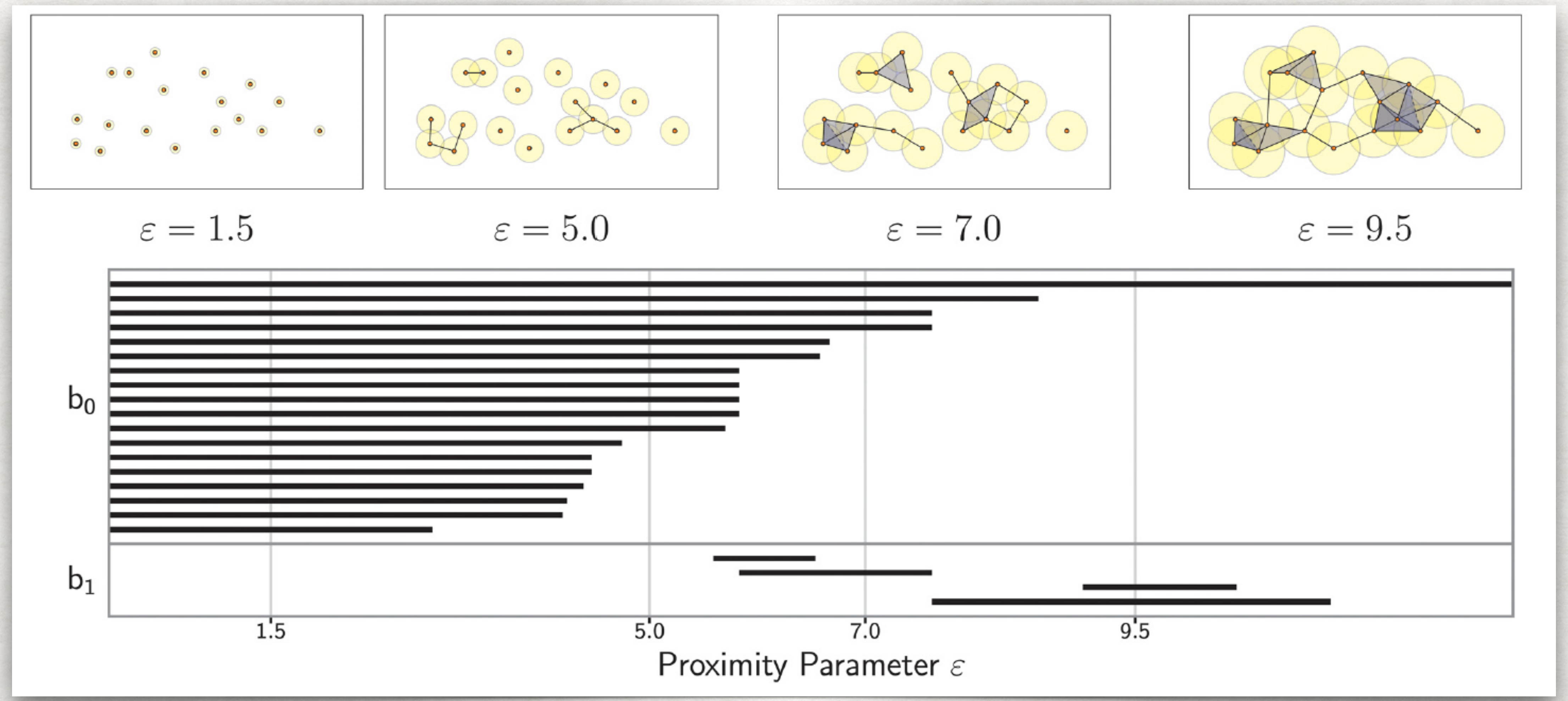
TOPOLOGICAL DATA ANALYSIS

BARCODES

- Ghrist , Barcodes: The persistent topology of data
- Topaz, Ziegelmeier, and Halverson 2015: Topological Data Analysis of Biological Aggregation Models



Questions in data analysis:

- How to infer high dimensional structure from low dimensional representations?
- How to assemble discrete points into global structures?

FIGURE 1. Determining the global structure of a noisy point cloud is not difficult when the points are in $\mathbb{E}^2$, but for clouds in higher dimensions, a planar projection is not always easy to decipher.

Themes in topological data analysis[1] :

1. Replace a set of data points with a family of **simplicial complexes**, indexed by a proximity parameter.
2. View these topological complexes using the novel theory of **persistent homology**.
3. Encode the persistent homology of a data set as a parameterized version of a Betti number (a **barcode**).

[1]Work of Carlsson, de Silva, Edelsbrunner, Harer, Zomorodian

CLOUDS OF DATA

FIGURE 1. Determining the global structure of a noisy point cloud is not difficult when the points are in $\mathbb{E}^2$, but for clouds in higher dimensions, a planar projection is not always easy to decipher.

