Report from Dagstuhl Seminar 23192

Topological Data Analysis and Applications

Ulrich Bauer*[1], Vijay Natarajan*[2], and Bei Wang*[3]

1 **TU München, DE.** mail@ulrich-bauer.org
2 **Indian Institute of Science – Bangalore, IN.** vijayn@iisc.ac.in
3 **University of Utah – Salt Lake City, US.** beiwang@sci.utah.edu

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 23192 'Topological Data Analysis and Applications'. The seminar brought together researchers with backgrounds in mathematics, computer science, and different application domains with the aim of identifying and exploring emerging directions within computational topology for data analysis. This seminar was designed to be a followup event to two successful Dagstuhl Seminars (17292, July 2017; 19212, May 2019). The list of topics and participants were updated to reflect recent developments and to engage wider participation. Close interaction between the participants during the seminar accelerated the convergence between mathematical and computational thinking in the development of theories and scalable algorithms for data analysis, and the identification of different applications of topological analysis.

Seminar May 7–12, 2023 – https://www.dagstuhl.de/23192
2012 ACM Subject Classification Human-centered computing → Visualization; Information systems → Data analytics; Mathematics of computing → Algebraic topology; Theory of computation → Computational geometry
Keywords and phrases algorithms, applications, computational topology, topological data analysis, visualization
Digital Object Identifier 10.4230/DagRep.13.5.71

1 Executive Summary

Vijay Natarajan (Indian Institute of Science – Bangalore, IN)
Ulrich Bauer (TU München, DE)
Bei Wang (University of Utah – Salt Lake City, US)

This Dagstuhl Seminar titled "Topology, Computation, and Data Analysis" brought together researchers in mathematics, computer science, and visualization to engage in active discussions on theoretical, computational, practical, and application aspects of topology for data analysis.

Context

Topology is considered one of the most prominent research fields in mathematics. It is concerned with the properties of a space that are preserved under continuous deformations and provides abstract representations of the space and functions defined on the space. The modern field of topological data analysis (TDA) plays an essential role in connecting mathematical theories to practice. It uses stable topological descriptors as summaries

* Editor / Organizer

Topological Data Analysis and Applications, *Dagstuhl Reports*, Vol. 13, Issue 5, pp. 71–95
Editors: Ulrich Bauer, Vijay Natarajan, and Bei Wang
Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

of data, separating features from noise in a robust way. The seminar brought together researchers from mathematics, computer science, and application domains (e.g., materials science, neuroscience, and biology) to accelerate emerging research directions and inspire new ones in the field of TDA.

Goals

The Dagstuhl Seminars 17292 (July 2017) and 19212 (May 2019) were successful in enabling close interaction between researchers from diverse backgrounds. The attendees consistently remarked about the benefits of building bridges between the two communities. The goals from the previous seminars were to strengthen existing ties, establish new ones, identify challenges that require the two communities to work together, and establish mechanisms for increased communication and transfer of results from one to the other. A key goal of the current seminar was to additionally bring in experts from a few application domains to provide the necessary context for identifying research problems in topological data analysis and visualization. Furthermore, we also encouraged interaction between researchers who worked within the same community to identify challenging problems in the area and to establish new collaborations.

Topics

The research topics, listed below, reflect highly active and emerging areas in TDA. They were chosen to span topics in theory, algorithms, and applications.

Multivariate data analysis. Topics include theoretical studies of multivariate topological descriptors (including multiparameter persistence), efficient algorithms for computing and comparing them, formal guarantees for data analysis based on such comparisons, and the development of practical tools based on such analysis. Combining topological analysis together with statistical learning-based methods were also of interest.

Geometry and topology of metric spaces. A cornerstone of TDA is the study of metric and geometric data sets by means of filtrations of geometric complexes, formed by connecting subsets of the data points according to some proximity parameter. The study of such filtrations using homology leads to a multi-scale descriptor of the data that combines geometric and topological aspects of its shape. Besides their use in TDA, geometric complexes also play an important role in geometric group theory and metric geometry. The results and insights from both areas carry great promise for mutual interactions, leading to a unified view on computational and theoretical aspects.

Applications. TDA is an emerging area in exploratory data analysis and has received growing interest and notable successes with an expanding research community. The application of topological techniques to large and complex data has opened new opportunities in science, engineering, and business intelligence. This seminar focused on a few key application areas, including material sciences, neuroscience, and biology.

Parallel and distributed computation. The computational challenges in TDA call for the use of advanced techniques of high-performance computing, including parallel, distributed, and GPU-based software. Many of the core methods of TDA, including persistent homology,

mapper, merge trees, and contour trees, have received implementations beyond serial computing, and the interest in utilizing modern state-of-the-art techniques continues unabated. The task of optimizing algorithms in TDA is not only a question of engineering. Many of the key insights leading to breakthrough improvements are based on a careful utilization of theoretical properties and insights.

Participants, schedule, and organization

The invitees were chosen based on their background in mathematics, computer science, and application domains. We also ensured diversity in terms of gender, country or region of workplace, and experience.

While welcoming theoretical talks, the attendees were strongly encouraged to prepare a talk that is rooted in applications. The aim was to foster discussions on topics and projects related to practical applications of topological analysis and visualization. The program for the week consisted of talks of different lengths, breakout sessions, and summary / discussion sessions with all participants. We scheduled a total of six long talks (35 minutes + 10 minutes Q&A) on Day-1 and the morning session of Day-2, each providing an introduction either to one of the four chosen topics of the seminar or to a specific application domain. The talks were given by Yasu Hiraoka (Curse of dimensionality in persistence diagrams), Manish Saggar (Precision dynamical mapping to anchor psychiatric diagnosis into biology), Andreas Ott (Topological data analysis and coronavirus evolution), Kelin Xia (Mathematical AI for molecular data analysis), Gunther Weber (Topological analysis for exascale computing: challenges & approaches), and Facundo Mémoli (Some recent results about Vietoris-Rips persistence). Short research talks (16 total, 15 minutes + 10 minutes Q&A) were scheduled during the morning sessions of Day-2 through Day-5.

The afternoon sessions were devoted to discussions, working groups, and interactions. On Monday, we led an open problem session where participants identified different open problems and future directions for research. This initial discussion helped identify working groups and topics for discussion during the week. We organized breakout sessions on Tuesday and Thursday. On Tuesday, after a quick discussion regarding discussion topics, we identified four topics of interest. Participants chose one of the four groups> the curse of dimensionality, distances on Morse and Morse-Smale complexes, computation of generalized persistence diagrams, Codistortion and Gromov–Hausdorff distance. After a quick discussion, we decided to continue discussions on the four topics on Thursday, and some participants chose to join a different group.

We organized an excursion to Trier on Wednesday afternoon followed by dinner at a restaurant. Many participants attended the guided tour and the dinner.

