Robert Ghrist
Department of Mathematics and Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana-Champaign
USA
ghrist@math.uiuc.edu

Research

Three examples of applied and computational homology

Computational algebraic topology has already existed for some decades, with as its main objective the generation of examples. Nowadays, the field is rapidly changing into an applied branch of mathematics that is important in its own right. Robert Ghrist, topologist at the University of Illinois and one of the winners of the 2007 Scientific American 50 award, gives us three examples that illustrate this development, each with a different origin.

Mathematics is limitless in its dual capacity for abstraction and incarnation. To a large degree, many of the modern revolutions in technology and information rest on piers of mathematics that assist, inform, or otherwise catalyze progress. It appears that those branches of mathematics which are most easily understood and communicated are precisely those which find greatest applicability in the modern world. To conclude from this that deeper or more difficult fields are inherently less applicable would be premature.

Consider for example the utility of algebraic topology. Long cloistered behind formal and categorical walls, this branch of mathematics has been the source of little in the way of concrete applications, as compares with its more analytic or combinatorial cousins. In this author's opinion, this is not due to a fundamental lack of applicability so much as to

1. the lack of a motivating exposition of the tools for practitioners; and
2. an historical lack of emphasis on computational features of the theory.

These two issues are coupled. Advances which demonstrate the utility of a topological theory spur the need for good computation. Good algorithms for computing topological data spur the search for further applications.

Algebraic topology is the mathematics that arises in the attempt to describe the *global* features of a space via *local* data. That such tools have utility in applied problems concerning large data sets is not difficult to argue. To give a sense of what is possible, we sketch three recent examples of specific applications of homological tools. This list is neither inclusive nor ranked: these examples were chosen for concreteness, simplicity, and timeliness. This brief and woefully incomplete sketch is meant as an appetizer, for which the truncated bibliography serves as a menu for the second course.

On Homology.
Homology is a machine that converts local data about a space into global algebraic structure. In its simplest form, homology takes as its argument simple pieces of a topological space X and returns a sequence of *abelian groups* $H_k(X)$, $k \in \mathbf{N}$. Homology is a *functor*, which in practice means:

1. topologically equivalent spaces (homotopic) have algebraically equivalent (isomorphic) homology groups; and
2. topological maps between spaces

$$f : X \to Y$$

induce algebraic maps (homomorphisms) on homology groups

$$f_* : H_*(X) \to H_*(Y).$$

Numerous homology theories exist, fine-tuned for different classes of spaces (simplicial, cellular, singular, etc.).

Roughly speaking, homology groups count and collate *holes* in a space. The simplest example of a homological invariant is the number of connected components of a space — $\dim H_0$ — the type of 'holes' that a zero-dimensional instrument can measure. A less trivial example of a homological invariant is the *Euler characteristic*. The Euler characteristic χ of a triangulated surface is the alternating sum of the number of simplices — vertices minus edges plus faces — and that this quantity is a topological invariant of the surface. For more general (but tame) spaces, $\chi(X)$ can be expressed either as the alternating sum of the number of k-dimensional cells of X, or, as

$$\sum_{k=0}^{\infty}(-1)^k \dim H_k(X).$$

This quantity, being based on homology, is an invariant. It is a signal example of a homological device, being both computable and invariant. Our first example of applied algebraic topology relies on this invariant.

Looking Forward

The three examples here surveyed are all applications of homological tools to problems of large and often noisy data sets. However, there are numerous other examples of a different nature under the same aegis of applied algebraic topology. Many of these are *obstruction-theoretic* in nature — topological measures of complexity of coordinating robots, synchronizing a network, or performing distributed asynchronous computation.

The list of mathematical ideas which were once erroneously derided as useless abstractions (uniform convergence, matrix algebra, group theory, etc.) is sufficiently long and embarrassing so as to suggest patience in the case of applied algebraic topology. Given that the (hard) work of generating good algorithms for computing topological invariants for realistic systems is so recent [4], it can be successfully argued that the current spate of advances in applied algebraic topology is neither coincidental nor terminal.

How many people are in the building?

Problem: Target Enumeration. Consider a store whose ceiling tiles, walls, and carpet are embedding with people-counting sensors. How can these local sensors collaborate to determine the number of customers in the store?