Point cloud data coming from physical objects in 3-d

CLOUDS TO COMPLEXES

Cech complex

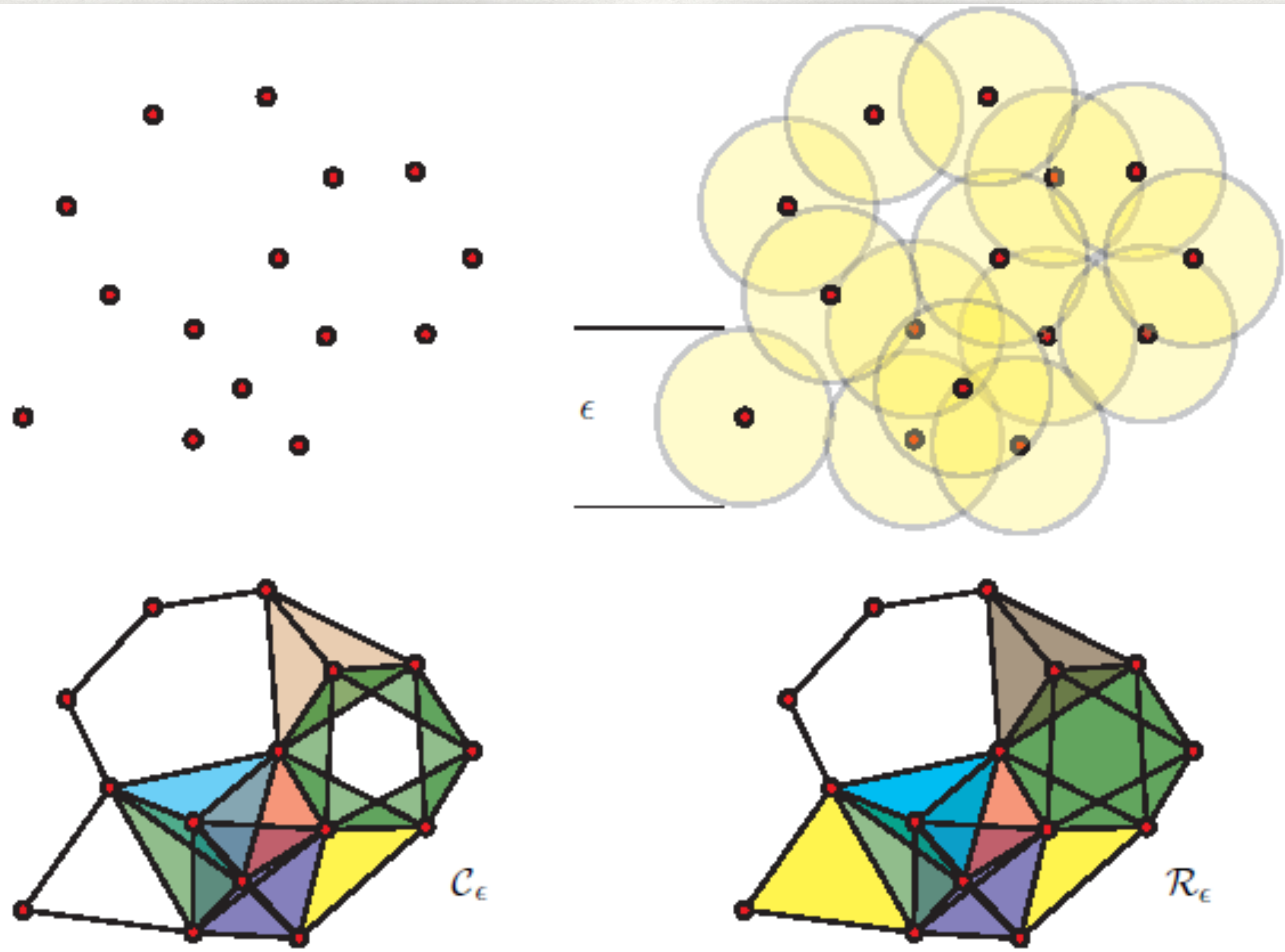


FIGURE 2. A fixed set of points [upper left] can be completed to a a Čech complex $\mathcal{C}_\epsilon$ [lower left] or to a Rips complex $\mathcal{R}_\epsilon$ [lower right] based on a proximity parameter ϵ [upper right]. This Čech complex has the homotopy type of the $\epsilon/2$ cover $(S^1 \vee S^1 \vee S^1)$, while the Rips complex has a wholly different homotopy type $(S^1 \vee S^2)$.

Rips complex

CHOICE OF PARAMETER ϵ?

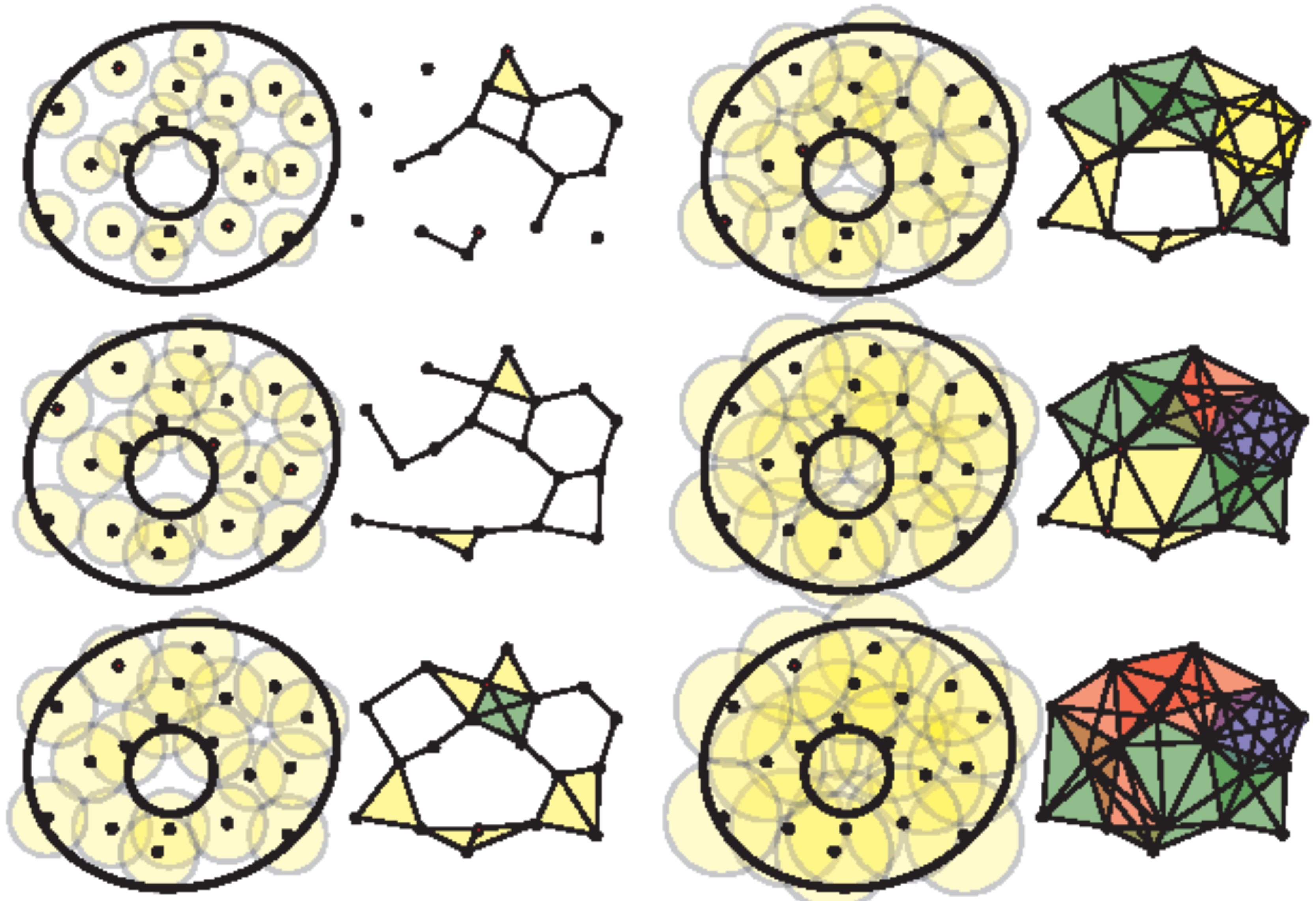


FIGURE 3. A sequence of Rips complexes for a point cloud data set representing an annulus. Upon increasing ϵ, holes appear and disappear. Which holes are real and which are noise?