All working groups summarized the discussion during their breakout sessions and presented it to all participants on Thursday evening and Friday morning. These summary sessions were also interactive and resulted in follow-up discussions between smaller groups of participants. We organized a final discussion and feedback session on Friday morning to close the seminar and to make future plans.

Results and reflection

The schedule for the first day helped initiate interaction between participants and continue the discussions during the week. While the introductory talks provided sufficient details on interesting application domains, the open problem session allowed many participants to quickly pose topics of interest. In particular, the format of this session extended beyond proposing specific stated open problems, asking also for contributions, discussion points, and thoughts that would not typically be brought up in such a session. This resulted in a very lively and engaging discussion that encouraged participants to share their perspectives on important current and future research directions.

In summary, we think that the seminar was successful in achieving the objective of encouraging discussions and interaction between researchers with backgrounds in mathematics, computer science, and application domains who are interested in the areas of topological data analysis and visualization. It helped identify new directions for research and has hopefully sparked the engagement of researchers from one community into the activities and research workshops and venues of the other. We strongly believe that the seminar provided a highly valuable contribution to bridging the gap between theory and applications in TDA.

The participants were highly appreciative of the balance between theoretical and applied topics and between the participants and those who presented during the week. They highlighted that the diverse group of participants sharing a strong interest in novel perspectives and exchange of ideas made the workshop an exceptional experience. Several felt that the discussions helped them identify topics for future research or introduced them to new collaboration possibilities.

2 Table of Contents

3 Overview of Talks

3.1 Quantifying and tracking inter-feature separation

Talha Bin Masood (Linköping University, SE)

Topological descriptors such as merge trees and extremum graphs have proven to be very useful for multiscale feature-based analysis of scalar field data. However, in some applications extraction of features is not enough, understanding the separation/topological distance between the extracted features is also important. Intrinsic tree distance can be used to quantify this topological distance. I will talk about two different applications where quantifying feature separation is useful and has physically interpretable meaning. One of the challenges that arise in the context of time-varying or ensemble scalar field data is tracking and visualization of the change in inter-feature separation with the change in time or input parameters. I will present some preliminary ideas and results in this direction.

3.2 Density-based Riemannian metrics and persistent homology

Ximena Fernández (Durham University, GB)

Joint work of Ximena Fernández, Eugenio Borghini, Gabriel Mindlin, Pablo Groisman
URL https://ximenafernandez.github.io/reveal.js-presentations/slides/FermatDistance_Dagstuhl.html#/

Several methods for geometric inference relies in the choice of an appropriate metric in the sample point cloud. Consider a scenario where the data is a noisy sample of a manifold embedded in Euclidean space, drawn according to a positive density over the manifold. I propose to learn a metric directly from the data (called *Fermat distance*) that turns out to be an estimator of an intrinsic density-based metric over the underlying manifold. I will show some convergence results, robustness properties of the use of this metric in the computation of persistent homology and some applications in real data. I will also discuss a couple of open questions.

3.3 Modified Finsler metrics for vector field visualization

Hans Hagen (RPTU – Kaiserslautern, DE)

Visualizing vector fields and their impact on free-form surface modelling like car hoods or airplane wings is a hot topic. We can “use” these vector fields to “deform” the metric of these surfaces, generating a Finsler metric. Can such a Finsler metric be useful for vector field visualization?

3.4 Persistent homology of a periodic filtration

Teresa Heiss (IST Austria – Klosterneuburg, AT)

Joint work of Teresa Heiss, Herbert Edelsbrunner, Chiara Martyka, Dmitriy Morozov

Persistent homology is well-defined and well studied for tame filtrations, for example various ones arising from finite point sets. However, periodic filtrations – for example used to study periodic point sets, like the atom positions of a crystal – are not tame, because there are infinitely many periodic copies of a homology class appearing at the same filtration value. We therefore extend the definition of persistent homology to periodic filtrations, which is a surprisingly difficult endeavor. In contrast to related work, we quantify how fast the multiplicities of persistence pairs tend to infinity with increasing window size, in a way that is stable under perturbations and invariant under different finite representations of the infinite periodic filtration. This project is still ongoing research, but I'll explain what we already know and what we don't know yet.

3.5 Curse of dimensionality in persistence diagrams

Yasuaki Hiraoka (Kyoto University, JP)

It is well known that persistence diagrams stably behave under small perturbations to the input data. This is the consequence of stability theorems, firstly proved by Cohen-Steiner, Edelsbrunner, and Harer (2007), and then extended by several researchers. On the other hand, if the input data is realized in a high-dimensional space with a small noise, the curse of dimensionality (CoD) causes serious adverse effects on data analysis, especially leading to inconsistency of distances. In this talk, I will show several examples of CoD appearing in persistence diagrams (e.g., from single-cell RNA sequencing data in biology). Those examples demonstrate that the classical stability theorems are not sufficient to guarantee stable behaviors of persistence diagrams for high-dimensional data. Then I will show several mathematical results about the existence and the (partial) resolution of CoD in persistence diagrams. This is a joint work with Liu Enhao, Yusuke Imoto and Shu Kanazawa.

3.6 Topological feature tracking in visualization applications

Ingrid Hotz (Linköping University, SE)

Topology in visualization – balance between beautiful concepts and practical needs

Tracking of features is a fundamental task in visual data analysis. In our work, we use topological descriptors as an abstraction for tracking. An essential step thereby is the choice of appropriate similarity measures to detect structural changes and establish a correspondence between individual features respecting their spatial embedding. One way to approach both demands is to consider labeled merge trees as the feature descriptor. In this talk, some examples of such approaches for tracking features are discussed.

3.7 The Density-Delaunay-Cech bifiltration

Michael Kerber (TU Graz, AT)

The density-Rips bifiltration is a standard construction in multi-parameter persistence, but suffers from the size explosion, as its single-parameter counterpart. On the other hand, it is well-known that at least in low Euclidean dimensions, alpha filtrations are much faster to compute and also geometrically more accurate. There are two major challenges to define and compute alpha-filtrations for two parameters. I will propose a way how to handle them. This is (very) ongoing work with Angel Alonso (TU Graz).

3.8 Towards a theory of persistence for gradient-like Morse-Smale vector fields

Claudia Landi (University of Modena, IT)

Joint work of Claudia Landi, Clemens Luc Bannwart

In topological data analysis, a function $f : M \longrightarrow R$ is often studied through the homology of its sublevel sets. One can obtain a topological summary of f in the form of a persistence barcode [3]. By a result of Morse theory, if M is a closed manifold and f is nice enough, then M is homotopy equivalent to a CW-complex with one k-cell for each critical point of index k [5]. Persistent homology and Morse theory are closely related, since the values of f at the critical points are equal to the start- and endpoints of the bars in the barcode. The gradient of f induces a chain complex, where the boundary operator is defined by counting the flow lines between critical points (see e.g. [1]). This process works more generally for gradient-like Morse-Smale vector fields and also for combinatorial vector fields in the sense of Forman [4]. However, for gradient-like Morse-Smale vector fields, there does not yet exist a persistence barcode such as for functions.

