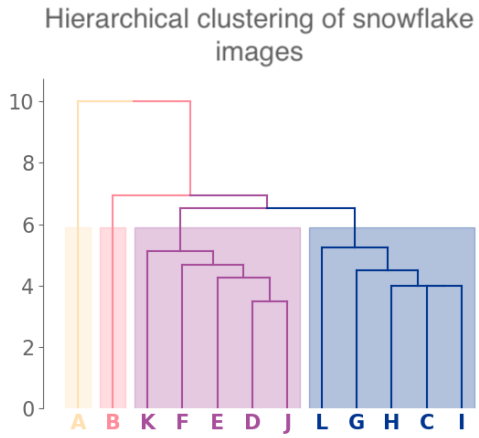


Figure 4.23: Dendrogram from clustering the snowflakes in Figure 2.5.

The images of snowflakes consist of edges (the black lines in our images) and cells (the white spaces that are bounded by the edges). We refer to the outer areas that extend from the center of the snowflakes as their "points". The twelve snowflakes have fairly regular crystalline structures, so our computation of PH predominantly records information about the distribution of cell sizes in a snowflake. The inherent hexagonal nature of snowflakes and the regularity of their crystalline structures largely overwhelms our ability to use PH to glean information about their spatial relationships and irregularities.

Examining the clusters (see Figure 4.23) reveals that snowflake A and snowflake B each reside in their own cluster, and the remaining snowflakes split into two more clusters. Snowflake A's PD (see Figure 4.22a) is dominated by a feature in H_1 with an early birth and a late death. (See the point at the top-left corner of the diagram.) This arises from the large feature that is formed by the bold ring around the center of the snowflake. None of the other snowflakes have a bold central ring. More generally, we observe few features in the PD for snowflake A. By contrast, snowflake B's features are largely close to the diagonal (see Figure 4.22b) because the initial manifold of the snowflake does not have large holes. Notably,

we do not observe any points in the top-left region of its PD. The cell sizes in snowflake B are smaller than those in most of the other snowflakes, and even its central ring structure includes a large number of small holes. The remaining snowflakes either have more large holes than snowflake B, or they do not have a bold central ring like the one in snowflake A. Note that Snowflake B has a PD that is much closer than that of Snowflake A to those of the other snowflakes.

CHAPTER 5

Using topology to study neighborhood segregation

This chapter contains research from my work at the 2018 Voting Rights Data Institute. I worked in collaboration with Eion Blanchard (University of Illinois Urbana-Champaign), Moon Duchin (Tufts University), Austin Eide (University of Nebraska, Lincoln), and Patrick Girardet (University of California, San Diego). The Voting Rights Data Institute is a six-week workshop held by the Metric Geometry and Gerrymandering Group that focuses on collaborations around a variety of topics related to gerrymandering and voting access. Our project focused on using topological tools to attempt to detect and understand residential segregation. This work is not complete, but my collaborators and I explored several ideas for novel methods over the course of the project. I present them in the following sections. Much of the text in the following sections is extracted from my write-up on the project.

5.1 Segregation

Racial segregation is an important phenomenon with far-reaching implications across various facets of society. In this work, we focus on racial segregation in America. Because of its continued relevance in policymaking and social justice, racial segregation has been a topic of broad study in social science, demography, economics, and various other fields [LPG07, LS84, DC20]. While these studies have yielded many insights into specific factors that lead to segregation and its effects on social outcomes for various communities, there continue to be many questions about the mechanisms that drive neighborhood segregation. In our study, we present a topological framework with which we can study racial segregation and show

that it is able to capture properties of racial segregation in various United States cities. This framing of racial segregation allows us to apply powerful topological tools to the study of neighborhood formation and community development.

Racial segregation has been studied for decades in social science and history. Some studies have examined the causes of racial segregation [Sei98, Rot17], and others have examined the impacts of segregation on education [Moo01], economic outcomes [RM10], crime [PK93], and more. Various computational tools have been developed for measuring and understanding segregation [MD88, LPG07]. In addition to statistics and measures of residential segregation, various models have been developed to study the dynamics of segregation [Zha11, CMR08, Cha03, Sch71] While these measures and models have contributed to understanding the mechanisms of racial segregation, we believe that there is value in studying segregation and neighborhood formation in terms of topological obstructions in an appropriate space.

In the context of neighborhood segregation, we can interpret the boundaries between neighborhoods as obstructions. These obstructions have both demographic and spatial elements. For example, in the city of Chicago, the border between Hyde Park and Kenwood separates two adjacent neighborhoods with very different demographics (due mostly to the location of University of Chicago in Hyde Park). If we can build an appropriate topological-space approximation of the South Side of Chicago, we can detect this obstruction as a topological feature. In this project, our goal is to find general ways of building topological-space approximations whose topological features correspond to neighborhood boundaries. This places the problem of racial segregation into a topological setting, allowing us in turn to apply topological tools like homology to studying the structure of racial segregation.

We focus on building topological spaces that capture demographic and spatial properties of interest. We then probe these spaces with PH to detect topological obstructions. We propose two novel constructions, which we then demonstrate are capable of capturing boundaries of racial segregation. We then perform a cross-sectional study to various U.S. cities to demonstrate the viability of our methods.

5.2 Methods

To apply PH to a data set, the data must first be transformed into a suitable filtered complex. We use filtered simplicial complexes (in which the base units are collections of n-simplices) because of the availability of PH packages that take filtered simplicial complexes as an input. We must carefully select the construction of filtered simplicial complex to obtain meaningful results from a PH computation. I discuss the difficulty of constructing filtered simplicial complexes in Section 3.4.1, and I study the impact of making an appropriate choice in Section 4.3. In the following subsections, we describe two novel methods that we use to build these filtered simplicial complexes to capture the properties that we seek. Because of the reliance of neighborhood formation on both demographic composition and spatial factors, our methods incorporate both existing demographic and spatial data into our construction of filtered simplicial complexes.

5.2.1 Demographic edges

We use spatial data to build a graph for the area that we study, where graph adjacency is determined by rook adjacency of two census tracts. As discussed in Section 4.3 in the context of precincts, two census tracts are rook adjacent if they share an edge. This graph is nearly planar (non-planarity can result from a variety of map drawing degeneracies, but it affects a relatively small fraction of census tracts), and we compute the approximate planar faces of the graph [DCV08].

To construct a filtered simplicial complex, we begin by adding a 0-simplex (i.e., a vertex) for every census tract. We then add 1-simplices (i.e., edges) according to demographic similarity. We add 2-simplices (i.e., faces) whenever all of the edges of a planar face in the graph is in the simplicial complex. By adding 1-simplices according to decreasing demographic similarity, we create a filtered simplicial complex whose filtration parameter is demographic distance.

This particular construction yields what is essentially an approximation of demographic space, with a spatial component that helps determine some of the connectedness properties of the space. Accordingly, one should interpret the features of the homology of this filtered simplicial complex as obstructions in demography. In Section 5.3.1 we will discuss how to interpret this simplicial complex and its homology.

5.2.2 Spatial edges

In our second method for constructing a filtered simplicial complex, we also begin by building a network from a map and computing its planar faces. Once again, the 0-simplexes of this filtered simplicial complex are census tracts. However, we now determine the edges by spatial adjacency, and we fill in the "planar" faces only when all vertices of the face are sufficiently demographically similar. Our filtration parameter is again given by demographic similarity.

The primary difference between this method and the one in Section 5.2.1 is that this filtered simplicial complex is an approximation of physical space, rather than demographic space, with the demographic similarities determining connectedness. In a similar vein to our demographic-edge construction, we now interpret obstructions as existing in physical space. We further discuss this interpretation in Section 5.3.2.

5.3 Results

To illustrate the differences between our two choices of filtered simplicial complexes and their homologies, we begin by examining Washington, D.C. as a case study. In Section 5.3.3, we will discuss results on Chicago, but we begin with Washington, D.C. because of its small size (and ensuing fast computation time) and well-studied segregation patterns. In Figure 5.1, we highlight the majority black neighborhoods on the eastern side of the city. We use the ACS category "Black or African American" in the racial breakdown of each neighborhood. These segregation patterns were largely the result of housing discrimination and several waves of

large-scale migration into the city [DC20].

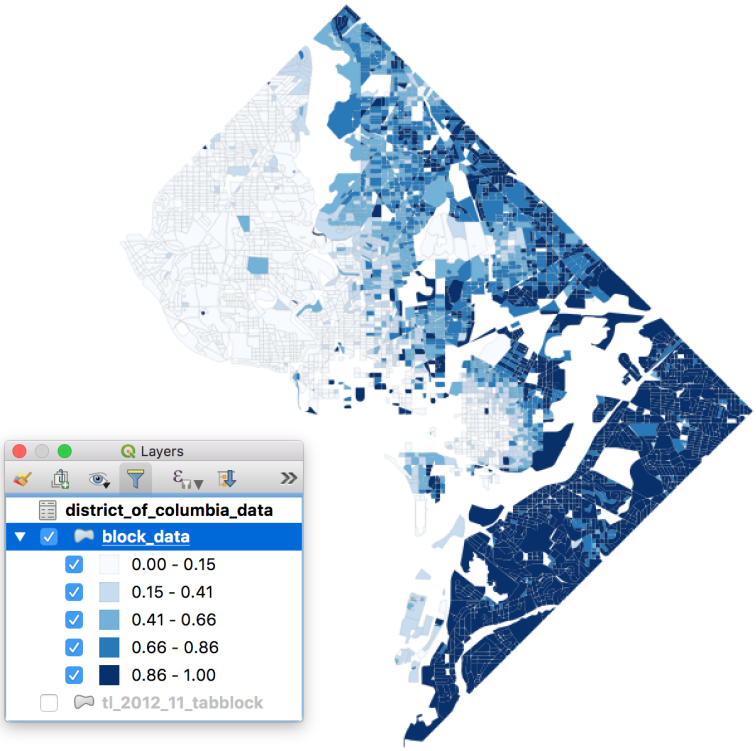
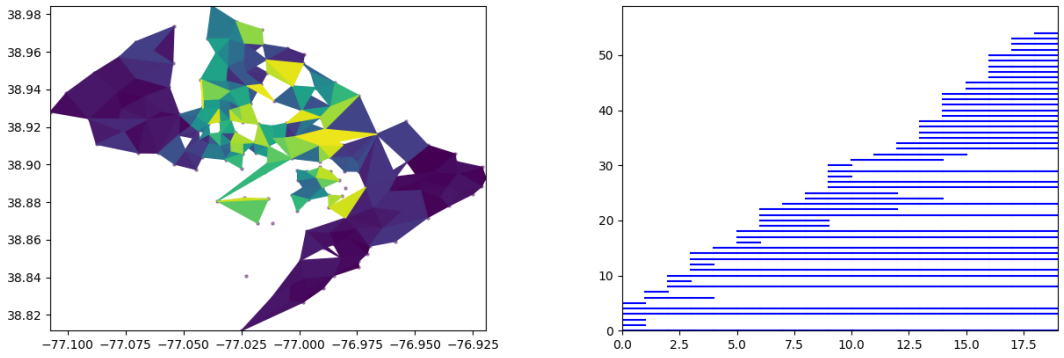


Figure 5.1: City of Washington, D.C. colored by the proportion of Black or African American residents. [This image was created by Eion Blanchard.]

