ESANN 2020 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Online event, 2-4 October 2020, i6doc.com publ., ISBN 978-2-87587-074-2.
Available from http://www.i6doc.com/en/.

ASAP - A Sub-sampling Approach for Preserving Topological Structures

Abolfazl Taghribi[1], Kerstin Bunte[1], Michele Mastropietro[3],
Sven De Rijcke[3], and Peter Tino[2]

1- University of Groningen, Faculty of Science and Engineering, The Netherlands
2- University of Birmingham, School of Computer Science, UK
3- Ghent University, Department of Physics & Astronomy, Belgium

Abstract. Topological data analysis tools enjoy increasing popularity in a wide range of applications. However, due to computational complexity, processing large number of samples of higher dimensionality quickly becomes infeasible. We propose a novel sub-sampling strategy inspired by Coulomb's law to decrease the number of data points in d-dimensional point clouds while preserving its Homology. The method is not only capable of reducing the memory and computation time needed for the construction of different types of simplicial complexes but also preserves the size of the voids in d-dimensions, which is crucial e.g. for astronomical applications. We demonstrate and compare the strategy in several synthetic scenarios and an astronomical particle simulation of a dwarf galaxy for the detection of superbubbles (supernova signatures).

1 Introduction

Topological data analysis (TDA) provides exploration tools for increasingly diverse applications in various domains, ranging from biology, networking, natural images, geometry and material science [1]. Persistent homology (PH) is a TDA technique for computing the properties of shapes of a finite metric space (also called point-cloud data set) and can capture these features in an extended range of scales. Nonetheless, as the number of points or the dimensions of a dataset increases, the computation of the PH soon becomes impractical. Numerous methods and toolboxes provide novel approaches to tackle the problem of computational costs. Sparse Rips filtration [2], which builds an ϵ-net on top of the point set then associates weights to each node, is a provably good approximation of the full data Rips filtration. In [3] two new atomic operations for efficient computation of PH are suggested and SimBa [4] combines these two strategies to reach a higher sparsity increasing the efficiency for computation of Rips filtration. The toolbox Ripser [5] decreases the computational costs by avoiding to build the complete coboundary matrix, building and storing only the parts needed, which improves the memory consumption leading to a decline in computational time. These methods are limited to Rips and are not extendible to other types of filtration. A general concept for scaling down the computation was reported in [6] proposing to sub-sample the data randomly repeatedly and construct an average landscape for the point cloud. Although their approach can be applied for constructing all types of filtration, it is sensitive to the distribution of the data on the structures as a consequence of random sampling.

Physical particle simulations are one way of investigating astronomical phenomena such as galaxies and supernovas. Radiation and winds from massive stars at the end of their life can greatly affect the dynamics of gas in the interstellar medium (ISM) and in turn, change the structure of the galaxy and its ability to create new stars. Dwarf galaxies are very sensitive to the physical processes determining their evolution due to their low mass and are therefore used as probes to characterize, study and isolate them in simulations. Similar to real dwarfs simulated irregular galaxies have a very clumpy ISM and holes due to supernovae as visible in the gas density distribution [7, 8]. The characterization of the distribution of supernova shells in the ISM (so-called superbubbles), and the energies of the expanding shells [9, 10], can shed light on the feedback physical processes. Superbubbles are of great astronomical interest but typically measured by eye in available catalogues and automatic quantitative tools are highly desirable. In this contribution we propose ASAP[1], a sub-sampling approach for preserving topological structure, that reduces the computational cost suitable for different types of PH filtration, general d-dimensional point clouds and large number of samples. In the following the strategy is explained in detail, furthermore compared in several controlled experiments and finally used to investigate a snapshot of a particle simulation by computing the number and size of superbubbles within a jellyfish-like dwarf galaxy.

2 Method

Computing the persistent homology for the analysis of the evolution of shapes across different resolutions is often prohibitive due to the combinatorial nature of existing algorithms complexity, in both time and space. Therefore, we propose a two-stage strategy based on subsampling and Coulomb's law[11]. As described before, we first subsample points from the point-cloud data set P (finite metric space) to reduce the amount of computation time and memory. The subset $N_r \subset P$ aims to contain fewer points $s \in N_r$ for which the persistent diagram $D(N_r)$ approximates the persistent diagram of the full data $D(P)$. Therefore the set N_r has to satisfy the following two conditions [2] checked in every step:

(1) covering $\quad d(\boldsymbol{p}, \boldsymbol{s}) \leq r \quad \forall \boldsymbol{p} \in P, \exists s \in N_r$ and

(2) packing $\quad d(\boldsymbol{s}_i, \boldsymbol{s}_j) > r \quad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in N_r$ with $i \neq j$.

