FP7-ICT-2011.8
Contract no.: 318493

www.toposys.org

Deliverable D2.2

Progress and Activity Report on WP2

Deliverable Nature:	Report (R)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	M24
Actual Delivery Date:	M24
Version:	1.0

Contents

1 Summary

Our main goal in this work package is to bring the strength of statistics to topological tools. Our first task is to gain a better understanding of topological noise. Most results in applied topology for approximating the underlying persistence or homology of a space are of the form, given a sufficient sample, there exists some finite simplicial complex we can build on top of our sample such that the answer we obtain will be correct or approximately correct in the appropriate sense.

In particular for persistence, stability gives an upper bound on the magnitude of the noise. Importantly however, these are only upper bounds often in a form which are unsuitable for statisitcal tests. Little is known how the "topological noise" behaves. This is important for constructing a null hypothesis as well as constructing algorithms which deal with outliers in a more robust way than current topological methods.

We present several results:

- Central limit theorems for Betti numbers on stationary point processes. We also note that currently in preparation are central limit theorems for persistent Betti numbers.
- Simplification of 2D vector fields with sufficient and necessary conditions for simplification - with a version in preparation for 3D vector fields.
- Gaussian kernels on persistence diagrams. This is in the spirit of *persistence landscapes* which raises persistence diagrams into a Hilbert space, vastly speeding up the computation of distances between diagrams (as opposed to the Hungarian algorithm for computing bottleneck distance).

Much of this work is in preparation or submission but we highlight the work which is available:

- D. Yogeshwaran, E. Subag and R.J. Adler, "Random geometric complexes in the thermodynamic regime," accepted
- P. Skraba, B. Wang, G. Chen and P. Rosen. "2D Vector Field Simplification Based on Robustness". IEEE Pacific Visualization (PacificVis) , 2014.Best Paper)

We do not go into the details but also highlight the work in the paper

- Adler, R. J., Bartz, K., Kou, S. C., Monod, A. "Estimating Thresholding Levels for Random Fields via Euler Characteristics." submitted.

which is important for understanding how to use topological information for determining parameters. This type of work is also complementary to the semi-supervised framework for learning we present in WP3. Furthermore, TOPOSYS had a poster appear at this years' ECCS

- Adler, R. J., Skraba P. "Topological Detection of Heavy Tailed Distributions," ECCS 2014

Finally, we highlight an upcoming result on long bar in uniform point processes. This is currently work in progress which will be presented at a workshop shortly and is in preparation for submission. We highlight the result here however.

Let $\mathcal{P}_n$ be a homogeneous Poisson process on the unit cube $Q = [0,1]^d$ with rate $\lambda = n$, and let $\mathcal{U}(\mathcal{P}_n, r) := \bigcup_{p \in \mathcal{P}_n} B_r(p)$. For every i-cycle σ we denote $B(\sigma), D(\sigma), P(\sigma)$ the birth-time, death-time, and persistence (ratio), respectively. We want to argue that

$$P_{\max} = \max_{\sigma} P(\sigma) = \Theta\left(\left(\frac{\log n}{\log\log n}\right)^{1/i}\right).$$

2 Stationary Point Processes

We consider the topology of simplicial complexes with vertices the points of a random point process and faces determined by distance relationships between the vertices. In particular, we study the Betti numbers of these complexes as the number of vertices becomes large, obtaining limit theorems for means, strong laws, concentration inequalities and central limit theorems.

As opposed to most prior papers treating random complexes, the limit with which we work is in the so-called 'thermodynamic' regime (which includes the percolation threshold) in which the complexes become very large and complicated, with complex homology characterised by diverging Betti

numbers. The proofs combine probabilistic arguments from the theory of stabilizing functionals of point processes and topological arguments exploiting the properties of Mayer-Vietoris sequences. The Mayer-Vietoris arguments are crucial, since homology in general, and Betti numbers in particular, are global rather than local phenomena, and most standard probabilistic arguments are based on the additivity of functionals arising as a consequence of locality.

This paper is concerned with structures created by taking (many) random points and building the structure based on neighbourhood relations between the points. Perhaps the simplest way to describe this is to let $\Phi = \{x_1, x_2, \dots\}$ be a finite or countable, locally finite, subset of points in $\mathbb{R}^d$, for some $d > 1$, and to consider the set

$$\mathcal{C}_B(\Phi, r) \triangleq \bigcup_{x \in \Phi} B_x(r), \tag{1}$$

where $0 < r < \infty$, and $B_x(r)$ denotes the d-dimensional ball of radius r centred at $x \in \mathbb{R}^d$.

When the points of Φ are those of a stationary Poisson process on $\mathbb{R}^d$, this union is a special case of a 'Boolean model', and its integral geometric properties – such as volume, surface area, Minkowski functionals – have been studied in the setting of stochastic geometry since the earliest days of that subject. Our interest, however, lies in the homological structure of $\mathcal{C}_B(\Phi, r)$, in particular, as expressed through its Betti numbers. Thus our approach will be via the tools of algebraic topology, and, to facilitate this, we shall generally work not with $\mathcal{C}_B(\Phi, r)$ but with a homotopically equivalent abstract simplical complex with a natural combinatorial structure. This will be the Čech complex with radius r built over the point set Φ, denoted by $\mathcal{C}(\Phi, r)$.

The first, and perhaps most natural topological question to ask about these sets is how connected are they. This is more a graph theoretic question than a topological one, and has been well studied in this setting, with [30] being the standard text in the area. There are various 'regimes' in which it is natural to study these questions, depending on the radius r. If r is small, then the balls in (1) will only rarely overlap, and so the topology of both $\mathcal{C}_B(\Phi, r)$ and $\mathcal{C}(\Phi, r)$ will be mainly that of many isolated points. This is known as the 'dust regime'. However, as r grows, the balls will tend to overlap, and so a large, complex structure will form, leading to the notion of 'continuum percolation', for which the standard references are [15] and [24]. The percolation transition occurs within what is known as the 'thermodynamic', regime, and is typically the hardest to analyse. The third and final regime arises as r continues to grow, and (loosely speaking) $\mathcal{C}_B(\Phi, r)$ merges into a single large set with no empty subsets and so no interesting topology.

There has been considerable recent interest in the topological properties of $\mathcal{C}_B(\Phi, r)$ and $\mathcal{C}(\Phi, r)$ that go beyond mere connectivity or the volumetric measures provided by integral geometry. These studies were initiated by Matthew Kahle in [18], in a paper which studied the growth of the expected Betti numbers of these sets when the underlying point process Φ was either a Poisson process or a random sample from a distribution satisfying mild regularity properties. Shortly afterwards, more sophisticated distributional results were proven in [20]. An extension to more general stationary point processes Φ on $\mathbb{R}^d$ can be found in [40], while, in the Poisson and binomial settings, [3] looks at these problems from the point of view of the Morse theory of the distance function. Recently [4] has established important – from the point of view of applications – extensions to the results of [3, 18, 20] in which the underlying point process lies on a manifold of lower dimension than an ambient Euclidean space in which the balls of (1) are defined.

However, virtually all of the results described in the previous paragraph (with the notable exception of some growth results for expected Betti numbers in [18] and numbers of critical points in [3]) deal with the topology of the dust regime. What is new in the current paper is a focus on

the thermodynamic regime, and new results that go beyond the earlier ones about expectations. Moreover, because of the long range dependencies in the thermodynamic regime, proofs here involve considerably more topological arguments than is the case for the dust regime.

2.1 Summary of Results

Throughout we shall assume that all our point processes are defined over $\mathbb{R}^d$ for $d \geqslant 2$. Denoting Betti numbers of a set $A \subset \mathbb{R}^d$ by $\beta_k(A)$, $k = 1, \ldots, d-1$, we are interested in $\beta_k(\mathcal{C}_B(\Phi, r))$ for point processes $\Phi \subset \mathbb{R}^d$. Since the Betti numbers for $k \geqslant d$ are identically zero, these values of k are uninteresting. On the other hand, $\beta_0(A)$ gives the number of connected components of A. While this is clearly interesting and important in our setting, it has already been studied in detail from the point of view of random graph theory, as described above. Indeed, (sometimes stronger) versions of virtually all our results for the higher Betti numbers already exist for β_0 (cf. [1, 30]), and so this case will appear only peripherally in what follows. A summary of our results, grouped according to the underlying point processes involved:

1. *General stationary point processes:* For a stationary point process Φ and $r \in (0, \infty)$, we study the asymptotics of $\beta_k(\mathcal{C}_B(\Phi \cap W_l, r))$ as $l \to \infty$ and where $W_l = [-\frac{l}{2}, \frac{l}{2})^d$. We show convergence of expectations (Lemma 2.5) and, assuming ergodicity, we prove strong laws for all the Betti numbers and a concentration inequality for β_0.

2. *Stationary Poisson point processes:* Retain the same notation as above, but take $\Phi = \mathcal{P}$, a stationary Poisson point process on $\mathbb{R}^d$. In this setting we prove a central limit theorem for the Betti numbers of $\mathcal{C}_B(\mathcal{P} \cap W_l, r)$ and $\mathcal{C}(\mathcal{P} \cap W_l, r)$, for any $r \in (0, \infty)$, as $l \to \infty$. We also treat the case in which l points are chosen uniformly in W_l and obtain a similar result, although in this case we can only prove the central limit theorem for $r \notin I_d$, where the interval I_d will be defined in Section 2.7. Informally, I_d is the interval of radii where both $\mathcal{C}_B(\mathcal{P}, r)$ and its complement have unbounded components a.s.. We only remark here that $I_2 = \varnothing$ and I_d is a non-degenerate interval for $d \geqslant 3$.

3. *Inhomogeneous Poisson and binomial point processes:* Now, consider either the Poisson point process $\mathcal{P}_n$ with non-constant intensity function nf, for a 'nice', compactly supported, density f, or the binomial process of n iid random variables with probability density f. In this case the basic set-up requires a slight modification, and so we consider asymptotics for $\beta_k(\mathcal{C}_B(\mathcal{P}_n, r_n))$ as $n \to \infty$ and $nr_n^d \to r \in (0, \infty)$. We derive an upper bound for variances and a weak law. In the Poisson case, we also derive a variance lower bound for the top homology. For the corresponding binomial case we prove a concentration inequality (and use this to prove a strong law for both cases.

A few words on our proofs: In the case of stationary point processes, we shall use the nearly-additive properties of Betti numbers along with sub-additive theory arguments. In the Poisson and binomial cases, the proofs center around an analysis of the so-called add-one cost function,

$$\beta_k(\mathcal{C}_B(\mathcal{P} \cup \{O\}, r)) - \beta_k(\mathcal{C}_B(\mathcal{P}, r)),$$

where O is the origin in $\mathbb{R}^d$. While simple combinatorial topology bounds with martingale techniques suffice for strong laws, weak laws, and concentration inequalities, a more careful analysis via the Mayer-Vietoris sequence is required for the central limit theorems.

Our central limit theorems rely on similar results for stabilizing Poisson functionals (cf. [32]), which in turn were based upon martingale central limit theory. As for variance bounds, while upper bounds can be derived via Poincaré or Efron-Stein inequalities, the more involved lower bounds exploit the recent bounds developed in [22] using chaos expansions of Poisson functionals.

One of the difficulties in analyzing Betti numbers that will become obvious in the proof of the central limit theorem is their global nature. Most known examples of stochastic geometric functionals satisfy both the notions of stabilization (cf. [32]) known as 'weak' and 'strong' stabilization. However, we shall prove that higher Betti numbers satisfy weak stabilization but satisfy strong stabilization only for certain radii regimes. We are unable to prove strong stabilization of higher Betti numbers for all radii regimes because of the global dependence of Betti numbers on the underlying point process.

Beyond the Čech complex Although this paper concentrates on the Čech complex as the basic topological object determined by a point process, this is but one of the many geometric complexes that could have been chosen. There are various other natural choices including the Vietoris-Rips, alpha, witness, cubical, and discrete Morse complexes (cf. [13, Section 7], [41, Section 3]) that are also of interest. In particular, the alpha complex is homotopy equivalent to the Čech complex ([41, Section 3.2]), as is an appropriate discrete Morse complex ([13, Theorem 2.5]). This immediately implies that all the limit theorems for Betti numbers in this paper also hold for these complexes. Moreover, since our main topological tools can be shown to hold for all the complexes listed above, most of our arguments should easily extend to obtain similar theorems for these cases as well.

2.2 Preliminaries

We refer readers to the standard texts such as [16, 25] for more details on the topology we need, while [34, 38] covers the point process material. In this report we omit proofs but refer the reader to the paper described in the summary.

Our two main topological tools are collected in the following two lemmas. The first is needed for obtaining various moment bounds on Betti numbers of random simplicial complexes, and the second will replace the role that additivity of functionals usually plays in most probabilistic limit theorems. Because the arguments underlying these lemmas are important for what follows, and will be unfamiliar to most probabilistic readers, we shall prove them both. However both contain results that are well known to topologists.

Lemma 2.1. *Let $\mathcal{K}, \mathcal{K}^1$ be two finite simplicial complexes such that $\mathcal{K} \subset \mathcal{K}^1$ (i.e., every simplex in $\mathcal{K}$ is also a simplex in $\mathcal{K}^1$). Then, for every $k \geqslant 1$, we have that*

$$\big|\beta_k(\mathcal{K}^1) - \beta_k(\mathcal{K})\big| \;\leqslant\; \sum_{j=k}^{k+1} \#\big\{j\text{-simplices in } \mathcal{K}^1 \backslash \mathcal{K}\big\}.$$

With a little more work, one can go further than the previous lemma and derive an explicit equality for differences of Betti numbers. This is again a classical result in algebraic topology which is derived using the Mayer-Vietoris sequence (see [10, Corollary 2.2]). However we shall state it here as it is important for our proof of the central limit theorem.

A little notation is needed before we state the lemma. A sequence of Abelian groups $G_1, \ldots, G_l$ and homomorphisms $\eta_i : G_i \to G_{i+1}$, $i = 1, \ldots, l-1$ is said to be *exact* if $\operatorname{im} \eta_i = \ker \eta_{i+1}$ for all $i = 1, \ldots, l-1$. If $l = 5$ and G_1 and G_5 are trivial, then the sequence is called *short exact*.