We present a pipeline that takes as an input a gradient-like Morse-Smale vector field on a surface, produces a parameterized epimorphic chain complex, and encodes it as a barcode. More precisely, we produce a sequence of chain complexes, such that the first one is the chain complex induced by the vector field and after that, each one is a quotient of the previous one. These quotients correspond to topological simplifications of the vector field by certain moves (introduced in [2]), and the times of taking the quotients depend on the value of a parameter measuring the local robustness of the vector field. In the end we are left with a vector field that has a very simple topological structure. Geometrically, for each move that is applied, we extract a topological feature. Algebraically, for each quotient, we split off an indecomposable contractible summand from the initial chain complex. Remembering the times when the moves were applied then yields a barcode. Similarly to the usual persistent homology construction for real valued functions, this pipeline paves the way for the development of a theory of persistence for vector fields.

References

1 A. Banyaga and D. Hurtubise, Lectures on Morse homology, Texts in the Mathematical Sciences, Springer Netherlands, 2004.
2 M. J. Catanzaro, J. Curry, B. T. Fasy, J. Lazovskis, G. Malen, H. Riess, B. Wang, and M. J. Zabka, Moduli spaces of Morse functions for persistence, Journal of Applied and Computational Topology (2019), 1–33.
3 H. Edelsbrunner and J. Harer, Persistent homology–a survey, Discrete & Computational Geometry – DCG 453 (2008).
4 R. Forman, Combinatorial vector fields and dynamical systems, Mathematische Zeitschrift 228 (1998), 629–681.
5 J. W. Milnor, Morse theory, Annals of mathematics studies, Princeton University Press, 1963.

3.9 The (not so) mysterious rhomboid bifiltration

Michael Lesnick (University at Albany, US)

The multicover bifiltration is a density-sensitive extension of the union-of-balls bifiltration commonly considered in TDA. It is robust to outliers, in a strong sense, and doesn't depend on any extra parameters. These properties make the multicover bifiltration a natural candidate for applications, if it can be computed. With this in mind, Edelsbrunner and Osang introduced a polyhedral bifiltration called the rhomboid bifiltration and gave a polynomial time algorithm for computing it. Corbet et al. showed that this bifiltration is topologically equivalent to the multicover bifiltration. In this talk, I'll give a poset-theoretic definition of the rhomboid tiling which is different from (but equivalent to) the one given by Edelsbrunner and Osang. With this as inspiration, I'll sketch a new proof of topological equivalence of the multicover and rhomboid bilfiltrations.

3.10 A spontaneous demo of the Topology ToolKit (TTK)

Joshua A. Levine (University of Arizona – Tucson, US)

In this short talk, I'll give a brief of overview of some of the features of the Topology ToolKit, a software package for topological data analysis of scalar fields. Rather than diving into the implementation details, this presentation will focus on ease of use and applications. To demonstrate, I'll walk through a surprise demo.

TTK comes shipped with Kitware's ParaView, and it can also be built from source. Many more examples are available at `https://topology-tool-kit.github.io/examples/index.html`.

3.11 Some recent results about Vietoris-Rips persistence

Facundo Mémoli (Ohio State University – Columbus, US)

Persistence barcodes provide computable signatures for datasets (metric spaces). These signatures absorb both geometric and topological information from metric spaces in a stable manner.

One question that motivated our work is: how strong are these signatures? A related question is that of ascertaining their relationship to other more classical invariants such as curvature.

In this talk I will describe some results about characterizing metric spaces via persistence barcodes arising from Vietoris-Rips filtrations. Of particular interest is a relationship which we established linking persistence barcodes to Gromov's filling radius.

Another aspect I will mention is the determination of the Gromov-Hausdorff distance between spheres (when endowed with their geodesic distance). In this case, VR-barcodes do permit telling spheres apart, but 1/2 of the bottleneck distance does not match the exact value of the GH-distance.

This work is joint with Sunhyuk Lim, Osman Okutan, and Zane Smith.

3.12 Topological optimization with big steps

Dmitriy Morozov (Lawrence Berkeley National Laboratory, US)

Joint work of Dmitriy Morozov, Arnur Nigmetov

Using persistent homology to guide optimization has emerged as a novel application of topological data analysis. Existing methods treat persistence calculation as a black box and backpropagate gradients only onto the simplices involved in particular pairs. We show how the cycles and chains used in the persistence calculation can be used to prescribe gradients to larger subsets of the domain. In particular, we show that in a special case, which serves as a building block for general losses, the problem can be solved exactly in linear time. We present empirical experiments that show the practical benefits of our algorithm: the number of steps required for the optimization is reduced by an order of magnitude.

3.13 Topological data analysis and coronavirus evolution

Andreas Ott (KIT – Karlsruher Institut für Technologie, DE)

Joint work of Michael Bleher, Lukas Hahn, Maximilian Neumann, Juan Ángel Patiño-Galindo, Mathieu Carrière, Ulrich Bauer, Raúl Rabadán, Andreas Ott, KIT Steinbuch Centre for Computing
Main reference Michael Bleher, Lukas Hahn, Maximilian Neumann, Juan Ángel Patiño-Galindo, Mathieu Carrière, Ulrich Bauer, Raúl Rabadán, Andreas Ott: "Topological data analysis identifies emerging adaptive mutations in SARS-CoV-2", medRxiv, Cold Spring Harbor Laboratory Press, 2023.
URL https://doi.org//10.1101/2021.06.10.21258550

Topological methods have in recent years found applications in the life sciences. In this talk, I will present an application of persistent homology to the surveillance of critical mutations in the evolution of the coronavirus SARS-CoV-2. I will explain the underlying geometric idea, how it connects with biology, its implementation in the CoVtRec pipeline, and some concrete results from the analysis of current pandemic data.

3.14 Precision dynamical mapping to anchor psychiatric diagnosis into biology

Manish Saggar (Stanford University, US)

URL https://braindynamicslab.github.io/projects/dp2/

Understanding the neurobiological underpinnings of psychiatric disorders has long been a challenge in the field of neuroscience. This talk aims to address this issue by exploring how noninvasive neuroimaging, despite its inherent limitations, can be leveraged to anchor psychiatric disorders into neurobiology. Two main challenges in this endeavor are identified: (a) the inherent noise in noninvasive neuroimaging devices, and (b) the limited utilization of biophysical models.

To tackle the first challenge, we propose the application of Topological Data Analysis (TDA), specifically Mapper, as a novel approach. I present some promising results on how Mapper can capture evoked transitions during tasks, intrinsic transitions during resting states, and changes in the landscape or shape associated with psychiatric disorders such as Major Depressive Disorder (MDD), Attention Deficit Hyperactivity Disorder (ADHD), as well as various pharmacological interventions (e.g., Methylphenidate, Psilocybin) and neuromodulation techniques (e.g., sp-TMS, rTMS).

