Statistical Topological Data Analysis (TDA)

Lecture 1

Wolfgang Polonik

Department of Statistics, UC Davis

IHP, Sept. 15, 2022 - Dec. 1, 2022

Acknowledgements

Students: Eunseong Bae, Rui Hu, Irene Kim, Benjamin Roycraft (UC Davis), Olympio Hacquard, Vincent Divol (ENS, Paris)

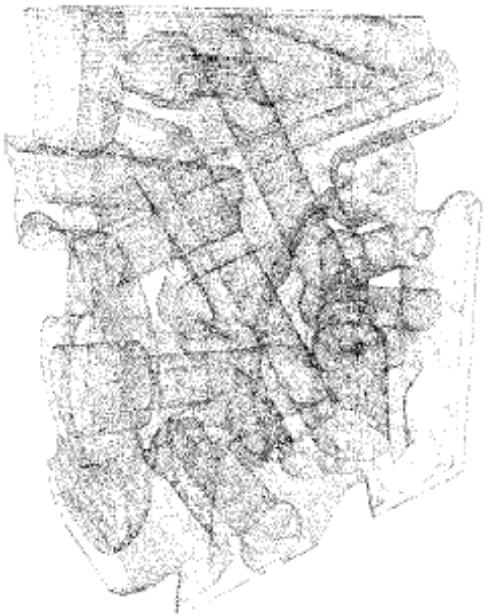
PostDoc: Johannes Krebs (U. Heidelberg)

Colleagues: Frederic Chazal (INRIA, Paris), Bertrand Michel (Nantes), Dietmar Kültz, Javier Arsuaga, Krishna Balasubramanian (UC Davis),...

Other contributors to the field (see below), Wikipedia, Google, ...

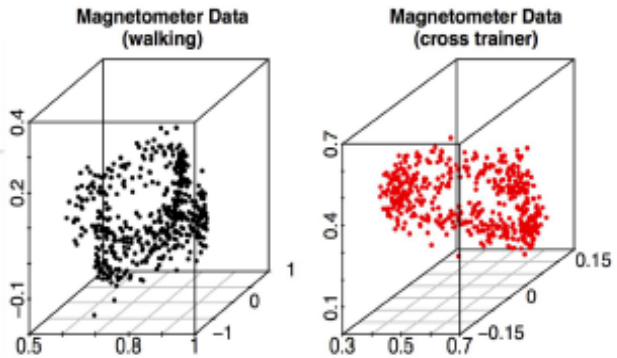
⋮

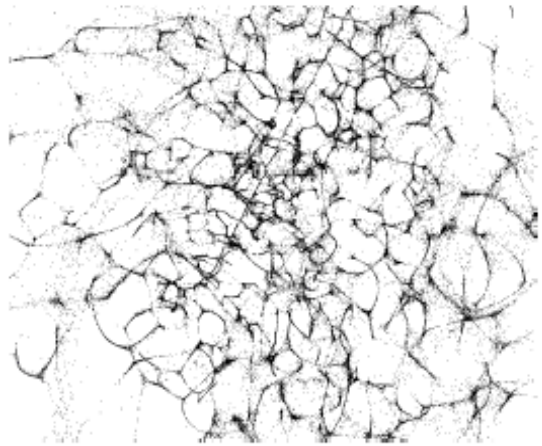
What is statistical topological data analysis (TDA)



[Scanned 3D object]

[Shape database]





[Galaxies data]

Graphic courtesy of F. Chazal

What is statistical topological data analysis (TDA)

Tools and ideas arise from various areas, such as

- Applied/computational topology
- probability theory
- computational geometry
- statistics

What is statistical topological data analysis (TDA)

GOAL: Extract qualitative, topological *and geometric* information about underlying distribution point clouds in a metric space

- topological information:
  number of k-dimensional holes (Betti numbers);
- 0-dim holes: connected components $\rightsquigarrow$ clustering
- major tool: persistent homology (includes clustering, but goes far beyond)

- other topological information:
  Morse complex of a density $\rightsquigarrow$ clustering

- related, more geometrically motivated objects of interest: critical points, ridges, level set clusters, intrinsic dimension, reach,...