Lemma 2.2 (Mayer-Vietoris Sequence)**.** *Let $\mathcal{K}_1$ and $\mathcal{K}_2$ be two finite simplicial complexes and $\mathcal{L} = \mathcal{K}_1 \cap \mathcal{K}_2$ (i.e., $\mathcal{L}$ is the complex formed from all the simplices in both $\mathcal{K}_1$ and $\mathcal{K}_2$). Then the following are true:*

1. *The following is an exact sequence, and, furthermore, the homomorphisms λ_k are induced by inclusions:*

$$\cdots \to H_k(\mathcal{L}) \stackrel{\lambda_k}{\to} H_k(\mathcal{K}_1) \oplus H_k(\mathcal{K}_2) \to H_k(\mathcal{K}_1 \cup \mathcal{K}_2) \\ \to H_{k-1}(\mathcal{L}) \stackrel{\lambda_{k-1}}{\to} H_{k-1}(\mathcal{K}_1) \oplus H_{k-1}(\mathcal{K}_2) \to \cdots$$

2. *Furthermore,*

$$\beta_k(\mathcal{K}_1 \bigcup \mathcal{K}_2) \;=\; \beta_k(\mathcal{K}_1) + \beta_k(\mathcal{K}_2) + \beta(N_k) + \beta(N_{k-1}) - \beta_k(\mathcal{L}),$$

where $\beta(G)$ denotes the rank of a vector space G and $N_j = \ker \lambda_j$.

2.3 Point Process Preliminaries

A point process Φ is formally defined to be a random, locally-finite (Radon), counting measure on $\mathbb{R}^d$. More formally, let $\mathcal{B}_{\lfloor}$ be the σ-ring of bounded, Borel subsets of $\mathbb{R}^d$ and let $\mathbb{M}$ be the corresponding space of non-negative Radon counting measures. The Borel σ-algebra $\mathcal{M}$ is generated by the mappings $\mu \to \mu(B)$ for all $B \in \mathcal{B}_{\lfloor}$. A point process Φ is a random element in $(\mathbb{M}, \mathcal{M})$, i.e. a measurable map from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{M}, \mathcal{M})$. The distribution of Φ is the measure $\mathbb{P}\Phi^{-1}$ on $(\mathbb{M}, \mathcal{M})$.

We shall typically identify Φ with the positions $\{x_1, x_2, \dots\}$ of its atoms, and so for Borel $B \subset \mathbb{R}^d$ it we shall allow ourselves to write

$$\Phi(B) \;=\; \sum_i \delta_{x_i}(B) \;=\; \#\{i : x_i \in B\} \;=\; \#\{\Phi \cap B\},$$

where $\#$ denotes cardinality and δ_x the single atom measure with mass one at x. The intensity measure of Φ is the non-random measure defined by $\mu(B) = \mathbb{E}\{\Phi(B)\}$, and, when μ is absolutely continuous with respect to Lebesgue measure, the corresponding density is called the intensity of Φ.

For Borel $A \subset \mathbb{R}^d$, we write Φ_A for both the restricted random measure given by $\Phi_A(B) := \Phi(A \cap B)$ (when treating Φ itself as a measure) and the point set $\Phi \cap A$ (when treating Φ as a point set). To save space, we shall write Φ_l for Φ_{W_l}, where W_l is the 'window' $[-l/2, l/2)^d$, for all $l \geqslant 0$.

For a measure $\phi \in \mathbb{M}$, let $\phi_{(x)}$ be the translate measure given by $\phi_{(x)}(B) = \phi(B - x)$ for $x \in \mathbb{R}^d$ and $B \in \mathcal{B}_{\lfloor}$. A point process is said to be *stationary* if the distribution of $\Phi_{(x)}$ is invariant under such translation, i.e. $\mathbb{P}\Phi_{(x)}^{-1} = \mathbb{P}\Phi^{-1}$ for all $x \in \mathbb{R}^d$. For a stationary point process in $\mathbb{R}^d$, $\mu(B) = \lambda|B|$ for all $B \in B_b$, where $|B|$ denotes the Lebesgue measure of B, and the constant of proportionality λ is called the *intensity* of the point process.

For a set of measures $\Theta \in \mathcal{M}$, let the translate family be $\Theta_x := \{\phi_{(x)} : \phi \in \Theta\}$. A point process Φ is said to be *ergodic* if

$$\mathbb{P}\{\Phi \in \Theta\} \;\in\; \{0, 1\}$$

for all $\Theta \in \mathcal{M}$ for which

$$\mathbb{P}\left\{\Phi \in (\Theta \backslash \Theta_x) \cup (\Theta_x \backslash \Theta)\right\} = 0$$

for all $x \in \mathbb{R}^d$.

Finally, we say that Φ *has all moments* if, for all bounded Borel $B \subset \mathbb{R}^d$, we have

$$\mathbb{E}\Big\{[\Phi(B)]^k\Big\} < \infty, \quad \text{for all } k \geqslant 1. \tag{2}$$

2.4 Limit theorems for stationary point processes

This section is concerned with the Čech complex $\mathcal{C}(\Phi_l, r)$, where Φ is a stationary point process on $\mathbb{R}^d$ with unit intensity and, as above, Φ_l is the restriction of Φ to the window $W_l = [-l/2, l/2)^d$. The radius r is arbitrary but fixed.

It is natural to expect that, as a consequence of stationarity, letting $l \to \infty$, $l^{-d}\mathbb{E}\{\beta_k(\mathcal{C}(\Phi_l, r))\}$ will converge to a limit. Furthermore, if we also assume ergodicity for Φ, one expects convergence of $l^{-d}\beta_k(\mathcal{C}(\Phi_l, r))$ to a random limit. All this would be rather standard fare, and rather easy to prove from general limit theorems, if it were only true that Betti numbers were additive functionals on simplicial complexes, or, alternatively, the Betti numbers of Čech complexes were additive functionals of the underlying point processes. Although this is not the case, Betti numbers are 'nearly additive', and a correct quantification of this near additivity is what will be required for our proofs.

As hinted before Lemma 2.1, the additivity properties of Betti numbers are related to simplicial counts $S_j(\mathcal{X}, r)$, which, for $j \geqslant 0$, denotes the number of j-simplices in $\mathcal{C}(\mathcal{X}, r)$, and $S_j(\mathcal{X}, r; A)$, which denotes the number of j-simplices with at least one vertex in A.

Our first results are therefore limit theorems for these quantities.

Lemma 2.3. *Let Φ be a unit intensity stationary point process on $\mathbb{R}^d$, possessing all moments. Then, for each $j \geqslant 0$, there exists a constant $c_j := c(\mathcal{L}_\Phi, j, d, r)$ such that*

$$\mathbb{E}\{S_j(\Phi_A, r)\} \leqslant \mathbb{E}\{S_j(\Phi, r; A)\} \leqslant c_j|A|.$$

Lemma 2.4. *Let Φ be a unit intensity, ergodic, point process on $\mathbb{R}^d$ possessing all moments. Then, for each $j \geqslant 0$, there exists a constant, $\widehat{S}_j := \widehat{S}(\mathcal{L}_\Phi, j, d, r)$, such that, with probability one,*

$$\lim_{l\to\infty} \frac{S_j(\Phi, r; W_l)}{l^d} = \lim_{l\to\infty} \frac{S_j(\Phi_l, r)}{l^d} = \widehat{S}_j(\mathcal{L}_\Phi, r).$$

2.5 Strong Law for Betti numbers

In this section we shall start with a convergence result for the expectation of $\beta_k(\mathcal{C}(\Phi_l, r))$ when Φ is a quite general stationary point process, and then proceed to a strong law. We treat these results separately, since convergence of expectations can be obtained under weaker conditions than the strong law. In addition, seeing the proof for expectations first should make the proof of strong law easier to follow.

From [40, Theorem 4.2] we know that

$$\mathbb{E}\{\beta_k(\mathcal{C}(\Phi_l, r))\} = O(l^d).$$

The following lemma strengthens this result.

Lemma 2.5. *Let Φ be a unit intensity stationary point process possessing all moments. Then, for each $0 \leqslant k \leqslant d-1$, there exists a constant $\widehat{\beta}_k := \widehat{\beta}_k(\mathcal{L}_\Phi, r) \in [0, \infty)$ such that*

$$\lim_{l \to \infty} \frac{\mathbb{E}\{\beta_k(\mathcal{C}(\Phi_l, r))\}}{l^d} \;=\; \widehat{\beta}_k.$$

Remark 2.6. *The lemma is interesting only in the case when $\widehat{\beta}_k > 0$, and this does not always hold. However, it can be guaranteed for negatively associated point processes (including Poisson processes, simple perturbed lattices and determinantal point processes) under some simple conditions on void probabilities, cf. [40, Theorem 3.3].*

thm 2.7. *Let Φ be a unit intensity ergodic point process possessing all moments. Then, for $0 \leqslant k \leqslant d-1$, and $\widehat{\beta}_k$ as in Lemma 2.5,*

$$\frac{\beta_k(\mathcal{C}(\Phi_l, r))}{l^d} \;\stackrel{a.s.}{\to}\; \widehat{\beta}_k.$$

The following concentration inequality is an easy consequence of the general concentration inequality of [29].

thm 2.8. *Let Φ be a unit intensity stationary determinantal point process. Then for all $l \geqslant 1$, $\epsilon > 0$, and $a \in (\frac{1}{2}, 1]$, we have that*

$$\mathbb{P}\left\{\left|\beta_0(\mathcal{C}(\Phi_{l^{\frac{1}{d}}}, r)) - \mathbb{E}\Big\{\beta_0(\mathcal{C}(\Phi_{l^{\frac{1}{d}}, r))}\Big\}\right| \;\geqslant\; \epsilon l^a\right\} \;\leqslant\; 5 \exp\left(-\frac{\epsilon^2 l^{2a-1}}{16 K_d(\epsilon l^{a-1} + 2K_d)}\right),$$

where K_d is the maximum number of disjoint unit balls that can be packed into $B_0(2)$.

2.6 Poisson and binomial point processes

Since there is already an extensive literature on $\beta_0(\mathcal{C}(\mathcal{X}.r))$ for Poisson and binomial point processes, albeit in the language of connectedness of random graphs (e.g. [30]), in this section we shall restrict ourselves only to β_k for $1 \leqslant k \leqslant d-1$.

The models we shall treat start with a Lebesgue-almost everywhere continuous probability density f on $\mathbb{R}^d$, with a compact, convex support that (for notational convenience) includes the origin, and such that

$$0 \;<\; \inf_{x \in \mathrm{supp}(f)} f(x) \stackrel{\Delta}{=} f_* \;\leqslant\; f^* \stackrel{\Delta}{=} \sup_{x \in \mathbb{R}^d} f(x) < \infty. \tag{3}$$

The models are $\mathcal{P}_n$, the Poisson point process on $\mathbb{R}^d$ with intensity nf, and the binomial point process $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, where the X_i are i.i.d. random vectors with density f. From [18], we know that for both $\mathcal{P}_n$ and $\mathcal{X}_n$ the thermodynamic regime corresponds to the case $nr_n^d \to r \in (0, \infty)$, so that for such a radius regime we have that

$$\mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{P}_n, r_n))\} = \Theta(n), \qquad \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{X}_n, r_n))\} = \Theta(n).$$

In proving limit results for Betti numbers in these cases, much will depend on moment estimates for the add-one cost function. The add-one cost function for a real-valued functional F defined over finite point-sets $\mathcal{X}$ is defined by

$$D_x F(\mathcal{X}) \;\stackrel{\Delta}{=}\; F(\mathcal{X} \cup \{x\}) - F(\mathcal{X}), \qquad x \in \mathbb{R}^d. \tag{4}$$

Our basic estimate follows. For notational convenience, we write

$$\beta_k^n(\mathcal{X}) \stackrel{\Delta}{=} \beta_k(\mathcal{C}(\mathcal{X}, r_n)),$$

where $\{r_n\}_{n\geqslant 1}$ is a sequence of radii to be determined.

Lemma 2.9. *Let $1 \leqslant k \leqslant d-1$. For the Poisson point process $\mathcal{P}_n$ and binomial point process $\mathcal{X}_n$, with $nr_n^d \to r \in (0,\infty)$, we have that*

$$\Delta_k \stackrel{\Delta}{=} \max\left(\sup_{n\geqslant 1}\sup_{x\in\mathbb{R}^d} \mathbb{E}\{|D_x\beta_k^n(\mathcal{P}_n)|^4\}, \sup_{n\geqslant 1}\sup_{x\in\mathbb{R}^d} \mathbb{E}\{|D_x\beta_k^n(\mathcal{X}_n)|^4\}\right) \tag{5}$$

is finite

2.6.1 Strong laws

We begin with a lemma giving variance inequalities, which, en passant, establish weak laws for Betti numbers.

Lemma 2.10. *For the Poisson point process $\mathcal{P}_n$ and binomial point process $\mathcal{X}_n$, with $nr_n^d \to r \in (0,\infty)$, and each $1 \leqslant k \leqslant d-1$, there exists a positive constant c_1 such that for all $n \geqslant 1$,*

$$\mathsf{VAR}(\beta_k(\mathcal{C}(\mathcal{P}_n, r_n))) < c_2 n, \qquad \mathsf{VAR}(\beta_k(\mathcal{C}(\mathcal{X}_n, r_n))) < c_2 n. \tag{6}$$

Thus, as $n \to \infty$,

$$n^{-1}\left[\beta_k(\mathcal{C}(\mathcal{P}_n, r_n)) - \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{P}_n, r_n))\}\right] \stackrel{\mathbb{P}}{\to} 0,$$

and

$$n^{-1}\left[\beta_k(\mathcal{C}(\mathcal{X}_n, r_n)) - \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{X}_n, r_n))\}\right] \stackrel{\mathbb{P}}{\to} 0.$$

Thanks to the recent bound of [22, Theorem 5.2], we can also give a lower bound for the Poisson point process in the case of $k = d-1$.