5.3.1 Demographic edges and interpretations for a case study of Washington, D.C.

We first consider a filtered simplicial complex that we build with demographic edges, as described in Section 5.2.1. By using this filtered simplicial complex as a starting point for the PH package DIONYSUS [The17], we obtain the barcode showed in Figure 5.2. In this barcode, we observe a large number of topological features of various lengths. In general, the length of a particular bar does not necessarily indicate that the corresponding feature is an important one (as discussed in Section 3.4). However, our choice of filtered simplicial complex yields a specific interpretation for barcodes of various length. Longer bars correspond to larger

obstructions; in the demographic space that we examine, we can interpret these bars as describing the mixing properties of populations.



(a) (b)

Figure 5.2: We show (a) Filtered simplicial complex with demographic edges constructed from Washington, D.C. racial data and (b) its associated barcode. In (a), we color simplices that enter the filtration by time of entry, with the latest filtrations in the lightest colors. In (b), we observe many features in the PH. The longest bars correspond to simplices that form around the middle region of the city. From Figure 5.1, we observe that the Black or African American population of Washington, D.C. is concentrated mostly around the eastern portions of the city. The filtered simplicial complex captures the east/west divide in the demographics of the city. The highest-birth-time simplices (colored in bright yellow) span the East/West divide in areas where the proportion of Black or African American residents changes gradually across census tracts.

Each feature in the barcode in Figure 5.2 corresponds to a topological cycle that is not a boundary. Therefore, there is some cycle in the 1-skeleton of the filtered simplicial complex. That is, we can move through the demographic space around that cycle by moving from one census tract to another, where the only allowable steps are through census tracts that are sufficiently similar. However, because the cycle is not a boundary, the path through physical space that corresponds to this cycle in demographic space does not surround a

meaningful region of the map. For example, it may not pass through adjacent census tracts, or it may surround a census tract whose demographics do not match those of the cycle. Because we allow cycles through demographically similar regions that might be very far apart physically, many of these cycles persist throughout the entire filtration. We identify these infinite-persistence cycles in Figure 5.2 using yellow bars. We focus our attention on those features that do eventually die.

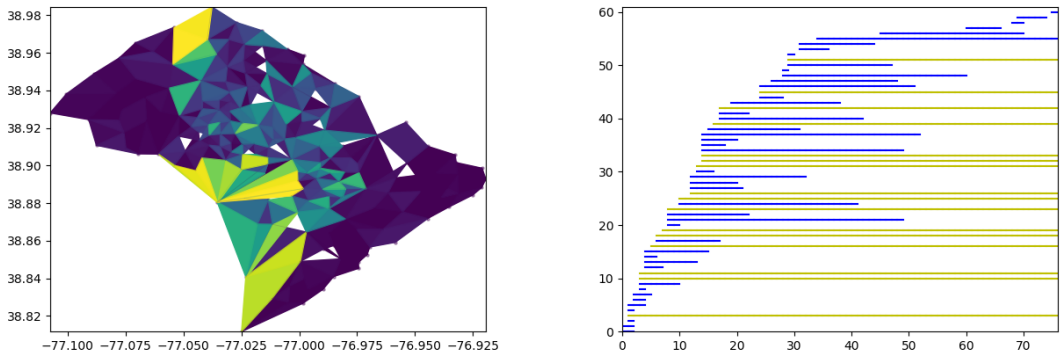
For this construction, we use demographic edges to approximate demographic space. As we increase the filtration parameter ϵ, the number of 1-simplices increases very quickly, resulting in a large simplicial complex and slow computations.

5.3.2 Spatial edges and interpretations for a case study of Washington, D.C.

If we instead consider a filtered simplicial complex with spatial edges, as described in Section 5.2.2, we can obtain information about spatial obstructions. In Figure 5.3, we show the barcode for the PH of the filtered simplicial complex that we construct for Washington, D.C.. Our choice of construction implies that longer barcodes correspond to physical obstructions.

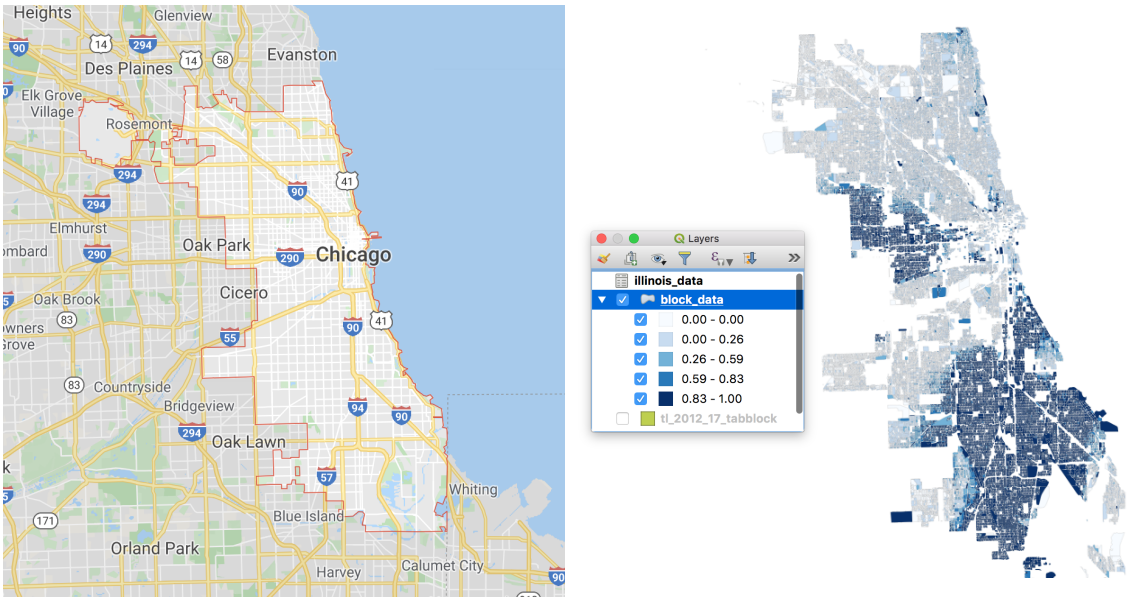
5.3.3 A second case study: Chicago

As another case study, we also construct both types of filtered simplicial complex for the city of Chicago. Segregation patterns in Chicago were informed by and crystallized by the construction of superhighways that cut through the city [Tom17]. Historical records indicate that these highways were used specifically to break up neighborhoods with large Black or African American populations. Highways were also used to create barriers between neighborhoods with different demographic makeups during "city beautification initiatives". The presence of these highways as physical obstructions has continued to lock relatively poor Black or African American populations into certain neighborhoods of the city, and the pattern of highways is visible in racial choropleths of Chicago.



(a) (b)

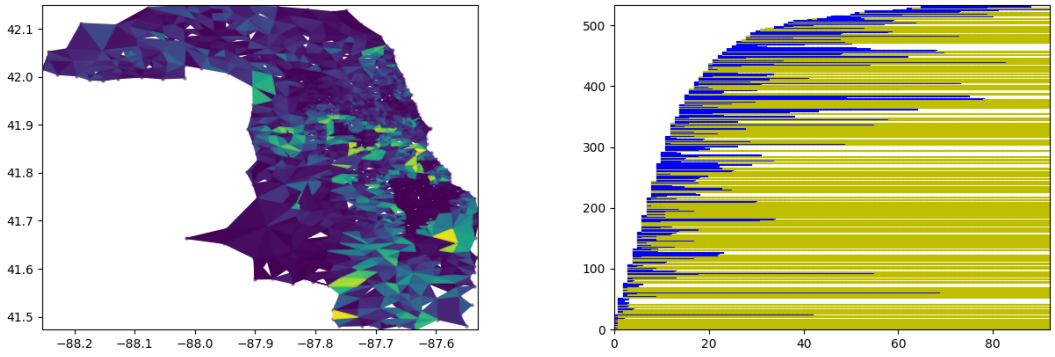
Figure 5.3: We show (a) the filtered simplicial complex with spatial edges that we construct from Washington, D.C. racial data and (b) its associated barcode. In (a), we color simplices that enter the filtration by time of entry, with the latest filtrations in the lightest colors. In (b), we observe many features in the PH. The longest bars correspond to simplices that form around the middle region of the city. We highlight infinite persistence bars in yellow for ease of reading the barcode. The filtered simplicial complex captures the east/west divide in the demographics of the city. The highest-birth-time simplices connect adjacent census tracts with large differences in their proportion of Black or African American residents.



(a) (b)

Figure 5.4: (a) Highways of Chicago from Google Maps. (b) Census tracts of Chicago colored by their proportion of Black or African American residents. Comparing (a) and (b), the patterns of highways are visible in the choropleth in (b).

We constructed only the filtered simplicial complex with spatial edges for the city of Chicago because it has many more census tracts than Washington, D.C. Because every possible pair of census tracts contributes a demographic edge, it was intractable for us to compute the filtered simplicial complex with demographic edges in the Chicago case for our project. In Figure 5.5, I show the filtered simplicial complex and its associated barcode for the spatial-edges filtered simplicial complex of Chicago. The longest features straddle the locations of the highways. [Image in (b) was created by Eion Blanchard.]



(a) (b)

Figure 5.5: We show (a) the filtered simplicial complex with spatial edges that we from Chicago racial data and (b) itsassociated barcode. In (a), we color simplices that enter the filtration by time of entry, with the latest filtrations in the lightest colors. In (b), we observe many features in the PH. We highlight infinite length features in yellow to increase readability of the barcode. The filtered simplicial complex follows some of the major highway lines in Chicago. We observe a variety of brighter simplices that are on the South Side, an area known for containing many segregated neighborhoods.