What is statistical topological data analysis (TDA)

A word of caution:

Topology deals with invariants; topological structure is independent of a coordinate system.

This is not necessarily the case for the methods in topological data analysis! They usually depend on the choice of the metric used to measure distances.

What is statistical topological data analysis (TDA

Covers a broad range:

- abstract ideas from various fields
- computational ideas
- heuristic geometric ideas $\rightsquigarrow$ very appealing for applications(!)

Remarks on early history of persistence homology

- TDA started to exist as a field in early 2000 (with some precursors)
- origins in computational topology
- early milestones:
  - Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002): Topological persistence and simplification. *Discrete Comput. Geom.*
  - Carlsson, G., Collins, A., Guibas, L. and Zomorodian, A. (2004): Persistence barcodes for shapes. In *Proc. 2nd Sympos. Geometry Process.*

Remarks on early history of persistence homology

- TDA started to exist as a field in early 2000 (with some precursors)
- origins in computational topology
- early milestones:
  - Edelsbrunner, H., Letscher, D., and Zomorodian, A. (2002): Topological persistence and simplification. *Discrete Comput. Geom.*
  - Carlsson, G., Collins, A., Guibas, L. and Zomorodian, A. (2004): Persistence barcodes for shapes. In *Proc. 2nd Sympos. Geometry Process.*
- at first, no randomness involved; just "point cloud data" (discrete)
- randomness (and some statistics) started to enter with
  - Bubenik, P., and Kim, P.T. (2007): A statistical approach to persistent homology. *Homology, Homotopy and Applications*.
  - Niyogi, P., Smale, S., and Weinberger, S. (2008): Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*

Topological data analysis today

Some major contributors:

DataShape group (INRIA, France)

Boissonnat, J.-D.
Chazal, F. (PI)
Glisse, M.
Michel, B. (Nante, France)
Oudot, S.Y.

Carnegie Mellon group lead by:

Wassermann, L.
Genovese, C.
Rinaldo, A.

Group at IST (Austria)
lead by Edelsbrunner, H.

Other major contributors:
Carlsson, G. (Stanford, Ayasdi, Unbox AI),
Bubenik, P. (U. Florida)
Adler, R. (Technion)
Bobrowski, O. (Technion)
Kahle, M. (U. Washington)
Bauer, U. (TUM)
Bendich, P. (Duke, GDA)
Mukherjee, S. (Duke)

⋮

Introductory texts/books

- Edelsbrunner, H. and Harer, J. (2010): Computational topology: An introduction. *AMS*

- Wasserman, L. (2018): Topological data analysis. *Annual Reviews of Statistics and Its Applications*, **5**, 501-532. (arXiv:1609.08227)

- Boissonnat, J.-D., Chazal, F. and Yvinec, M. (2018): Geometric and Topological Inference. *Cambridge Texts In Appl. Math.*

- Rabadán, R. and Blumberg, A.J. (2019): Topological Data Analysis for Genomics and Evolution *Cambridge University Press*

- Chazal, F. and Michel, B. (2021): An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, **4** 4:667963.

- Virk, Z. (2022): Introduction to Persistent Homology.

- Dey, T.K. and Wang, Y. (2022): Computational topology for data analysis. *Cambridge University Press*

Goals of these lectures

- understand persistence diagrams (PDs) and barcodes and how they depend on filtrations
  - emphasis on challenges with interpreting and analyzing PDs
- study of some existing related statistical inference methods
- along the way: introduce things like minimax theory, aspects of empirical processes (covering numbers, metric entropy); bootstrap, construction of confidence regions,...

Question: What are barcodes and persistence diagrams?