USE PERSISTENT HOMOLOGY

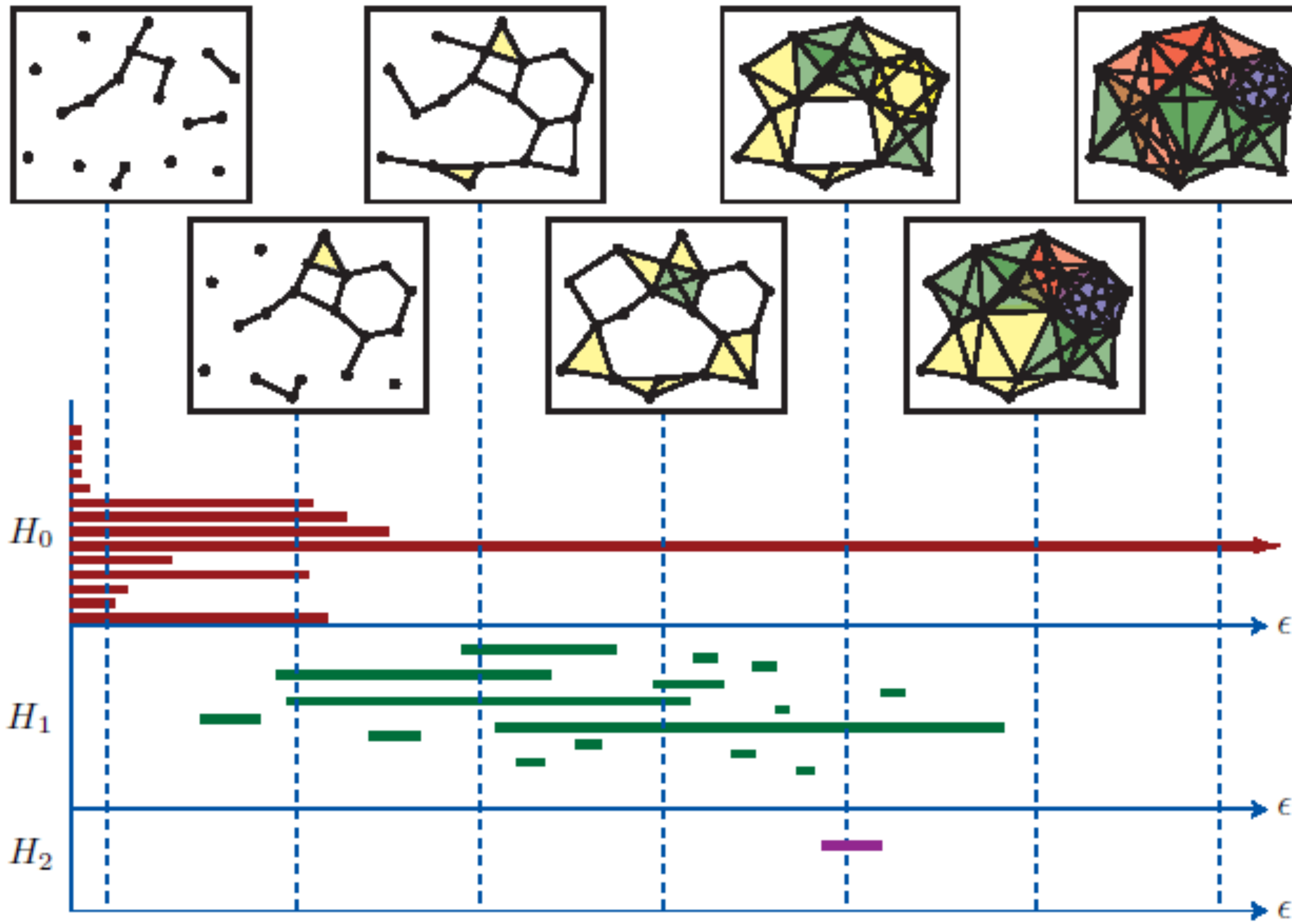


FIGURE 4. [bottom] An example of the barcodes for $H_*(\mathsf{R})$ in the example of Figure 3. [top] The rank of $H_k(\mathcal{R}_{\epsilon_i})$ equals the number of intervals in the barcode for $H_k(\mathsf{R})$ intersecting the (dashed) line $\epsilon = \epsilon_i$.

APPLICATION (TOPAZ ET AL.): BIOLOGICAL AGGREGATION MODELS

- Use two math models of biological aggregation (bird flocks, fish schools, etc.): Vicsek et al., D'Orsogna et al.
- Generate point clouds in position-velocity space at different times.
- Analyze the topological structure of these point clouds by calculating the first few Betti numbers.
- Visualize of Betti numbers using contours plots.
- Compare results to other measures/parameters used to quantify global behavior of aggregations.

TDA AND PERSISTENT HOMOLOGY

FORMING A SIMPLICIAL COMPLEX

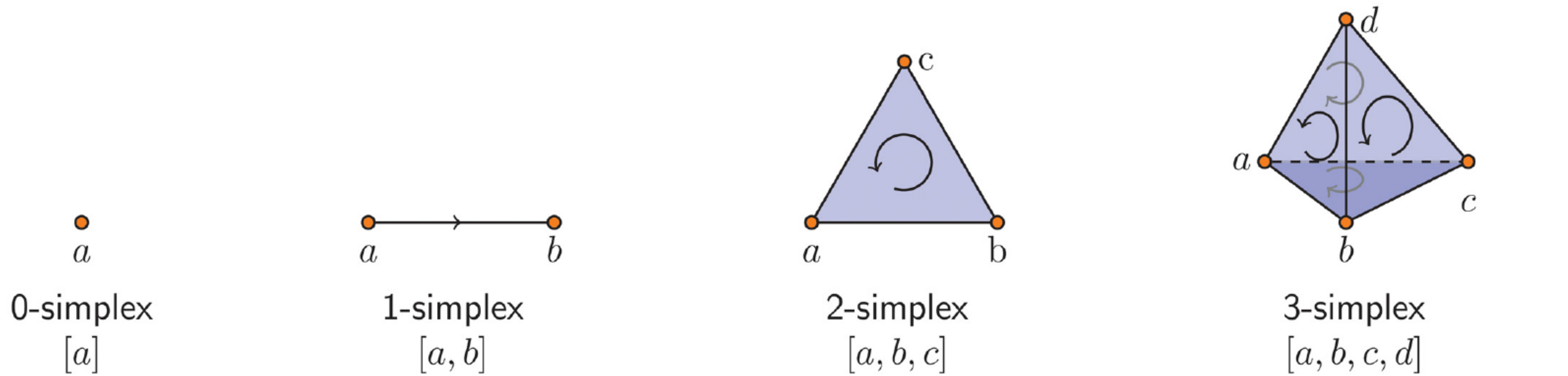


Fig 1. Oriented *k*-simplices for *k* = 0, 1, 2, 3. These *k*-simplices are the building blocks used to construct a simplicial complex from a point cloud of data.

k-simplices

- 1-simplex (edge): 2 points are within ϵ of each other.
- 2-simplex (triangle): 3 points are pairwise within ϵ from each other

EXAMPLE OF A VIETORIS-RIPS COMPLEX

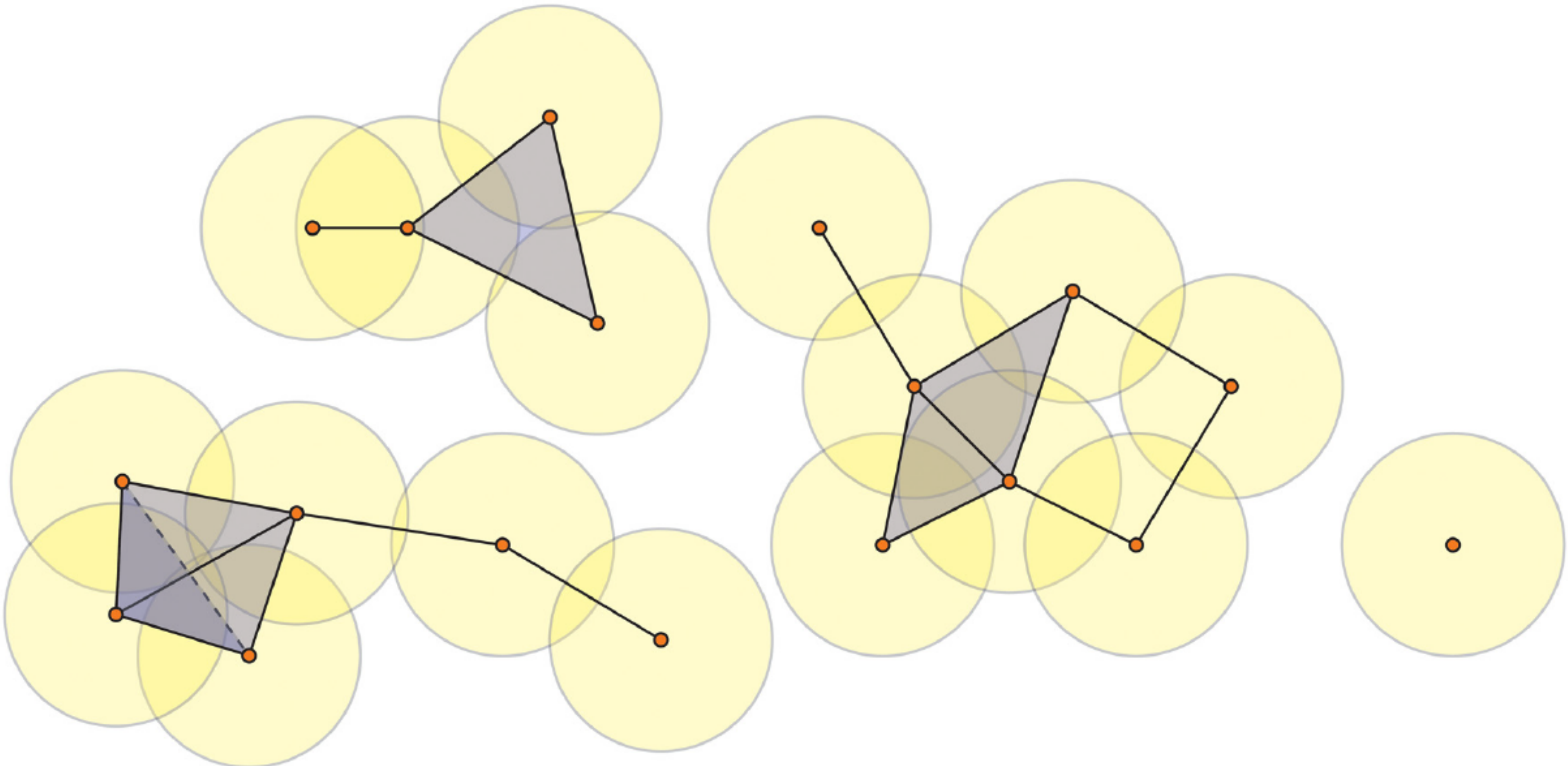


Fig 2. Example of a Vietoris-Rips complex. The 18 points are 0-simplices. Two 0-simplices form a 1-simplex (an edge) if their $\varepsilon/2$-neighborhoods (yellow circles) intersect. Three vertices form a 2-simplex (a triangle) if they are pairwise connected by edges. Four vertices form a 3-simplex (a tetrahedron) if they are pairwise connected by edges.