We satisfy (1) by selecting a random point $\boldsymbol{s}_i$, insert it to N_r and remove all points $\{\boldsymbol{p}_j\}$ from P belonging to an open ball centered around $\boldsymbol{s}_i$ with radius r:

$$B(\boldsymbol{s}_i, r) = \{\boldsymbol{p}_j \in P : d(\boldsymbol{s}_i, \boldsymbol{p}_j) \leq r\} \Rightarrow P = P \backslash \{\{\boldsymbol{p}_j\}, \boldsymbol{s}_i\} \text{ and } N_r = N_r \cap \boldsymbol{s}_i. \quad (1)$$

This process is repeated until the point set P is depleted implicating that all points are covered by at least one open ball of a sample point in N_r. Due to the removal of points in every step also the packing condition is fulfilled for all remaining points in P. Their distance must be larger than r from $\boldsymbol{s}_i$.

The sub-sampling strategy fulfils both necessary conditions, but the result is not completely uniform and the pairwise distance of any sample point pair is between r and $2r$. However, it is more desirable to have sample points equidistant from each other forming a uniform grid. As a result, we expect when all points on its boundary connect to each other it coincides with the birth time of the void. Moreover, in astronomical applications it is crucial to measure the size of the cycles, cavities and streams as accurately as possible, for which N_r needs to contain the borders of the data. Therefore we propose an extension to the sampling inspired by the movement of identical electrical particles, such as electrons, on the surface of a conductive sphere. The electrons will repel each other based on Coulomb's law and form a uniform distribution. To take advantage of this physical repulsion force each sample is repelled by neighbouring samples by

$$\boldsymbol{m}_i = \mathrm{disp}(\boldsymbol{s}_i) = \sum_{\boldsymbol{s}_j \in \mathcal{N}_i} \frac{\boldsymbol{s}_j - \boldsymbol{s}_i}{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|} \cdot \frac{\gamma}{\|\boldsymbol{s}_j - \boldsymbol{s}_i\|^2} \quad , \quad (2)$$

where the set $\mathcal{N}_i$ consists of sample points in $2r$ radius of $\boldsymbol{s}_i$ and γ denotes the learning rate. If neighbouring points are far from $\boldsymbol{s}_i$ the force will be low, and the learning rate controls the strength of the movement. The appropriate range for the displacement is between $(0.1r, r)$, since the effect of smaller movements is negligible and larger movements result in $\boldsymbol{s}_i$ intruding positions already covered

[1] The code and the synthetic datasets are available in https://github.com/abst0603/ASAP

Algorithm 1: ASAP a sub-sampling approach preserving topological structures

```
Input : data $P$, radius $r$, learning rate constant $\tau$        Output: $N_r$
Initialize: $P_{\text{tmp}} = P$, $N_r = \emptyset$, $\gamma = 1$, and $t = 1$
while ($P_{tmp} \neq \emptyset$)
    Select a random point $\boldsymbol{s}_i$ from $P_{\text{tmp}}$
    $N_r = N_r \cap \boldsymbol{s}_i$ and remove points from $P_{\text{tmp}}$ following Eq. (1)
while ($\gamma > 0.1r^3$)                                   /* repulsion forces */
    Calculate $\gamma$ based on Eq. (3)
    forall ($\boldsymbol{s}_i \in N_r$)
        Compute $\boldsymbol{m}_i$ Eq. (2) and $\hat{\boldsymbol{s}_i}$ using Eq. (4)
        if ($d(\hat{\boldsymbol{s}}_i, \boldsymbol{s}_j) > r \forall \boldsymbol{s}_j \in N_r \ AND \ \boldsymbol{s}_j \neq \boldsymbol{s}_i$)
            $\boldsymbol{s}_i = \hat{\boldsymbol{s}}_i$
    $P_{\text{tmp}} = P$
    forall ($\boldsymbol{s}_i \in N_r$)                     /* fulfil covering condition */
        Remove all points which belongs to $B(\boldsymbol{s}_i, r)$ from $P_{\text{tmp}}$
    while ($P_{tmp} \neq \emptyset$)
        Select a random point $\boldsymbol{s}_i$ from $P_{\text{tmp}}$
        $N_r = N_r \cap \boldsymbol{s}_i$ and remove points from $P_{\text{tmp}}$ following Eq. (1)
    $t++$
```