CHAPTER 6

Ongoing projects

In this chapter, I discuss several current projects with social-science applications. The first project, which I discuss in Section 6.1, focuses on adaptations of bounded-confidence models. It is in collaboration with Heather Zinn Brooks (UCLA), Yacoub H. Kureh (UCLA), and Mason A. Porter (UCLA). We aim to explore the behavior of bounded-confidence models on generalizations of network structures like multilayer networks. In Section 6.2, I discuss a project that is in collaboration with Yacoub H. Kureh (UCLA), Alexis Piazza (ACLU Los Angeles), and Victor Leung (ACLU Los Angeles). We seek to model homelessness in California schools to improve identification of schools and school districts that undercount homeless students.

6.1 Adaptations of bounded-confidence models

Along with my collaborators Heather Zinn Brooks, Yacoub H. Kureh, and Mason A. Porter, I am interested in exploring opinion dynamics in networks that are not graphs. In this section, I describe our current research in this area. In Chapter 7, I will describe a planned future project for exploring a simplicial version of a bounded-confidence model. We have briefly discussed model formulation, but subsequent research on this model is future work.

6.1.1 Multilayer bounded-confidence models

In today's world, people consume and digest information from a large variety of sources. As a result, our opinions can be informed by information that spreads over multiple networks (which one can encode as "layers" in a multilayer network [KAB14]), that we may treat differently. For example, different types of news may spread on Facebook versus Twitter versus traditional news outlets, or an individual may have different groups of contacts that spread different types of information. In this project, we examine dynamical systems on multilayer networks to model how information that spreads over different layers can change individuals opinions. Specifically, we wonder whether multilayer networks will lead to different spreading properties than the single-layer networks that were explored in [MVP18, BP20].

6.1.1.1 Model statement

We begin by proposing multilayer versions of a Hegselmann–Krause bounded-confidence model [HK02] (defined in Section 3.5.1). In this work, we study a simple model in which there are only two layers and an individual's opinion on each layer is the same. As a future extension, we may suppose that an individual can express different opinions on different layers. Individuals can be affected by both intralayer connections and interlayer connections. In our model, individuals are connected between layers only to themselves. Therefore, our multilayer networks are examples of multiplex networks [KAB14]. An individual updates their opinion according to a synchronous update rule as follows:

$$x_v(t+1) = \frac{\sum_{\alpha,\beta\in\mathbf{L}} \sum_{w\in V} \mathcal{A}_{vw\alpha\beta} x_w(t) \mathbb{1}_{|x_v(t)-x_w(t)|<c}}{\sum_{\alpha,\beta\in\mathbf{L}} \sum_{w\in V} \mathcal{A}_{vw\alpha\beta} \mathbb{1}_{|x_v(t)-x_w(t)|<c}} \,. \tag{6.1}$$

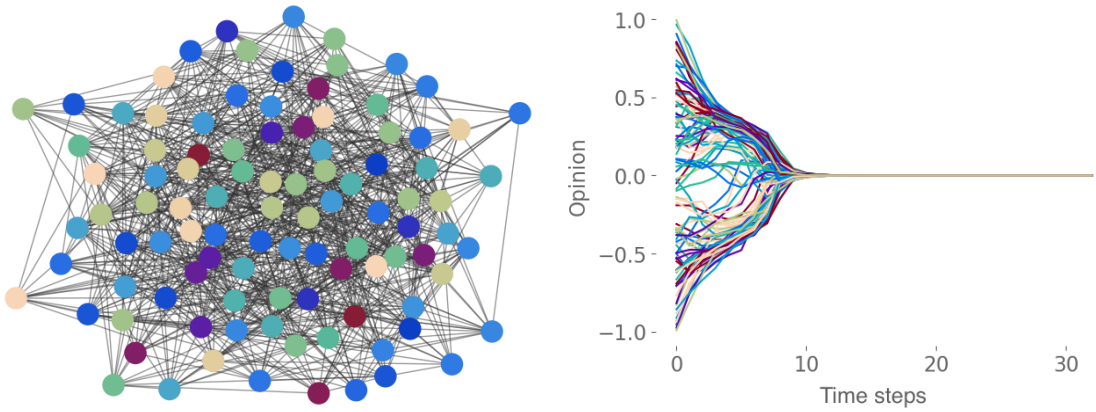
where our notation follows the multilayer network notation from Section 3.6.1. Additionally, $x_v(t)$ is the opinion of individual v at time t, the adjacency tensor $\mathcal{A}$ encodes a multilayer network, and c is the confidence bound. If we allow $\mathcal{A}$ to be a weighted adjacency tensor, we can incorporate individuals weighting some layers (in our context, some social-media

platforms) as more important than others.

The update rule in Equation (6.1) does not specify a number of layers. We will perform experiments on two-layer networks and leave experiments on networks with more layers as future work. We observe that in our model, the higher the number of layers, the more layers an individuals are connected to themselves on. As a result, individuals in networks with more layers will weight their own opinion more highly.

6.1.1.2 Results on networks with $G(N, p)$ layers

We begin by considering two-layer networks in which each layer is a $G(N, p)$ network (defined in Section 3.1.1.3) with n nodes and an independent, homogeneous probability p for each edge to exist. In Figure 6.1, we show one such network and the opinion history of each node on that network from time $t = 0$ until convergence.



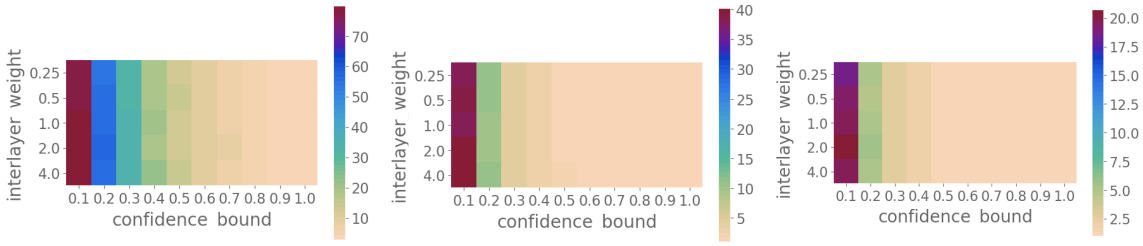
(a) (b)

Figure 6.1: (a) One layer of a two-layer network. We construct the network in each layer using a $G(N, p)$ model with $N = 100$ and $p = 0.1$ in each layer. We color nodes by initial opinion. (b) Opinion dynamics of the network in (a). The nodes in this network converge to a single opinion.

We vary N, p, c, and the "interlayer weight", which is the weight that we assign to

interlayer edges. For each set of parameters, we generate 1000 networks and initialize the opinions of all nodes in the network uniformly at random from the interval $[-1, 1]$. We then update the opinions of nodes according to our update rule in Equation (6.1) until convergence or until we exceed some maximum run time. In our experiments, this maximum was 10000 time steps.

After each simulation has converged (or exceeded maximum run time), we compute the number of clusters (which we take to be the number of connected components after the simulation finishes). We then generate heatmaps based on the mean and median number of clusters across our simulations. Previous work on multilayer networks has indicated that interlayer edge weights can impact measures of network structure and dynamical processes on such networks [BP18, PWC19], so we illustrate the effects of changing this parameter. In Figure 6.2, we show results for a selection of these parameters.



(a) $N = 100$ and $p = 0.02$ (b) $N = 100$ and $p = 0.06$ (c) $N = 100$ and $p = 0.1$

Figure 6.2: Heatmaps of the mean number of clusters at convergence for two-layer networks in which each layer is an instance of a $G(N, p)$ network with $N = 100$. Interlayer edge weight is on the vertical axis, and confidence bound is on the horizontal axis. The number of clusters changes very little with respect to interlayer edge weight. The number of clusters decreases as we increase p. This is consistent with results on single-layer $G(N, p)$ networks [MVP18].

Varying the interlayer weight (which we have taken to be homogeneous for all nodes in our examples) seems to have little effect on the number of clusters, indicating that this particular form of multilayer network does not behave markedly differently from single-layer

$G(N, p)$ networks. However, this result is not entirely unexpected. Our model is equivalent to running a Hegselmann–Krause bounded-confidence model on a network that is the sum of two $G(N, p)$ networks in the following sense.

If we take the adjacency matrices of two $G(N, p)$ networks with no self-edges and add them componentwise, we obtain all of the interlayer edges and edge weights of our two-layer $G(N, p)$ network. If we then add self-edges with weight equivalent to the intralayer edge weight in our two-layer model, we account for all intralayer edges. In a single-layer Hegselmann–Krause model, individuals place a weight on their own opinions as well. Therefore, with the exception of any edges that occur in both layers of a network, our two-layer network is equivalent to a single-layer network that includes self-weightings. For small values of p, few edges occur in both layers of the network. As a result, as $p \to 0$, the two-layer Hegselmann–Krause model behaves like a single-layer Hegselmann–Krause model on a $G(N, 2p)$ network. For very high values of p, many edges occur in both layers of the network, so that the limiting behavior as $p \to 1$ is equivalent to the behavior of a single-layer network with half the self-weighting (because at $p = 1$, every intralayer edge occurs in both layers, doubling the weights of those edges).

6.1.1.3 Possible alternative multilayer networks

We expect that networks that have more structure (e.g., cycle graphs, lattice graphs, or stochastic block models) than $G(N, p)$ networks may yield results that are more distinctive than those on single-layer networks. To this end, we will proceed by examining networks with two layers of ring networks, lattice networks, and others. We are also considering networks in which the two layers do not have the same parameters (e.g., $G(N, p)$ networks with different values of p in the two layers) or where the two layers do not have the same type of network structure. Alternatively, we are considering varying the interlayer edges either by not connecting all individuals to themselves across layers or by allowing individuals to have different opinions on the different layers.

6.2 Homelessness underreporting

Under the 1987 McKinney–Vento Homeless Assistance Act [Mck87], federal law requires that a variety of services be provided to homeless people. Specifically, the McKinney–Vento Act is a conditional funding act that provides federal grants to states in exchange for compliance with the act. Under the McKinney–Vento Act, schools must ensure that homeless children have free transportation to and from school. Additionally, homeless children must be allowed to attend their school of origin (the school they were attending before they became homeless) regardless of current residence. It also requires schools to register homeless children even if they cannot provide required documents (such as immunization records or proof of residence). To implement the terms of the McKinney–Vento Act, states must have a designated statewide homeless coordinator and local school distracts much appoint Local Education Liaisons. Liaisons are responsible for notifying homeless families of their rights under the act, facilitating access to their school of origin, and ensuring that school staff are aware of homeless students' rights under the act.