HOMOLOGY

BOUNDARIES

For $k \geq 0$, create an abstract vector space C_k with basis consisting of the set of k-simplices in S_ϵ. The elements of C_k are called k-chains.

- k-chains: linear combinations of simplices.
- boundary of k-simplex: union of (k-1)-subsimplices.

$$\partial_k([v_0, v_1, \ldots, v_k]) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

For example,

$$\partial_1([v_0, v_1]) = [v_1] - [v_0] \qquad \text{and} \qquad \partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_0, v_1].$$

SUBSPACES

$$k\text{-cycles}: \quad \mathsf{Z}_k := \ker(\partial_k : \mathsf{C}_k \to \mathsf{C}_{k-1}),$$

$$k\text{-boundaries}: \quad \mathsf{B}_k := \operatorname{im}(\partial_{k+1} : \mathsf{C}_{k+1} \to \mathsf{C}_k).$$

Note: a k-cycle is a k-chain with boundary 0.

HOMOLOGOUS CLASSES

Two cycles are homologous (equivalent) if they differ by a boundary.

HOMOLOGOUS CLASSES

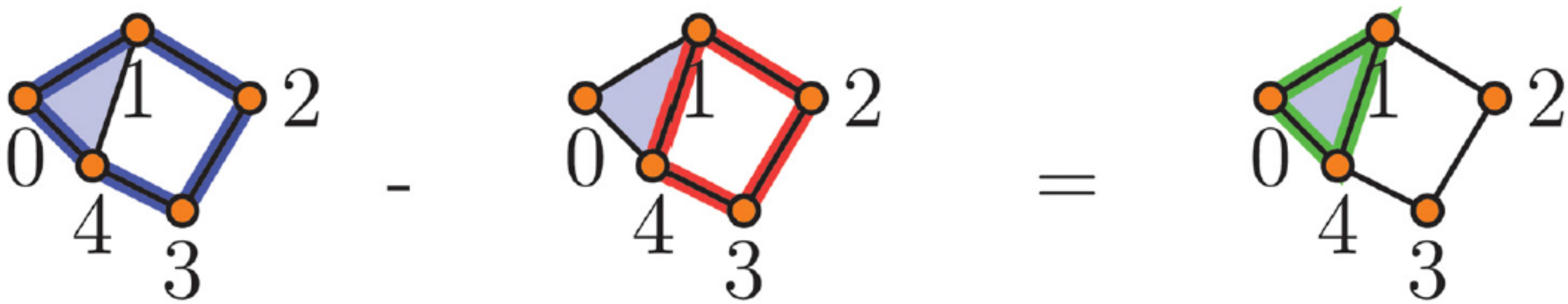


Fig 3. Example of homologous cycles. The blue 1-cycle and the red 1-cycle are homologous (equivalent), because their difference is the boundary of a triangle, shown in green; see text for a detailed explanation.

Define kth homology as the set of homology classes:

$$\mathsf{H}_k := \{\ [z]\ \big|\ z\ \in \mathsf{Z}_k\ \}.$$

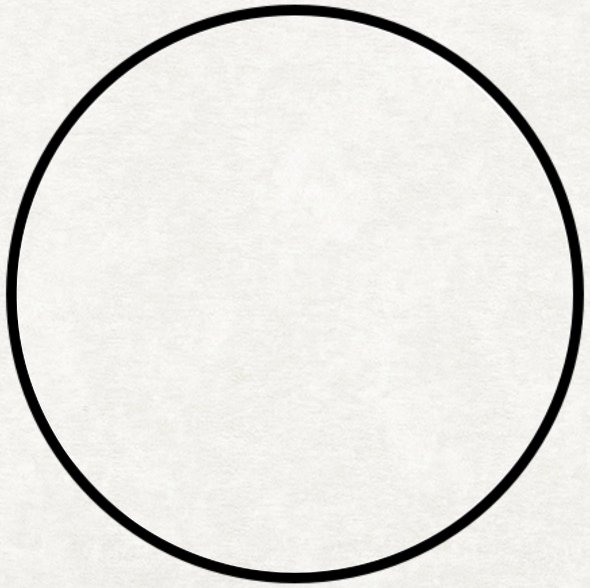
The kth Betti number is defined as the dimension:

$$\mathsf{b}_k = \dim(\mathsf{H}_k) = \dim(\mathsf{Z}_k) - \dim(\mathsf{B}_k).$$

BETTI NUMBERS

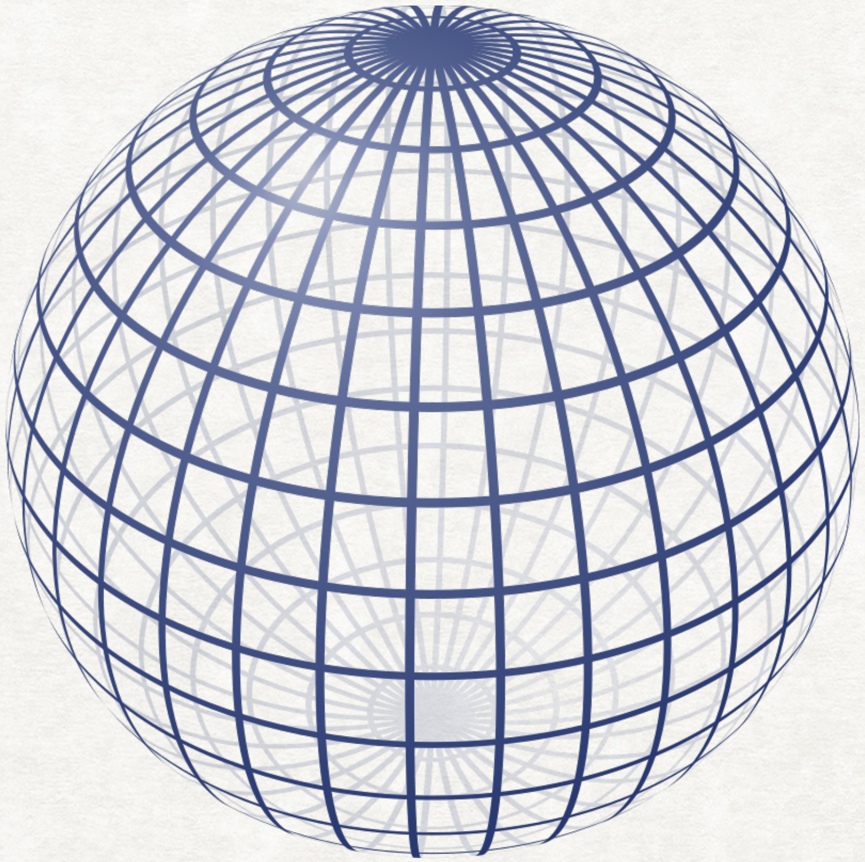
Betti numbers are topological invariants measuring the number of k-dimensional holes in an object:

- b_0: number of connected components.
- b_1: number of topological circles.
- b_2: number of trapped volumes, etc.