by other samples. The learning rate is gradually reduced in every step t following

$$\gamma = r^3 \exp(-t/\tau) \tag{3}$$

such that the samples converge to the new positions. τ is a constant which determines the decay rate of the learning rate. Instead of moving the samples itself we take the closest point in the original set $\hat{\boldsymbol{s}}_i \in P$ to the displacement position as substitute for $\boldsymbol{s}_i$

$$\hat{\boldsymbol{s}}_i = \arg\min_{\boldsymbol{p}_j}(d(\boldsymbol{p}_j, \boldsymbol{s}_i + \boldsymbol{m}_i)) \quad \forall \boldsymbol{p}_j \in P \tag{4}$$

if it is not contained in an open ball of any other sample point. Algorithm 1 details the complete procedure of the extended sampling strategy and Fig. 1 shows the result on a simple two-dimensional example. Panel (a) depicts the point cloud P consisting of a line and a square with a circular hole in the centre and the open ball cover after the random sampling in panel (b). The balls of N_r after the update using the repulsion force is illustrated in (c) resulting in a more uniform grid that covers all boundaries as desired.

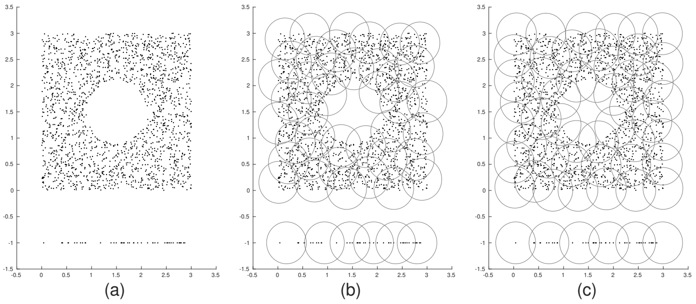

Fig. 1: Panel (a) points P distributed on a line and a square with a hole in the centre, (b) ball cover after random sub-sampling and (c) after repulsive selection.

3 Experiments

In this section, we address the comparison between ASAP as a preprocessing step for several types of point cloud filtration and state-of-the-art TDA methods. To

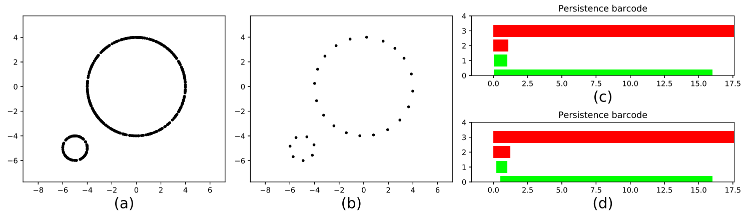

Fig. 2: 2 circles: (a) point set, (b) samples after applying ASAP, (c) persistence barcode of the point set and (d) the sub-sample.

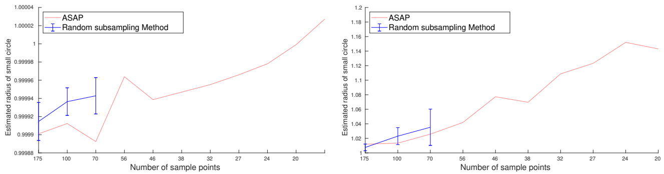

Fig. 3: The variation of the radius of smaller circle of the 2 circles data set when reducing the number of samples for Alpha (left) and Rips filtration (right).

compute the PH of point sets we mainly use GUDHI [12], which is faster and more memory efficient due to the structure of the simplex tree than many other toolboxes [1]. In order to obtain the Rips filtration on datasets with larger number of points, we build the simplicial complex using Ripser [5]. We first discuss controlled experiments with known ground truth, followed by the results of ASAP on real-world data from an astronomical galaxy particle simulation.

