Cambridge University Press
978-1-009-09816-8 — Computational Topology for Data Analysis
Tamal Krishna Dey , Yusu Wang
Frontmatter
More Information

Computational Topology for Data Analysis

Topological data analysis (TDA) has emerged recently as a viable tool for analyzing complex data, and the area has grown substantially in both its methodologies and applicability. Providing a computational and algorithmic foundation for techniques in TDA, this comprehensive, self-contained text introduces students and researchers in mathematics and computer science to the current state of the field. The book features a description of mathematical objects and constructs behind recent advances, the algorithms involved, computational considerations, as well as examples of topological structures or ideas that can be used in applications. It provides a thorough treatment of persistent homology together with various extensions – like zigzag persistence and multiparameter persistence – and their applications to different types of data, like point clouds, triangulations, or graph data. Other important topics covered include discrete Morse theory, the mapper structure, optimal generating cycles, as well as recent advances in embedding TDA within machine learning frameworks.

TAMAL KRISHNA DEY is Professor of Computer Science at Purdue University. Before joining Purdue, he was a faculty member in the CSE Department of the Ohio State University. He has held academic positions at Indiana University–Purdue University at Indianapolis, Indian Institute of Technology Kharagpur, and Max Planck Institute. His research interests include computational geometry, computational topology, and their applications to geometric modeling and data analysis. He has (co)authored two books, *Curve and Surface Reconstruction: Algorithms with Mathematical Analysis* (Cambridge University Press) and *Delaunay Mesh Generation* (CRC Press), and (co)authored more than 200 scientific articles. Tamal is a fellow of the IEEE, ACM, and Solid Modeling Association.

YUSU WANG is Professor in the Halıcıoğlu Data Science Institute at the University of California, San Diego. Prior to joining UCSD, she was Professor of Computer Science and Engineering at the Ohio State University and postdoctoral fellow at Stanford University. Yusu primarily works in topological and geometric data analysis, developing effective and theoretically justified algorithms for data analysis using geometric and topological ideas, as well as in applying them to practical domains. She received the DOE Early Career Principal Investigator Award in 2006 and NSF Career Award in 2008.

“A must-have up-to-date computational account of a vibrant area connecting pure mathematics with applications.”

- Herbert Edelsbrunner, IST Austria

“This book provides a comprehensive treatment of the algorithmic aspects of topological persistence theory, both in the classical one-parameter setting and in the emerging multi-parameter setting. It is an excellent resource for practitioners within or outside the field, who want to learn about the current state-of-the-art algorithms in topological data analysis.”

- Steve Oudot, Inria and École Polytechnique

Computational Topology for Data Analysis

TAMAL KRISHNA DEY
Purdue University

YUSU WANG
University of California, San Diego

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of
education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009098168
DOI: 10.1017/9781009099950

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Names: Dey, Tamal K. (Tamal Krishna), 1964- author. | Wang, Yusu, 1976- author.
Title: Computational topology for data analysis / Tamal Krishna Dey,
Department of Computer Science, Purdue University, Yusu Wang, Halıcıoğlu
glu Data Science Institute University of California, San Diego.
Description: First edition. | New York : Cambridge University Press, 2022.
| Includes bibliographical references and index.
Identifiers: LCCN 2021041859 (print) | LCCN 2021041860 (ebook) | ISBN
9781009098168 (hardback) | ISBN 9781009099950 (ebook)
Subjects: LCSH: Topology.
Classification: LCC QA611 .D476 2022 (print) | LCC QA611 (ebook) | DDC
514/.7–dc23/eng/20211029
LC record available at https://lccn.loc.gov/2021041859
LC ebook record available at https://lccn.loc.gov/2021041860