I will also highlight methodological advances in TDA that enhance its applicability in the context of noninvasive neuroimaging studies. By harnessing the power of TDA, we can gain deeper insights into the complex dynamics of brain activity and its relation to psychiatric disorders.

Finally, the talk concludes by posing open questions that warrant further investigation. These questions touch upon the potential integration of TDA with other analytical approaches, the optimization of experimental protocols, and the translation of findings into clinical practice. By addressing these open questions, we can foster a greater understanding of the neurobiological basis of psychiatric disorders and pave the way for innovative therapeutic strategies.

3.15 Persistent homology of the multiscale clustering filtration

Dominik Schindler (Imperial College London, GB)

Joint work of Dominik Schindler, Mauricio Barahona
Main reference Dominik J. Schindler, Mauricio Barahona: "Persistent Homology of the Multiscale Clustering Filtration", CoRR, Vol. abs/2305.04281, 2023.
URL https://doi.org//10.48550/arXiv.2305.04281

In many applications in data clustering, it is desirable to find not just a single partition but a sequence of partitions that describes the data at different scales, or levels of coarseness, leading naturally to Sankey diagrams as descriptors of the data. The problem of multiscale clustering then becomes how to to select robust intrinsic scales, and how to analyse and compare the (not necessarily hierarchical) sequences of partitions. Here, we define a novel filtration, the Multiscale Clustering Filtration (MCF), which encodes arbitrary patterns of cluster assignments across scales. We prove that the MCF is a proper filtration, give an equivalent construction via nerves, and show that in the hierarchical case the MCF reduces to the Vietoris-Rips filtration of an ultrametric space. We also show that the zero-dimensional persistent homology of the MCF provides a measure of the level of hierarchy in the sequence of partitions, whereas the higher-dimensional persistent homology tracks the emergence and resolution of conflicts between cluster assignments across scales. We briefly illustrate numerically how the structure of the persistence diagram can serve to characterise multiscale data clusterings.

3.16 Persistence diagrams and Mobius inversion

Primoz Skraba (Queen Mary University of London, GB & Jožef Stefan Institute – Ljubljana, SI)

Joint work of Primoz Skraba, Amit Patel

There are many ways of defining persistence diagrams. In this talk I will discuss the definition based on the Mobius inversion function which was introduced by Amit Patel under the name Generalized Persistence Diagrams. I will cover how this approach has appeared implicitly and explicitly in various results on persistence as well as various implications of this approach and (very) new developments. In particular, I will cover a surprising connection between Euler characteristics and persistence diagrams and discuss the many questions and directions which arise.

3.17 Betti matching

Nico Stucki (TU München, DE)

Main reference Nico Stucki, Johannes C. Paetzold, Suprosanna Shit, Bjoern H. Menze, Ulrich Bauer: "Topologically faithful image segmentation via induced matching of persistence barcodes", CoRR, Vol. abs/2211.15272, 2022.
URL https://doi.org//10.48550/arXiv.2211.15272

Segmentation models predominantly optimize pixel-overlap-based loss, an objective that is actually inadequate for many segmentation tasks. In recent years, their limitations fueled a growing interest in topology-aware methods, which aim to recover the topology of the segmented structures. However, so far, existing methods only consider global topological properties, ignoring the need to preserve topological features spatially, which is crucial for accurate segmentation. We introduce the concept of induced matchings from persistent homology to achieve a spatially correct matching between persistence barcodes in a segmentation setting. Based on this concept, we define the Betti matching error as an interpretable, topologically and feature-wise accurate metric for image segmentation, which resolves the limitations of the Betti number error. The Betti matching error is differentiable and efficient to use as a loss function. We demonstrate that it improves the topological performance of segmentation networks significantly across six diverse datasets while preserving the performance with respect to traditional scores.

3.18 TGDA for graph learning?

Yusu Wang (University of California, San Diego – La Jolla, US)

In recent years, graph neural networks have emerged as a power family of ML architectures for graph learning and optimization. Nevertheless, various limitations and challenges remain. In this talk, I will briefly introduce the message passing graph neural networks (MPNN), and describe a few results in aiming to provide better understanding of GNNs or to enhance their power using geometric and topological ideas. My goal is to stimulate further discussions / interests / new perspectives in this interesting direction of TGDA + GNN.

3.19 Topological analysis for exascale computing: challenges and approaches

Gunther Weber (Lawrence Berkeley National Laboratory, US)

Simulation has quickly evolved to become the "third pillar of science" and supercomputing centers provide the computational power needed for accurate simulations. The Exascale Computing Project (ECP) is a concentrated effort to cross the next barrier and build a supercomputer that can run simulations at quintillion calculations per second. Exascale computing exacerbates the already existing I/O-bottleneck that makes it impossible to write

all simulation results to disk. To mitigate this problem, in situ approaches perform data analysis and visualization while the simulation is running. This talk provides an overview over how topological data analysis enables automated choice of visualization parameters like isovalue for isosurface extraction. It furthermore outlines the challenges that current developments in supercomputer architecture pose to efficient algorithm design for topological data analysis and presents solution approaches.

3.20 A distance for geometric graphs via labeled merge tree interleavings

Erin Moriarty Wolf Chambers (St. Louis University, US)

Geometric graphs appear in many real world datasets, such as road networks, sensor networks, and molecules. We investigate the notion of distance between graphs and present a semi-metric to measure the distance between two geometric graphs via the directional transform combined with the labeled merge tree distance. Our distance is not only reflective of the information from the input graphs, but also can be computed in polynomial time. We illustrate its utility by implementation on a Passiflora leaf dataset.

3.21 Mathematical AI for molecular data analysis

Kelin Xia (Nanyang TU – Singapore, SG)

Artificial intelligence (AI) based molecular data analysis has begun to gain momentum due to the great advancement in experimental data, computational power and learning models. However, a major issue that remains for all AI-based learning models is the efficient molecular representations and featurization. Here we propose advanced mathematics-based molecular representations and featurization (or feature engineering). Molecular structures and their interactions are represented as various simplicial complexes (Rips complex, Neighborhood complex, Dowker complex, and Hom-complex), hypergraphs, and Tor-algebra-based models. Molecular descriptors are systematically generated from various persistent invariants, including persistent homology, persistent Ricci curvature, persistent spectral, and persistent Tor-algebra. These features are combined with machine learning and deep learning models, including random forest, CNN, RNN, GNN, Transformer, BERT, and others. They have demonstrated great advantage over traditional models in drug design and material informatics.