Lemma 2.11. *For the Poisson point process $\mathcal{P}_n$ with $nr_n^d \to r \in (0,\infty)$, there exists a positive constant c_1 such that for all $n \geqslant 1$,*

$$\mathsf{VAR}(\beta_{d-1}(\mathcal{C}(\mathcal{P}_n, r_n))) > c_1 n. \tag{7}$$

Remark 2.12. *Note that from the universal coefficient theorem ([25, Theorem 45.8]) and Alexander duality ([36, Theorem 16]), we have that*[1]

$$\tilde{H}_k(\mathcal{C}_B(\mathcal{P}_n, r)) \cong \tilde{H}_{d-k-1}(\mathbb{R}^d \backslash \mathcal{C}_B(\mathcal{P}_n, r)).$$

Thus

$$\beta_{d-1}(\mathcal{C}_B(\mathcal{P}_n, r)) = \beta_0(\mathbb{R}^d \backslash \mathcal{C}_B(\mathcal{P}_n, r)) - 1.$$

$\beta_0(\mathbb{R}^d \backslash \mathcal{C}_B(\mathcal{P}_n, r))$ is nothing but the number of components of the vacant region of the Boolean model, which is easier to analyse and this shall play a crucial role in our proof.

[1] The $\tilde{H}_k$ are the reduced homology groups and it suffices to note that $\tilde{H}_k \cong H_k$ for $k \neq 0$ and $H_0 \cong \tilde{H}_0(.) \oplus \mathbb{F}$.

Our next main result is a concentration inequality for $\beta_k(\mathcal{C}(\mathcal{X}_n, r_n)$.

thm 2.13. *Let $1 \leqslant k \leqslant d-1$, $\mathcal{X}_n$ be a binomial point process, and assume that $nr_n^d \to r \in (0,\infty)$. Then, for any $a > \frac{1}{2}$ and $\epsilon > 0$, for n large enough,*

$$\mathbb{P}\left\{\left|\beta_k(\mathcal{C}(\mathcal{X}_n, r_n) - \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{X}_n, r_n)\}\right| \geqslant \epsilon n^a\right\} \;\leqslant\; \frac{C}{\epsilon} n^{2k+2-a} \exp(-n^\gamma),$$

where $\gamma = (2a-1)/4k$ and $C > 0$ is a constant depending only on a, r, k, d and the density f.

The proof, close to that of [30, Theorem 3.17], is based on a concentration inequality for martingale differences.

We now finally have the ingredients needed to lift the weak laws of Lemma 2.10 to the promised strong convergence.

thm 2.14. *For the Poisson point process $\mathcal{P}_n$ and binomial point process $\mathcal{X}_n$, with $nr_n^d \to r \in (0,\infty)$, and each $1 \leqslant k \leqslant d-1$, we have, with probability one,*

$$\lim_{n\to\infty} n^{-1}\left[\beta_k(\mathcal{C}(\mathcal{P}_n, r_n)) - \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{P}_n, r_n)\}\right] \;=\; 0,$$

and

$$\lim_{n\to\infty} n^{-1}\left[\beta_k(\mathcal{C}(\mathcal{X}_n, r_n)) - \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{X}_n, r_n)\}\right] \;=\; 0.$$

2.7 Central Limit Theorem

We have finally come to the main result: central limit theorems for Betti numbers.

We start with some definitions from percolation theory for the Boolean model on Poisson processes ([24]) needed for the proof of the Poisson central limit theorem. Recall firstly that we say that a subset A of $\mathbb{R}^d$ *percolates* if it contains an unbounded connected component of A.

Now let $\mathcal{P}$ be a stationary Poisson point process on $\mathbb{R}^d$ with unit intensity. (Unit intensity is for notational convenience only. The arguments of this section will work for any constant intensity.) We define the critical (percolation) radii for $\mathcal{P}$ as follows:

$$r_c(\mathcal{P}) \;\triangleq\; \inf\{r : \mathbb{P}\{C(\mathcal{P}, r) \text{ percolates}\} > 0\},$$

and,

$$r_c^*(\mathcal{P}) \;\triangleq\; \sup\{r : \mathbb{P}\left\{\mathbb{R}^d \backslash C(\mathcal{P}, r) \text{ percolates}\right\} > 0\}.$$

By Kolmogorov's zero-one law, it is easy to see that the both of the probabilities inside the infimum and supremum here are either 0 or 1. The first critical radius is called the *critical radius for percolation of the occupied component* and the second is the *critical radius for percolation of the vacant component.*

We define the *interval of co-existence*, $I_d(\mathcal{P})$, for which unbounded components of both the Boolean model and its complement co-exist, as follows:

$$I_d(\mathcal{P}) \;=\; \begin{cases} (r_c, r_c^*] & \text{if } \mathbb{P}\{C(\mathcal{P}, r_c) \text{ percolates}\} = 0, \\ [r_c, r_c^*] & \text{otherwise.} \end{cases}$$

From [24, Theorem 4.4 and Theorem 4.5], we know that $I_2(\mathcal{P}) = \varnothing$ and from [33, Theorem 1] we know that $I_d(\mathcal{P}) \neq \varnothing$ for $d \geqslant 3$. In high dimensions, it is known that $r_c \notin I_d(\mathcal{P})$ (cf. [39]).

We now need a little additional notation. Let $\{B_n\}_{n\geqslant 1}$ be a sequence of bounded Borel subsets in $\mathbb{R}^d$ satisfying the following four conditions:

(A) $|B_n| = n$, for all $n \geqslant 1$.

(B) $\bigcup_{n\geqslant 1} \bigcap_{m\geqslant n} B_m \;=\; \mathbb{R}^d$.

(C) $|(\partial B_n)^{(r)}|/n \to 0$, for all $r > 0$.

(D) There exists a constant b_1 such that $\mathrm{diam}(B_n) \leqslant b_1 n^{b_1}$, where $\mathrm{diam}(B)$ is the diameter of B.

In a moment we shall state and prove a central limit theorem for the sequences of the form $\beta_k(\mathcal{C}(\mathcal{P} \cap B_n, r))$, when the B_n are as above. Setting up the central limit theorem for the binomial case requires a little more notation.

In particular, we write $\mathcal{U}_n$ to denote the point process obtained by choosing n points uniformly in B_n, and call this the *extended binomial point process*. This is a natural binomial counterpart to the Poisson point process $\mathcal{P} \cap B_n$.

We finally have all that we need to formulate the main central limit theorem.

thm 2.15. *Let $\{B_n\}$ be a sequence of sets in $\mathbb{R}^d$ satisfying conditions (A)–(D) above, and let $\mathcal{P}$ and $\mathcal{U}_n$, $n \geqslant 1$, respectively, be the unit intensity Poisson process and the extended binomial point process described above. Take $k \in \{1, \ldots, d-1\}$ and $r \in (0, \infty)$. Then there exists a constant $\sigma^2 > 0$ such that, as $n \to \infty$,*

$$n^{-1}\mathsf{VAR}(\beta_k(\mathcal{C}(\mathcal{P} \cap B_n, r)) \;\to\; \sigma^2,$$

and

$$n^{-1/2}\left(\beta_k(\mathcal{C}(\mathcal{P} \cap B_n, r)) - \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{P} \cap B_n, r))\}\right) \;\Rightarrow\; N(0, \sigma^2).$$

Furthermore, for $r \notin I_d(\mathcal{P})$, there exists a τ^2 with $0 < \tau^2 \leqslant \sigma^2$ such that

$$n^{-1}\mathsf{VAR}(\beta_k(\mathcal{C}(\mathcal{U}_n, r)) \to \tau^2,$$

and

$$n^{-1/2}\left(\beta_k(\mathcal{C}(\mathcal{U}_n, r)) - \mathbb{E}\{\beta_k(\mathcal{C}(\mathcal{U}_n, r))\}\right) \;\Rightarrow\; N(0, \tau^2).$$

The constants σ^2 and τ^2 are independent of the sequence $\{B_n\}$.

Remark 2.16. *The condition $r \notin I_d(\mathcal{P})$, needed for the binomial central limit theorem, is rather irritating, and we are not sure if it is necessary or an artefact of the proof. It is definitely not needed for the case $k = d-1$. To see this, note that from the duality argument of Remark 2.12, we have that*

$$\beta_{d-1}(\mathcal{C}(\mathcal{P} \cap B_n, r)) \;=\; \beta_0(\mathbb{R}^d \backslash \mathcal{C}(\mathcal{P} \cap B_n, r)) - 1.$$

However, $\mathbb{R}^d \backslash \mathcal{C}(\mathcal{P} \cap B_n, r)$ is nothing but the vacant component of the Boolean model, and central limit theorems for $\beta_0(\mathbb{R}^d \backslash \mathcal{C}(\mathcal{X} \cap B_n, r))$ for both Poisson and binomial point processes are given in [32, p1040] for all $r \in (0, \infty)$. By the above duality arguments, this proves both the central limit theorems of Theorem 2.15, when $k = d-1$, and without the requirement that $r \notin I_d(\mathcal{P})$.

3 Robustness-based Simplification of 2D Vector Fields

This work describes the simplification of vector fields (primarily for visualization). Note that this work does not assume the vector field is a gradient vector filed of a scalar field. It gives a constructive algorithm for constructing the simplification as well as shows that in a given metric no smaller perturbation simplification exists. It ultimately makes use of Laplacian smoothing, which shows how this can be altered such that it will always work if simplification is possible.

A considerable amount of research has been focused on vector field simplification based on the notion of a topological skeleton [60, 62]. A topological skeleton consists of critical points connected by special streamlines called *separatrices*, which provide a condensed representation of the flow by dividing the domain into regions of uniform flow behavior. However, existing simplification techniques rely on the stable extraction of the topological skeleton, which can be difficult due to instability in numerical integration, especially when processing highly rotational flows, e.g. Figure 1. Furthermore, the distance and area-based relevance measures that are commonly used to determine the cancellation ordering of critical points typically rely on geometric proximities and do not consider the flow magnitude, an important physical property of the flow.

Here we show new vector field simplification scheme derived from the recently introduced notion of *robustness*. Intuitively, the robustness of a critical point is the minimum amount of perturbation,

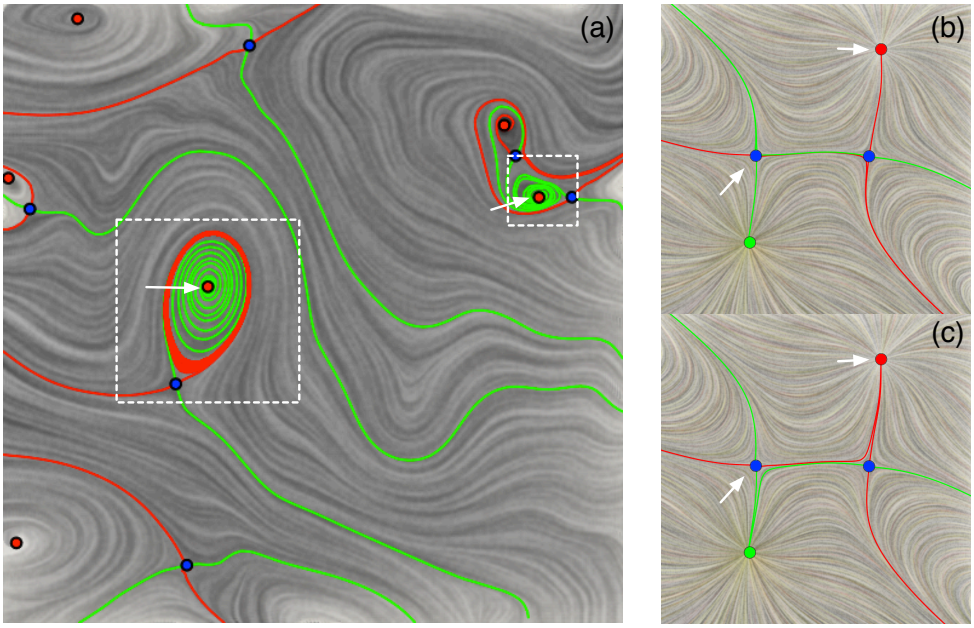

Figure 1: Topological skeleton: Sinks (and saddle-sink separatrices) are red, sources (and saddle-source separatrices) green, and saddles blue. (a) A highly rotational flow field where the pointed critical points are close to Hopf-bifurcations. Numerical inaccuracies may accumulate during integration and separatrices may intersect or switch. (b)-(c) Instability of separatrices under a small perturbation: The upper right sink is not connected with the saddle on the left in (b), but is after a small perturbation in (c).

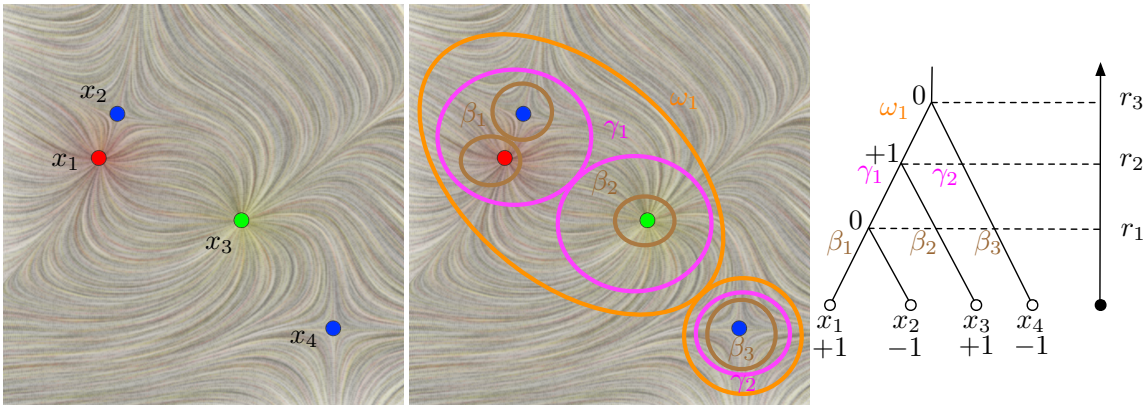

Figure 2: Figure adapted from [75]. Suppose the vector field is continuous, where sinks are red, sources are green, and saddles are blue. From left to right: vector fields f, relations among components of $\mathbb{F}_r$, and augmented merge trees. f contains four critical points, a sink x_1, a source x_3, and two saddles x_2 and x_4. We use β, γ, ω, etc. to represent components of certain sublevel sets.

with respect to a metric encoding flow magnitude, that is required to cancel it within a local neighborhood. Our method finds, in the space of all vector fields, the one that is closest to the original vector field with a particular set of critical points removed, according to a metric based on the L_∞ norm (the maximum point-wise modification to the vector field). Our results are optimal in this norm, that is, there exists no simplification with a smaller perturbation.