ISBN 978-1-009-09816-8 Hardback

Contents

Preface

In recent years, the area of topological data analysis (TDA) has emerged as a viable tool for analyzing data in applied areas of science and engineering. The area started in the 1990s with the computational geometers finding an interest in studying the algorithmic aspect of the classical subject of algebraic topology in mathematics. The area of computational geometry flourished in the 1980s and 1990s by addressing various practical problems and enriching the area of discrete geometry in the course of doing so. A handful of computational geometers felt that, analogous to this development, computational topology has the potential to address the area of shape and data analysis while drawing upon and perhaps developing further the area of topology in the discrete context; see, for example, [26, 116, 119, 188, 292]. The area gained momentum with the introduction of persistent homology in early 2000 followed by a series of mathematical and algorithmic developments on the topic. The book by Edelsbrunner and Harer [149] presents these fundamental developments quite nicely. Since then, the area has grown in both its methodology and applicability. One consequence of this growth has been the development of various algorithms which intertwine with the discoveries of various mathematical structures in the context of processing data. The purpose of this book is to capture these algorithmic developments with the associated mathematical guarantees. It is appropriate to mention that there is an emerging sub-area of TDA which centers more around statistical aspects. This book does not deal with these developments, though we mention some of them in the last chapter where we describe the recent results connecting TDA and machine learning.

We have 13 chapters in the book listed in the table of contents. After developing the basics of topological spaces, simplicial complexes, homology groups, and persistent homology in the first three chapters, the book is then devoted to presenting algorithms and associated mathematical structures in various contexts of topological data analysis. These chapters present materials mostly not covered in any book on the market. To elaborate on this claim, we briefly give an overview of the topics covered by the present book. Chapter 4 presents a generalization of the persistence algorithm to extended settings such as to simplicial maps (instead of inclusions), and zigzag sequences with both inclusions and simplicial maps. Chapter 5 covers algorithms on computing optimal generators for both persistent and nonpersistent homology. Chapter 6 focuses on algorithms that infer homological information from point cloud data. Chapter 7 presents algorithms and structural results for Reeb graphs. Chapter 8 considers general graphs, including directed ones. Chapter 9 focuses on various recent results on characterizing nerves of covers, including the well-known mapper and its multiscale version. Chapter 10 is devoted to the important concept of discrete Morse theory, its connection to persistent homology, and its applications to graph reconstruction. Chapters 11 and 12 introduce multiparameter persistence. The standard persistence is defined over a one-parameter index set such as $\mathbb{Z}$ or $\mathbb{R}$. Extending this index set to a poset such as $\mathbb{Z}^d$ or $\mathbb{R}^d$, we get d-parameter or multiparameter persistence. Chapter 11 focuses on computing indecomposables for multiparameter persistence that are generalizations of bars in the one-parameter case. Chapter 12 focuses on various definitions of distances among multiparameter persistence modules and their computations. Finally, we conclude with Chapter 13, which presents some recent developments of incorporating persistence into the machine learning (ML) framework.

This book is intended for an audience comprising researchers and teachers in computer science and mathematics. Graduate students in both fields will benefit from learning the new materials in topological data analysis. Because of the topics, the book plays the role of a bridge between mathematics and computer science. Students in computer science will learn the mathematics in topology that they are usually not familiar with. Similarly, students in mathematics will learn about designing algorithms based on mathematical structures. The book can be used for a graduate course in topological data analysis. In particular, it can be part of a curriculum in data science which has been/is being adopted in universities. We are including exercises for each chapter to facilitate teaching and learning.

There are currently a few books on computational topology/topological data analysis on the market to which our book will be complementary. The materials covered in this book predominantly are new and have not been covered in any of the previous books. The book by Edelsbrunner and Harer [149] mainly focuses on early developments in persistent homology and do not cover the materials in Chapters 4–13 in this book. The recent book of Boissonnat et al. [39] focuses mainly on reconstruction, inference, and Delaunay meshes. Other than Chapter 6, which focuses on point cloud data and inference of topological properties, and Chapters 1–3, which focus on preliminaries about topological persistence, there is hardly any overlap. The book by Oudot [250] mainly focuses on algebraic structures of persistence modules and inference results. Again, other than the preliminary Chapters 1–3 and Chapter 6, there is hardly any overlap. Finally, unlike ours, the books by Tierny [286] and by Rabadán and Blumberg [260] mainly focus on applying TDA to specific domains of scientific visualizations and genomics, respectively.