Synthetic data with ground truth. We first experiment on a simple two-dimensional dataset which was introduced in [6]. 500 points are distributed uniformly on two circles with radius 1 and 4. Fig. 2 shows the barcodes of the complete point cloud and samples which are selected based on ASAP, respectively. Each bar illustrates the birth-death interval of a topological feature of the point cloud. The barcodes for features of homology groups H0, H1, and H2 are presented in red, green, and blue, respectively. Barcodes were denoised by removing bars with minimum persistence smaller than threshold 0.5. The proposed method reduces the original point set to only 27 points, which not only depict a similar persistence barcode and preserves the death-times for all bars, but the death time of the bars for betti number 1 shows the correct value for the radii of both circles. We also compared both sub-sampling methods reducing the number of points consecutively while observing the resulting radius estimate of the small circle (see Fig. 3). Note that the circle with radius 4 is robust for both sub-sampling method. We repeat the random sub-sampling 10 times as suggested in [6], and the number of points in each sub-sample of ASAP corresponds to hyperparameter r ranging from $[0.1, 1]$. For both alpha filtration (panel (a)) and Rips filtration (panel (b)) ASAP preserves the radius of the smaller circle with fewer number of samples.

To compare the methods in higher dimensions we spread points nonuniformly and unevenly on two hyper-spheres in $\mathbb{R}^5$ with radius 1 and 2 (referred to as 2Spheres dataset). Even though the data consists only of 1200 points the computation of Rips filtration is very memory consuming for this dataset. Note that the code of Simba [4] provided by the authors only returns the betti numbers up to 3 dimensions. We sub-sample the point cloud based on ASAP and [6] respectively and construct the alpha complex on both resulting sub-sets. As proposed in [6] we iterate the sampling procedure for 10 times. We show the mean value in the barcode plot and results of both methods in Fig. 4. Since the data is not

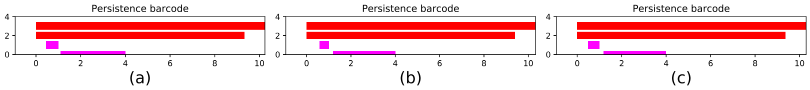

Fig. 4: Persistence barcode of 2Spheres data: panel (a) spheres in $\mathbb{R}^5$, (b) samples extracted by ASAP and (c) samples extracted by [6].

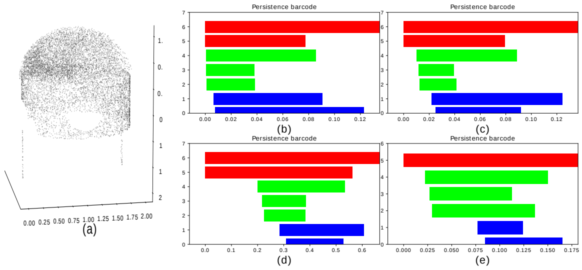

Fig. 5: Synthetic dwarf galaxy(a). Barcode for: Alpha filtration of the point set(b) and 558 sub-samples(c); Rips filtration 558 sub-samples(d); and SimBa(e).

uniform, random sub-sampling as proposed by [6] cannot preserve its homology. If we reduce the number of samples to less than 1000 points the PH changes significantly in several iterations. ASAP preserves not only the homology of the data in $\mathbb{R}^5$ with only 655 sub-sampled points, but also the death to birth times of the barcodes for betti number 4, corresponding to the radii of the spheres.

Lastly, we create a more realistic synthetic example representative for our astronomical application. Namely 9656 non-uniformly distributed points in $\mathbb{R}^3$ forming a synthetic dwarf galaxy containing 2 cavities, 3 cycles in a half spherical head, a connected and a separated stream (see panel (a) in Fig. 5). Unfortunately, GUDHI and Ripser for the computation of the Rips filtration fail due to high memory usage. The persistence barcode for alpha filtration is shown in panel (b). SimBa can compute the Rips filtration, as depicted in panel (e) Fig. 5, but looses one bar related to betti number 0 and the death times do not conform with the true size of the cycles and cavities within the head. We also evaluate [6] sub-sampling a random selection of half of the points repeated 10 times. Since the data is not uniformly distributed on the structure this strategy alters the homology. Sub-sampling with ASAP $r = 0.15$, on the other hand, leaves only 558 points (5.7% of the original set) while preserving the homology and the radii of cycles and cavities almost perfectly as illustrated in the corresponding persistence barcode using Alpha filtration (Fig. 5(c)). Ripser can compute the Rips filtration on the smaller subsets acquired by ASAP $r \geq 0.1$ (panel (d)).