3.1 Background

We provide relevant background in degree theory and robustness by reviewing previous work [45, 75] with minimal algebraic definitions and illustrating the related concepts through an example (Figure 2 adapted from [75]). We also provide introductory descriptions of isolating neighborhoods and Laplacian smoothing [47, 78].

Degrees. For a critical point x in 2D, its degree $\deg(x)$ equals its *(Poincaré) index*, that is, the number of field rotations while traveling along a closed curve centered at x counter-clockwise. Sources, sinks, centers, and saddles have indices $+1$, $+1$, $+1$ and -1, respectively. Furthermore, for a (path-)connected component C that encloses several critical points, its degree $\deg(C)$ is the sum of the respective degrees of those critical points [45]. For our robustness-based simplification strategy, we rely on a corollary of the Poincaré-Hopf theorem (which is also employed by topological-skeleton-based simplification, e.g., [72]), which states that if a connected component C in 2D has degree zero, then it is possible to replace the vector field inside C with a vector field free of critical points.

Merge tree. To analyze a continuous 2D vector field $f : \mathbb{R}^2 \to \mathbb{R}^2$, we define a corresponding scalar function (referred to as the flow magnitude function) $f_0 : \mathbb{R}^2 \to \mathbb{R}$ which assigns for each point the magnitude (Euclidean norm) of the corresponding vector, $f_0(x) = ||f(x)||_2$. We use $\mathbb{F}_r = f_0^{-1}(-\infty, r]$ to denote the *sublevel set* of f_0 for some $r \geqslant 0$. $\mathbb{F}_0$ is precisely the set of critical

points of f.

Increasing r from 0, the space $\mathbb{F}_r$ evolves and we can construct a graph that tracks the (connected) components of $\mathbb{F}_r$ as they appear and merge. This is called a *merge tree* (or *join tree* as described in [43]). The root represents the entire domain of f_0 and the leaves represent the creation of a component at a local minimum. An internal node represents the merging of two or more components. We further record an integer at each node, which is the degree of the corresponding component in the sublevel set, and refer to the result as an augmented merge tree. An initial computation of the degrees of critical points is sufficient to determine the degree of any component of any sublevel set by computing the sum of the degrees of the critical points lying in it [45]. An example is shown in Figure 2[2]. The merge tree on the right shows how the components of the sublevel sets $\mathbb{F}_r$ evolve. At $r = 0$ there are four components that correspond to the four critical points, each with nonzero degree. At $r = r_1$, components that contain x_1 and x_2 merge into a single component β_1, which has zero degree. When $r = r_2$, components β_1 and β_2 merge into a single component γ_1 with degree +1, while β_3 grows into γ_2. Finally at $r = r_3$, the single component ω_1 has zero degree.

Static robustness and its properties. The *(static) robustness* of a critical point is the height of its lowest degree zero ancestor in the merge tree [44, 75]. The static robustness quantifies the stability of a critical point with respect to perturbations of the vector fields through the following lemmas explicitly stated in [75].

We first define the concept of *perturbation*. Let $f, h : \mathbb{R}^2 \to \mathbb{R}^2$ be two continuous 2D vector fields. Define the distance between the two mappings as $\mathrm{d}(f, h) = \sup_{x \in \mathbb{R}^2} ||f(x) - h(x)||_2$. A continuous mapping h is an *r-perturbation* of f, if $\mathrm{d}(f, h) \leqslant r$.

Lemma 3.1 (Critical Point Cancellation [75])**.** *Suppose a critical point x of f has robustness r. Let C be the connected component of $\mathbb{F}_{r+\delta}$ containing x, for an arbitrarily small $\delta > 0$. Then, there exists an $(r + \delta)$-perturbation h of f, such that $h^{-1}(0) \cap C = \varnothing$ and $h = f$ except possibly within the interior of C.*

Lemma 3.2 (Degree & Critical Point Preservation [75])**.** *Suppose a critical point x of f has robustness r. Let C be the connected component of $\mathbb{F}_{r-\delta}$ containing x, for some $0 < \delta < r$. For any ϵ-perturbation h of f where $\epsilon \leqslant r - \delta$, the sum of the degrees of the critical points in $h^{-1}(0) \cap C$ is $\deg(C)$. If C contains only one critical point x, we have $\deg(h^{-1}(0) \cap C) = \deg(x)$. That is, x is preserved as there is no ϵ-perturbation that could cancel it.*

Revisiting the example in Figure 2, the robustness of the critical points x_1, x_2, x_3, and x_4 is r_1, r_1, r_3, and r_3, respectively. Since the robustness of x_3 is r_3, for any $\delta > 0$, we consider a component $C \subseteq \mathbb{F}_{r_3+\delta}$ that is slightly larger than ω_1 and contains x_3 (in fact, ω_1 contains all four critical points). Lemma 3.1 implies the existence of an $(r_3 + \delta)$-perturbation that cancels x_3 by locally modifying the component C. Now consider another component $C' \subseteq \mathbb{F}_{r_3-\delta}$ where $r_2 < r_3 - \delta < r_3$, then C' has degree +1. Lemma 3.2 states that any $(r_3 - \delta)$-perturbation preserves the degree of C'.

Isolating neighborhood and Laplacian smoothing. Previously, topology-based simplification has focused on cancelling pairs of critical points that are connected by separatrices. Zhang et al. [78] and Chen et al. [47] propose to compute an *isolating neighborhood* surrounding a pair of critical points, where a critical-point-free vector field can be found by solving a constrained optimization problem, referred to as a vector-valued *Laplacian smoothing* [78].

[2] We do not show any components that appear after $r = 0$ as they have zero degrees and do not correspond to critical points of the vector field.

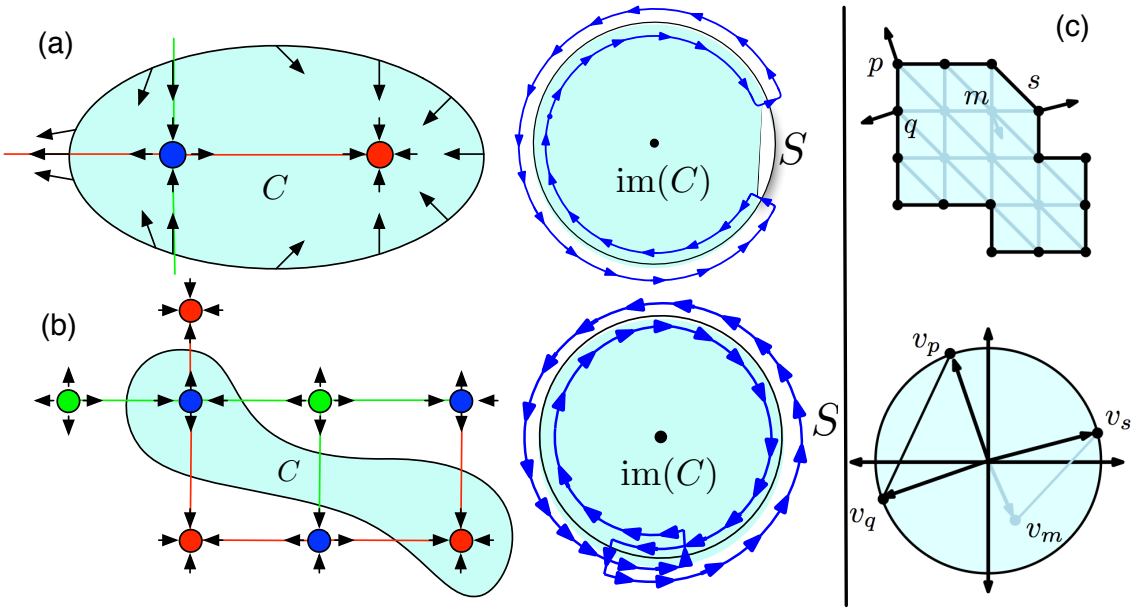

Figure 3: (a)-(b): Illustrative examples for uncovered (a) and covered (b) boundaries of im (C). (c): A component and its image space with a few mappings highlighted.

Based on Conley index theory, every boundary point of an isolating neighborhood can be classified as either an *entrance* or *exit* point. If an isolating region C in the domain contains multiple critical points and has a trivial Conley index, the flow inside C can be replaced with a new field free of critical points [78]. A typical situation for C to have a trivial Conley index is when its boundary ∂C consists of a single inflow and a single outflow component. As shown in the later examples, such an isolating neighborhood is not always easy to construct. The robustness-based method has no such constraint and only requires the degree of C to be zero.

3.2 Robustness-Based Simplification Algorithms

In robustness-based simplification, we first locate sets of critical points that share the lowest zero-degree ancestors in the merge tree and sort them based on their robustness values. For each set with a common robustness r, we compute the corresponding component of the sublevel set $C \subseteq \mathbb{F}_r$. Since by construction $\deg(C) = 0$, our strategy can simplify C, whereas the distance-based strategy requires an isolating neighborhood with trivial Conley index.

Given a 2D vector field restricted to a degree-zero component C, $f : C \to \mathbb{R}^2$, we define the *image space* of C, im (C). For each point $p \in C$, we have a vector $v_p = f(p) \in \mathbb{R}^2$. im $(C) \subset \mathbb{R}^2$ is constructed by mapping p to its vector coordinates v_p. The origin in im (C) corresponds to the critical points (0 vectors) in C. Since $C \subseteq \mathbb{F}_r$, it follows that $\forall p \in C$, $||v_p||_2 \leqslant r$, therefore im (C) is contained within a disc of radius r in $\mathbb{R}^2$. We denote the boundary of this disc by S.

Now suppose the boundary of C, denoted as ∂C, is a simple closed curve[3]. Note that the above maps ∂C to S, obtaining the image, $\mathrm{im}\,(\partial C)$. We refer to the boundary of $\mathrm{im}\,(C)$ as *uncovered*, if $\mathrm{im}\,(\partial C) \subset S$, otherwise, as *covered*. Figures 3(a)-(b) illustrate these concepts. Note that both examples have zero degree. In 3(a), the region C encloses a saddle-sink pair connected by a separatrix. By traversing counter-clockwise along ∂C and observing how its image $\mathrm{im}\,(\partial C)$ wraps around S, we see that the boundary of $\mathrm{im}\,(C)$ is uncovered. In 3(b), the region C encloses a saddle-sink pair not connected by separatrix and the boundary of $\mathrm{im}\,(C)$ is covered.

In the PL setting, the vector field f is restricted to a triangulation K of C, $f : K \to \mathbb{R}^2$, where the support of K, $|K| = C$. We construct the image of C by mapping each vertex $p \in K$ to its vector coordinates $v_p = f(p)$. Through linear interpolation, this construction also maps edges and triangles in K to edges and triangles in $\mathrm{im}\,(C)$ (Figure 3(c)). The concept of *covered* and *uncovered* boundaries of $\mathrm{im}\,(C)$ can be defined similarly up to a small additive constant.

Algorithm Overview Our simplification strategy consists of four operations:

- **Smoothing**(C): Perform Laplacian smoothing on C;
- **Cut**(C): Deform the vector field in its image space $\mathrm{im}\,(C)$ to remove critical points in C;
- **Unwrap**(C): Modify the vector field in its image space $\mathrm{im}\,(C)$ so part of its boundary is uncovered;
- **Restore**(C): Set the boundary to its original value.

Three cases are classified by the Conley index of C, denoted as $\mathrm{CH}_*(C)$. The operations to simplify each case are:

(a) If $\mathrm{CH}_*(C)$ is trivial, return $C_1 =$ **Smoothing**(C).

(b) If $\mathrm{CH}_*(C)$ is nontrivial and the boundary of $\mathrm{im}\,(C)$ is uncovered, then $C_1 =$ **Cut**(C), and return $C_2 =$ **Smoothing**(C_1).

(c) If $\mathrm{CH}_*(C)$ is nontrivial and the boundary of $\mathrm{im}\,(C)$ is covered, then $C_1 =$ **Unwrap**(C), $C_2 =$ **Cut**(C_1), $C_3 =$ **Restore**(C_2) and return $C_4 =$ **Smoothing**(C_3).

By construction, $deg(C) = 0$ in all three cases. Indeed, $deg(C) \neq 0$ is a sufficient condition such that there exists no simplification.

For the details on **Cut** and **Unwrap**, we direct the reader to the paper.

3.3 Synthetic Examples

We illustrate our robustness-based simplification strategy on three PL synthetic examples, highlighting the three different cases.

SyntheticA (Figure 4) corresponds to the example in Figure 2. It involves pairs of critical points connected by separatrices. At r_1, we have a component that contains critical points x_1 and x_2 and at r_3 we have a component that contains all four critical points x_1 to x_4. The simplification hierarchy involves two steps ranked by robustness values: first x_1 and x_2 are simplified, and then

[3]This is not needed, but it simplifies the algorithm and exposition.

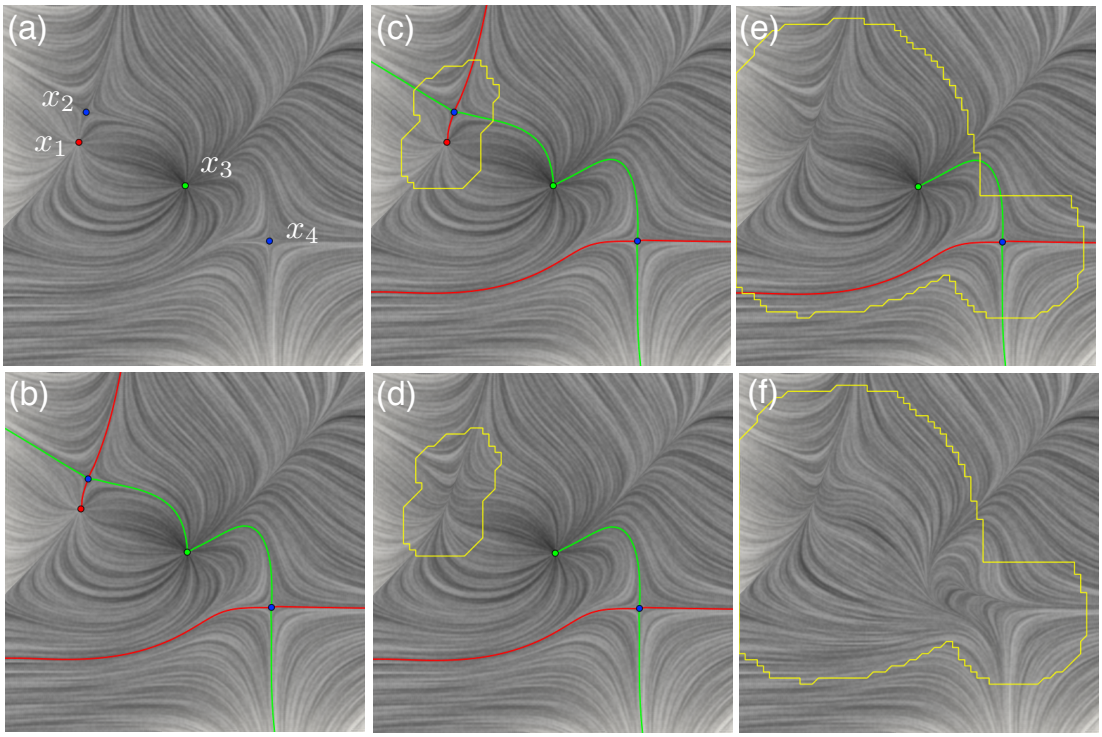

Figure 4: SyntheticA. (a) The original vector field: sinks are red, sources are green and saddles are blue. (b) The topological skeleton: saddle-sink separatrices are red and saddle-source separatrices are green. (c)-(d) 1st level simplification: before (c) and after (d) **Smoothing**. (e)-(f) 2nd level simplification: before (e) and after (f) **Smoothing**.