This book, as any other, is not created in isolation. Help coming from various corners contributed to its creation. It was seeded by the class notes that we developed for our introductory course on Computational Topology and Data Analysis which we taught at the Ohio State University. During this teaching, the class feedback from students gave us the hint that a book covering the increasingly diversified repertoire of topological data analysis was necessary at this point. We thank all those students who had to bear with the initial disarray that was part of freshly gathering coherent material on a new subject. This book would not have been possible without our own involvement with TDA, which was mostly supported by grants from the National Science Foundation (NSF). Many of our PhD students worked through these projects, which helped us consolidate our focus on TDA. In particular, Tao Hou, Ryan Slechta, Cheng Xin, and Soham Mukherjee gave their comments on drafts of some of the chapters. We thank all of them. We thank everyone from the TGDA@OSU group for creating one of the best environments for carrying out research in applied and computational topology. Our special thanks go to Facundo Mémoli, who has been a great colleague (who has collaborated with us on several topics) as well as a wonderful friend at OSU. We also acknowledge the support of the Department of CSE at the Ohio State University where a large amount of the contents of this book were planned and written. The finishing came to fruition after we moved to our current institutions.

Finally, it is our pleasure to acknowledge the support of our families who kept us motivated and engaged throughout the marathon of writing this book, especially during the last stretch overlapping the 2020–2021 Coronavirus pandemic. Tamal recalls his daughter Soumi and son Sounak asking him continually about the progress of the book. His wife Kajari extended all the help necessary to make space for extra time needed for the book. Despite suffering from the reduced attention to family matters, all of them offered their unwavering support and understanding graciously. Tamal dedicates this book to his family and his late parents Gopal Dey and Hasi Dey without whose encouragement and love he would not have been in a position to take up this project. Yusu thanks her husband Mikhail Belkin for his never-ending support

and encouragement throughout writing this book and beyond. Their two children Alexander and Julia contributed in their typical ways by making every day delightful and unpredictable for her. Without their support and love, she would not have been able to finish this book. Finally, Yusu dedicates this book to her parents Qingfen Wang and Jinlong Huang, who always gave her space to grow and encouraged her to do her best in life, as well as to her great aunt Zhige Zhao and great uncle Humin Wang, who kindly took her into their care when she was 13. She can never repay their kindness.

Prelude

We make sense of the world around us primarily by understanding and studying the "shape" of the objects that we encounter in real life or in a digital environment. Geometry offers a common language that we usually use to model and describe shapes. For example, the familiar descriptors such as distances, coordinates, angles, and so on from this language assist us to provide detailed information about a shape of interest. Not surprisingly, people have used geometry for thousands of years to describe objects in their surroundings.

However, there are many situations where detailed geometric information is not needed and may even obscure the really useful structure that is not so explicit. A notable example is the *Seven Bridges of Königsberg* problem, where, in the city of Königsberg, the Pregel river separates the city into four regions, connected by seven bridges, as shown in Figure 1 (map and description taken from the Wikipedia page for "Seven bridges of Königsberg"). The question is to find a walk through the city that crosses each bridge exactly once. The story goes that the mathematician Leonhard Euler observed that factors such as the precise shape of these regions and the exact path taken are not important. What is important is the *connectivity* among the different regions of the city as connected by the bridges. In particular, the problem can be modeled abstractly using a graph with four nodes, representing the four regions in the city of Königsberg, and seven edges representing the bridges connecting them. The problem then reduces to what's later become known as finding the Euler tour (or Eulerian cycle) in this graph, which can be easily solved.

For another example, consider animation in computer graphics, where one wants to develop software that can continuously deform one object into another (in the sense that one can stretch and change the shape, but cannot break and add to the shape). Can we continuously deform a frog into a prince this way?[1]

[1] Yes, according to Disney movies.

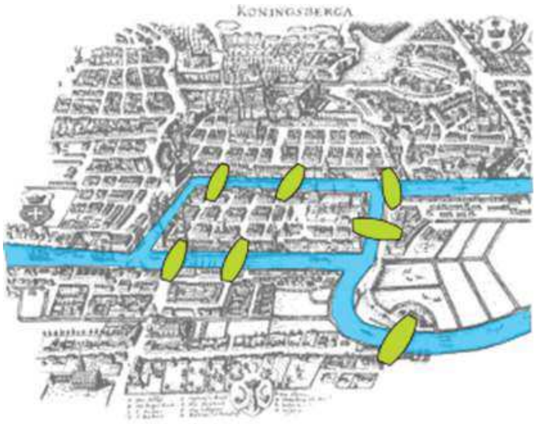

Figure 1 "Map of Königsberg in Euler's time showing the actual layout of the seven bridges, highlighting the river Pregel and the bridges" (the drawing by Bogdan Giuşcă is licensed under CC BY-SA 3.0).