Tool: Euler Characteristic Integration. One of the fundamental difficulties in large-scale sensor networks is *data aggregation*. A sufficiently dense collection of nodes will sample an environment redundantly. The goal of sensing is to compress this redundant local data into a global description of the environment. The operation of stitching local information over patches is the fundamental defining property of a *sheaf*, a means of assigning an algebraic object to open subsets of a space in such a manner that restrictions and overlaps are respected.

As an example, consider the problem of counting a collection of *targets*. Some fixed but unknown number N of targets lie in a domain D. The domain is filled with sensors, each of which can determine how many targets are nearby. It matters not how the sensors operate (*e.g.*, via infrared, acoustic, or optical sensing). Assume simple sensors which merely detect the number of nearby targets, with no information about target identity, distance, or bearing. In the continuum limit (where one has a sensor at each point in D), this yields a counting function $h : D \to \mathbf{N}$. The problem is to determine the number of targets, given only h.

The solution lies in an elegant integration theory which uses Euler characteristic as a *measure*. For compact sets A, B, the Euler characteristic satisfies $\chi(A \cup B) = \chi(A) + \chi(B) - \chi(A \cap B)$. Note the similarity of this to the definition of a measure. Indeed, χ is a type of scale-invariant topological volume, as was known going back to Hadwiger and Blaschke at least. It is straightforward to construct a measure $d\chi$ against which one can integrate certain functions. The type of piecewise-constant or *constructible* function $h : D \to \mathbf{N}$ that a sensor field returns is integrable in this theory.

Recent work of Baryshnikov et al. [1] gives a simple formula for computing the number of targets as $\int_D h \, d\chi$, in the setting where each target is detected by sensors on a topologically trivial (*e.g.*, convex) neighborhood. Because this is a topological integration theory, there are no geometric restrictions. Sensors can, *e.g.*, count the number of vehicles driving over a domain laced with vibration sensors, counting subcompacts and SUV's as equals.

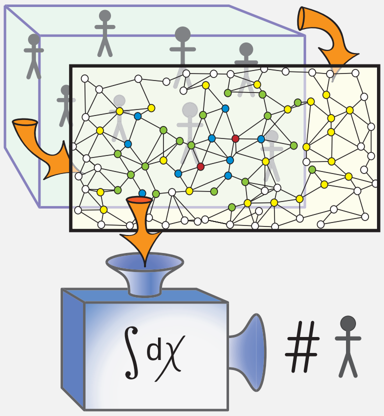

Figure 1 Integration with respect to Euler characteristic enumerates redundant data over a sensor network.

This is the starting point for a broad array of applications which rely on constructible sheaves and the sheaf-theoretic properties of $d\chi$. Precisely because the answer is expressed in terms of an integration theory, one can do the following:

- For a sparse network of sensor nodes, determining the number of targets becomes the numerical problem of approximating the topological integral via a discrete sampling.
- Thanks to a version of the *Fubini theorem* for $d\chi$, one can count moving targets over time without the need to embed clocks on the sensor nodes.
- Because integration is a local operation, target-counting can be performed by the network itself with a *distributed*, local computation.

Moral: "Data aggregation is a topological integration."

It looks chaotic to me!

Problem: Experimental Verification of Chaotic Dynamics. An experiment (physical or numerical) yields data that looks chaotic. Is it rigorously chaotic, or just noisy?

Tool: Conley Index Theory. One of the great scientific lessons of the 20^{th} century was that when a physical system exhibits erratic temporal behaviour, it may not be due to randomness or poor measurement — deterministic systems can exhibit well-defined *chaos*. However, it is a persistent challenge to demonstrate that a given system is chaotic. The *Lorenz equations* — themselves a cartoon model of fluid flow — were only recently shown to be rigorously chaotic, after more than thirty years' inquiry. Still more intractable remain data coming from physical experiments, in which system noise and instrument errors conspire to frustrate analysis. There seems to be little recourse for the experimentalist beyond saying: it looks chaotic to me.

A prime feature of topological methods is that, being global, they are typically impervious to the *noise* inherent in physical systems. Such is the case here. Work of Mischaikow et al. [5] uses a homological invariant of dynamics combined with *a priori* bounds on the noise amplitudes to determine the rigorous dynamics of an experimental system based on noisy time-series data.