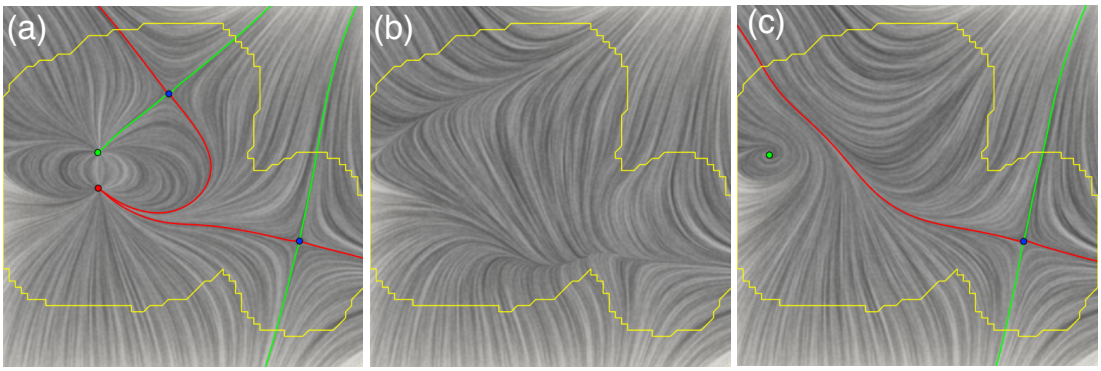

Figure 5: SyntheticB. (a) the original vector field with its topological skeleton. (a)-(b): Single level simplification before (a) and after (b) by **Cut** and **Smoothing**. (c) Only applying **Smoothing** does not make the region a critical point free field.

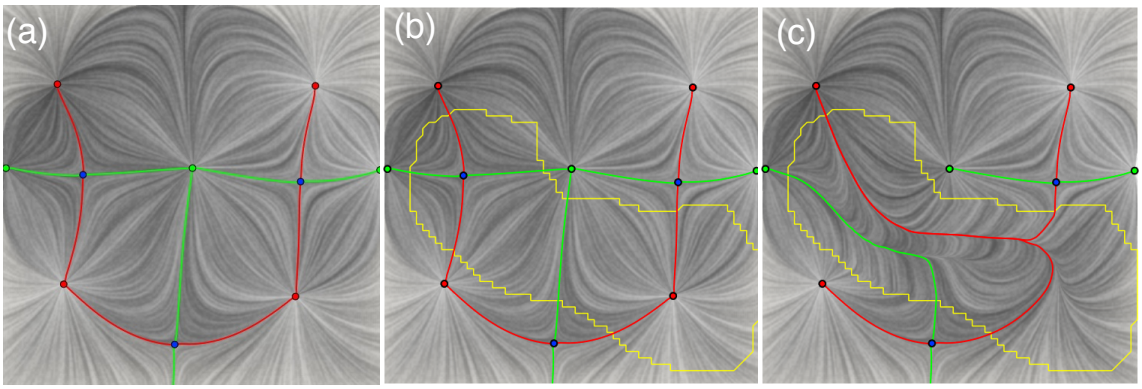

Figure 6: SyntheticC. (a) the original vector field with topological skeleton. (b)-(c) Before (b) and after (c) simplification by combining **Unwrap**, **Cut** and **Smoothing**.

x_3 and x_4. Since both components (marked by yellow boundary) have a trivial Conley index, this corresponds to case (a), where only **Smoothing** operations are needed. SyntheticB (Figure 5) involves a group of four critical points that are interconnected by separatrices, which could be simplified in a single level using a robustness-based strategy. Since the component of interest has a nontrivial Conley index, directly applying Laplacian smoothing fails (as shown in Figure 5(c)). The component's boundary is uncovered, so we apply case (b) of our simplification by combining **Cut** with **Smoothing**.

SyntheticC (Figure 6) corresponds to case (c) of our algorithm. This is an untypical case involving a pair of critical points not directly connected by a separatrix. In this case, the component of interest C has nontrivial Conley index, and the boundary of its image is covered. The robustness-based strategy cancels the critical point pair without any issue by combining **Unwrap**, **Cut** and **Smoothing** operations. We further focus on this example by illustrating the image space of C, $\mathrm{im}\,(C)$, during various steps of simplification in Figure 7. In Figure 7(a), the entire boundary and disk are covered. However, from the left phase plot in Figure 8, we can see that the degree is 0. Once the optimal unwrapping point is computed, we perform the **Unwrap** operation, giving the right phase plot in Figure 8 and the image space in Figure 7(b), leaving the boundary S uncovered. The effect of the **Cut** operation in image space is shown in Figure 7(c), creating a void surrounding the origin. Lastly, in Figure 7(d), the boundary is restored for the final output.

3.4 Results

We demonstrate our robustness-based simplification strategy on a number of real-world datasets. When possible, we compare our method with distance-based simplification. Ths is only a subset of the obtained results.

We identified a number of scenarios where the distance-based and robustness-based methods disagree. Two pairs of critical points are studied(see the full paper). Even though the pairing of these four critical points is consistent with both metrics, their actual simplification orderings are different. The distance-based method cancels the pair in the middle-right of the domain first, while the robustness-based method cancels the lower-middle pair first. Figure 9 provides an example that shows the discrepancy of the two approaches in determining the simplification ordering of critical

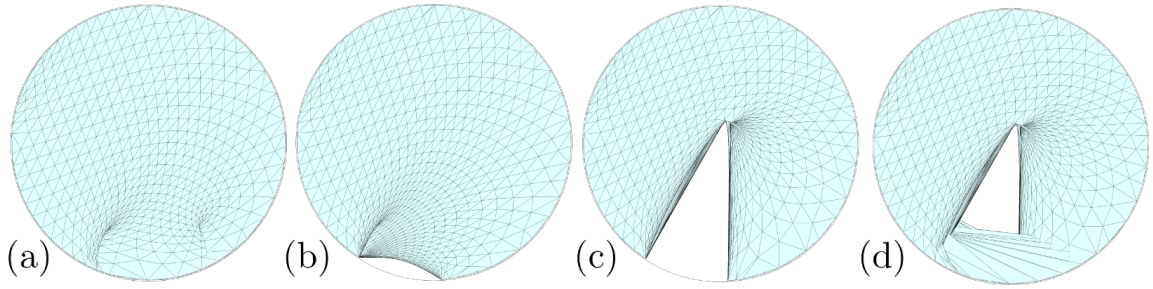

Figure 7: SyntheticC. The image space is shown through the different steps: (a) original, (b) after **Unwrap**, (c) after **Cut**, and (d) final output after **Restore**.

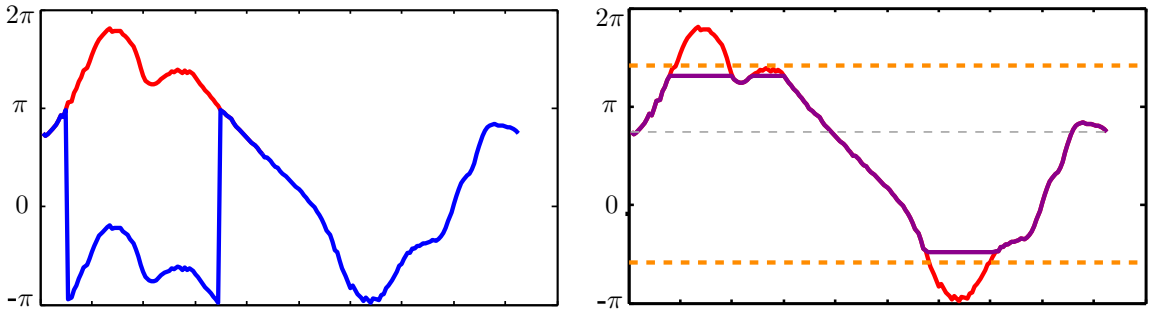

Figure 8: SyntheticC. Left: The phase plot, original version (blue), and the phase-unwrapped version (red). Right: The phase plot with optimal unwrap point (orange) and the modified phase plot with boundary uncovered (purple).

point pairs in the time-varying setting. In this example, we look at consecutive time steps from the OceanD dataset. Figure 9(a) highlights the critical points of interest. The pairings of these four critical points again agree with each other using both topological-skeleton and robustness metrics. We perform a per-slice simplification using the two approaches. The results are shown in the second (distance-based) and third (robustness-based) columns in (b)-(c), respectively. From the results, we see that the cancellation orderings change over time using the distance-based metric. This is due to an increased distance between the two critical points near the upper-right corner, resulting in a change of the simplification order. On the other hand, the robustness for these two pairs is stable. Therefore, the robustness-based simplification returns a consistent outcome in this example.

There are a number of cases where the topological-skeleton-based metric combined with the Laplacian smoothing technique is incapable of simplifying the given vector field. For example, for the SyntheticB dataset shown in Figure 5, it is impossible to find an isolating neighborhood with a trivial Conley index that encloses all the critical points due to the boundary condition. Therefore, even though the obtained local region is guaranteed to be zero degree, Laplacian smoothing fails to solve for a critical point free field. On the other hand, the simplification algorithm introduced in Section 3.2 successfully simplifies the field. The boundary configuration of this region is rather complex and does not satisfy a trivial Conley index. The Laplacian smoothing based on this boundary configuration fails (bottom), but the proposed simplification method succeeds. These

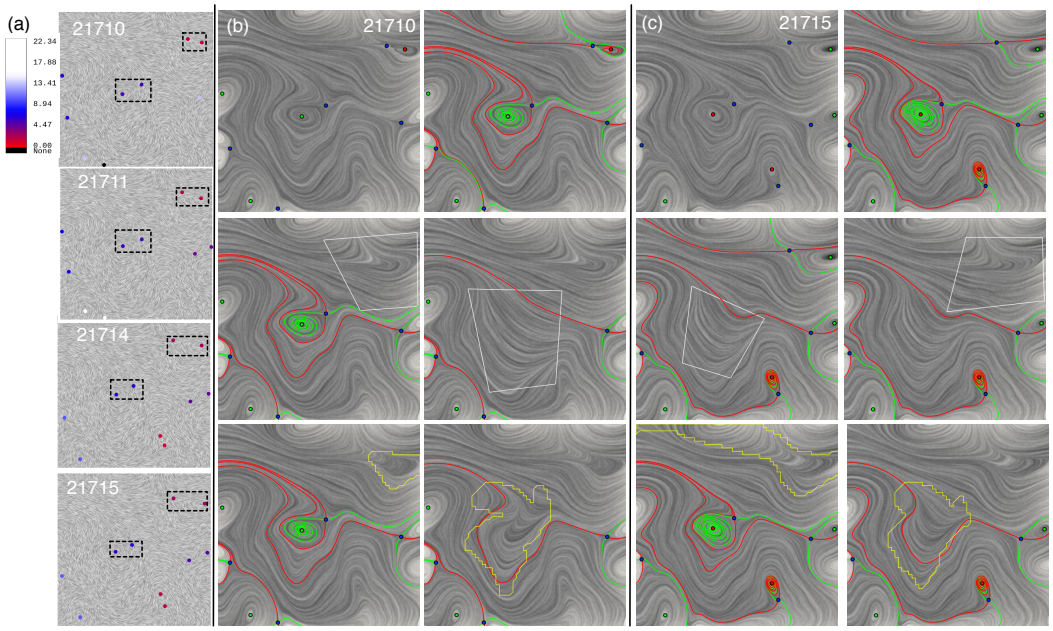

Figure 9: The OceanD dataset. (a) A sampled time series with pairs of critical points highlighted, where white numbers indicate time stamps. (b) #21710. (c) #21715. For each subfigure (b)-(c), Top Row: The original vector field (left) and with the separatrices (right). Middle Row: The simplification ordering for the distance-based strategy. Bottom Row: The simplification ordering for the robustness-based strategy. Orderings for distance and robustness-based methods are consistent in (b) and different in (c).

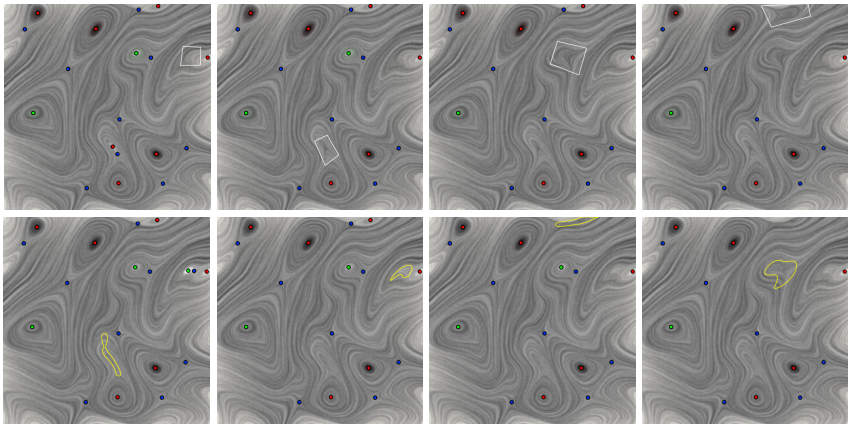

Figure 10: The Combustion dataset. The bottom-up hierarchical simplifications (Top) from the distance-based strategy and (Bottom) from the robustness-based strategy.

two examples showcase the utility of the proposed algorithm in solving a critical point free field within any given regions with zero degree. This relieves the requirement of the trivial Conley index whose corresponding isolating neighborhood is sometimes difficult to obtain.