Is it possible to continuously deform a tea cup into a bunny? It turns out the latter is not possible.

In these examples, the core structure of interest behind the input object or space is characterized by the way the space is connected, and the detailed geometric information may not matter. In general, topology intuitively models and studies properties that are invariant as long as the connectivity of space does not change. As a result, topological language and concepts can provide powerful tools to characterize, identify, and process essential features of both spaces and functions defined on them. However, to bring topological methods to the realm of practical applications, we need not only new ideas to make topological concepts and the resulting structures more suitable for modern data analysis tasks, but also algorithms to compute these structures efficiently. In the past two decades, the field of applied and computational topology has developed rapidly, producing many fundamental results and algorithms that have advanced both fronts. This progress has further fueled the significant growth of *topological data analysis* (TDA), which has already found applications in various domains such as computer graphics, visualization, materials science, computational biology, neuroscience, and so on.

In Figure 2, we present some examples of the use of topological methodologies in applications. The topological structures involved will be described later in the book.

An important development in applied and computational topology in the past two decades centers around the concept of *persistent homology*, which

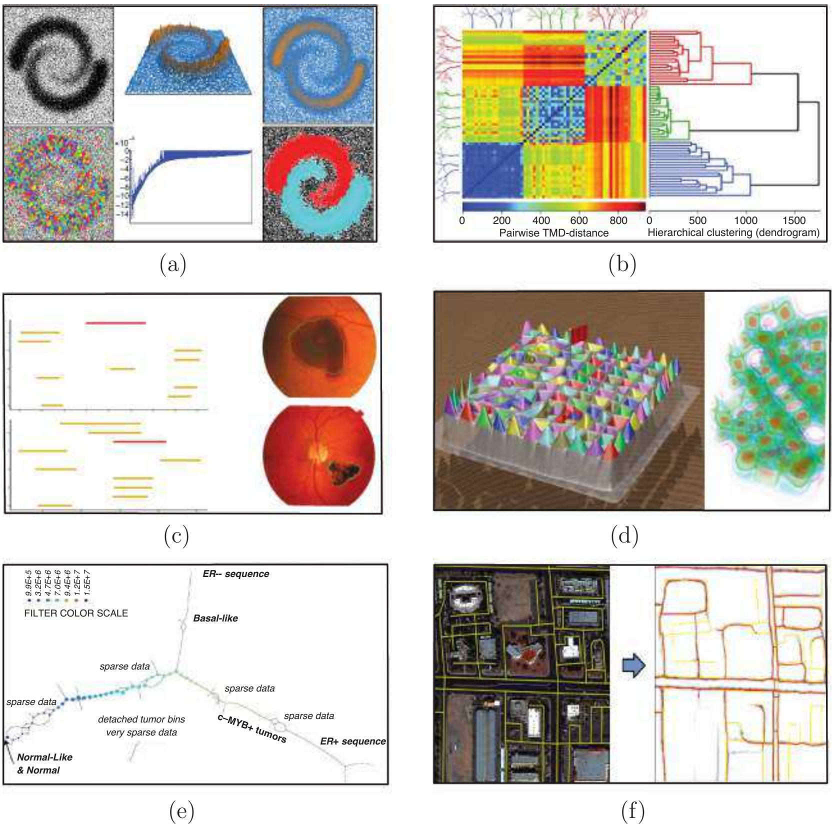

Figure 2 Examples of the use of topological ideas in data analysis. (a) A persistence-based clustering strategy. The persistence diagram of a density field estimated from an input noisy point cloud (shown in the top row) is used to help group points into clusters (bottom row). Reprinted by permission from Springer Nature: Springer Nature, Discrete and Computational Geometry, "Analysis of scalar fields over point cloud data," Chazal et al. [86], © 2011. (b) Using persistence diagram summaries to represent and cluster neuron cells based on their tree morphology. Image taken from [206] licensed by Kanari et al. (2018) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/). (c) Using the optimal persistent 1-cycle corresponding to a bar (red) in the persistence barcode, defects in diseased eyes are localized. Image taken from [127]. (d) Topological landscape (left) of the 3D volumetric silicium dataset from [299]. A volume rendering of the silicium dataset is on the right. However, note that it is hard to see all the structures forming the lattice of the crystal, while the topological landscape view shows clearly that most of them have high function values and are of similar sizes. Image taken from [299], reprinted by permission from IEEE: Weber et al. (2007). (e) Mapper structure behind the high-dimensional cell gene expression dataset can show not only the cluster of different tumor or normal cells, but also their connections. Image taken from [245], reprinted by permission from Nicolau et al. (2011, figure 3). (f) Using a discrete Morse-based graph skeleton reconstruction algorithm to help reconstruct road networks from satellite images even with few labeled training data. Image taken from [138].

generalizes the classic algebraic structure of homology groups to the multiscale setting aided by the concept of so-called *filtration* and *persistence modules* (discussed in Chapters 2 and 3). This helps significantly to broaden the applications of homological features to characterizing shapes/spaces of interest.