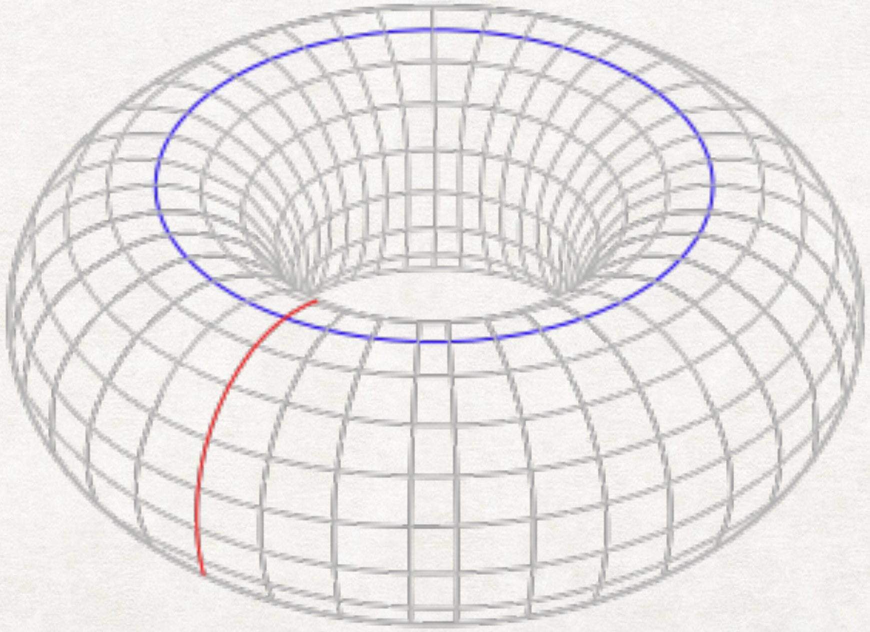
$$b = (b_0, b_1, b_2, ...) = (1, 1, 0, ...)$$

MORE EXAMPLES



$$b = (b_0, b_1, b_2, ...) = (1, 0, 1, 0, ...)$$

$$b = (b_0, b_1, b_2, ...) = (1, 2, 1, 0, ...)$$



For simplicial complexes, Betti numbers b_k are the number of homologically distinct k-cycles after discarding k-boundaries.

Still have the dependence on ϵ!

To resolve this, note that for $\epsilon_1 \leq \epsilon_2 \leq ... \leq \epsilon_M$

$$S_{\varepsilon_1} \subseteq S_{\varepsilon_2} \subseteq \cdots \subseteq S_{\varepsilon_{M-1}} \subseteq S_{\varepsilon_M}.$$

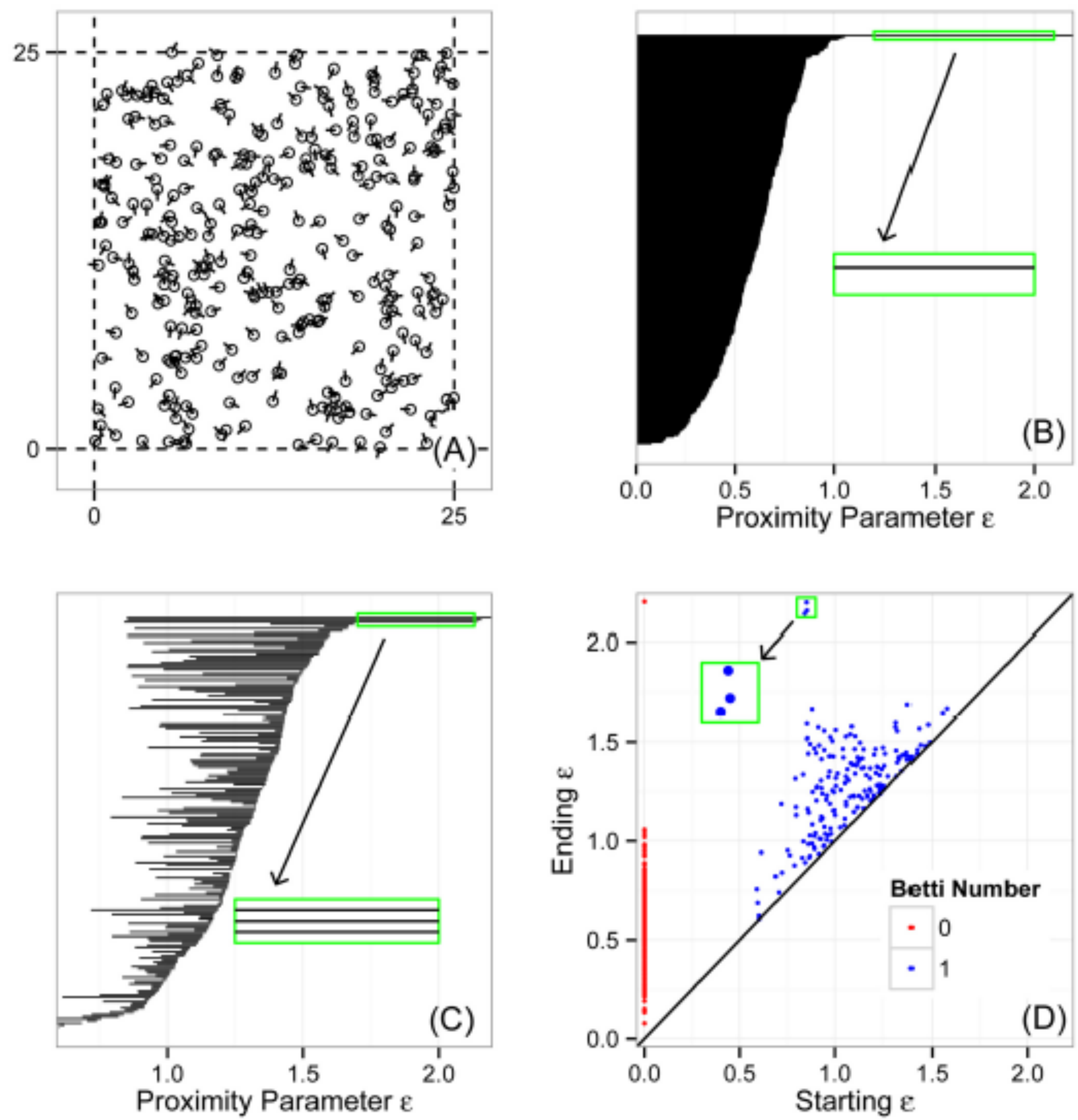
Persistent homology tracks topological features which persist across a range of ϵ.

How to visualize? See barcodes in Chad's slides.

VISUALIZATION OF BETTI NUMBERS

Vicsek model

b_0



b_1

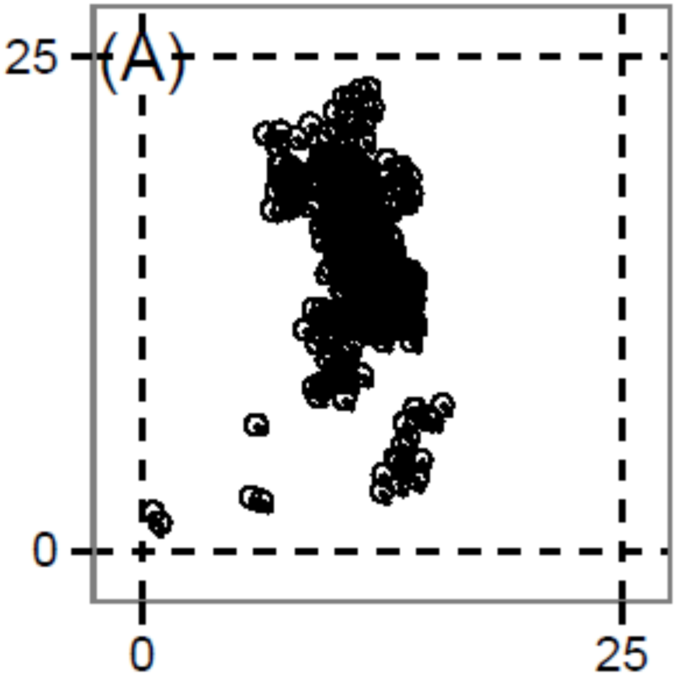
Persistence plot

Fig 5. A random initial condition used to simulate the Vicsek model (9) and topological analysis of this initial state. (A) Random initial positions (x, y) and headings θ of $N = 300$ particles in a square of size $\ell = 25$ with periodic boundary conditions. The underlying space in which the data lives is a three-torus T^3 which has Betti numbers $b = (1, 3, 3, 1, 0, \ldots)$. (B) Barcode for Betti number $b_0(\varepsilon, 0)$, showing topological connected components. The zoomed box shows a single persistent bar, corresponding to the entire ensemble of particles. (C) Barcode for Betti number $b_1(\varepsilon, 0)$, showing topological circles. The zoomed box shows three persistent bars, representing the three circles comprising the three-torus. (D) Persistence plot, which displays the information in (B) and (C) by encoding each bar's starting and ending value of ε as a point in the Cartesian plane. Red points show b_0 and blue points show b_1. The zoomed box shows the three points representing the three persistent topological circles of the random initial condition in (A).