Figure 6 shows a nontypical case that involves the cancellation of a pair of critical points not directly connected by separatrix. It is impossible for the topological-skeleton-based method to compute an isolating neighborhood that encloses two critical points (but not the others) not connected by separatrix [47]. Nonetheless, the robustness metric derives a local region that encloses only these two critical points with total degree equal to zero under a certain configuration of the flow magnitude. Hence, these two critical points may be cancelled. Whereas this may rarely occur in the real-world data, it illustrates the flexibility and generality of the proposed method. In practice, a simpler but similar situation may occur.

3.5 Discussions

The algorithm comes with theoretical guarantees on the amount of perturbation we introduce. The motivation for Laplacian smoothing is to produce more visually appealing results. However, to the best of our knowledge, no nontrivial bounds exist on the amount of perturbation introduced by such a smoothing. In practice, smoothing only marginally increases the amount of perturbation
Scalability: Our method should scale to very large datasets. The robustness computation and the simplification steps (e.g., **Cut** and **Unwrap**) run in linear time in the size of the mesh. For example, for a region of 21k vertices and 64k edges, **Cut** required 2 seconds in MATLAB and 0.03 seconds in C++.
Generality: The simplification procedure requires only that the degree of the boundary be zero and so applies to a wide range of cases. It can deal with highly rotational data (e.g., centers) as well as cases where critical points are not connected by separatrices.
Other metrics: We use robustness and the L_∞ norm (the maximum over the domain), but using other metrics such as the L_2-norm, which incorporates both the magnitude of the vectors and the area to capture a quantity closer to the energy of a perturbation, would be interesting. The simplification requires only degree-zero components and any metric could be used to construct a hierarchy. It is an open question to find degree-zero regions under different metrics.
Time-varying and 3D vector fields: The main challenge in simplifying time-varying 2D vector fields is to achieve consistency across time-slices, e.g., obtaining critical points correspondences at a given simplification level. Finally, the prospect of extending our framework to 3D vector fields (currently in preparation) is promising. Whereas there remain technical obstacles, certain operations (such as cutting and smoothing) in our pipeline readily extend to higher dimensions.

4 A Multi-Scale Kernel for Topological Machine Learning

Visual data is typically piped through complex processing chains in order to extract information that can be used to solve high-level inference problems, such as recognition, detection or segmentation. This information might be in the form of low-level appearance descriptors, eg, SIFT [99], or of higher-level nature, eg, responses of batteries of object detectors [98] or features extracted at specific layers of deep convolutional networks [79]. In many problems, the consolidated data is then fed to some discriminant classifier such as the popular SVMs.

While there has been substantial progress on the encoding of low-level visual information, only recently have people started looking into the *topological structure* of the visual data as an additional

source of information. With the emergence of *topological data analysis (TDA)* [84], tools for efficiently identifying topological structure have become broadly available. Using these tools, several authors have recently demonstrated that TDA can capture characteristics of the data that other methods often fail to provide, [105, 97].

Along these lines, studying persistent homology [92] is particularly popular, since it allows to capture, roughly speaking, the birth and death times of topological features, eg, connected components, holes, etc., at multiple scales. This information is then represented in the form of *persistence diagrams (PD)*. However, using PDs as input to machine learning algorithms requires a notion of similarity or distance between PDs. Popular distance measures, such as the bottleneck or Wasserstein distance are prohibitively expensive to compute when the diagrams contain a large number of points (eg, when the input data is noisy, as it is the case in most vision problems). A popular strategy to facilitate machine learning when the input data does not have a vector space structure is to design a *kernel* that maps the data into a reproducing kernel Hilbert space (RKHS).

We propose a multi-scale L_2 embedding of persistence diagrams, based on the principles of heat diffusion. The L_2 norm between embedded PDs is then used to define a distance measure that is bounded, from above, by the degree-1 Wasserstein distance. We show that the negative of this squared norm is a valid (conditionally) positive definite kernel and can thus be used within most kernel-based learning techniques. The heat parameter of our kernel controls its robustness to noise and can be tuned to the data (just like the parameters of a RBF kernel).

From a conceptual point of view, Bubenik's concept of *persistence landscapes* [83] is possibly the closest to ours, since an embedding of persistence diagrams is proposed. While persistence landscapes were not explicitly designed for use in machine learning algorithms, we will draw the connection to our work in §4.3 and show that they in fact admit the definition of a valid (conditionally) positive definite kernel. In summary, persistence landscapes as well as our approach represent computationally attractive alternatives to the bottleneck or Wasserstein distance.

4.1 Background

There is a natural metric associated to persistence diagrams, called the *bottleneck distance.* Loosely speaking, the distance of two diagrams is expressed by minimizing the largest distance of any two corresponding points, over all bijections between the two diagrams. Let F and G be two persistence diagrams, each augmented by adding each point (t, t) on the diagonal with countably infinite multiplicity. The *bottleneck distance* is

$$d_B(F, G) := \inf_{\gamma} \sup_{x \in F} \|x - \gamma(x)\|_\infty \tag{8}$$

where γ ranges over all bijections from the individual points of F to the individual points of G. As shown by Cohen-Steiner [90], persistence diagrams are stable with respect to the bottleneck distance.

The bottleneck distance embeds into a more general class of distances, called *Wasserstein distances.* For any positive real number p, the *p-Wasserstein distance* is

$$d_{W,p}(D_f, D_g) := \left(\inf_{\gamma} \sum_{x \in D_f} \|x - \gamma(x)\|_\infty^p \right)^{\frac{1}{p}} , \tag{9}$$

where again γ ranges over all bijections from the individual elements of D_f to the individual elements of D_g.

4.2 Embedding persistence diagrams into L_2

In this section, we propose an explicit *multi-scale* embedding of persistence diagrams into L_2. The Hilbert space structure of L_2 will then be used to define a distance between diagrams and a kernel for topological machine learning.

Note that a persistence diagram D, ie, a multi-set of points in $\mathbb{R}^2$, does not possess a Hilbert space structure per se. However, D can be uniquely represented as a sum of Dirac delta distributions[4] by replacing each point in D with a Dirac delta centered at that point. Since Dirac deltas are elements of the Hilbert space $H^{-2}(\mathbb{R}^2)$ [96, Chapter 7], we obtain a canonical Hilbert space structure for persistence diagrams by adopting this point of view.

Unfortunately, the induced metric on the space of persistence diagrams does *not* take account of the distance of the points to the diagonal, and therefore it cannot be robust against perturbations of the diagrams. To address this issue, we propose to use the sum of Dirac deltas as an initial condition of a heat diffusion problem with a Dirichlet boundary condition on the diagonal. The solution of this partial differential equation is an L_2 function for any chosen scale parameter $\sigma > 0$. In the following paragraphs we will (1) formally define this multi-scale L_2 embedding of persistence diagrams, (2) describe a simple algorithm that exactly evaluates the inner product in this Hilbert space and (3) prove its robustness against perturbations.

Definition 4.1. *Let $\Omega = \{(x, y) \in \mathbb{R}^2 \colon y \geqslant x\}$ denote the space above the diagonal and let δ_p denote a Dirac delta centered at the point p. For a given persistence diagram D, we now consider the solution $u \colon \Omega \times \mathbb{R}_{\geqslant 0} \to \mathbb{R}$ of the partial differential equation*[5]

$$\partial_{xx} u + \partial_{yy} u = \partial_t u \qquad \text{in } \Omega \times \mathbb{R}_{>0}, \tag{10}$$

$$u = 0 \qquad \text{on } \partial\Omega \times \mathbb{R}_{\geqslant 0}, \tag{11}$$

$$u = \sum_{p \in D} \delta_p \qquad \text{on } \Omega \times \{0\}. \tag{12}$$

The $L_2(\Omega)$-embedding at scale $\sigma > 0$ of a persistence diagram D is now defined as

$$\Psi_\sigma(D) = u(\cdot, \cdot, \sigma) \ . \tag{13}$$

The solution of the partial differential equation (13) can be obtained by extending the domain from Ω to $\mathbb{R}^2$ and replacing (12) with

$$u = \sum_{p \in D} \delta_p - \delta_{\overline{p}} \qquad \text{on } \mathbb{R}^2 \times \{0\}, \tag{14}$$

where $\overline{p} = (b, a)$ is $p = (a, b)$ mirrored at the diagonal. It can be shown that the restriction to Ω of the solution to this extended problem solves the original equation. It is given by convolving the initial condition (14) with a Gaussian kernel:

$$u(x, y, t) = \frac{1}{4\pi t} \sum_{p \in D} e^{-\frac{\|(x,y)-p\|^2}{4t}} - e^{-\frac{\|(x,y)-\overline{p}\|^2}{4t}} \tag{15}$$

[4] A Dirac delta distribution is a function that evaluates a given smooth function at 0.

[5] Since the initial conditon (12) is not an L_2 function, this equation only makes sense in a distributional setting. For a rigorous treatment of existence and uniqueness of the solution see [96, Chapter 7].

Given two diagrams F and G, we can compute the $L_2(\Omega)$ inner product of their embeddings explicitly:

$$\begin{aligned}(\Psi_\sigma(F), \Psi_\sigma(G)) &= \int_\Omega \Psi_\sigma(F)\Psi_\sigma(G) \\ &= \frac{1}{8\pi\sigma} \sum_{\substack{p\in F \\ q\in G}} e^{-\frac{\|p-q\|^2}{8\sigma}} - e^{-\frac{\|p-\overline{q}\|^2}{8\sigma}}\end{aligned} \tag{16}$$

We show that our kernel is robust against perturbations of the input data by proving that the L_2-distance of the embeddings of two persistence diagrams is upper bounded by their Wasserstein distance. This allows us to leverage current and future stability theorems from TDA, see §4.1. The precise the statement of the theorems are omitted from this report but will be present in the submitted version of this work.

4.3 Kernel construction

The Hilbert space structure of $L_2(\Omega)$ allows us to construct kernels in different ways, eg, using the distance measure $d_{\mathrm{PSS}}(F,G) = \|\Psi_\sigma(\cdot) - \Psi_\sigma(\cdot)\|_{L_2(\Omega)}$. In particular, we have ([81, Exercise 1.20]):

thm 4.2. *Let $\mathcal{X}$ be the space of persistence diagrams and let $F, G \in \mathcal{X}$. Then the kernel $k_{PSS} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, defined by*

$$\begin{aligned}k_{PSS}(F,G) &= -d^2_{PSS}(F,G) \\ &= -\|\Psi_\sigma(F) - \Psi_\sigma(G)\|^2_{L_2(\Omega)}\end{aligned} \tag{17}$$

is conditionally positive definite.

Note that by saying *conditionally positive definite* we adhere to the commonly used terminology of [103]. Further, as a consequence of [81, Theorem 2.2], the kernel $e^{-t\cdot k_{\mathrm{PSS}}(\cdot,\cdot)}$ is *positive definite* for all $t > 0$. In the remainder we will refer to the kernel defined in (17) as the *persistence scale space (PSS) kernel*.

4.4 Evaluation

To evaluate the embedding proposed in §4.2, we (1) investigate conceptual differences to persistence landscapes and then (2) address performance in the context of shape classification/retrieval and texture recognition problems.

Comparison to persistence landscapes In [83], Bubenik introduced *persistence landscapes*, a representation of persistence diagrams as functions in the Banach space $L_p(\mathbb{R}^2)$. This construction was mainly intended for statistical computations, enabled by the vector space structure of L_p. Additionally, for $p = 2$ we can exploit the Hilbert space structure of $L_2(\mathbb{R}^2)$ to construct a kernel analogously to (17). For the purpose of this work, we refer to this kernel as the *persistence landscape (PLS)* kernel and denote by Ψ^L the corresponding $L_2(\mathbb{R}^2)$ embedding of persistence diagrams.

For the first experiment, let $F_t = \{-t, t\}$ and $G_t = \{-t+1, t+1\}$ be two diagrams with one point each and $t \in \mathbb{R}_{\geqslant 0}$. The two points move away from the diagonal with increasing t, while maintaining the same distance to each other. Consequently, $d_{W,q}$ and the PSS distance asymptotically approach

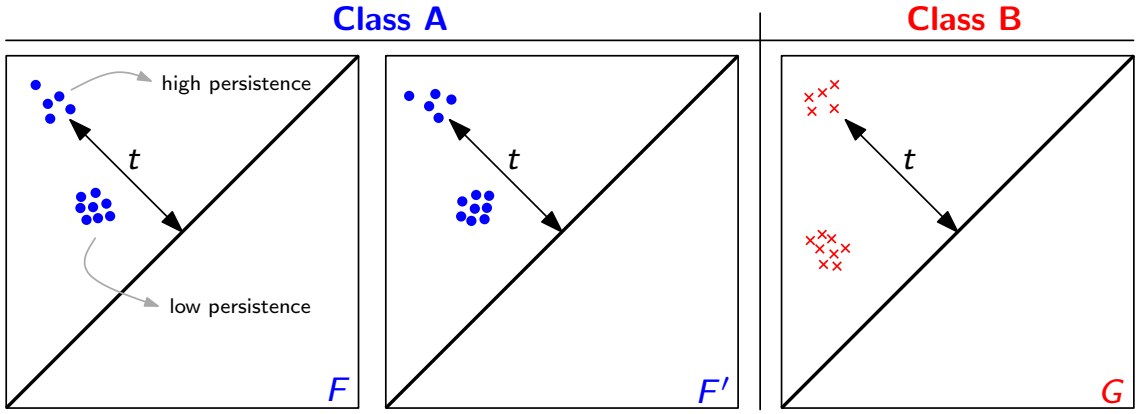

Figure 11: Two instances of persistence diagrams F, F' from class A and one diagram G from class B. The classes only differ in their points of low-persistence (ie, points closer to the diagonal).

a constant as $t \to \infty$. In contrast, the PLS distance grows in the order of $\sqrt{t}$ and, in particular, is unbounded. This means that the PLS distance emphasizes points of high persistence in the diagrams.