Figure 2(a) gives an example where persistent homology of a density field is used to develop a clustering strategy for the points [86]. In particular, at the beginning, each point is in its own cluster. Then, these clusters are grown using persistent homology, which identifies their importance and merges them according to this importance. The final output captures key clusters which may look like "blobs" or "curvy strips" – intuitively, they comprise dense regions separated by sparse regions.

Figure 2(b) gives an example where the resulting topological summaries from persistent homology have been used for clustering a collection of neurons, each of which is represented by a rooted tree (as neuron cells have tree morphology). We will see in Chapter 13 that persistent homology can serve as a general way to vectorize the features of such complex input objects.

In Figure 2(c), diseased parts of retinal degeneracy in two eyes are localized from image data. Algorithms for computing optimal cycles for bars in the persistent barcode as described in Chapter 5 are used for this purpose.

In Figure 2(d), we present an example where the topological object of a contour tree (the special loop-free case of the so-called Reeb graph as discussed in Chapter 7) has been used to give a low-dimensional terrain metaphor of a potentially high-dimensional scalar field. To illustrate further, suppose that we are given a scalar field $f: X \to \mathbb{R}$ where X is a space of potentially high dimension. To visualize and explore X and f in $\mathbb{R}^2$ and $\mathbb{R}^3$, just mapping X to $\mathbb{R}^2$ can cause significant geometric distortion, which in turn leads to artifacts in the visualization of f over the projection. Instead, we can create a 2D terrain metaphor $f': \mathbb{R}^2 \to \mathbb{R}$ for f which preserves the contour tree information as proposed in [299]; intuitively, this preserves the valleys/mountain peaks and how they merge and split. In this example, the original scalar field is in $\mathbb{R}^3$. However, in general, the idea is applicable to higher-dimensional scalar fields (e.g., the protein energy landscape considered in [184]).

In Figure 2(e), we give an example of an alternative approach of exploring a high-dimensional space X or functions defined on it via the mapper methodology (introduced in Chapter 9). In particular, the mapper methodology constructs a representation of the essential structure behind X via a pullback of a covering of Z through a map $f: X \to Z$. This intuitively captures the continuous structure of X at coarser level via the discretization of Z. See Figure 2(e), where the one-dimensional skeleton of the mapper structure behind a breast cancer microarray gene expression dataset is shown [245]. This continuous space representation not only shows "clusters" of different groups of tumors and of normal cells, but also how they connect in the space of cells, which are typically missing in standard cluster analysis.

Finally, Figure 2(f) shows an example of combining topological structures from the discrete Morse theory (Chapter 10) with convolutional neural networks to infer road networks from satellite images [138]. In particular, the so-called 1-unstable manifolds from discrete Morse theory can be used to extract hidden graph skeletons from noisy data.

We conclude this prelude by summarizing the aim of this book: introduce recent progress in applied and computational topology for data analysis with an emphasis on the algorithmic aspect.