In practice, many homeless children who are enrolled in schools are not reported as homeless. The McKinney–Vento Act defines homeless children as those who "lack a fixed, regular, and adequate nighttime residence" [Mck87], but this definition does not necessarily align with the general population's perception of what it means to be homeless. Additionally, according to the 2019 California audit of Youth Experiencing Homelessness (YEH) [How19], available data suggests that California Local Education Agencies (LEAs) are not doing enough to identify youth experiencing homelessness. This is defined to be youth who are currently homeless. While homeless education experts generally estimate that 5–10% of economically disadvantaged youth experience homelessness [How19], four of the six LEAs in the audit reported 3% or fewer of economically disadvantaged students as experiencing homelessness.

While it is unclear precisely why homeless students are underidentified, LEAs frequently do not sufficiently train school staff to identify homeless youth, and they may not disseminate

information to local schools about homeless education programs and services [How19]. State homeless education programs suffer from low staffing rates, leading to difficulty overseeing homeless education programs and ensuring that they are properly implemented.

The 2019 YEH audit concluded that LEAs that coordinate more with homeless service organizations do a better job of identifying youth experiencing homelessness. In the audit, LEAs that coordinated with homeless service organizations had lower rates of absenteeism, suspension, and dropping out than statewide averages, indicating that intervention at the LEA level can help homeless students achieve better educational performance. Identifying homeless youth is therefore a critical step for providing them with access to the support that they need.

In this work, my collaborators and I seek to develop models to aid in the identification of homeless youth. Specifically, because we believe that the majority of LEAs are underreporting (potentially severely) the number of enrolled homeless students, one of our major goals is to develop a model to identify which LEAs and schools are the most likely to need intervention.

6.2.1 Data exploration

One of the challenges in data about homelessness is its lack of reliability. Based on [How19], we suspect that the vast majority of LEAs are underreporting severely, but we do not have any way of knowing which LEAs are reliable. Because we believe that income and cost of living are likely to be correlated with the number of students experiencing homelessness, we have been searching for correlations between American Community Survey (ACS) income data [Bur16a] and the Calfornia Longitudinal Pupil Achievement Data System (CALPADS) [Edu19] data set, which records the number of enrolled homeless students at each school.

Initially, my focus has been to look for geographical patterns in homelessness or income level at the LEA level, but it is difficult to find geographical patterns that seem meaningful.

My collaborator Yacoub Kureh has been searching for correlations between the CALPADS homelessness data set and ACS data, but we have been unable to find meaningful correlations. (We suspect this is because of the inherent unreliability of the homelessness data.) In Section 6.2.2, I describe our plans for moving forward on the project. As these plans are currently in progress, we do not yet have results, but I include them in this thesis because we have put a significant amount of thought into model formulation.

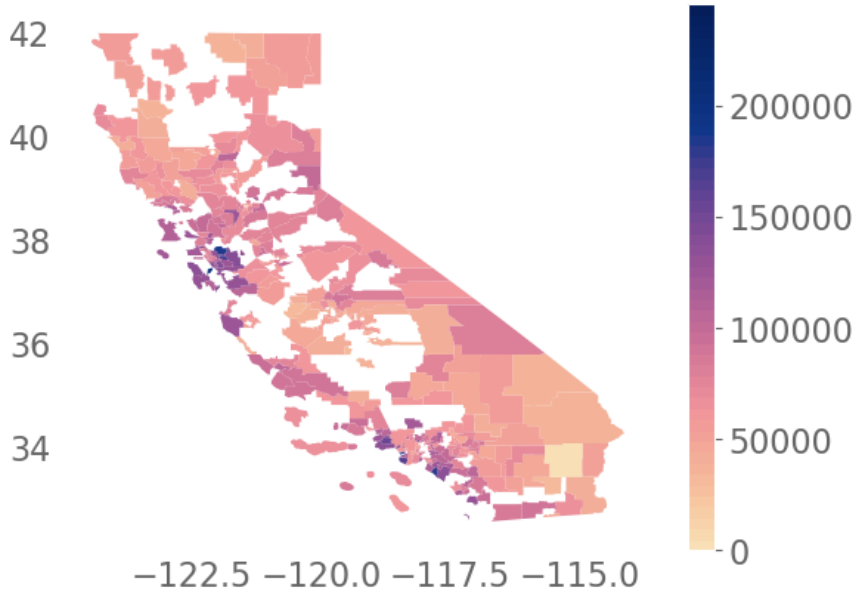


Figure 6.3: California LEAs colored by median income in dollars. Aside from higher income levels in urban areas, we are not able to discern any patterns.

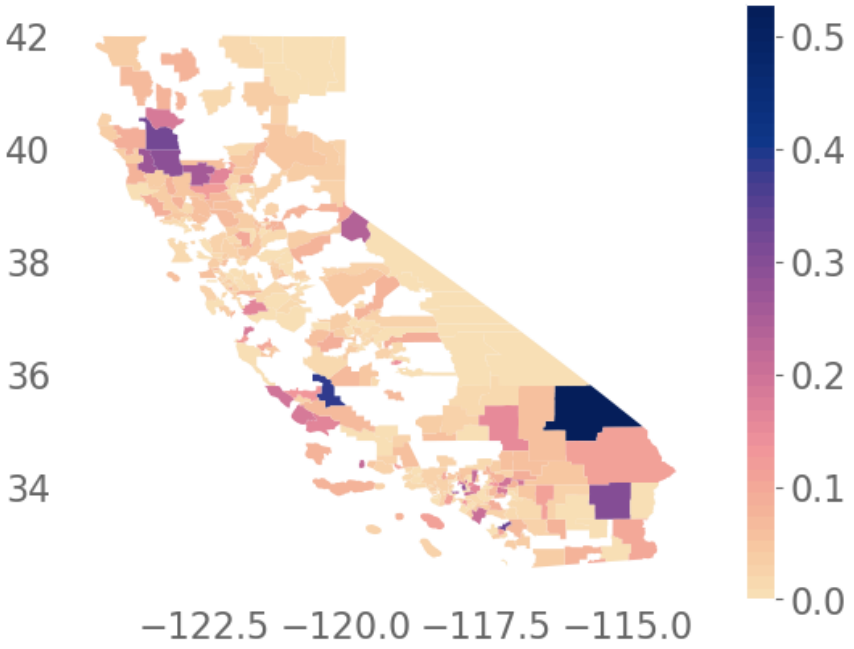


Figure 6.4: California LEAs colored by fraction of enrolled students who are experiencing homelessness. Similar to the choropleth for income in Figure 6.3, we do not observe any geographical patterns.

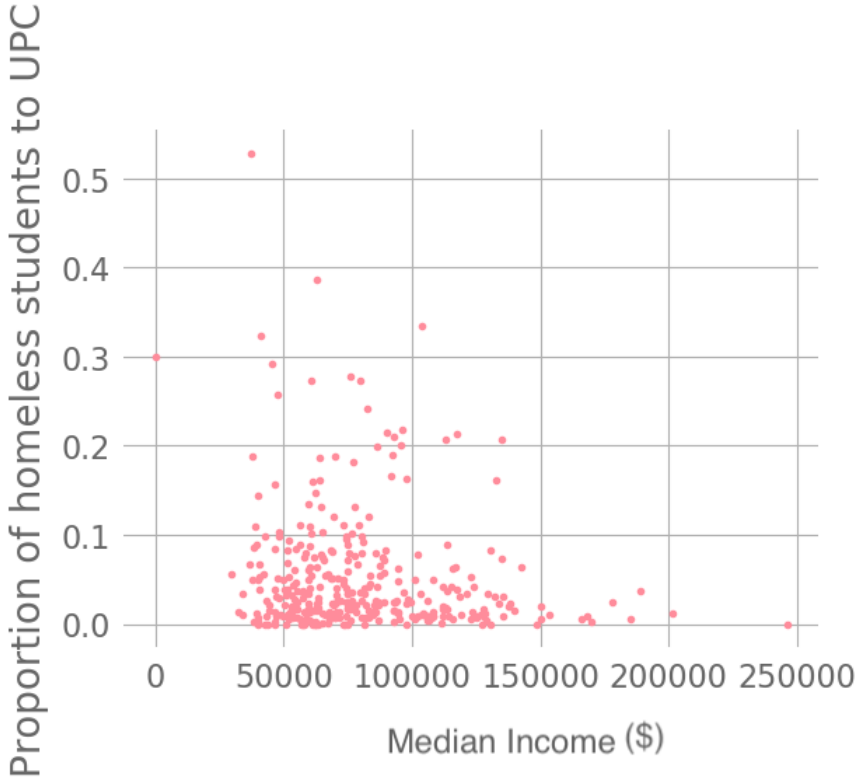


Figure 6.5: Scatter plot of median income versus ratio of homeless students to enrolled students. Here, we use the Unduplicated Pupil Count (UPC) from CALPADS to determine number of enrolled students. Each point of the scatter plot represents one California LEA. Although the scatter plot appears to be bounded above by a line with negative slope, we observe no clear correlation. It is possible that if underreporting were corrected, we would observe a negative correlation, as one may expect.

6.2.2 Plans for modeling

Because of the difficulty of working with the full CALPADs homelessness data set, our collaborators at the ACLU are hoping to conduct surveys of some LEAs. We hope that these surveys will be able to serve as a training set. Because we have limited surveying resources, one important question that we need to consider is how to obtain a representative

sample to survey.

Experts on homelessness in schools agree that overreporting does not happen [Pia19]. To this end, one sampling strategy that we have considered is to target the LEAs that we think are most likely to be underreporting with surveys, because these are the LEAs that provide the least information. Experts have estimated that the homeless student population of an LEA or school should be roughly 5% of its "free and reduced lunch" population [Pia19], so we will assume that schools that report more than 5% of its free-and-reduced-lunch population as homeless are reporting correctly.

We propose several approaches to incorporate these assumptions into our models. The first is to use only those schools that we assume (based on the 5% mark) to be correctly reporting to build a model for inferring rates in the remaining schools. For example, we can use income data from the ACS to perform linear regression. This method is very simple, but it will likely be biased toward schools that report the largest numbers. For example, some schools serve only homeless populations, so they report 100% homelessness.

Another approach is to attempt to use statistical inference to identify underreporting schools, based on priors that we obtain using the assumptions on underreporting and homelessness rates.

Let X be the unobserved variable that represents the true homelessness variable, and let $\tilde{X}$ be the observed proxy homelessness variable. We use x to denote a realization of X (and $\tilde{x}$ a realization of $\tilde{X}$), defined as follows:

$$x = \frac{h}{l}\,, \quad \tilde{x} = \frac{\tilde{h}}{l}\,, \tag{6.2}$$

where h is the true size of the homeless population, $\tilde{h}$ is the observed homeless population, and l is the free and reduced lunch population.