3.22 Minimal cycle representatives in persistent homology using linear programming

Lori Ziegelmeier (Macalester College – St. Paul, US)

Main reference Lu Li, Connor Thompson, Gregory Henselman-Petrusek, Chad Giusti, Lori Ziegelmeier: "Minimal Cycle Representatives in Persistent Homology Using Linear Programming: An Empirical Study With User's Guide", Frontiers Artif. Intell., Vol. 4, p. 681117, 2021.
URL https://doi.org//10.3389/frai.2021.681117

Cycle representatives of persistent homology classes can be used to provide descriptions of topological features in data. However, the non-uniqueness of these representatives creates ambiguity and can lead to many different interpretations of the same set of classes. One approach to solving this problem is to optimize the choice of representative against some measure that is meaningful in the context of the data. In this work, we provide a study of the effectiveness and computational cost of several $\ell 1$-minimization optimization procedures for constructing homological cycle bases for persistent homology with rational coefficients in dimension one, including uniform-weighted and length-weighted edge-loss algorithms as well as uniform-weighted and area-weighted triangle-loss algorithms. We conduct these optimizations via standard linear programming methods, applying general-purpose solvers to optimize over column bases of simplicial boundary matrices.

Our key findings are: (i) optimization is effective in reducing the size of cycle representatives, (ii) the computational cost of optimizing a basis of cycle representatives exceeds the cost of computing such a basis in most data sets we consider, (iii) the choice of linear solvers matters a lot to the computation time of optimizing cycles, (iv) the computation time of solving an integer program is not significantly longer than the computation time of solving a linear program for most of the cycle representatives, using the Gurobi linear solver, (v) strikingly, whether requiring integer solutions or not, we almost always obtain a solution with the same cost and almost all solutions found have entries in -1, 0, 1 and therefore, are also solutions to a restricted $\ell 0$ optimization problem, and (vi) we obtain qualitatively different results for generators in Erdős-Rényi random clique complexes.

4 Working groups

4.1 Codistortion and Gromov–Hausdorff distance

Ulrich Bauer (TU München, DE) and Facundo Mémoli (Ohio State University – Columbus, US)

Let $\mathcal{M}$ be the collection of compact metric spaces. The *Gromov–Hausdorff distance* between (X, d_X) and (Y, d_Y) in $\mathcal{M}$ is defined as

$$d_{GH}(X,Y) = \frac{1}{2} \inf_{\phi: X \leftrightarrow Y: \psi} \max(\mathrm{dis}(\phi), \mathrm{dis}(\psi), \mathrm{codis}(\phi, \psi)),$$

where

$$\mathrm{dis}(\phi) = \sup_{x,x' \in X} |d_X(x,x') - d_Y(\phi(x), \phi(x'))|,$$
$$\mathrm{codis}(\phi,\psi) = \sup_{x \in X, y \in Y} |d_X(x, \psi(y)) - d_Y(\phi(x), y)|$$

are the *distortion* of a map and the *codistortion* of a pair of maps between metric spaces, respectively. Separating the distortion and codistortion terms in this formula for the Gromov–Hausdorff distance, we obtain the variants

$$\hat{d}_{GH}(X,Y) = \frac{1}{2} \inf_{\phi:X\leftrightarrow Y:\psi} \max(\operatorname{dis}(\phi), \operatorname{dis}(\psi)),$$
$$\check{d}_{GH}(X,Y) = \frac{1}{2} \inf_{\phi:X\leftrightarrow Y:\psi} \operatorname{codis}(\phi,\psi).$$

▶ **Example 1.** If $*$ denotes the one point metric space, then we have $\check{d}_{GH}(X,*) = \frac{1}{2}\operatorname{rad}(X)$.

Clearly $\hat{d}_{GH}, \check{d}_{GH} \leq d_{GH}$. The following facts about the *distortion distance* $\hat{d}_{GH}$ are known. Below $\cong$ denotes the equivalence relation of isometry on $\mathcal{M}$.

1. $\hat{d}_{GH}$ is a legit distance on the set of isometry classes of compact metric spaces $\mathcal{M}/\cong$.
2. $\hat{d}_{GH}$ and d_{GH} generate the same topology.
3. $\hat{d}_{GH} \neq d_{GH}$.
4. $\hat{d}_{GH}$ can be computed via curvature sets.

Less is known about the *codistortion distance* $\check{d}_{GH}$. We state a few interesting questions.

1. Is $\check{d}_{GH}$ a distance on $\mathcal{M}/\cong$?
2. Is $\check{d}_{GH}$ bi-Lipschitz equivalent to d_{GH}?

In our discussion group, we answered these questions to the affirmative.

▶ **Proposition 2.**

$$\check{d}_{GH} \leq d_{GH} \leq 2\check{d}_{GH}.$$

▶ Remark. The inequality $d_{GH} \leq 2\check{d}_{GH}$ is tight. To see this, consider the finite metric spaces X consisting of two points at distance 4 and Y consisting of the three points $\{0,2,3\}$ on the real line (with the usual metric).

▶ **Theorem 3.** *$\check{d}_{GH}$ a legitimate distance on $\mathcal{M}/\cong$.*

The following lemma is key to relating distortion and codistortion.

▶ **Lemma 4.** *Consider a pair of maps $\phi : X \leftrightarrow Y : \psi$ between metric spaces. Then*

- $\operatorname{codis}(\phi,\psi) \geq \sup_{x\in X} d_X(x, \psi\circ\phi(x))$.
- $2\operatorname{codis}(\phi,\psi) \geq \max(\operatorname{dis}(\phi), \operatorname{dis}(\psi))$,

Further insights and questions

1. $\hat{d}_{GH}$ is not bi-Lipschitz equivalent to d_{GH}. There is a family of pairs of finite ultrametric spaces $(X_k, Y_k)_k$ such that $d_{GH}(X_k,Y_k) \geq \frac{k}{2}$ but $\hat{d}_{GH}(X_k,Y_k) \leq 1$.
2. For all compact metric spaces X and Y we have

   $$\check{d}_{GH}(X,Y) \geq \frac{1}{4} d_B(\operatorname{dgm}(X), \operatorname{dgm}(Y)),$$

   where $\operatorname{dgm}(X)$ denotes the usual persistence diagram of the Vietoris-Rips filtration of X. Can this be improved to

   $$\check{d}_{GH}(X,Y) \geq \frac{1}{2} d_B(\operatorname{dgm}(X), \operatorname{dgm}(Y))?$$

   The stronger bound, when combined with the fact that $\check{d}_{GH} \leq d_{GH}$, would imply an improvement upon the usual Gromov–Hausdorff stability theorem for persistence diagrams arising from Vietoris-Rips filtrations.

3. Is it true that $\check{d}_{GH} \geq \check{d}_{GH} \circ H^{sl}$, where H^{sl} is single-linkage clustering, taking a finite metric space to a finite ultrametric space?
4. Is there a case where $\check{d}_{GH} < d_{GH}$?
5. What are natural lower bounds for $\check{d}_{GH}$? One of them is half the the difference of the respective radii of the spaces:

$$\check{d}_{GH}(X,Y) \geq \frac{1}{2}|\mathrm{rad}(X) - \mathrm{rad}(Y)|.$$

6. If $\mathrm{codis}(\phi,\psi) < \delta$, then $\mathrm{codis}(\phi \circ \psi \circ \phi, \psi \circ \phi \circ \psi) < 2\delta$.
7. If X and Y are ultrametric, do we have $\check{d}_{GH}(X,Y) = d_{GH}(X,Y)$?
8. Is there a constant $C > 0$ such that $d_H^Y(\phi(X),Y), d_H^X(\psi(Y),X) \leq C \cdot \mathrm{codis}(\phi,\psi)$?