- Persistence diagram: Edelsbrunner, Letscher, Zomorodian (2002)
- Barcodes: Carlsson, Collins, Guibas and Zomorodian (2004)

Barcodes

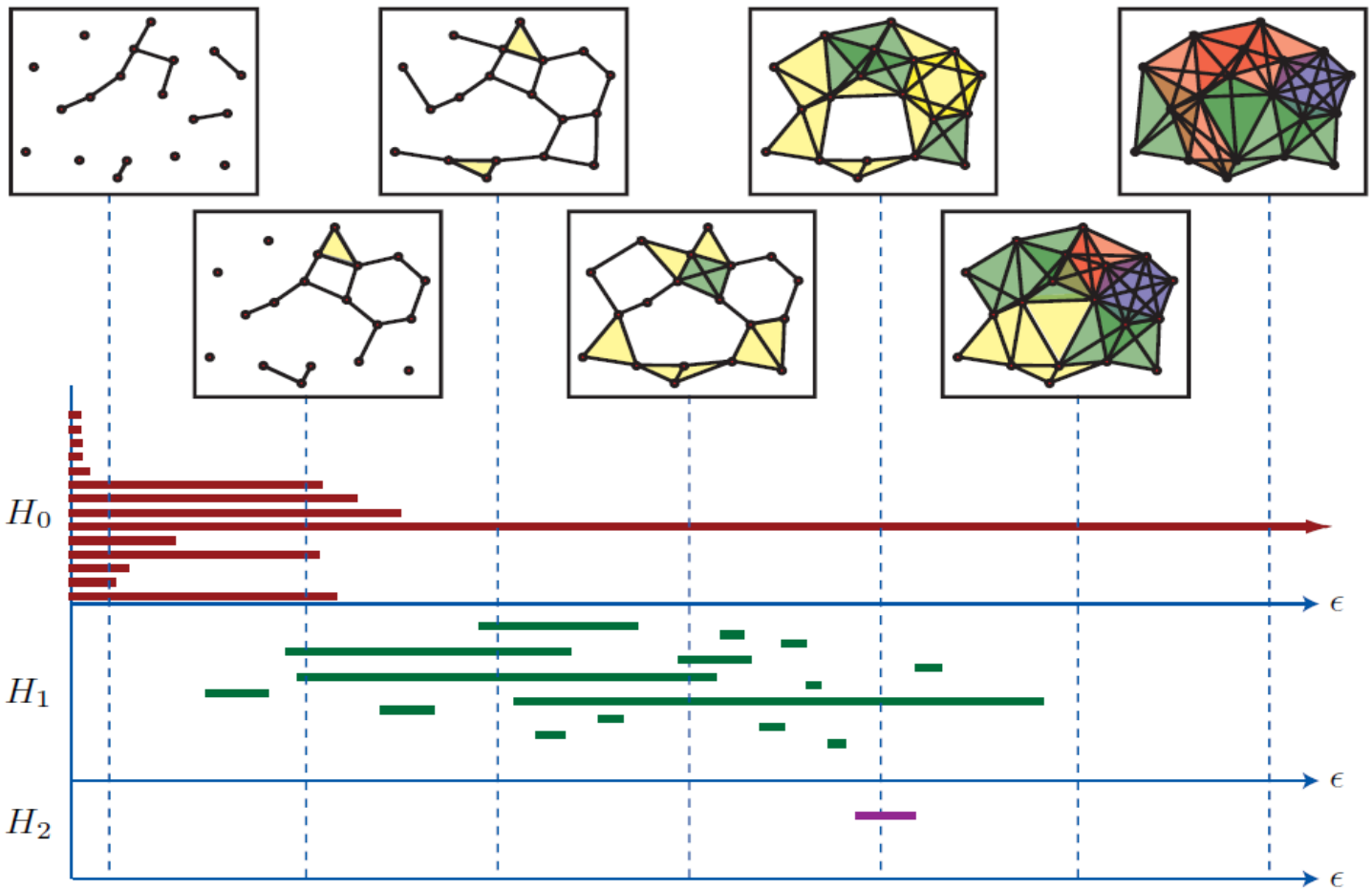
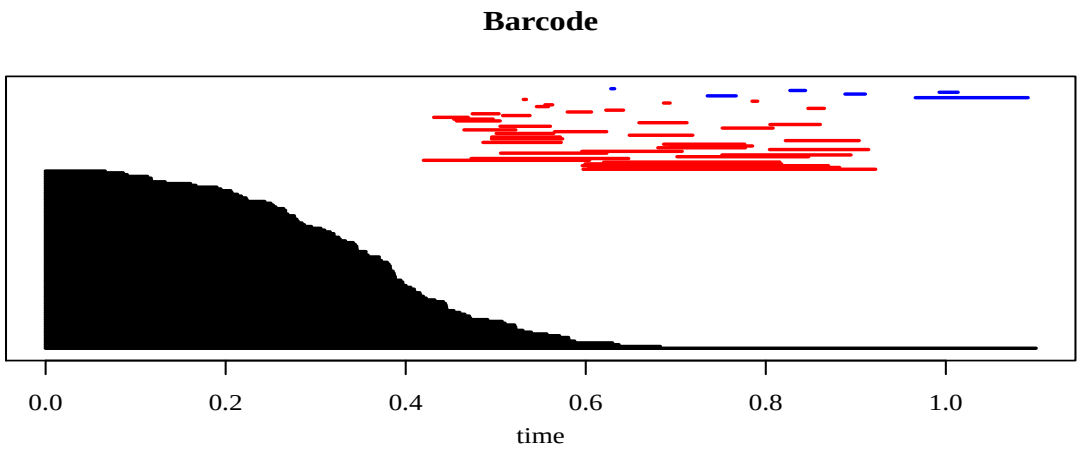
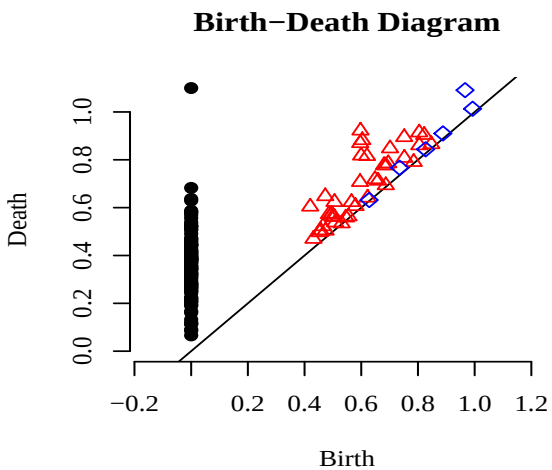
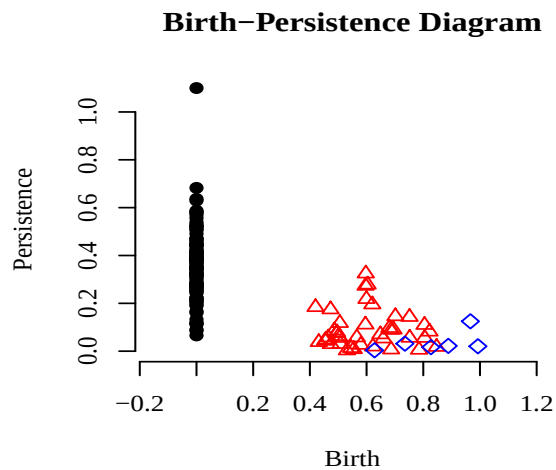