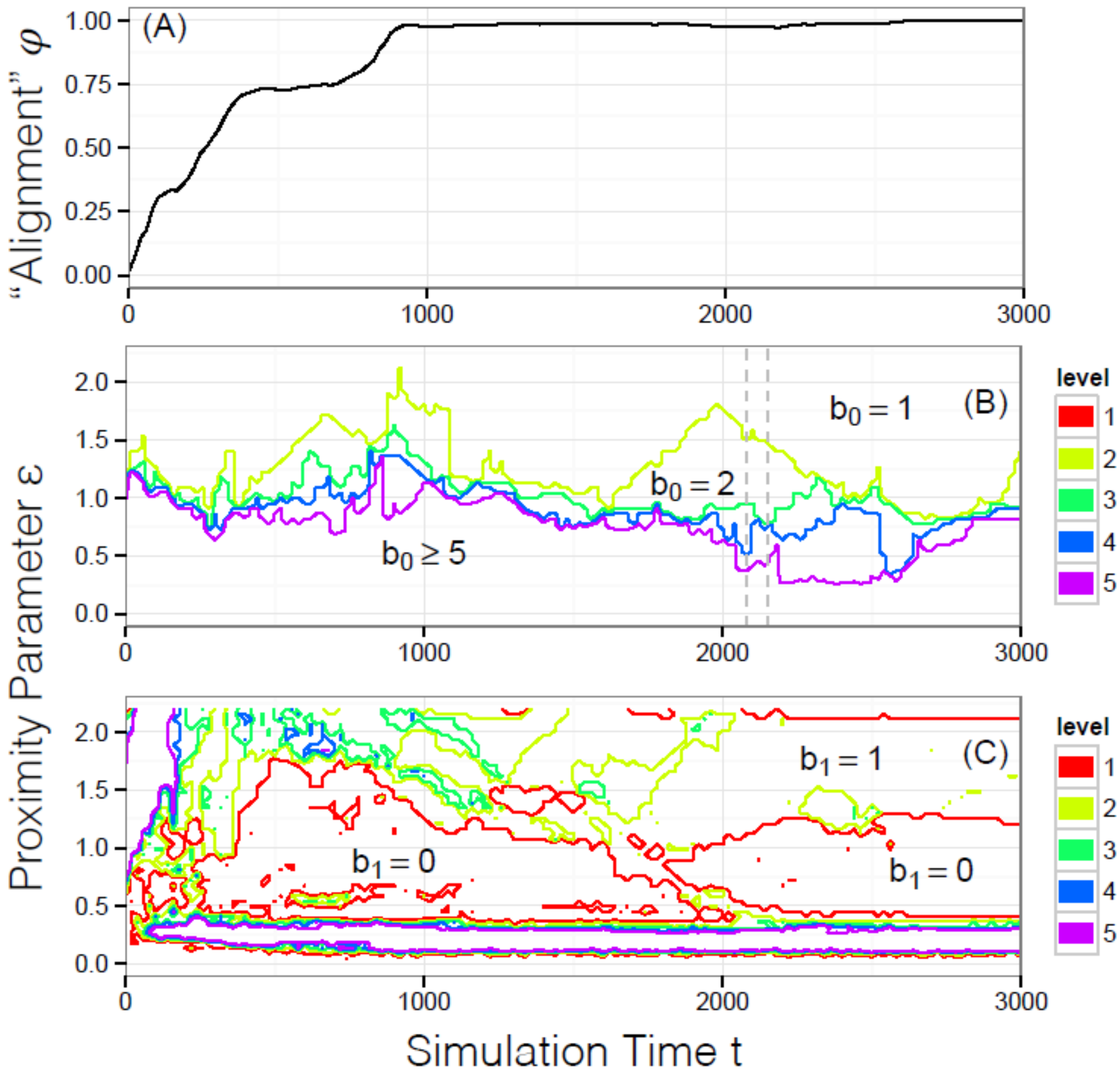
CROCKER

CONTOUR REALIZATION OF COMPUTED K-DIMENSIONAL HOLE EVOLUTION IN THE RIPS COMPLEX

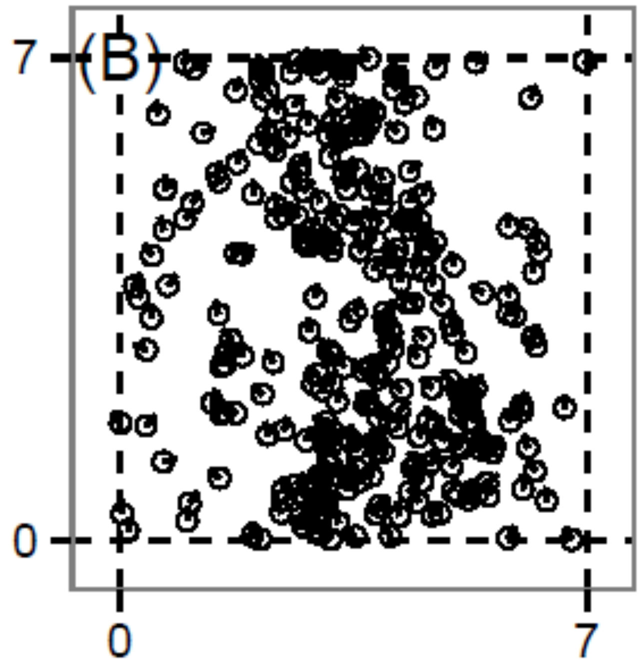
VICSEK, SIM #1



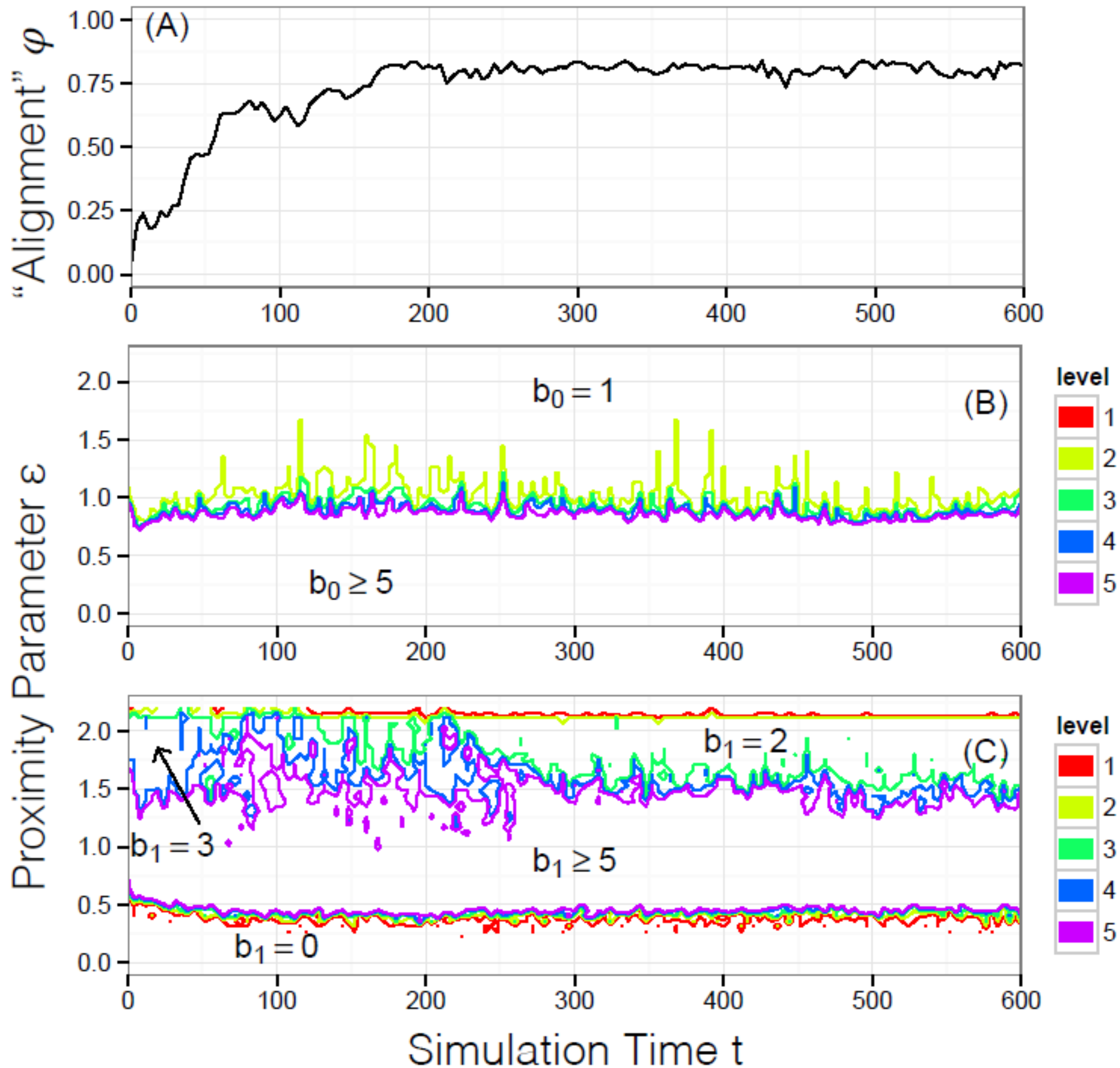