4.2 Curse of Dimensionality

Teresa Heiss (IST Austria – Klosterneuburg, AT), Ximena Fernández (Durham University, GB), Yasuaki Hiraoka (Kyoto University, JP), Claudia Landi (University of Modena, IT), Andreas Ott (KIT – Karlsruher Institut für Technologie, DE), Manish Saggar (Stanford University, US), and Dominik Schindler (Imperial College London, GB)

The stability result of Persistent Homology is not guaranteeing much in very high dimensions, when noise of at most ε is added in each dimension to the data. Indeed, the ℓ_2-distance between a point $x \in \mathbb{R}^d$ and its perturbed point $x+p$ with $||p||_{\ell_\infty} < \varepsilon$ is $||p||_{\ell_2} \leq \sqrt{d}\varepsilon$. Hence, the stability bound, namely the Hausdorff distance between the original and the perturbed point set, is $O(\sqrt{d}\varepsilon)$ as well.

We prove that this effect, the curse of dimensionality, cannot be circumvented in full generality, i.e., when an adversary is allowed to make the choices. This shows that we need some assumption on the data. We list some ideas for possible assumptions, and approaches that seem promising within these different assumptions.

Setting

The setting is for example motivated by gene expression data, with few (s) essential genes, many more ($d-s$) housekeeping genes, and a small measuring error for each gene. The persistence diagram of such data will, due to the curse of dimensionality, be very different than the desired persistence diagram of only the essential genes.

Given $\varepsilon > 0$, $s \in \mathbb{N}$, $A \subseteq \mathbb{R}^s$, an integer $d > s$, an affine linear map $L : \mathbb{R}^s \to \mathbb{R}^d$ with determinant 1, and for every $a \in A$, a vector p_a with $||p_a||_{\ell_\infty} < \varepsilon$. We denote the embedded point set $L(A)$ by X and the perturbed set $\{L(a) + p_a \mid a \in A\}$ by Y.

The Gromov-Hausdorff distance is $d_{GH}(A,Y) = d_{GH}(X,Y) \leq \max_{a\in A}||p_a||_{\ell_2} \leq \sqrt{d}\varepsilon$ and thus for $d >> s$ the stability theorem does not give a good bound for the Vietoris-Rips persistence diagrams:

$$d_B(PD(VR(A)), PD(VR(Y))) \leq 2d_{GH}(A,Y) = O(\sqrt{d}\varepsilon). \tag{1}$$

Note that since we consider getting $d_B(PD(VR(A)), PD(VR(Y))) = O(1)$ by matching everything to the diagonal as "cheating", one might want to consider a distance between persistence diagrams that does not allow matchings with the diagonal, like the Hausdorff distance. Another approach is to keep using the bottleneck distance and not be satisfied with $d_B(PD(VR(A)), PD(VR(Y))) = O(1)$, but insisting on wanting $d_B(PD(VR(A)), PD(VR(Y))) = O(1)\varepsilon$ or $o(1)$ as ε goes to 0.

We are searching for a modification $Z \subseteq \mathbb{R}^d$ of the observed data Y, such that $d_B(PD(VR(A)), PD(VR(Z))) = o(1)$. For example by dimensionality reduction.

When the Adversary Makes the Choices

There cannot be any fix to the curse of dimensionality in full generality, as the following argument shows. For every ε, and every $s \geq 1$, an adversary can choose

- the point set A as two points on the x-axis with distance 1 from each other,
- the embedding dimension $d > \frac{1}{\varepsilon^2}$,
- the affine linear map L with determinant 1 to map the x-axis to the direction spanned by the vector $(1, 1, \ldots, 1)$,
- and the two vectors p_1 and p_2 such that $L(a_1)+p_{a_1} = L(a_2)+p_{a_2}$, e.g. $p_{a_1} = \frac{1}{\sqrt{d}}(1, 1, \ldots, 1)$ and $p_{a_2} = 0$. As $\frac{1}{\sqrt{d}} < \varepsilon$, such a perturbation is allowed.

Then, the observed set Y consists of two points in the same spot, and thus does not have any non-essential homology, whereas A has a persistence pair with persistence 1. Hence, the distance $d_B(PD(VR(A)), PD(VR(Y))) = 1$ does not converge to zero when ε goes to zero. Furthermore, since all structure of A has been destroyed in Y, there is no hope to reconstruct an adequate modification Z to reconstruct the persistence of A, since when only given Y, we cannot know whether it has been created from a set A consisting of two points in the same position that have not been perturbed at all or from the above set A.

This shows that in order to have a chance against the curse of dimensionality, we need some assumptions on our data, instead of letting the adversary choose the data. Note that in the proof above, it was essential to let the adversary choose the embedding dimension d, the map L, and the perturbation vectors. Choosing A does not seem to be essential, it just makes the proof more convenient.

Possible Assumptions

Since we want to talk about the curse of dimensionality, we do not want to bound the embedding dimension d but instead keep letting the adversary choose d, or in other words imagine d as very large. Instead we can make assumptions on the affine linear map L or on the perturbation vectors:

1. A weak assumption would be assuming that the perturbation vectors have the form $p_a = w_a v_a$ with $w_a > 0$ an unknown constant depending on a, and v_a i.i.d. with an unknown distribution.
2. One can strengthen this by assuming a fixed known distribution for the v_a.
3. Or assuming that $w_a = 1$.
4. Another approach is assuming that the map L is axis-parallel or not too far from axis parallel.
5. Maybe assuming that the data is sampled very densely (for example $\frac{d}{n}$ converging to a constant).