image from Ghrist (2008)

Barcodes and persistence diagrams







Barcodes and persistence diagrams

- Barcodes consist of intervals
- Intervals can be identified with a tuple of real numbers (endpoints)
- Plot all these tuples into the plane $\rightsquigarrow$ PD

Barcodes and persistence diagrams

- Barcodes consist of intervals
- Intervals can be identified with a tuple of real numbers (endpoints)
- Plot all these tuples into the plane $\rightsquigarrow$ PD

As graphical tools: Barcode and PD contain the same basic information; two dual graphical representations

From a mathematical point of view: Persistence diagram is a point cloud (a multi-set) $\rightsquigarrow$ (perhaps) easier to analyze

Example of an application of persistence diagrams

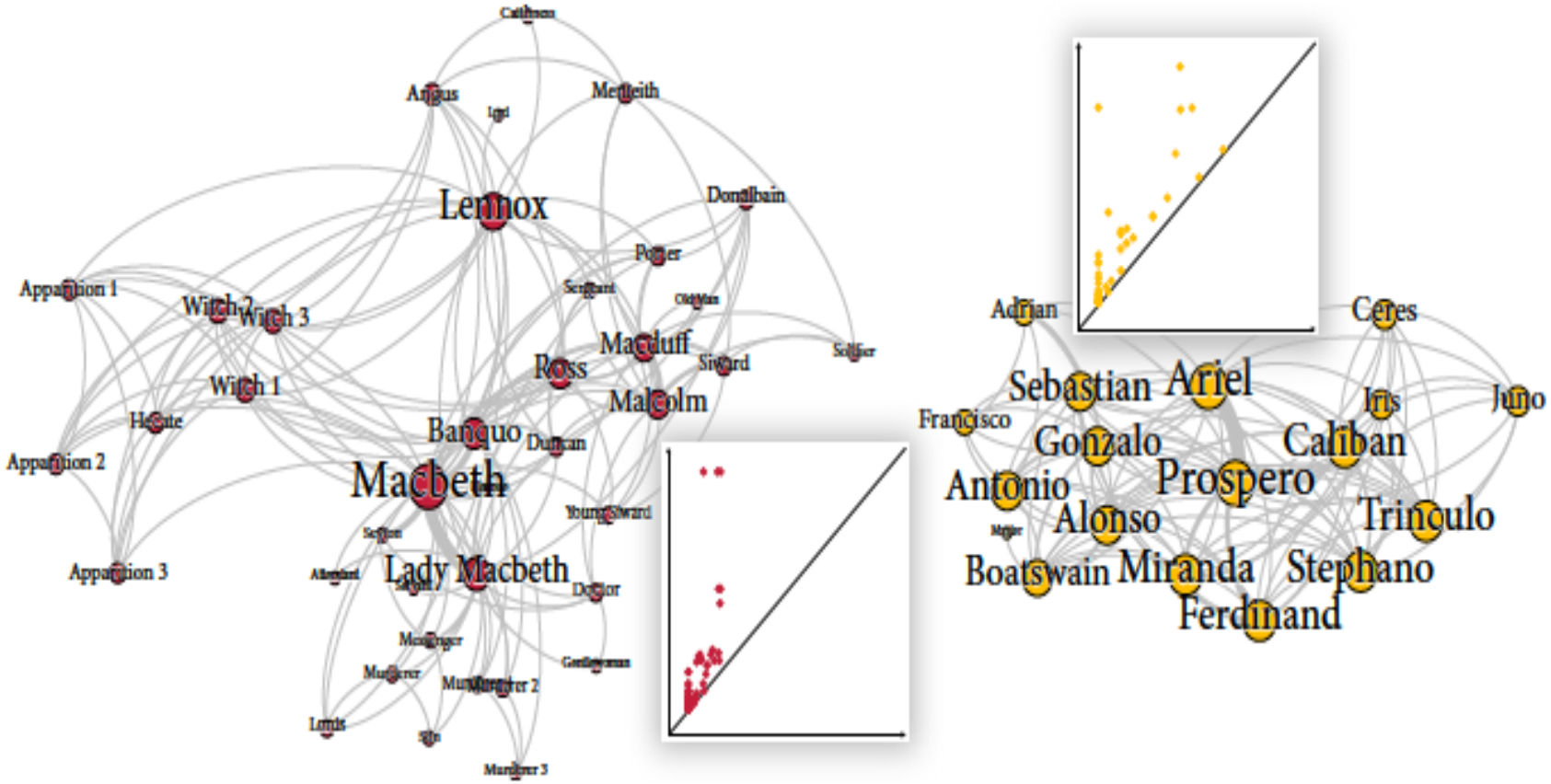


image from Rieck and Leitte (2015)