Let M_X be the misclassification variable, which controls whether X is correctly reported (in which case $M_x = 1$) or not ($M_x = 0$). We then assume that no district is overreporting.

That is, we assume

$$p(M_x = 1|\tilde{x} \geq 0.05) = 1\,, \quad p(M_x = 1|\tilde{x} < 0.05) = \theta\,,$$

where $\theta \in [0, 1]$ is a parameter.

We also assume that if $\tilde{x} = \tilde{\rho}$ is underreported, the true homeless variable x is uniformly distributed in $[\tilde{\rho}, 1]$. We use the uniform distribution for simplicity.

We are interested in estimating x given $\tilde{x}$. Using Bayes' Theorem, we can compute

$$\begin{aligned}
p(x = \rho|\,\tilde{x} = \tilde{\rho}) &= \frac{p(x = \rho, \tilde{x} = \tilde{\rho})}{p(\tilde{x} = \tilde{\rho})} \\
&= \frac{p(x = \rho, \tilde{x} = \tilde{\rho}|\,M_x = 0) + p(x = \rho, \tilde{x} = \tilde{\rho}|\,M_x = 1)}{p(\tilde{x} = \tilde{\rho})} \\
&= \frac{\mathbb{1}_{\tilde{\rho}<\rho}\mathbb{1}_{\rho<\tilde{0}.05}p(x = \rho|\,\tilde{x} = \tilde{\rho}) + \mathbb{1}_{\rho=\tilde{\rho}}}{p(\tilde{x} = \tilde{\rho})} \\
&= \frac{\mathbb{1}_{\tilde{\rho}<\rho}\mathbb{1}_{\rho<\tilde{0}.05}\frac{1}{1-\tilde{\rho}} + \mathbb{1}_{\rho=\tilde{\rho}}}{p(\tilde{x} = \tilde{\rho})} \\
&= \frac{\mathbb{1}_{\tilde{\rho}<\rho}\mathbb{1}_{\rho<\tilde{0}.05} + \mathbb{1}_{\rho=\tilde{\rho}}(1-\tilde{\rho})}{(1-\tilde{\rho})p(\tilde{x} = \tilde{\rho})}\,.
\end{aligned}$$

From this derivation, we have the conditional probability distribution

$$p(x = \rho|\,\tilde{x} = \tilde{\rho}) = \frac{\mathbb{1}_{\tilde{\rho}<\rho}\mathbb{1}_{\rho<\tilde{0}.05} + \mathbb{1}_{\rho=\tilde{\rho}}(1-\tilde{\rho})}{(1-\tilde{\rho})p(\tilde{x} = \tilde{\rho})}\,. \tag{6.3}$$

With the observed distribution $P(\tilde{X})$, Equation (6.3) allows us to maximize the likelihood that $p(x = \rho|\,\tilde{x} = \tilde{\rho})$, thereby inferring the value of the true homelessness variable x. This method depends heavily on our assumptions on the distributions of X and M_X.

We propose using a combination of these two methods to identify the LEAs that are most worth surveying. Once we have surveyed data, we will adjust our inference model by adding the newly surveyed LEAs to the training set.

CHAPTER 7

Conclusions and future work

In this thesis, I have discussed a variety of tools for exploring the structure of complex systems. In the following sections, I summarize key results from the thesis and describe future projects that extend some of the ideas in this thesis.

7.1 Key Results

With my advisor Mason A. Porter, I developed two novel methods for building simplicial complexes from map-based network data (see Chapter 4). Both methods can be applied more generally to spatial data sets. We used our new methods to characterize voting patterns, city street networks, scientific images, and random graphs. We also identified voting islands in California precincts, detected patterns in city street layouts (e.g., we quantified the sense that American cities are particularly gridlike), and categorized spiderwebs based on web regularity.

My collaborators Eion Blanchard, Moon Duchin, Austin Eide, and Patrick Girardet, and I suggested some methods for thinking about the interplay of various notions of space (see Chapter 5). We appllied our techniques to neighborhood segregation to identify neighborhood boundaries.

My collaborators Heather Zinn Brooks, Yacoub H. Kureh, and Mason A. Porter, and I proposed an extension (see Section 6.1) of the Hegselmann–Krause model for opinion dynamics to multilayer networks. We are currently investigating the convergence behaviors

of this model for different network structures.

My collaborators Yacoub H. Kureh, Victor Leung, Alexis Piazza, and I proposed methods (see Section 6.2) for identifying LEAS that are underreporting youth experiencing homelessness.

7.2 Future research

7.2.1 Topological tools for temporal networks

Much of my thesis work has focused on how to transform a data set into a topological object. In my future work, I intend to extend topological ideas from Morse theory to the study of data that varies in time. Morse theory allows one to compute changes in the topology of a space using Morse functions [MSW63]. This provides a computationally efficient approach for quantifying changes in topology.

I want to extend some of the ideas from my thesis research to build complex topological objects, potentially based on data sets on a variety of temporal scales. This will allow me to explore changes in network structure over time, which would meaningfully extend the current scope of topological data analysis.

7.2.2 Dynamical processes on simplicial networks

Another area that I seek to explore in the future is the concept of simplicial networks. Many social relationships are not pairwise, and there are a variety of applications (e.g., team-based competitions and peer pressure) that incorporate ideas of group influence. Framing networks as simplicial complexes provides one method of modeling these types of relationships [IPB19, HK20].

Because of the computational and conceptual complexity that is presented by the question of orientation in simplicial complexes, there remains a lot of work to be done on understand-

ing simplicial models of dynamical processes. In particular, I hope to bring some of my topological expertise to bear on the problem of opinion formation (e.g., in bounded-confidence models) in simplicial models. Specifically, I plan to work on models for understanding opinion formation in a simplicial network. How does peer pressure affect opinion formation? Can this lead to different types of consensus and polarization effects than in single-layer or multilayer networks?

7.3 Summary

Over the course of my Ph.D., I have helped develop and apply a variety of methods that incorporate topological notions of "closeness" into the study of complex systems, with a focus on social systems. I have performed case studies that demonstrate the value of topological tools for understanding global structure in networks and complex systems more generally, and some of these case studies have helped shed light on specific spatial applications (such as cities, voting, and spiderwebs). The methods that I have developed are computationally efficient for applications to 2D spatial networks, and they provide a new way of thinking about planar network data as topological spaces, rather than as graphs. I believe that applying this type of topological framework to applications can allow one to explore a variety of problem spaces in new and exciting ways, leading to new insights on the structure of complex systems.

REFERENCES

[AA17] Michał Adamaszek and Henry Adams. "The Vietoris–Rips complexes of a circle." *Pacific Journal of Mathematics*, **290**(1):1–40, 2017.

[AAN04] Réka Albert, István Albert, and Gary L. Nakarado. "Structural vulnerability of the North American power grid." *Physical Review E*, **69**(2):025103, 2004.

[ADK08] Alex Arenas, Albert Díaz-Guilera, Jurgen Kurths, Yamir Moreno, and Changsong Zhou. "Synchronization in complex networks." *Physics Reports*, **469**(3):93–153, 2008.

[AEK17] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier. "Persistence images: A stable vector representation of persistent homology." *Journal of Machine Learning Research*, **18**(8):1–35, 2017.

[AEM07] Dominique Attali, Herbert Edelsbrunner, and Yuriy Mileyko. "Weak witnesses for Delaunay triangulations of submanifolds." In *Proceedings - SPM 2007: ACM Symposium on Solid and Physical Modeling*, pp. 143–150, New York, NY, 2007. ACM Press.

[AM19] Alberto Aleta and Yamir Moreno. "Multilayer networks in a nutshell." *Annual Review of Condensed Matter Physics*, **10**(1):45–62, 2019.

[Bar11] Marc Barthélemy. "Spatial networks." *Physics Reports*, **499**(1–3):1–101, 2011.

[Bar17] Marc Barthelemy. "From paths to blocks: New measures for street patterns." *Environment and Planning B: Urban Analytics and City Science*, **44**(2):256–271, 2017.

[Bar18] Marc Barthelemy. *Morphogenesis of Spatial Networks*. Springer International Publishing, Cham, 2018.

[Bar19] Marc Barthelemy. "The statistical physics of cities." *Nature Reviews Physics*, **1**(6):406–415, 2019.

[Bat17] Michael Batty. *The New Science of Cities*. The MIT Press, Cambridge, MA, 2017.

[BBB94] Daniel Beavon, Patricia Brantingham, and Paul Brantingham. "The influence of street networks on the patterning of property offenses." *Crime Prevention Studies*, **2**:115–148, 1994.

[BBG18] Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. "Human mobility: models and applications." *Physics Reports*, **734**:1–74, 2018.

[BCV14] Francois Baccelli, Avhishek Chatterjee, and Sriram Vishwanath. "Stochastic bounded confidence opinion dynamics." In *Proceedings of the 53rd IEEE Conference on Decision and Control*, pp. 3408–3413, Los Angeles, CA, 2014. Institute of Electrical and Electronics Engineers Inc.

[Bel15] Mariana Belgiu. "UIA Latitude/Longitude Graticules and World Countries Boundaries." ArcGIS, 2015. available at https://www.arcgis.com/home/item.html?id=a21fdb46d23e4ef896f31475217cbb08.

[BEM10] Paul Bendich, Herbert Edelsbrunner, Dmitriy Morozov, and Amit Patel. "The robustness of level sets." In Mark de Berg and Ulrich Meyer, editors, *"Algorithms — ESA 2010"*, volume 6346 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2010.

[BHM19] Helen M. Byrne, Heather A. Harrington, Ruth Muschel, Gesine Reinert, Bernadette J. Stolz-Pretzer, and Ulrike Tillmann. "Topology characterises tumour vasculature." *Mathematics Today*, **5**(5), 2019.

[BHO18] Mickaël Buchet, Yasuaki Hiraoka, and Ippei Obayashi. "Persistent homology and materials informatics." In Isao Tanaka, editor, *Nanoinformatics*, pp. 75–95. Springer-Verlag, Heidelberg, 2018.

[BHP20] Peter Bubenik, Michael Hull, Dhruv Patel, and Benjamin Whittle. "Persistent homology detects curvature." *Inverse Problems*, **36**(2):025008, 2020.

[BK87] Dmitri V. Boulatov and Vladimir A. Kazakov. "The ising model on a random planar lattice: The structure of the phase transition and the exact critical exponents." *Physics Letters B*, **186**(3–4):379–384, 1987.