In the second experiment, we specifically focus on the previous observation. Fig. 11 illustrates persistence diagrams from data samples of two fictive classes A (F,F') and B (G). We first consider the PLS distance between F and F'. As we have seen in the previous experiment, their PLS distance will be dominated by variations in the points of high persistence. Similarly, the PLS distance between F and G will also be dominated by these points as long as t is sufficiently large. Consequently, instances of classes A and B would be inseparable in a nearest neighbor setup, even though they could easily be distinguished by looking at low-persistence features. This is in contrast to the bottleneck, Wasserstein or our PSS distance.

4.5 Empirical results

We report results on two vision tasks where topological data analysis has already been shown to provide valuable discriminative information ([97]): *shape classification/retrieval* and *texture image classification*. The purpose of the experiments is *not* to outperform the state-of-the-art on these problems – which would be rather challenging by only using topological information – but to demonstrate the advantages of the proposed PSS kernel/distance with respect to the alternative PLS kernel/distance.

Datasets. For shape classification/retrieval, we use the SHREC 2014 [102] benchmark dataset, see Fig. 12 (top). It consists of both *synthetic* and *real* shapes, given as 3D meshes. The synthetic part contains 300 meshes of humans in 20 different poses; the *real* part contains 400 meshes of humans (male and female) in 10 different poses. We use the meshes in full resolution, ie, without any mesh decimation.

For the texture recognition experiments, we use the `Outex_TC_0000` test suite of the OuTeX database [100], downsampled to 32×32 patches. The test suite provides 100 predefined training/testing splits and each of the 24 classes is equally represented by 10 images during training and

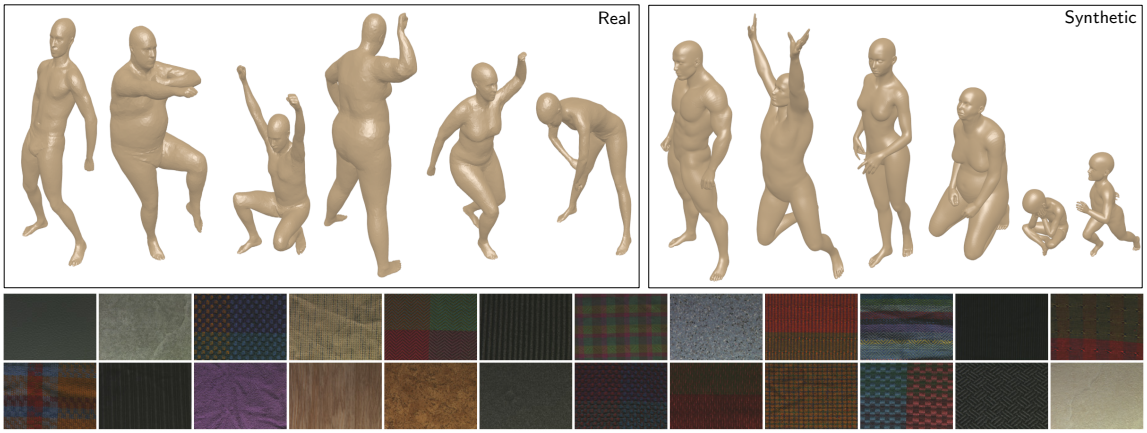

Figure 12: Examples from SHREC 2014 [102] (top) and the 24 texture classes of the OuTeX `Outex_TC_0000` test suite [100] (bottom).

testing.

4.5.1 Shape classification

Tables 1 and 2 list the classification results for the PLS and PSS kernel on the 20-class (synthetic) and 40-class (real) problem of SHREC 2014. All results are averaged over ten cross-validation runs using random 70/30 training/testing splits with a roughly equal class distribution.

In summary, for both types of data, we observe a consistent improvement in classification accuracy by switching from the PLS to the PSS kernel. For some choices of t_i, the gains even range up to $> 30\%$ while, in other cases, the improvements are rather small. This can be explained by the fact that varying t_i essentially varies the smoothness of the input data. While smoothness is a basic requirement for TDA in general, the PSS scale σ allows to address unfavourable smoothness settings to a certain extent. The PLS kernel, on the other hand, does not have this capability and essentially relies suitably preprocessed input data. As we can see, some choices of t_i do in fact lead performance close to the PSS kernel.

The important point, though, is that the PSS kernel allows tuning at the *classification stage*, eg, using cross-validation on the training parts of each split. This is the prevalent strategy for tuning RBF kernel parameters for instance. With the PLS kernel, we have to adjust the HKS time parameter which corresponds to changing the input data. This is undesirable in most situations, since HKS computation for meshes with a large number of vertices can be quite time-consuming and sometimes we might not even have access to the input data directly.

4.5.2 Shape retrieval

In addition to the classification experiments of the previous section, we also report retrieval results using standard performance measures (see [104, 102]). This allows us to evaluate the quality of the PSS distance directly.

For brevity, only the nearest-neighbor (NN) performance is shown in Table 3 (for all measures, see supplementary material). To study the effect of tuning the PSS scale σ, the PSS column lists

HKS t_i	PLS	PSS	Δ
t_1	68.0 ± 3.2	$\mathbf{94.7} \pm 5.1$	+26.7
t_2	88.3 ± 3.3	$\mathbf{99.3} \pm 0.9$	+11.0
t_3	61.7 ± 3.1	$\mathbf{96.3} \pm 2.2$	+34.7
t_4	81.0 ± 6.5	$\mathbf{97.3} \pm 1.9$	+16.3
t_5	84.7 ± 1.8	$\mathbf{96.3} \pm 2.5$	+11.7
t_6	70.0 ± 7.0	$\mathbf{93.7} \pm 3.2$	+23.7
t_7	73.0 ± 9.5	$\mathbf{88.0} \pm 4.5$	+15.0
t_8	81.0 ± 3.8	$\mathbf{88.3} \pm 6.0$	+7.3
t_9	67.3 ± 7.4	$\mathbf{88.0} \pm 5.8$	+20.7
t_{10}	55.3 ± 3.6	$\mathbf{91.0} \pm 4.0$	+35.7

Table 1: Shape classification on SHREC 2014 (synthetic).

HKS t_i	PLS	PSS	Δ
t_1	45.2 ± 5.8	$\mathbf{48.8} \pm 4.9$	+3.5
t_2	31.0 ± 4.8	$\mathbf{46.5} \pm 5.3$	+15.5
t_3	30.0 ± 7.3	$\mathbf{38.2} \pm 8.9$	+8.3
t_4	41.2 ± 2.2	$\mathbf{51.2} \pm 4.9$	+10.0
t_5	46.2 ± 5.8	$\mathbf{62.0} \pm 4.9$	+15.7
t_6	33.2 ± 4.1	$\mathbf{58.0} \pm 3.4$	+24.7
t_7	31.0 ± 5.7	$\mathbf{64.2} \pm 2.9$	+33.2
t_8	51.7 ± 2.9	$\mathbf{58.8} \pm 2.7$	+7.0
t_9	36.0 ± 5.3	$\mathbf{41.2} \pm 4.9$	+5.2
t_{10}	2.8 ± 0.6	$\mathbf{27.8} \pm 5.8$	+25.0

Table 2: Shape classification on SHREC 2014 (real).

the *maximum* NN performance that can be achieved over a range scales σ.

As we can see, the results are comparable to the classification experiment. However, at a few specific settings of the HKS time parameter t_i, the PLS distance performs on par, or better than the PSS distance. As indicated in §4.5.1 this can be explained by the changes in the smoothness of the input, as a function of t_i. Another observation is that NN performance of the PLS distance is quite unstable around the optimal choice of t_i, eg, it drops from 91% to 51.3% and 76.7% on SHREC 2014 (synthetic) and from 60.5% to 38.25% and 42.25% on SHREC 2014 (real). In contrast, the PSS distance exhibits relatively stable performance.

To put these results into context with existing works in shape retrieval, Table 3 also lists the top three entries (out of 22) of [102] on the same benchmark. On both real and synthetic datasets, we rank among the top five entries. This is interesting, since it indicates that TDA is a rich source of discriminative information for t his particular problem. In addition, since we only assess one HKS time parameter at a time, performance could potentially be improved further by more elaborate fusion strategies.

4.6 Texture Recognition

For texture recognition, all results are averaged over the 100 training/testing splits of the `Outex_TC_0000` test suite. Table 4 lists the performance of the PSS and the PLS kernel for 0-dimensional features (ie, connected components). Higher-dimensional features did not lead to any useful results on this problem. For comparison, we also list the performance of a simple nearest neighbour (NN) and a SVM classifier, trained on normalized histograms of CLBP responses. The NN classifier uses the

HKS t_i	PLS	PSS	Δ	PLS	PSS	Δ
t_1	53.3	**88.7**	+35.4	**24.0**	23.7	−0.3
t_2	91.0	**94.7**	+3.7	20.5	**25.7**	+5.2
t_3	76.7	**91.3**	+14.6	16.0	**18.5**	+2.5
t_4	84.3	**93.0**	+8.7	26.8	**33.0**	+6.2
t_5	85.0	**92.3**	+7.3	28.0	**38.7**	+10.7
t_6	63.0	**77.3**	+14.3	28.7	**36.8**	+8.1
t_7	65.0	**80.0**	+15.0	43.5	**52.7**	+9.2
t_8	73.3	**80.7**	+7.4	**70.0**	58.2	−11.8
t_9	73.0	**83.0**	+10.0	45.2	**56.7**	+11.5
t_{10}	51.3	**69.3**	+18.0	3.5	**44.0**	+40.5
Top-3 [102]	99.3 – 92.3 – 91.0			68.5 – 59.8 – 58.3		

Table 3: Nearest neighbor retrieval performance. *Left:* SHREC 2014 (synthetic); *Right:* SHREC 2014 (real).

CLBP Operator	PLS	PSS	Δ
`CLBP-S`	58.0 ± 2.29	$\mathbf{69.1 \pm 2.6}$	+11.1
`CLBP-M`	45.2 ± 2.48	$\mathbf{55.0 \pm 2.5}$	+9.8
`CLBP-S` (NN-χ^2)	68.5 ± 2.3		
`CLBP-M` (NN-χ^2)	71.0 ± 2.4		
`CLBP-S` (SVM-χ^2)	76.1 ± 2.2		
`CLBP-M` (SVM-χ^2)	76.7 ± 1.8		

Table 4: Classification results on the `Outex_TC_0000` test suite (downsampled to 32×32) .

χ^2 distance, the SVM uses the corresponding χ^2 kernel.

As in the previous experiments, the PSS kernel performs better than the PLS kernel by a large margin, with gains up to 12% in accuracy. The results using PSS kernel are also substantially higher than the results reported in [97] using the PSL distance on the same dataset in a comparable setup.

Finally, we remark that tuning the PLS kernel is less straightforward in this experiment. While we could artificially smooth the input images, CLBP responses or even tweak the radius of the CLBP operator, all choices require changes at the beginning of the processing pipeline. In contrast, adjusting the PSS scale σ via cross-validation is done at the end of the pipeline during classifier training.

4.7 Discussion

The proposed multi-scale L_2 embedding of persistence diagrams opens the pathway to using topological data analysis in the framework of kernel-based learning techniques. While the kernel enables us to leverage topological information, eg, when training discriminant classifiers for shapes or textures, we have also seen that this does not immediately lead to state-of-the-art performance. In fact, TDA should be considered a complementary source of information that can now be integrated with existing approaches, eg, using well-established algorithms from kernel learning [94].

References

[1] M.A. Akcoglu and U. Krengel. Ergodic theorems for superadditive processes. *J. Reine Angew. Math.*, 323:53–67, 1981.

[2] A. Björner. Topological methods. In *Handbook of Combinatorics*, volume 2, pages 1819–1872. Elsevier, Amsterdam, 1995.

[3] O. Bobrowski and R.J. Adler. Distance functions, critical points, and topology for some random complexes. arXiv:1107.4775, 2011.

[4] O. Bobrowski and S. Mukherjee. The topology of probability distributions on manifolds. arXiv:1307.1123, 2013.

[5] G. Carlsson. Topology and data. *Bull. Am. Math. Soc. (N.S.)*, 46(2):255–308, 2009.

[6] G. Carlsson. Topological pattern recognition for point cloud data. 2013. math.stanford.edu/∼gunnar/actanumericathree.pdf.

[7] T. K. Chalker, A. P. Godbole, P. Hitczenko, J. Radcliff, and O. G. Ruehr. On the size of a random sphere of influence graph. 31(3):596–609, 09 1999.

[8] A.E. Costa, M. Farber, and T. Kappeler. Topics of stochastic algebraic topology. *Electronic Notes in Theoretical Computer Science*, 283(0):53 – 70, 2012. Proceedings of the workshop on Geometric and Topological Methods in Computer Science (GETCO).

[9] L. Decreusefond, E. Ferraz, H. Randriam, and A. Vergne. Simplicial homology of random configurations. *Adv. Appl. Prob.*, 46:1–20, 2014.

[10] C.J.A. Delfinado and H. Edelsbrunner. An incremental algorithm for Betti numbers of simplicial complexes. In *Proceedings of the Ninth Annual Symposium on Computational Geometry*, SCG '93, pages 232–239, New York, NY, USA, 1993. ACM.

[11] H. Edelsbrunner and J.L. Harer. *Computational Topology, An Introduction.* American Mathematical Society, Providence, RI, 2010.

[12] B. Efron and C. Stein. The jackknife estimate of variance. *The Ann. of Stat.*, 9(3):pp. 586–596, 1981.

[13] R. Forman. A user's guide to discrete Morse theory. *Sém. Lothar. Combin.*, 48, 2002.

[14] R. Ghrist. Barcodes : the persistent topology of data. *Bull. Am. Math. Soc. (N.S.)*, 45(1):61–75, 2008.

[15] P. Hall. *Introduction to the Theory of Coverage Processes.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York, 1988.

[16] A. Hatcher. *Algebraic Topology.* Cambridge University Press, Cambridge, New York, 2002.

[17] D. Hug, G. Last, and M. Schulte. Second order properties and central limit theorems for geometric functionals of Boolean models. arXiv:1308.6519, 2013.

[18] M. Kahle. Random geometric complexes. *Discrete Comput. Geom.*, 45(3):553–573, 2011.

[19] M. Kahle. Topology of random simplicial complexes: a survey, 2013. arXiv:1301.7165.