The mathematical tool used is the *Conley index*, an algebraic-topological extension of the Morse index. Consider the flow of rainwater falling on a mountainous terrain D: this flow is that of $-\nabla h$, where $h : D \to \mathbf{R}$ is the height function of the terrain. The *Morse index* of a critical point of h is an integer that classifies the type of critical point: minima have index 0, saddle-passes have index 1, and maxima have index 2. The homological Conley index enriches the Morse index from integers to (homology types of) spaces: for a Morse function, the Conley index of a critical point is a *sphere* of dimension the Morse index.

The Conley index, unlike the Morse index, applied to non-gradient and non-smooth vector fields, as well as to discrete-time dynamics. It is efficacious, even to the point of detecting chaotic dynamics. This index is computable for realistic systems, thanks to recent progress in computational homology [4]. Work of Mischaikow et al. takes (noisy) *time-series* data and represents the dynamics as a *multi-valued map* on a cubical complex. By adapting the Conley index to this setting and computing the homological index, it is possible to verify the underlying dynamics, so long as the noise tolerances respect the discretization assumptions. Rigorous results about experimental or numerical data include the following:

- For experimentally-generated data on the dynamics of a *magneto-elastic ribbon* in an oscillating magnetic field, a Conley index approach proves that the experimental system is chaotic (has positive topological entropy) [5]. The method is robust, and works even when environmental noise alters the appearance of the data significantly.
- Numerical simulations of the *Kuramoto-Sivashinsky* partial differential equation indicate various stationary solutions. A Conley index computation [6] proves that these solutions exist, with a computational effort of the same order as a re-run of the numerical solution at a finer resolution.

Moral: "It's hardly more expensive to prove the dynamics than to simulate it."

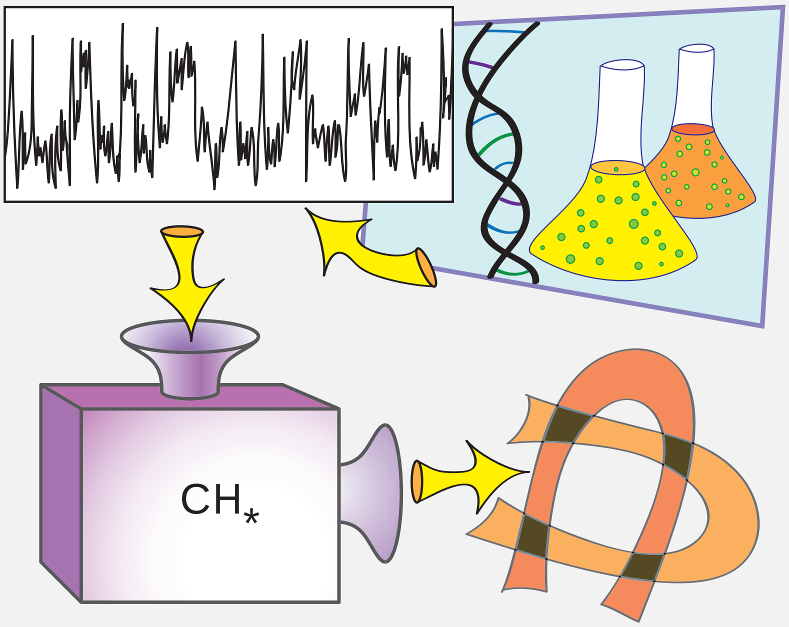

Figure 2 The Conley index CH_* of experimental time series data can rigorously verify chaotic dynamics.

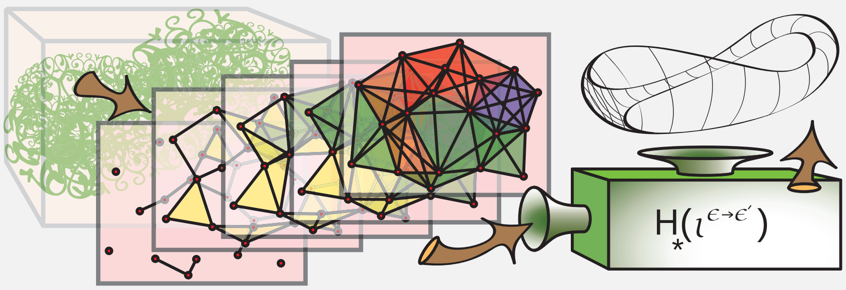

Figure 3 Persistent homology of a simplicial approximation finds hidden structures in large data sets.

What does the data look like?

Problem: High-Dimensional Data Analysis. Given a large, high-dimensional data set, how can one determine its shape and structure?