[BM13] Andrew J. Blumberg and Michael A. Mandell. "Quantitative homotopy theory in topological data analysis." *Foundations of Computational Mathematics*, **13**(6):885–911, 2013.

[BMM16] Paul Bendich, J. S. Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. "Persistent homology analysis of brain artery trees." *Annals of Applied Statistics*, **10**(1):198–218, 2016.

[Boe17] Geoff Boeing. "OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks." *Computers, Environment and Urban Systems*, **65**:126–139, 2017.

[Boe18] Geoff Boeing. "A multi-scale analysis of 27,000 urban street networks: every us city, town, urbanized area, and zillow neighborhood." *Environment and Planning B: Urban Analytics and City Science*, p. 2399808318784595, 2018. available at doi:10.1177/2399808318784595.

[Boe19a] Geoff Boeing. "Spatial information and the legibility of urban form: Big data in urban morphology." *International Journal of Information Management*, 2019. available at doi: 10.1016/j.ijinfomgt.2019.09.009.

[available at doi: 10. 1016/ j. ijinfomgt. 2019. 09. 009 .]

[Boe19b] Geoff Boeing. "Urban spatial order: Street network orientation, configuration, and entropy." *Applied Network Science*, **4**(1):67, 2019.

[BOP15] Danielle S. Bassett, Eli T. Owens, Mason A. Porter, M. Lisa Manning, and Karen E. Daniels. "Extraction of force-chain network architecture in granular materials using community detection." *Soft Matter*, **11**(14):2731–2744, 2015.

[BP18] Javier M. Buldú and Mason A. Porter. "Frequency-based brain networks: From a multiplex framework to a full multilayer description." *Network Neuroscience*, **2**(4):418–441, 2018.

[BP20] Heather Z. Brooks and Mason A. Porter. "A model for the influence of media on the ideology of content in online social networks." *Physical Review Research*, **2**:023041, 2020.

[BS18] Richard Barnes and Justin Solomon. "Gerrymandering and compactness: Implementation flexibility and abuse." *arXiv:1803.02857*, 2018.

[Bub15] Peter Bubenik. "Statistical topological data analysis using persistence landscapes." *Journal of Machine Learning Research*, **16**(1):77–102, 2015.

[Bur16a] U S Census Bureau. "Selected Economic Characteristics in the United States, Table DP02." data.census.gov, 2016.

[Bur16b] U S Census Bureau. "Selected Social Characteristics in the United States, Table DP02." data.census.gov, 2016.

[BZ18] Andrew Banman and Lori Ziegelmeier. "Mind the Gap: A Study in Global Development Through Persistent Homology." In Erin Wolf Chambers, Brittany Terese Fasy, and Lori Ziegelmeier, editors, *Research in Computational Topology*, volume 13 of *Association for Women in Mathematics*, pp. 125–144. Springer International Publishing, Cham, 2018.

[CAL15] Aaron Clauset, Samuel Arbesman, and Daniel B. Larremore. "Systematic inequality and hierarchy in faculty hiring networks." *Science Advances*, **1**(1):e1400005, 2015.

[CCK17] Ciro Cattuto, Kevin Chan, Marton Karsai, Nicola Perra, and Bruno Ribeiro. "Machine Learning in Network Science." NetSci17 Satellite, 2017.

[Cha03] Camille Zubrinsky Charles. "The dynamics of racial residential segregation." *Annual Review of Sociology*, **29**(1):167–207, 2003.

[CMR08] David Card, Alexandre Mas, and Jesse Rothstein. "Tipping and the dynamics of segregation." *The Quarterly Journal of Economics*, **123**(1):177–218, 2008.

[CSL06] Alessio Cardillo, Salvatore Scellato, Vito Latora, and Sergio Porta. "Structural properties of planar graphs of urban street patterns." *Physical Review E*, **73**(6):066107, 2006.

[DC20] Prologue DC. "Mapping Segregation in Washington DC.", 2020. available at https://www.mappingsegregationdc.org/.

[DCL16] Yucheng Dong, Xia Chen, Haiming Liang, and Cong Cong Li. "Dynamics of linguistic opinion formation in bounded confidence model." *Information Fusion*, **32, Part A**:52–61, 2016.

[DCV08] Mark De Berg, Otfried Cheong, Marc Van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag, Berlin, Heidelberg, 3rd edition, 2008.

[Dit01] Jan Christian Dittmer. "Consensus formation under bounded confidence." *Nonlinear Analysis, Theory, Methods and Applications*, **47**(7):4615–4621, 2001.

[DNA00] Guillaume Deffuant, David Neau, Frederic Amblard, and Gérard Weisbuch. "Mixing beliefs among interacting agents." *Advances in Complex Systems*, **03**(01n04):87–98, 2000.

[DSC17] Michela Del Vicario, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. "Modeling confirmation bias and polarization." *Scientific Reports*, **7**(1):40391, 2017.

[DT18] Moon Duchin and Bridget Eileen Tenner. "Discrete geometry for electoral geography." *arXiv:1808.05860*, 2018.

[Ebe97] Mark Ebers. *The Formation of Inter-Organizational Networks*. Oxford University Press, Oxford, 1997.

[Edu19] California Department of Education. "CALPADS Unduplicated Pupil Count (UPC) Source File (K–12)." California Longitudinal Pupil Achievement Data System (CALPADS), 2019. available at https://www.cde.ca.gov/ds/sd/sd/filescupc.asp.

[EEB11] Paul Expert, Tim S. Evans, Vincent D. Blondel, and Renaud Lambiotte. "Uncovering space-independent communities in spatial networks." *Proceedings of the National Academy of Sciences of the United States of America*, **108**(19):7663–7668, 2011.

[EH10] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction.* American Mathematical Society, Providence, RI, 2010.

[EKS83] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. "On the shape of a set of points in the plane." *IEEE Transactions on Information Theory*, **29**(4):551–559, 1983.

[ER59] Paul Erdős and Alfréd Rényi. "On random graphs I." *Publicationes Mathematicae*, **6**(6):290–297, 1959.

[FCP09] Damien A. Fair, Alexander L. Cohen, Jonathan D. Power, Nico U.F. Dosenbach, Jessica A. Church, Francis M. Miezin, Bradley L. Schlaggar, and Steven E. Petersen. "Functional brain networks develop from a "local to distributed" organization." *PLoS Computational Biology*, **5**(5):e1000381, 2009.

[FP20a] Michelle Feng and Mason A. Porter. "Persistent homology of geospatial data: A case study with voting." *arXiv:1902.05911*, 2020. *SIAM Review*, in press.

[FP20b] Michelle Feng and Mason A. Porter. "Quantifying "political islands" with persistent homology." *SIAM News*, (January/February 2020), 2020. available at https://sinews.siam.org/Details-Page/quantifying-political-islands-with-persistent-homology.

[FP20c] Michelle Feng and Mason A. Porter. "Spatial applications of topological data analysis: Cities, snowflakes, random structures, and spiders spinning under the influence." *arXiv:2001.01872*, 2020.

[FSP19] Kelly R. Finn, Matthew J. Silk, Mason A. Porter, and Noa Pinter-Wollman. "The use of multilayer network analysis in animal behaviour." *Animal Behaviour*, **149**:7–22, 2019.

[GB15] Riccardo Gallotti and Marc Barthelemy. "The multilayer temporal network of public transport in Great Britain." *Scientific Data*, **2**(1):140056, 2015.

[GFO18] Frederic Gibou, Ronald Fedkiw, and Stanley Osher. "A review of level-set methods and some recent applications." *Journal of Computational Physics*, **353**:82–109, 2018.

[GGB16] Chad Giusti, Robert Ghrist, and Danielle S. Bassett. "Two's company, three (or more) is a simplex." *Journal of Computational Neuroscience*, **41**(1):1–14, 2016.

[Ghr08] Robert Ghrist. "Barcodes: The persistent topology of data." In *Bulletin of the American Mathematical Society*, volume 45, pp. 61–75, 2008.

[Gil02] Rick Gillman. "Geometry and gerrymandering." *Math Horizons*, **10**(1):10–12, 2002.

[GMD18] Marko Gosak, Rene Markovič, Jurij Dolenšek, Marjan Slak Rupnik, Marko Marhl, Andraž Stožer, and Matjaž Perc. "Network science of biological systems at different scales: A review." *Physics of Life Reviews*, **24**:118–135, 2018.

[Goo19] Google. "Google Maps search for Shanghai.", 2019. available at https://www.google.com/maps/place/Shanghai,+China/data=!4m2!3m1!1s0x35b27040b1f53c33:0x295129423c364a1?sa=X&ved=2ahUKEwjSmom9nevmAhXNuZ4KHangDhIQ8gEwK3oECBkQBA.

[GW16] Sandra González-Bailón and Ning Wang. "Networked discontent: The anatomy of protest campaigns in social media." *Social Networks*, **44**:95–104, 2016.

[Hat02] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, Cambridge, 2002.

[HC69] Peter Haggett and Richard J. Chorley. *Network Analysis in Geography*. Edward Arnold, London, 1st edition, 1969.

[HDJ14] Sylvie Huet, Guillaume Deffuant, and Wander Jager. "Rejection mechanism in 2d bounded confidence provides more conformity." *arXiv:1404.7270*, 2014.

[HK02] Rainer Hegselmann and Ulrich Krause. "Opinion dynamics and bounded confidence models, analysis, and simulation." *Journal of Artificial Societies and Social Simulation*, **5**, 2002. available at http://jasss.soc.surrey.ac.uk/5/3/2.html.

[HK20] Leonhard Horstmeyer and Christian Kuehn. "Adaptive voter model on simplicial complexes." *Physical Review E*, **101**(2):022305, 2020.

[HMM19] Devon P. Humphreys, Melissa R. McGuirl, Michael Miyagi, and Andrew J. Blumberg. "Fast estimation of recombination rates using topological data analysis." *Genetics*, **211**(4):1191–1204, 2019.

[Hoc83] Ronald R. Hocking. "Developments in linear regression methodology: 1959–1982." *Technometrics*, **25**(3):219–230, 1983.

[HOG12] Luke Heaton, Boguslaw Obara, Vincente Grau, Nick Jones, Toshiyuki Nakagaki, Lynne Boddy, and Mark D. Fricker. "Analysis of fungal networks." *Fungal Biology Reviews*, **26**(1):12–29, 2012.

[Hol15] Petter Holme. "Modern temporal network theory: A colloquium." *European Physical Journal B*, **88**:234, 2015.

[How19] Elaine M. Howle. "Youth Experiencing Homelessness.", 2019. available at https://www.auditor.ca.gov/reports/2019-104/chapters.html.