- Intermittent clustering
- Loss of two topol. circles
- b = (2 - 4, 1, …)



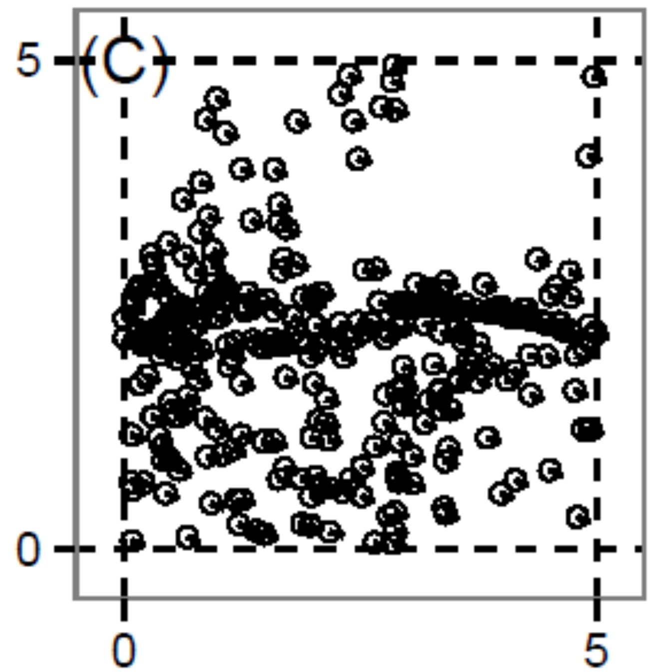
VICSEK, SIM #2



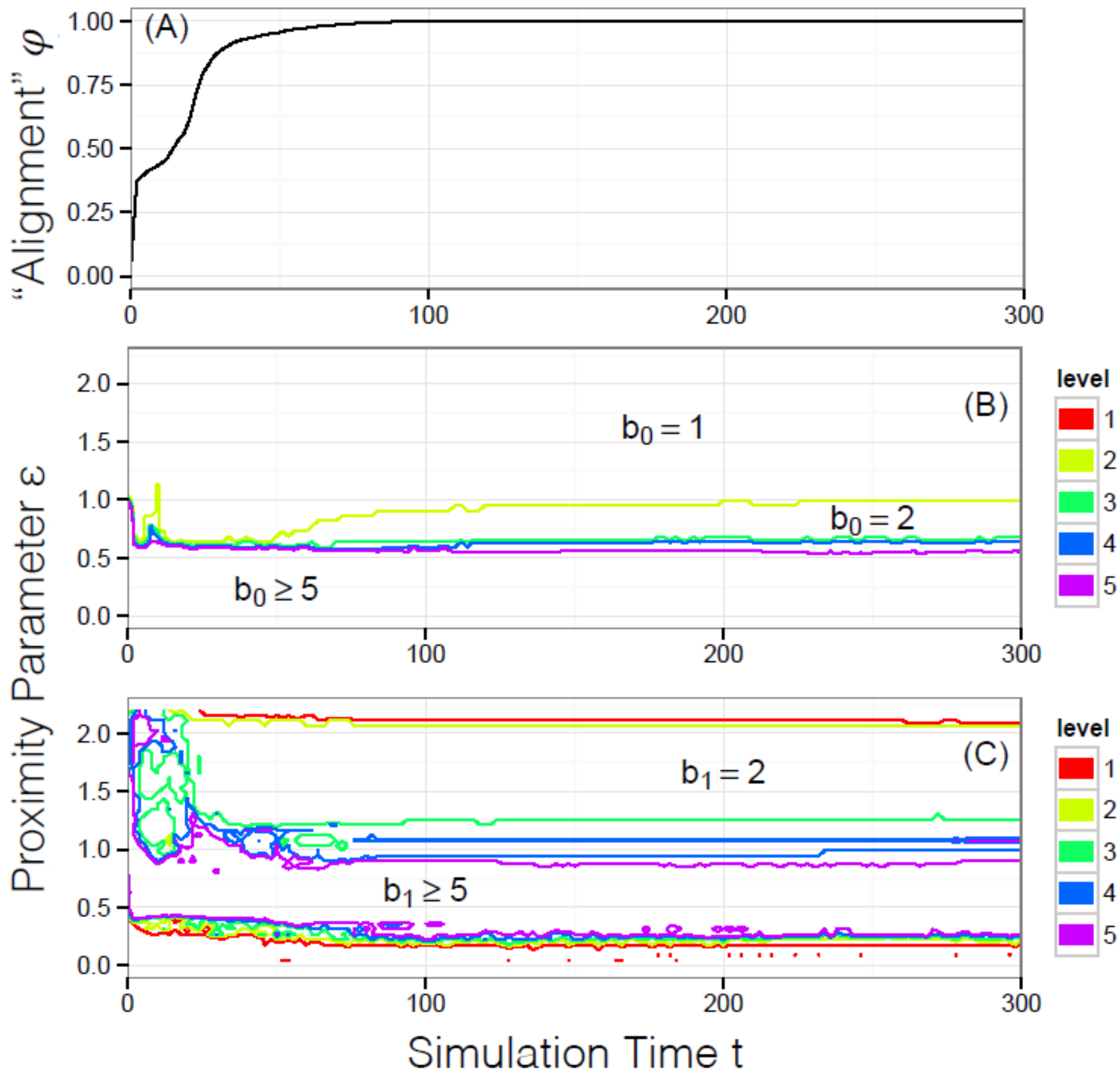
- One group
- Two persistent topol. circles
- b = (1, 2, 1, …)