Tool: Persistent Homology. Though the subject of topology is often introduced in terms of doughnuts, coffee cups, knots, or other visual icons, the true strength of topology is the ease with which it analyses high-dimensional objects. The impact of this strength is perhaps best asserted in data-analysis, where the incoming rate of large, high-dimensional data sets currently far exceeds statisticians' abilities to analyse and describe the data sets.

Assume for the sake of argument that one is given a data set that consists of a *sampling* (perhaps, though not necessarily random) of a reasonable subset $X \subset \mathbf{E}^n$ of Euclidean space. Nature has trained the human brain to reconstructing shapes from planar projections, but this works only for certain (small!) values of n.

Knowing the homology of X is a good basis for asserting the global features of the 'true' model X of the data. Several basic statistical ideas – *e.g.*, *clustering* – are readily seen to correspond to something homological, in this example $\dim H_0$. The natural question presents itself: how can one compute $H_*(X)$ from a discrete sampling of points $N \subset X$?

The work of Carlsson et al. employs the following strategy. Fix a parameter $\epsilon > 0$, and build a simplicial complex R_ϵ as follows: a k-simplex of R_ϵ is a collection of $k+1$ data points in N pairwise within distance ϵ. Fixing $X \subset \mathbf{E}^n$ a manifold, then for ϵ sufficiently small and N sufficiently dense, the complex R_ϵ has the same homotopy type (and thus homology) as X. However, one is given a fixed data set, and further refinement maybe be expensive or impossible. Thus, one is forced to vary ϵ. Which ϵ best captures the true topology of the underlying data set? For ϵ too small, R_ϵ is a discrete set; for ϵ too large, R_ϵ is a single simplex. In this context, the golden mean may not exist.

Algebraic-topology suggests a functional approach. One of the simplest and best insights of the Grothendieck programme is the notion that the topology of a given space is framed in the *mappings* to or from that space. With this perspective as guide, one considers the ordered sequence of spaces $\{R_\epsilon\}$ for $\epsilon > 0$, stitched together by *inclusion maps* $\iota^{\epsilon\to\epsilon'} : R_\epsilon \hookrightarrow R_{\epsilon'}$ for $\epsilon < \epsilon'$. The homology of the family of maps $\iota^{\epsilon\to\epsilon'}$ is the called the *persistent homology* of the data set: $\iota_*^{\epsilon\to\epsilon'}$ captures which homological features (holes in the data set) *persist* over the range of parameters $[\epsilon, \epsilon']$.

Carlsson et al. use the classification of *modules* over a polynomial ring (with field coefficients) to compute persistent homology and to correlate it with the birth and death of topological features in the data [7]. This allows a principled and automatic distillation of complex data sets into global features — a method that does not rely on projections or heuristics.

Specific successes of the method include the following.

- Persistent homology was used [2] to find significant features hidden in a large data set of pixellated *natural images* compressed onto a 7-dimensional sphere; most notable is a persistent *Klein bottle* in H_2, which in turn yields insights into the structure of the space of natural images.
- Recent work [3] uses persistent homology to find hidden structures in experimental data associated with the *V1 visual cortex* of certain primates.

Moral: "The shape of the data lies not in a single space, but in a diagram of spaces."

References

1 Y. Baryshnikov and R. Ghrist, 'Target enumeration via Euler characteristic integrals,' preprint (2007).

2 G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, 'On the local behavior of spaces of natural images,' *Intl. J. Comput. Vision*, 76(1), (2008), 1–12.

3 G. Carlsson, T. Ishkhanov, F. Mémoli, D. Ringach, G. Sapiro, 'Topological analysis of the responses of neurons in V1,' preprint (2007).

4 T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology*, Applied Mathematical Sciences 157, Springer-Verlag, 2004.

5 K. Mischaikow, M. Mrowzek, J. Reiss, and A. Szymczak, 'Construction of Symbolic Dynamics from Experimental Time Series,' *Phys. Rev. Lett.* 82, (1999) 1144 - 1147.

6 P. Zgliczynski and K. Mischaikow, 'Rigorous numerics for partial differential equations: the Kuramoto-Sivashinsky equation,' *Foundations of Comp. Math.*, 1, (2001), 255–288.

7 A. Zomorodian and G. Carlsson, 'Computing Persistent Homology,' *Disc. and Comp. Geom.*, 33, (2005), 249–274.