[HS12] Petter Holme and Jari Saramäki. "Temporal networks." *Physics Reports*, **519**(3):97–125, 2012.

[HSS08] Aric Hagberg, Pieter Swart, and Daniel Schult. "Exploring network structure, dynamics, and function using NetworkX." In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *7th Python in Science Conference (SciPy 2008)*, pp. 11–15, Pasadena, CA, 2008.

[HSY20] Baorui Han, Dazhi Sun, Xiaomei Yu, Wanlu Song, and Lisha Ding. "Classification of urban street networks based on tree-like network features." *Sustainability (Switzerland)*, **12**(2):628, 2020.

[ID19] Paul Samuel P. Ignacio and Isabel K. Darcy. "Tracing patterns and shapes in remittance and migration networks via persistent homology." *EPJ Data Science*, **8**:1, 2019.

[IPB19] Iacopo Iacopini, Giovanni Petri, Alain Barrat, and Vito Latora. "Simplicial models of social contagion." *Nature Communications*, **10**(1):2485, 2019.

[IW06] John Iceland and Rima Wilkes. "Does socioeconomic status matter? Race, class, and residential segregation." *Social Problems*, **53**(2):248–273, 2006.

[Jal03] Marika Jalovaara. "The joint effects of marriage partners' socioeconomic positions on the risk of divorce." *Demography*, **40**(1):67–81, 2003.

[KAB14] Mikko Kivelä, Alex Arenas, Marc Barthelemy, James P. Gleeson, Yamir Moreno, and Mason A. Porter. "Multilayer networks." *Journal of Complex Networks*, **2**(3):203–271, 2014.

[KBC14] Florian Klimm, Danielle S. Bassett, Jean M. Carlson, and Peter J. Mucha. "Resolving structural variability in network models and the brain." *PLoS Computational Biology*, **10**(3):e1003491, 2014.

[KBN16] Violeta Kovacev-Nikolic, Peter Bubenik, Dragan Nikolić, and Giseon Heo. "Using persistent homology and dynamical distances to analyze protein binding." *Statistical Applications in Genetics and Molecular Biology*, **15**(1):19–38, 2016.

[KDS16] Lida Kanari, Paweł Dłotko, Martina Scolamiero, Ran Levi, Julian Shillcock, Kathryn Hess, and Henry Markram. "Quantifying topological invariants of neuronal morphologies." *arXiv:1603.08432*, 2016.

[KGK13] Miroslav Kramá, Arnaud Goullet, Lou Kondic, and Konstantin Mischaikow. "Persistence of force networks in compressed granular media." *Physical Review E*, **87**(4):042207, 2013.

[KOA18] Heetae Kim, David Olave-Rojas, Eduardo Álvarez-Miranda, and Seung Woo Son. "In-depth data on the network structure and hourly activity of the central chilean power grid." *Scientific Data*, **5**:180209, 2018.

[Kol09] Eric D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, New York, NY, 2009.

[KPM12] Michael W. Kraus, Paul K. Piff, Rodolfo Mendoza-Denton, Michelle L. Rheinschmidt, and Dacher Keltner. "Social class, solipsism, and contextualism: How the rich are different from the poor." *Psychological Review*, **119**(3):546–572, 2012.

[KS13] Michael Kerber and R. Sharathkumar. "Approximate Čech complex in low and high dimensions." In Leizhen Cai, Siu-Wing Cheng, and Tak-Wah Lam, editors, *Algorithms and Computation. ISAAC 2013*, volume 8283 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2013.

[LB14a] Rémi Louf and Marc Barthelemy. "How congestion shapes cities: From mobility patterns to scaling." *Scientific Reports*, **4**(1):5561, 2014.

[LB14b] Rémi Louf and Marc Barthelemy. "A typology of street patterns." *Journal of the Royal Society Interface*, **11**(101):20140924, 2014.

[LBK89] Ivan Light, Parminder Bhachu, and Stavros Karageorgis. "Migration networks and immigrant entrepreneurship." In Parminder Bhachu, editor, *Immigration and Entrepreneurship: Culture, Capital, and Ethnic Networks*, volume V. Routledge, New York, NY, 1989.

[LCP14] Sang Hoon Lee, Mihai Cucuringu, and Mason A. Porter. "Density-based and transport-based core-periphery structures in networks." *Physical Review E*, **89**(3):032810, 2014.

[LEF16] Louis David Lord, Paul Expert, Henrique M. Fernandes, Giovanni Petri, Tim J. Van Hartevelt, Francesco Vaccarino, Gustavo Deco, Federico Turkheimer, and Morten L. Kringelbach. "Insights into brain architectures from the homological scaffolds of functional connectivity networks." *Frontiers in Systems Neuroscience*, **10**:85, 2016.

[LFP17] Sang Hoon Lee, Mark D. Fricker, and Mason A. Porter. "Mesoscale analyses of fungal networks as an approach for quantifying phenotypic traits." *Journal of Complex Networks*, **5**(1):145–159, 2017.

[LHS19] Jesse S. F. Levitt, Mustafa Hajij, and Radmila Sazdanovic. "Big data approaches to knot theory: Understanding the structure of the Jones polynomial." *arXiv:1912.10086*, 2019.

[Lib07] Kenneth G. Libbrecht. *The Art of the Snowflake : A Photographic Album.* Voyageur Press, Minneapolis, MN, 2007.

[Lib16] Kenneth G. Libbrecht. *Field Guide to Snowflakes : Identifying Crystal Types, the Science Behind Snowflakes, Observation Tools and Tips, a Close-Up Look at Nature's Art.* Voyageur Press, Minneapolis, MN, 2016.

[Lib19] Kenneth G. Libbrecht. "Snow crystals." *arXiv:1910.06389*, 2019.

[Lor07] Jan Lorenz. "Continuous opinion dynamics under bounded confidence: A survey." *International Journal of Modern Physics C*, **18**(12):1819–1838, 2007.

[LP19] Andrew Liu and Mason A. Porter. "Spatial strength centrality and the effect of spatial embeddings on network architecture." *arXiv:1910.01174*, 2019. *Physical Review E*, in press.

[LPG07] Daniel T. Lichter, Domenico Parisi, Steven Michael Grice, and Michael C. Taquino. "National estimates of racial segregation in rural and small-town America." *Demography*, **44**(3):563–581, 2007.

[LRS19] Renaud Lambiotte, Martin Rosvall, and Ingo Scholtes. "From networks to optimal higher-order models of complex systems." *Nature Physics*, **15**(4):313–320, 2019.

[LS84] John R. Logan and Mark Schneider. "Racial segregation and racial change in American suburbs, 1970–1980." *American Journal of Sociology*, **89**(4):874–888, 1984.

[Mal17] Timothy Malacarne. "Rich friends, poor friends: Inter-socioeconomic status friendships in secondary school." *Socius*, **3**:237802311773699, 2017.

[Man18] Antony S.R. Manstead. "The psychology of social class: How socioeconomic status impacts thought, feelings, and behaviour." *British Journal of Social Psychology*, **57**(2):267–291, 2018.

[Mar09] John Levi Martin. *Social Structures.* Princeton University Press, Princeton, NJ, 2009.

[Mck87] "The McKinney–Vento Homeless Assistance Act.", 1987. Title 42 USC Chapter 119, Subchapter VI, Part B: Education for Homeless Children and Youths.

[MD88] Douglas S. Massey and Nancy A. Denton. "The dimensions of residential segregation." *Social Forces*, **67**(2):281–315, 1988.

[Mee07] Ezra Meeker. *The Ox Team; Or, the Old Oregon Trail, 1852-1906; An Account of the Author's Trip Across the Plains, From the Missouri River to Puget Sound, at the Age of Twenty-Two, With an Ox and Cow Team in 1852, and of His Return With an Ox Team in the Year 1906, at the Age of Seventy-Six.* New York, NY, 1907.

[MG11] Wesley Earl Marshall and Norman W. Garrick. "Does street network design affect traffic safety?" *Accident Analysis and Prevention*, **43**(3):769–781, 2011.

[Mit69] J. Clyde Mitchell. *Social Networks in Urban Situations: Analyses of Personal Relationships in Central African Towns.* Manchester University Press, Manchester, 1969.

[Moo01] James Moody. "Race, school integration, and friendship segregation in America." *American Journal of Sociology*, **107**(3):679–716, 2001.

[MSW63] John W. Milnor, Michael Spivak, and Robert Wells. *Morse Theory.* Princeton University Press, Princeton, NJ, 1963.

[MTB15] Elizabeth Munch, Katharine Turner, Paul Bendich, Sayan Mukherjee, Jonathan Mattingly, and John Harer. "Probabilistic Fréchet means for time varying persistence diagrams." *Electronic Journal of Statistics*, **9**(1):1173–1204, 2015.

[MTP18] Barbara I. Mahler, Ulrike Tillmann, and Mason A. Porter. "Analysis of contagion maps on a class of networks that are spatially embedded in a torus." *arXiv:1812.09806*, 2018.

[MVP18] X. Flora Meng, Robert A. Van Gorder, and Mason A. Porter. "Opinion formation and distribution in a bounded-confidence model on various networks." *Physical Review E*, **97**(2):022312, 2018.

[MZ19] Facundo Mémoli and Ling Zhou. "Persistent homotopy groups of metric spaces." *arXiv:1912.12399*, 2019.

[NBP19] Silvia Nauer, Lucas Böttcher, and Mason A. Porter. "Random-graph models and characterization of granular networks." *Journal of Complex Networks*, p. doi:10.1093/comnet/cnz037, 2019.

[NCR95] David A. Noever, Raymond J. Cronise, and Rachna A. Relwani. "Using spider-web patterns to determine toxicity." *NASA Tech Briefs*, **19**(4):22148, 1995.

[New18] Mark E. J. Newman. *Networks.* Oxford University Press, Oxford, 2nd edition, 2018.

[OF03] Stanley Osher and Ronald Fedkiw. *Level Set Methods and Dynamic Implicit Surfaces*, volume 153. Springer-Verlag, Heidelberg, 2003.

[OPT17] Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, and Heather A. Harrington. "A roadmap for the computation of persistent homology." *EPJ Data Science*, **6**(1):17, 2017.

[PCV15] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. "Epidemic processes in complex networks." *Reviews of Modern Physics*, **87**:925, 2015.

[Pen07] Mathew Penrose. *Random Geometric Graphs*. Oxford University Press, Oxford, 2007.