[20] M. Kahle and E. Meckes. Limit theorems for Betti numbers of random simplicial complexes. *Hom., Hom. and Appl.*, 15(1):343–374, 2013.

[21] G. Last and M.D. Penrose. Poisson process Fock space representation, chaos expansion and covariance inequalities. *Probab. Th. Rel. Fields*, 150(3-4):663–690, 2011.

[22] G. Last, G. Peccatti and M. Schulte. Normal approximation on Poisson spaces: Mehler's formula, second order Poincaré inequality and stabilization, 2014. arXiv:1401.7568.

[23] N. Linial and R. Meshulam. Homological connectivity of random 2-complexes. *Combinatorica*, 26(4):475–487, 2006.

[24] R. Meester and R. Roy. *Continuum Percolation*. Cambridge University Press, Cambridge, 1996.

[25] J.R. Munkres. *Elements of Algebraic Topology*. Addison-Wesley, 1984.

[26] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008.

[27] P. Niyogi, S. Smale, and S. Weinberger. A topological view of unsupervised learning from noisy data. *SIAM J. Comput.*, 40(3):646–663, 2011.

[28] I. Nourdin and G. Peccati. *Normal Approximations with Malliavin Calculus*, volume 192 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2012. From Stein's method to universality.

[29] R. Pemantle and Y. Peres. Concentration of Lipschitz functionals of determinantal and other strong Rayleigh measures. *Combinatorics, Probability and Computing*, 23:140–160, 2013.

[30] M.D. Penrose. *Random Geometric Graphs*. Oxford University Press, New York, 2003.

[31] M.D. Penrose. Laws of large numbers in stochastic geometry with statistical applications. *Bernoulli*, 13(4):1124–1150, 2007.

[32] M.D. Penrose and J.E. Yukich. Central limit theorems for some graphs in computational geometry. *Ann. Appl. Probab.*, 11(4):1005–1041, 2001.

[33] A. Sarkar. Co-existence of the occupied and vacant phase in Boolean models in three or more dimensions. *Adv. in Appl. Probab.*, 29(4):878–889, 1997.

[34] R. Schneider and W. Weil. *Stochastic and Integral Geometry*. Probability and its Applications (New York). Springer-Verlag, Berlin, 2008.

[35] M. Schulte. *Malliavin-Stein Method in Stochastic Geometry*. PhD thesis, 2013. http://repositorium.uni-osnabrueck.de/handle/urn:nbn:de:gbv:700-2013031910717.

[36] E.H. Spanier. *Algebraic Topology*. McGraw-Hill Book Co., New York, 1966.

[37] J. M. Steele. An Efron-Stein inequality for nonsymmetric statistics. *Ann. Statist.*, 14(2):753–758, 1986.

[38] D. Stoyan, W. Kendall, and J. Mecke. *Stochastic Geometry and its Applications*. Wiley, Chichester, 1995.

[39] H. Tanemura. Critical behavior for a continuum percolation model. In *Probability Theory and Mathematical Statistics (Tokyo, 1995)*, pages 485–495. World Sci. Publ., River Edge, NJ, 1996.

[40] D. Yogeshwaran and R.J. Adler. On the topology of random complexes built over stationary point processes. arXiv:1211.0061, 2012.

[41] A. Zomorodian. Topological data analysis. In *Advances in Applied and Computational Topology*, volume 70 of *Proc. Sympos. Appl. Math.*, pages 1–39. Amer. Math. Soc., Providence, RI, 2012.

[42] A.J. Zomorodian. *Topology for Computing*, volume 16 of *Cambridge Monographs on Applied and Computational Mathematics*. Cambridge University Press, Cambridge, 2009.

[43] H. Carr, J. Snoeyink, and U. Axen. Computing contour trees in all dimensions. *ACM-SIAM Symp. of Discrete Algorithms*, pages 918–926, 2000.

[44] F. Chazal, A. Patel, and P. Skraba. Computing the robustness of roots. Manuscript, http://ailab.ijs.si/primoz_skraba/papers/fp.pdf, 2011.

[45] F. Chazal, A. Patel, and P. Skraba. Computing well diagrams for vector fields on R^n. *Applied Math. Letters*, 25(11):1725–1728, 2012.

[46] G. Chen, Q. Deng, A. Szymczak, R. Laramee, and E. Zhang. Morse set classification and hierarchical refinement using Conley index. *IEEE TVCG*, 18(5):767–782, 2012.

[47] G. Chen, K. Mischaikow, R. Laramee, P. Pilarczyk, and E. Zhang. Vector field editing and periodic orbit extraction using Morse decomposition. *IEEE TVCG*, 13(4):769–785, 2007.

[48] G. Chen, K. Mischaikow, R. Laramee, and E. Zhang. Efficient Morse decompositions of vector fields. *IEEE TVCG*, 14(4):848–862, 2008.

[49] W. de Leeuw and R. van Liere. Collapsing flow topology using area metrics. In *IEEE Vis*, pages 349–354, 1999.

[50] W. de Leeuw and R. van Liere. Visualization of global flow structures using multiple levels of topology. *Data Visualization*, pages 45–52, 1999.

[51] W. de Leeuw and R. van Liere. Multi-level topology for flow visualization. *Comp. & Graph.*, 24(3):325–331, 2000.

[52] T. Delmarcelle and L. Hesselink. The topology of symmetric, second-order tensor fields. *IEEE Vis*, pages 140–147, 1994.

[53] T. Dey and R. Wenger. Stability of critical points with interval persistence. *Disc. Comp. Geom.*, 38:479–512, 2007.

[54] H. Edelsbrunner, J. Harer, and A. Zomorodian. Hierarchical Morse-Smale complexes for piecewise linear 2-manifolds. *Discrete Comp. Geometry*, 30:87–107, 2003.

[55] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comp. Geometry*, 28:511–533, 2002.

[56] H. Edelsbrunner, D. Morozov, and A. Patel. The stability of the apparent contour of an orientable 2-manifold. In *Topo. M. in Data Anal. and Vis.*, pages 27–41. 2010.

[57] H. Edelsbrunner, D. Morozov, and A. Patel. Quantifying transversality by measuring the robustness of intersections. *Found. of Comp. Math.*, 11:345–361, 2011.

[58] A. G. Gyulassy, M. A. Duchaineau, V. Natarajan, V. Pascucci, E. M. Bringa, A. Higginbotham, and B. Hamann. Topologically clean distance fields. *IEEE TVCG*, 13(6):1432–1439, 2007.

[59] E. Hawkes, R. Sankaran, P. Pébay, and J. Chen. Direct numerical simulation of ignition front propagation in a constant volume with temperature inhomogeneities. *Combustion and Flame*, 145:145–159, 2006.

[60] J. Helman and L. Hesselink. Representation and display of vector field topology in fluid flow data sets. *IEEE Computer*, 22(8):27–36, 1989.

[61] T. Klein and T. Ertl. Scale-space tracking of critical points in 3D vector fields. *Topo. Meth. in Vis.*, pages 35–49, 2007.

[62] R. Laramee, H. Hauser, L. Zhao, and F. Post. Topology based flow visualization: the state of the art. In *Topo. Meth. in Vis.*, pages 1–19, 2007.

[63] M. Maltrud, F. Bryan, and S. Peacock. Boundary impulse response functions in a century-long eddying global ocean simulation. *Environmental Fluid Mechanics*, 10:275–295, 2010.

[64] K. Polthier and E. Preuß. Identifying vector fields singularities using a discrete hodge decomposition. *Vis. and Math. III*, pages 112–134, 2003.

[65] J. Reininghaus, J. Kasten, T. Weinkauf, and I. Hotz. Efficient computation of combinatorial feature flow fields. *IEEE TVCG*, 2011.

[66] J. Reininghaus, N. Kotava, D. Guenther, J. Kasten, H. Hagen, and I. Hotz. A scale space based persistence measure for critical points in 2d scalar fields. *IEEE TVCG*, 17(12):2045–2052, 2011.

[67] J. Reininghaus, C. Lowen, and I. Hotz. Fast combinatorial vector field topology. *IEEE TVCG*, 17(10):1433–1443, 2011.

[68] H. Theisel, C. Rössl, and H.-P. Seidel. Combining topological simplification and topology preserving compression for 2D vector fields. In *Pacific Graphics*, pages 419–423, 2003.

[69] H. Theisel, C. Rössl, and H.-P. Seidel. Compression of 2D vector fields under guaranteed topology preservation. *CGF*, 22(3):333–342, 2003.

[70] Y. Tong, S. Lombeyda, A. Hirani, and M. Desbrun. Discrete multiscale vector field decomposition. *ACM ToG*, 22:445–452, 2003.

[71] X. Tricoche, G. Scheuermann, and H. Hagen. A topology simplification method for 2D vector fields. In *IEEE Vis*, pages 359–366, 2000.

[72] X. Tricoche, G. Scheuermann, and H. Hagen. Continuous topology simplification of planar vector fields. *IEEE Vis*, pages 159–166, 2001.

[73] X. Tricoche, G. Scheuermann, and H. Hagen. Topology-based visualization of time-dependent 2D vector fields. *EuroVis*, pages 117–126, 2001.

[74] X. Tricoche, T. Wischgoll, G. Scheuermann, and H. Hagen. Topology tracking for the visualization of time-dependent two-dimensional flows. *Comp. & Graph.*, 26:249–257, 2002.

[75] B. Wang, P. Rosen, P. Skraba, H. Bhatia, and V. Pascucci. Visualizing robustness of critical points for 2D time-varying vector fields. *CGF*, 32(3):221–230, 2013.

[76] T. Weinkauf, H. Theisel, K. Shi, H.-C. Hege, and H.-P. Seidel. Extracting higher order critical points and topological simplification of 3d vector fields. In *Visualization, 2005. VIS 05. IEEE*, pages 559–566. IEEE, 2005.

[77] R. Westermann, C. Johnson, and T. Ertl. A level-set method for flow visualization. *IEEE Vis*, pages 147–154, 2000.

[78] E. Zhang, K. Mischaikow, and G. Turk. Vector field design on surfaces. *ACM ToG*, 25:1294–1326, 2006.

[79] A.Krizhevsk, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*, 2012.

[80] U. Bauer, M. Kerber, and J. Reininghaus. Distributed computation of persistent homology. In *ALENEX*, 2014.

[81] C. Berg, J.-P. Reus-Christensen, and P. Ressel. *Harmonic Analysis on Semi-Groups – Theory of Positive Definite and Related Functions*. Springer, 1984.

[82] M. Bronstein. Scale-invariant heat kernel signatures for non-rigid shape recognition. In *CVPR*, 2010.

[83] P. Bubenik. Statistical topological data analysis using persistence landscapes. arXiv, available at `http://arxiv.org/abs/1207.6437`, 2013.

[84] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc.*, 46:255–308, 2009.

[85] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2(3):1–27, 2011.

[86] F. Chazal, L. Guibas, S. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. In *SoSG*, 2011.

[87] C. Chen, D. Freedman, and C. Lampert. Enforcing topological constraints in random field image segmentation. In *CVPR*, 2013.

[88] M. Chung, P. Bubenik, and P. Kim. Persistence diagrams of cortical surface data. In *IPMI*, 2009.

[89] M. Chung, V. Singh, P. Kim, K. Dalton, and R. Davidson. Topological characterization of signal in brain images using min-max diagrams. In *MICCAI*, 2009.

[90] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007.

[91] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have L_p-stable persistence. *Found. Comput. Math.*, 10(2):127–139, 2010.

[92] H. Edelsbrunner and J. Harer. *Computational Topology. An Introduction.* AMS, 2010.

[93] M. Gao, C. Chen, S. Zhang, Z. Qian, D. Metaxas, and L. Axel. Segmenting the papillary muscles and the trabeculae from high resolution cardiac ct through restoration of topological handles. In *IPMI*, 2013.

[94] M. Gönen and E. Alpaydin. Multiple kernel learning algorithms. *J. Mach. Learn. Res.*, 12:2211–2268, 2011.

[95] Z. Guo, L. Zhang, and D. Zhang. A completed modeling of local binary pattern operator for texture classification. *IEEE Trans. Image Process.*, 19(6):16571663, 2010.

[96] R. J. j. Iorio and V. de Magalhães Iorio. *Fourier analysis and partial differential equations.* Cambridge Studies in Advanced Mathematics, 2001.

[97] C. Li, M. Ovsjanikov, and F. Chazal. Persistence-based structural recognition. In *CVPR*, 2014.

[98] L.-J. Li, H. Su, E. Xing, and L. Fei-Fei. Object Bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010.

[99] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[100] T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllonen, and S. Huovinen. OuTeX – new framework for empirical evaluation of texture analysis algorithms. In *ICPR*, 2002.

[101] D. Pachauri, C. Hinrichs, M. Chung, S. Johnson, and V. Singh. Topology-based kernels with application to inference problems in alzheimers disease. *IEEE Trans. Med. Imag.*, 30(10):1760–1770, 2011.

[102] D. Pickup, X. Sun, P. L. Rosin, R. R. Martin, Z. Cheng, Z. Lian, M. Aono, A. Ben Hamza, A. Bronstein, M. Bronstein, S. Bu, U. Castellani, S. Cheng, V. Garro, A. Giachetti, A. Godil, J. Han, H. Johan, L. Lai, B. Li, C. Li, H. Li, R. Litman, X. Liu, Z. Liu, Y. Lu, A. Tatsuma, and J. Ye. SHREC '14 track: Shape retrieval of non-rigid 3d human models. In *Proceedings of the 7th Eurographics workshop on 3D Object Retrieval*, EG 3DOR'14. Eurographics Association, 2014.

[103] B. Scholkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, USA, 2001.

[104] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser. The Princeton shape benchmark. In *Shape Modeling International*, 2004.

[105] P. Skraba, M. Ovsjanikov, F. Chazal, and L. Guibas. Persistence-based segmentation of deformable shapes. In *CVPR Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment*, 2010.

[106] J. Sun, M. Ovsjanikov, and L. Guibas. A concise and probably informative multi-scale signature based on heat diffusion. In *SGP*, 2009.

[107] H. Wagner, C. Chen, and E. Vuini. Efficient computation of persistent homology for cubical data. In *Topological Methods in Data Analysis and Visualization II*, Mathematics and Visualization, pages 91–106. Springer Berlin Heidelberg, 2012.