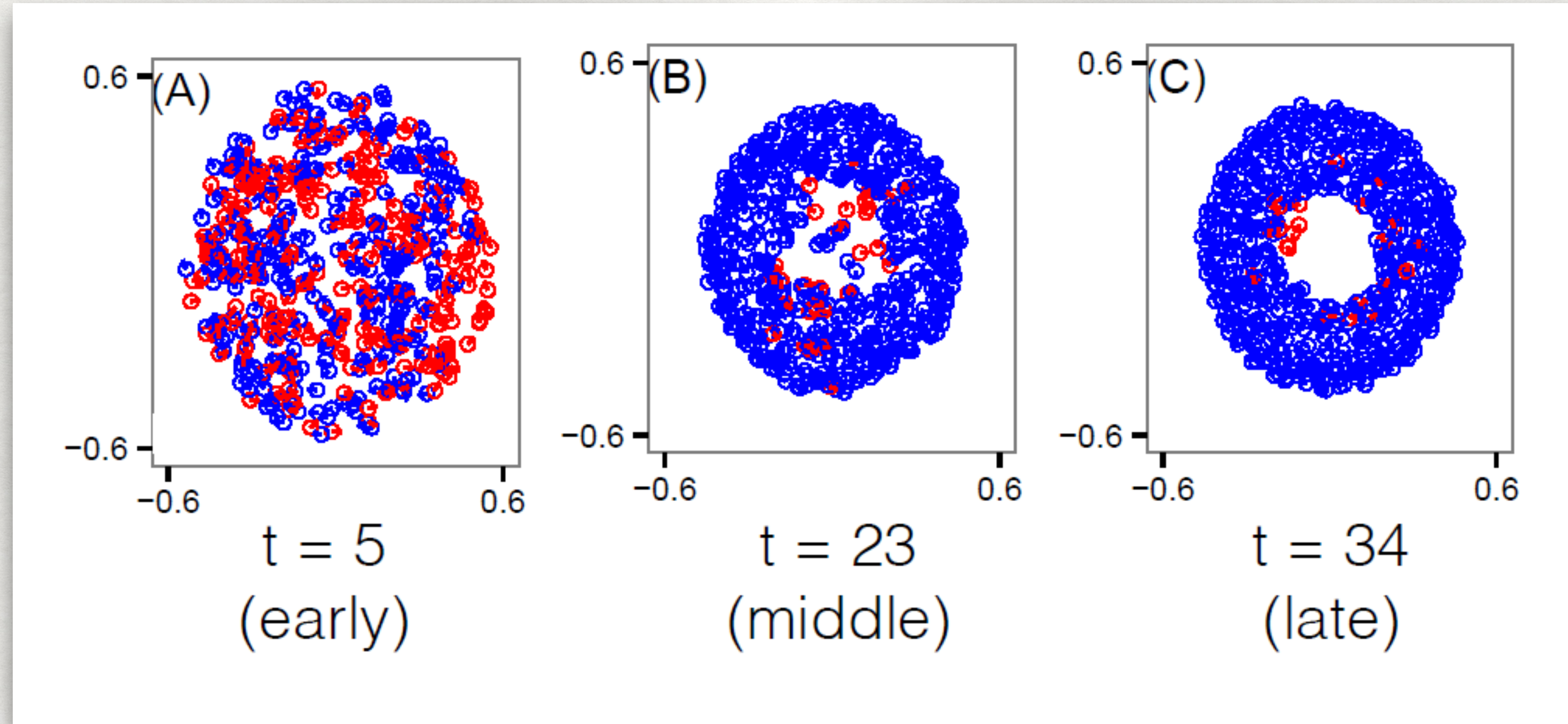
VICSEK, SIM #3



- One group, one rogue
- Two persistent topol. circles
- b = (1, 2, 0, …)
- Hole in the data



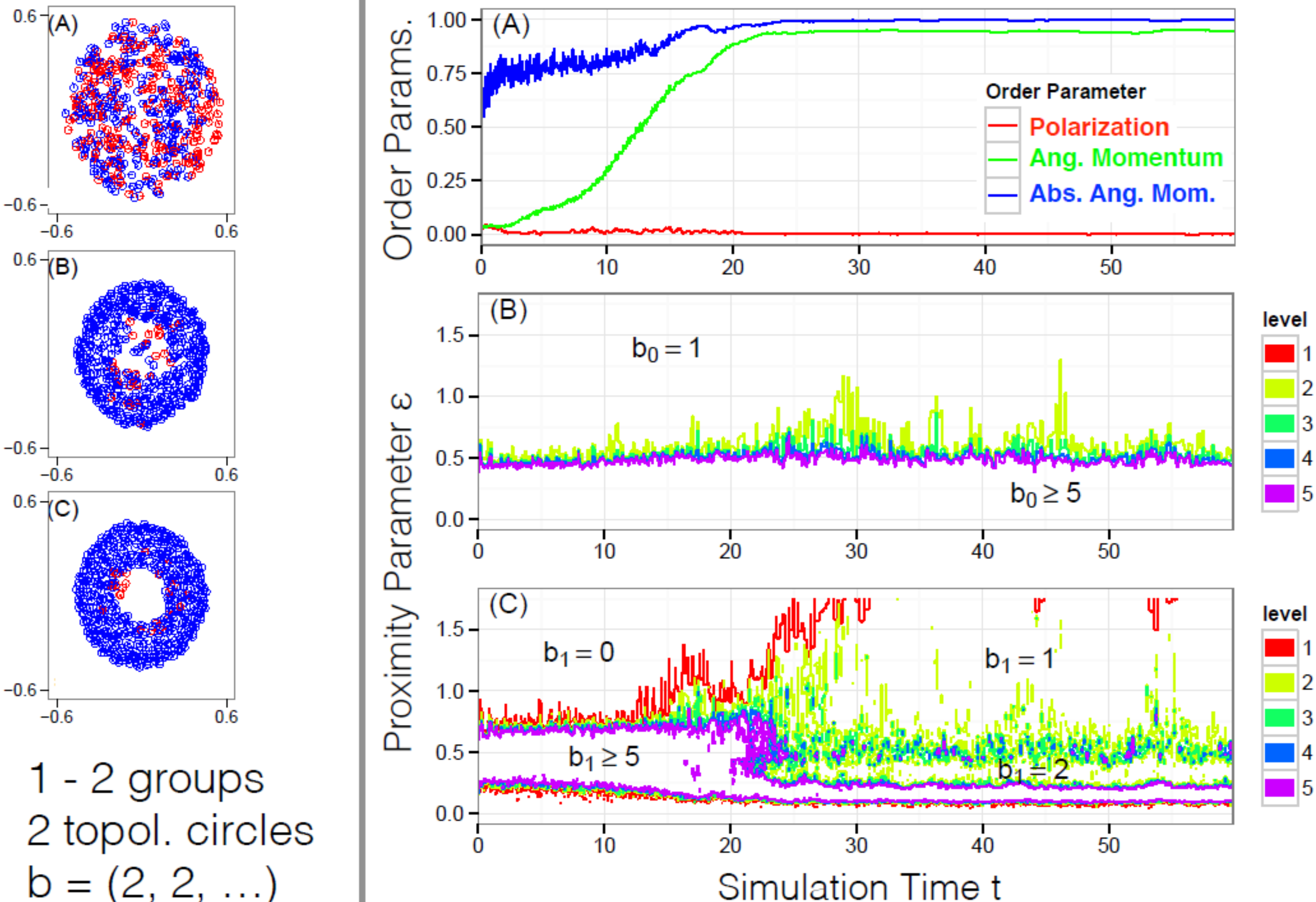
D'ORSOGNA MODEL



Attractive-repulsive interactions between particles

Produces many patterns including mills (rotating rings/swarms)

D'ORSOGNA



- 1 - 2 groups
- 2 topol. circles
- b = (2, 2, …)

CONCLUSIONS

- Introduced CROCKER visualization where features persisting over time and scale appear as large regions in a contour plot.
- Vicsek model: distinguished simulations where order parameter did not.
- D'Orsogna model: recognize the presence of double mill.

- **Limitations:** only first 2 Betti numbers with few exceptions.
- Over time, no math guarantee that the same components are consistent.