[PG16] Mason A. Porter and James P Gleeson. *Dynamical Systems on Networks: A Tutorial*, volume 4 of *Frontiers in Applied Dynamical Systems: Reviews and Tutorials*. Springer International Publishing, Cham, 2016.

[Pia19] Alexis Piazza. Personal communication, 2019.

[PK93] Ruth D. Peterson and Lauren J. Krivo. "Racial segregation and black urban homicide." *Social Forces*, **71**(4):1001–1026, 1993.

[PKC10] Paul K. Piff, Michael W. Kraus, Stéphane Côté, Bonnie Hayden Cheng, and Dacher Keltner. "Having less, giving more: The influence of social class on prosocial behavior." *Journal of Personality and Social Psychology*, **99**(5):771–784, 2010.

[Por12] Mason A. Porter. "Small-world network." *Scholarpedia*, **7**(2):1739, 2012.

[Por18] Mason A. Porter. "What is... a multilayer network?" *Notices of the American Mathematical Society*, **65**(11):1419–1423, 2018.

[Por20] Mason A. Porter. "Nonlinearity + Networks: A 2020 Vision." In Panayotis G. Kevrekidis, Jesús Cuevas-Maraver, and Avadh Behari Saxena, editors, *Emerging Frontiers in Nonlinear Science*, volume 32 of *Nonlinear Systems and Complexity*. Springer International Publishing, Cham, 2020. In press, available at arXiv:1911.03805.

[PPD18] Lia Papadopoulos, Mason A. Porter, Karen E. Daniels, and Danielle S. Bassett. "Network analysis of particles and grains." *Journal of Complex Networks*, **6**(4):485–565, 2018.

[PPP17] Shai Pilosof, Mason A. Porter, Mercedes Pascual, and Sonia Kéfi. "The multilayer nature of ecological networks." *Nature Ecology and Evolution*, **1**:0101, 2017.

[Pum20] Denise Pumain. *Theories and Models of Urbanization: Geography, Economics and Computing Sciences*, volume 24. Springer International Publishing, Cham, 2020.

[PWC19] Liming Pan, Wei Wang, Shimin Cai, and Tao Zhou. "Optimal interlayer structure for promoting spreading of the susceptible-infected-susceptible model in two-layer networks." *Physical Review E*, **100**(2):022316, 2019.

[RK16] Henrik Ronellenfitsch and Eleni Katifori. "Global optimization, local adaptation, and the role of growth in distribution networks." *Physical Review Letters*, **117**(13):138301, 2016.

[RLD15] Henrik Ronellenfitsch, Jana Lasser, Douglas C. Daly, and Eleni Katifori. "Topological phenotypes constitute a new dimension in the phenotypic space of leaf venation networks." *PLoS Computational Biology*, **11**(12):e1004680, 2015.

[RLK19] Jason W. Rocks, Andrea J. Liu, and Eleni Katifori. "Revealing structure-function relationships in functional flow networks via persistent homology." *arXiv:1901.00822*, 2019.

[RM10] Jacob S. Rugh and Douglas S. Massey. "Racial segregation and the American foreclosure crisis." *American Sociological Review*, **75**(5):629–651, 2010.

[Rot17] Richard Rothstein. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright, New York, NY, 1st edition, 2017.

[RPJ16] Francisco A. Rodrigues, Thomas K.DM. Peron, Peng Ji, and Jürgen Kurths. "The Kuramoto model in complex networks." *Physics Reports*, **610**:1–98, 2016.

[Sch71] Thomas C. Schelling. "Dynamic models of segregation." *The Journal of Mathematical Sociology*, **1**(2):143–186, 1971.

[Sei98] Marc Seitles. "The perpetuation of residential racial segregation in America: Historical discrimination, modern forms of exclusion, and inclusionary remedies." *Journal of Land Use & Environmental Law*, **14**(1):89–124, 1998.

[SFK16] Jon Schleuss, Joe Fox, and Priya Krishnakumar. "California 2016 Election Precinct Maps." https://github.com/datadesk/california-2016-election-precinct-maps, 2016.

[SHP17] Bernadette J. Stolz, Heather A. Harrington, and Mason A. Porter. "Persistent homology of time-dependent functional networks constructed from coupled time series." *Chaos*, **27**(4):047410, 2017.

[Sim19] SimpleMaps. "World Cities Database.", 2019. Available at https://simplemaps.com/data/world-cities.

[SJ15] Pramod Shinde and Sarika Jalan. "A multilayer protein-protein interaction network analysis of different life stages in caenorhabditis elegans." *Europhysics Letters*, **112**(5):58001, 2015.

[SLC16] Marta Sarzynska, Elizabeth A. Leicht, Gerardo Chowell, and Mason A. Porter. "Null models for community detection in spatially embedded, temporal networks." *Journal of Complex Networks*, **4**(3):363–406, 2016.

[SMC07] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson. "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition." In M Botsch, R Pajarola, B Chen, and M Zwicker, editors, *Eurographics Symposium on Point-Based Graphics*, pp. 91–100. The Eurographics Association, 2007.

[Son17] Esther Song. "Administrative district boundaries of city of Shanghai, People's Republic of China, 2017." ArcGIS, 2017. available at https://www.arcgis.com/home/item.html?id=105f92bd1fe54d428bea35eade65691b.

[SPG19] Alina Sîrbu, Dino Pedreschi, Fosca Giannotti, and János Kertész. "Algorithmic bias amplifies opinion fragmentation and polarization: A bounded confidence model." *PLoS ONE*, **14**(3):e0213246, 2019.

[SR51] Ray Solomonoff and Anatol Rapoport. "Connectivity of random nets." *The Bulletin of Mathematical Biophysics*, **13**(2):107–117, 1951.

[SRC08] Ricard V. Solé, Martí Rosas-Casals, Bernat Corominas-Murtra, and Sergi Valverde. "Robustness of the European power grids under intentional attack." *Physical Review E*, **77**(2):026102, 2008.

[Tea20] The GIMP Development Team. "GIMP.", 2020. available at https://www.gimp.org.

[The17] The Regents of the University of California. "Dionysus." GitHub, 2017. available at https://www.mrzv.org/software/dionysus2/.

[TKH15] Dane Taylor, Florian Klimm, Heather A. Harrington, Miroslav Kramár, Konstantin Mischaikow, Mason A. Porter, and Peter J. Mucha. "Topological data analysis of contagion maps for examining spreading processes on networks." *Nature Communications*, **6**:7723, 2015.

[TMD20] Sarah Tymochko, Elizabeth Munch, Jason Dunion, Kristen Corbosiero, and Ryan Torn. "Using persistent homology to quantify a diurnal cycle in hurricanes." *Pattern Recognition Letters*, **133**:137–143, 2020.

[Tom17] Tomasz. "How Chicago Built its "Superhighways"." Medium, 2017. available at https://medium.com/@freerangehuman/how-chicago-built-its-superhighways-2a3803ce919c.

[TSW20] Jason Thompson, Mark Stevenson, Jasper S. Wijnands, Kerry Nice, Gideon Aschwanden, Jeremy Silver, Mark Nieuwenhuijsen, Peter Rayner, Robyn Schofield,

Rohit Hariharan, and Christopher N. Morrison. "Injured by design: A global perspective on urban design and road transport injury." *The Lancet*, **4**(1):PE32–E42, 2020.

[Var14] Kush R. Varshney. "Bounded confidence opinion dynamics in a social network of bayesian decision makers." *IEEE Journal on Selected Topics in Signal Processing*, **8**(4):576–585, 2014.

[VC64] Vladimir N. Vapnik and Alexey Chervonenkis. "A note on one class of perceptrons." *Automation and Remote Control*, **25**(6):937–945, 1964.

[Vie27] Leopold Vietoris. "Über den höheren zusammenhang kompakter räume und eine klasse von zusammenhangstreuen abbildungen." *Mathematische Annalen*, **97**(1):454–472, 1927.

[Wat02] Duncan J. Watts. "A simple model of global cascades on random networks." *Proceedings of the National Academy of Sciences of the United States of America*, **99**(9):5766–5771, 2002.

[WF94] Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, Cambridge, 1994.

[Wit71] Peter N. Witt. "Drugs alter web-building of spiders: A review and evaluation." *Behavioral Science*, **16**(1):98–113, 1971.

[WS98] Duncan J. Watts and Steven H. Strogatz. "Collective dynamics of 'small-world' networks." *Nature*, **393**(6684):440–442, 1998.

[WSK17] Yu Wu, Gabriel Shindnes, Vaibhav Karve, Derrek Yager, Daniel B. Work, Arnab Chakraborty, and Richard B. Sowers. "Congestion barcodes: Exploring the topology of urban congestion using persistent homology." In *2017 IEEE Conference on Intelligent Transportation Systems, Proceedings (ITSC)*, pp. 1–6, Yokohama, 2017. Institute of Electrical and Electronics Engineers Inc.

[XW14] Kelin Xia and Guo Wei Wei. "Persistent homology analysis of protein structure, flexibility, and folding." *International Journal for Numerical Methods in Biomedical Engineering*, **30**(8):814–844, 2014.

[YB20] Gökhan Yalnlz and Nazmi Burak Budanur. "Inferring symbolic dynamics of chaotic flows from persistence." *Chaos*, **30**(3):033109, 2020.

[Yin13] Fabian M. Ying. "Dynamical processes on Random Geometric Graphs.", 2013. Available at https://www.math.ucla.edu/~mason/research/fabian-report-092913.pdf.

[YKA16] Jaejun Yoo, Eun Young Kim, Yong Min Ahn, and Jong Chul Ye. "Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages." *Journal of Neuroscience Methods*, **267**:1–13, 2016.

[YWB19] Fabian Ying, Alisdair O.G. Wallis, Mariano Beguerisse-Díaz, Mason A. Porter, and Sam D. Howison. "Customer mobility and congestion in supermarkets." *Physical Review E*, **100**(6):062304, 2019.

[ZC04] Afra Zomorodian and Gunnar Carlsson. "Computing persistent homology." *Proceedings of the Annual Symposium on Computational Geometry*, **33**(2):347–356, 2004.

[ZDZ19] Quanbo Zha, Yucheng Dong, Hengjie Zhang, Francisco Chiclana, and Enrique Herrera-Viedma. "A personalized feedback mechanism based on bounded confidence learning to support consensus reaching in group decision making." *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–11, 2019.

[Zha11] Junfu Zhang. "Tipping and residential segregation: A unified Schelling model." *Journal of Regional Science*, **51**(1):167–193, 2011.

[Zom10] Afra Zomorodian. "Fast construction of the Vietoris–Rips complex." *Computers and Graphics*, **34**(3):263–271, 2010.