Possible Solutions

Possible ideas how to pass from Y to Z:

- Assuming Assumption 4 above, there are $d - s$ coordinates that are pure noise. Hence, choose the s coordinates with the most variance and set the other coordinates to zero. In application where the noise p_a might be approximately linearly depending on the length $||a||_{\ell_2}$, one could for example replace the variance by the variance divided by the mean.
- Assuming at least Assumption 1 above: Use neural network auto-encoder (and afterwards possibly UMAP?) for dimensionality reduction from Y to Z.
- Assuming assumption 2: Mimic what RECODE does, namely, if I understand correctly, using a PCA technique that is designed by statisticians for weakening the curse of dimensionality for that particular distribution.
- Assuming at least Assumption 1 above: Dominik's idea: apply some variant of hierarchical clustering to observed data and obtain dendogram -> this leads to an ultrametric -> we can analyze the new ultrametric space with Vietoris Rips Persistent Homology. Alternatively: clustering and then MCF.
- Assume at least Assumptions 1 and 5 above: Hope that something like the law of large numbers would yield that the effects of the many strong (up to $O(\sqrt{d}\epsilon)$) perturbations average each other out, such that a degree-Rips / density-Rips approach or something similar to Ximena's work might filter out the noise.
- Assuming maybe Assumption 1 or 4 above: Some kind of bootstrap idea would be to subsample, say s (or a bit more) out of d, dimensions many times, knowing that most of the time one would mostly just get the noise, but maybe there is some way to distinguish the non-noise persistence diagrams from the purely-noise ones. The advantage would be that the diagram where the correct s dimensions are selected, would not have the curse of dimensionality. But it seems difficult to extract this useful information from the huge bag of persistence diagrams. Furthermore, $\binom{d}{s}$ is very large, so it does not seem feasible from a computational perspective.
- Additional to the other ideas, Primoz Skraba said that it might help to look at persistence rather as death divided by birth, rather than death – birth. However, that alone would not be enough of course.

4.3 Computation of Generalized Persistence Diagrams

Michael Lesnick (University at Albany, US), Teresa Heiss (IST Austria – Klosterneuburg, AT), Michael Kerber (TU Graz, AT), Dmitriy Morozov (Lawrence Berkeley National Laboratory, US), Primoz Skraba (Queen Mary University of London, GB & Jožef Stefan Institute – Ljubljana, SI), and Nico Stucki (TU München, DE)

One breakout session was focused loosely on understanding the problem of computing generalized persistence diagrams (GPDs), as defined by Kim and Mémoli.

Given a poset P, persistence module $M : P \to \mathrm{Vec}$, and a *generalized interval (a.k.a. spread)* $I \subset P$, the generalized rank of M over I is the rank of the map

$$\lim_I M \to \operatorname{colim}_I M.$$

The map sending each such I to its generalized rank is the *generalized rank invariant* (GRI) of M. Taking the Möbius inversion of the GRI yields the *generalized persistence diagram* (GPD), a kind of signed barcode for generalized persistence.

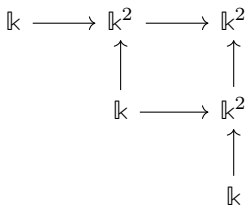
Signed barcodes have become a hot topic in TDA in the last few years. There are multiple ways to define a signed barcode, namely, by taking Möbius inversions of different functions, or by relative homological algebra with respect to different exact structures.

Among the various options, the GPD studied here is an appealing choice because it is a relatively rich invariant and also has a very simple interpretation in the case of *spread-decomposable modules*: On such modules, the GPD simply counts the number of copies of each spread in the decomposition. In contrast, other types of signed barcodes can be rather complicated on such modules. This makes the problem of computing the GPD interesting. This problem is mostly open, in spite of some interesting recent work by Dey, Kim, Mémoli on the related problem of computing the GRI at fixed indices.

Our group explored (in a very preliminary way), the following related questions:

- What does the GPD look like on specific examples of non-spread decomposable modules? How quickly does its size grow as the support of the module grows.
- In the special case that M has a small *encoding* in the sense of Ezra Miller's work (i.e., there exists a surjection of posets $f : P \to Q$ and a functor $N : Q \to \mathrm{Vec}$ with $f = N \circ g$ and $|Q|$ small), is efficient computation of M possible? How does the complexity of computing the GPD depend on $|Q|$? What bounds on $|Q|$ can be expected in practical 2-parameter persistence computations? How does one compute f?
- Can the ideas underlying recent work by Morozov and Patel on the output-sensitive computation of signed barcodes for 2-parameter persistence also be useful for computing GPDs?

There was some progress made. The first example we looked at was



We made an attempt to understand if a module was nearly interval indecomposable, how complex could the generalized rank invariant be. Under the appropriate choice of morphisms, the above decomposes into many non-trivial pieces in the generalized rank invariant.

Following up on this, there was also a discussion on whether such non-interval indecomposables occur in practice/arises in random settings, e.g. from random point clouds. Michael Kerber suggested a bifiltration example which was finite and the point positions were generic, i.e. a positive but small perturbation does not affect the decomposition. We then showed that this will occur in a uniform Poisson point process with probability 1 as the number of points goes to ∞. The idea behind the proof is that one can define a random variable of the event of such a configuration occurring, i.e. using the small neighborhood of the example. As the configuration is finite and generic, the probability of the event is strictly positive. As the expected number of such neighborhoods goes to ∞, the probability must go to 1. Michael Kerber also reported his experimental results which indeed show that non-interval indecomposables arise in many experiments.

While we did not make decisive progress, the discussions were illuminating and left us with a better understanding of invariants and their computation.

4.4 Distances on Morse and Morse-Smale complexes

Erin Moriarty Wolf Chambers (St. Louis University, US), Ulrich Bauer (TU München, DE), Talha Bin Masood (Linköping University, SE), Ximena Fernández (Durham University, GB), Hans Hagen (RPTU – Kaiserslautern, DE), Ingrid Hotz (Linköping University, SE), Claudia Landi (University of Modena, IT), Joshua A. Levine (University of Arizona – Tucson, US), Vijay Natarajan (Indian Institute of Science – Bangalore, IN), Dominik Schindler (Imperial College London, GB), Bei Wang (University of Utah – Salt Lake City, US), Yusu Wang (University of California, San Diego – La Jolla, US), Gunther Weber (Lawrence Berkeley National Laboratory, US), Kelin Xia (Nanyang TU – Singapore, SG), and Lori Ziegelmeier (Macalester College – St. Paul, US)

Given $f\colon \mathcal{M} \to \mathbb{R}$, the Morse complexes partition $\mathcal{M}$ into ascending/descending manifolds of minima/maxima. The Morse-Smale complex is the intersection of the two Morse complexes. Given f_1, f_2, and their Morse-Smale complexes MSC_1, MSC_2, how to define distances or metrics $d(MSC_1, MSC_2)$? This working group met on two days and focused on brainstorming likely lists of ideas worth pursuing, hopefully sparking ideas for future work in the participants when tackling this difficult and surprisingly open problem.

Summary from Day 1

We first discussed how the Morse and Morse-Smale complexes are defined, and how they can differ. Note that we often require that f has some 'niceness' assumptions about how the ascending/descending manifolds intersect, requires transversality, etc. These assumptions are fairly common, and ensure controlled behavior such as degree constraints on the graph and genericity of the resulting curves.

We then brainstormed a list of possible "objects" we could compute distances on (i.e. which piece of the complex) and what sorts of distances we could compute on that object. All of these objects are some structure given on the Morse complex, but some retain more structure (i.e. just keeping the graph versus using information about the 2-dimensional pieces of the complex).

1. First, we discussed just keeping a 1-dimensional skeleton (specifically, the Morse Graph of separatrices), rather than the full complex. With this information, we could consider any of the following distances:
   - Graph-based distances, e.g. interleaving and edit distances. There is a wealth of these in the literature, but it is unclear if they utilize the real structure of the Morse complex.
   - Distances on geometrically-embedded graphs / metric graphs (e.g. edge lengths + function values). These are well studied, but often computationally intractable.
   - Distances based on computational geometric measures (e.g. Frechet distances): These are well studied in computational geometry, but again unclear how they match with Morse graphs.
   - Distances based on optimal transport
   - Graph / graph kernel / spectral methods

2. Using a 2D complex: This retains more information from the Morse complex, but higher dimensional comparisons are plausibly more difficult. We considered the following options for this structure:
   - Information-theoretic distances (KL divergence)
   - Partition-based distances (Rand index)
   - Interleaving distance on the Reeb space: While interleavings on Reeb graphs are more well known, the basic idea should extend up a dimension to Reeb spaces as well.
   - Optimal transport: There is preliminary work on these in the viz community, so they may be more tractable.
   - Haussdorf distance / CG metrics / Frechet: Many of these become NP-Hard on two dimensional surfaces or even terrains or polygons with holes, but they are not well-studied on Morse complexes.
   - Distance on the Hasse graph: This yields a much different graph, which perhaps would be amenable to different types of computations but still captures much of the topology and adjacency information.
3. Finally, we also mentioned a few alternatives and other objects we could use which are based on the Morse complex as well:
   - One option was the extremal graph, which is a subset of the 1-skeleton, rather than the full skeleton. This perhaps is a simpler object than the full skeleton retaining the most 'interesting' information, although it is not clear what would be lost.
   - Dual graph of face adjacency: Again, this retains something interesting but flips the graph to the dual, which in some cases in computational geometry will allow different operations than the primal gives and/or can have nicer properties.

After discussing options of what objects we could to study, we then considered what desirable properties of interest would exist for such metrics, both theoretical and practical. These include:

1. Stability wrt small changes in scalar field, i.e. a bound on $d(MSC_1, MSC_2)$ vs $\|f_1 - f_2\|$
2. Stability wrt topological simplification of the field
3. Metric properties (i.e. triangle inequality, symmetric, etc)
4. Universality: This is an idea from topological data analysis, which looks for the most descriptive option amongst stable metrics. There has been recent work on Reeb graphs, which perhaps may be extended to the slightly more general Morse complex.
5. Discriminativity: Again drawing from Reeb graph metrics literature, there are many times when one distance is strictly more powerful (often at the cost of complexity). This may be a useful notion in order to compare the relative power of metrics on Morse complexes as well.
6. Computability (and/or heuristics to reduce computation time)
7. Interpretability / Locality (i.e., edit distance can tell us which edits cost what, so the cost has a discrete mapping which can be considered on its own)
8. Practicality, as opposed to worse case computational complexity: Which distances are actually feasible to implement and/or approximate?

Summary from Day 2

Thursday began with a brief review, but then we focused on a couple of specific possible directions, discussing how best to proceed in computing and/or using a distance computed in that manner. We outlined several promising approaches, which we would like to propose as likely directions for developing distances.

First, we began with a discussion of what could be done when focusing on the full complex. In this case, we considered the optimal transport approach, which seems most likely to succeed in practice, although there are some interesting challenges.

- We began by discussing how to compute optimal transport metrics, in terms of optimizing the matching to connectivity while also preserving associated properties stored on vertices (position, function value), edges (edge length, edge geometry), etc.
- We then determined several strategies towards extending this notion to Morse complexes. In particular, we see some complexity with extending the adjacency portion of these to a cell-based metric. The basic construction we considered most likely creates a bipartite graph of the adjacencies between cells of dimensions differing by 1. The challenge will then be managing this across all dimensions in a consistent way.

While there are significant challenges with optimal transport, it nonetheless seems a major alternative worth future study, given its success in other practical domains.

Our next portion of the discussion was based on the recent success of the study of Reeb graph metrics. While perhaps less practical, these have desirable theoretical properties, and so it seems worth investigating which might generalize to mroe general Morse or Morse-Smale complexes.

- Interleaving distances are well studied in topological data analysis, and in Reeb graphs have a nice combinatorial characterization via the thickening functor. In addition, interleavings are computable on general persistence modules and are fixed parameter tractable on simple classes of Reeb graphs. To the best of our group's knowledge, there is no notion of interleavings formally defined on Morse complexes, but the theoretical definitions would likely generalize.
- Edit distances are well studied on graphs and combinatorial objects, and appeal to computer scientists given their utility in other domains. On Reeb graphs, they have been generalized in an unusual way in order to prove stability and universality in quite recent work. We again are unaware of any work generalizing these edits distances to Morse complexes.
- Two more recently defined Reeb metrics are the functional distortion and contortion distances, which draw inspiration from Gromov-Hausdorff notions of metrics. One possible approach to generalize this to Morse complexes is to 'thicken' the space along the normal direction of separatrices, rather than along the function value space.

Finally, the group discussed complexity. Unfortunately, we suspect many if not all of these notions will be difficult to compute. There is perhaps hope of approximation or heuristics, but work remains even on Reeb graphs.

We concluded with a general discussion of other computational geometry and topology notions which have been used in simpler settings, such as Fréchet-based distances, local homology, and persistence distortion. Unfortunately, none seemed obvious candidates for study on Morse complexes.

Participants

- Ulrich Bauer
TU München, DE
- Talha Bin Masood
Linköping University, SE
- Ximena Fernández
Durham University, GB
- Hans Hagen
RPTU – Kaiserslautern, DE
- Teresa Heiss
IST Austria – Klosterneuburg, AT
- Yasuaki Hiraoka
Kyoto University, JP
- Ingrid Hotz
Linköping University, SE
- Michael Kerber
TU Graz, AT
- Claudia Landi
University of Modena, IT
- Michael Lesnick
University at Albany, US
- Joshua A. Levine
University of Arizona – Tucson, US
- Facundo Memoli
Ohio State University – Columbus, US
- Dmitriy Morozov
Lawrence Berkeley National Laboratory, US
- Vijay Natarajan
Indian Institute of Science – Bangalore, IN
- Andreas Ott
KIT – Karlsruher Institut für Technologie, DE
- Manish Saggar
Stanford University, US
- Dominik Schindler
Imperial College London, GB
- Primoz Skraba
Queen Mary University of London, GB & Jožef Stefan Institute – Ljubljana, SI
- Nico Stucki
TU München, DE
- Yusu Wang
University of California, San Diego – La Jolla, US
- Bei Wang Phillips
University of Utah – Salt Lake City, US
- Gunther Weber
Lawrence Berkeley National Laboratory, US
- Erin Moriarty Wolf Chambers
St. Louis University, US
- Kelin Xia
Nanyang TU – Singapore, SG
- Lori Ziegelmeier
Macalester College – St. Paul, US