Table 1 presents the total number of simplices arising in every filtration on all synthetic datasets investigated. SimBa can only compute the Rips filtration and although Ripser compute the Rips filtration on the synthetic dwarf dataset it does not provide any information about the size of the simplicial complex inside the structure indicated by '-' in the respective columns. We indicate with '∞' whenever the computation of the Rips filtration fails due to the memory complexity. For our proposed method and random subsampling by [6] (abbreviated by RSM), we report the results for the number of samples preserving the homology of the data after denoising.

Jellyfish-like dwarf galaxy particle simulation data. Fig. 6 panel (a) shows the point set corresponding to the position of 33500 gas particles at a specific point in time acquired from a real astronomical particle simulation of a dwarf galaxy. The distribution of points in this point cloud varies significantly and the points are dispersed on multiple separated parts. Hence we expect to see several bars linked with betti number 0 for this dataset. We sub-sample the

Table 1: Comparison of the number of simplices constructed by several methods.

		ASAP		**RSM**[6]		**SimBa**	**GUDHI**	
dataset	(n, d)	Alpha	RIPS	Alpha	RIPS	RIPS	Alpha	RIPS
2circles	(500,2)	81	1 640	309	37 876	1 031	2 345	13 752 927
2Spheres	(1 200,5)	349 541	∞	584 657	∞	-	718 531	∞
s.dwarf	(9 659,3)	13 801	-	250 991	-	63 004	250 991	∞

point set using ASAP with $r = 0.7$ reducing the set to $\approx 10\%$ of the total. Then the alpha simplicial complex was constructed on subset. Fig. 6 panel (b) shows the persistence barcode for the reduced set, denoised using a threshold of 0.6 for the minimum length of each bar. The data consists of 8 distinguished parts (red bars) and 4 cavities (blue bars) inside, which denote the size of each superbubble (5.7,3.98,1.66,1.48) inside the simulated jellyfish-like dwarf galaxy.

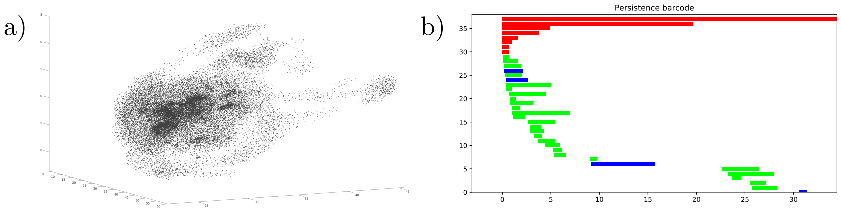

Fig. 6: Dwarf galaxy (a) point set and (b) Alpha filtration of ASAP subset.

4 Conclusion

We introduce the novel method ASAP for sub-sampling a point cloud that preserves the topological properties and reduces the memory consumption and computational cost for TDA analysis. The formulation is expandable for d-dimensions, is not limited to a specific type of filtration and its performance is shown for a variety of data sets. Since the approach preserves the size of topological features it is useful for domains where the accuracy of such information is indispensable, such as astronomy where it informs about physical processes.

Acknowledgement. We thank support from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 721 463.

References

[1] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1):1–38, December 2017.

[2] D. R. Sheehy. Linear-Size Approximations to the Vietoris–Rips Filtration. *Discrete & Computational Geometry*, 49(4):778–796, June 2013.

[3] T. K. Dey, F. Fan, and Y. Wang. Computing Topological Persistence for Simplicial Maps. In *Symposium on Computational Geometry (SoCG)*, March 2014.

[4] T. K. Dey, D. Shi, and Y. Wang. SimBa: An Efficient Tool for Approximating Rips-filtration Persistence via Simplicial Batch Collapse. *JEA*, 24(1):1.5:1–1.5:16, 2019.

[5] U. Bauer. Ripser: efficient computation of Vietoris-Rips persistence barcodes. *arXiv:1908.02518 [cs, math]*, August 2019. arXiv: 1908.02518.

[6] F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Subsampling Methods for Persistent Homology. In F. Bach and D. Blei, editors, *Proc. of the 32nd ICML*, pages 2143–2151, Lille, France, July 2015. PMLR.

[7] H.-X. Zhang, D. A. Hunter, and B. G. Elmegreen. Hi power spectra and the turbulent interstellar medium of dwarf irregular galaxies. *ApJ*, 754(1):29, Jul 2012.

[8] R. Verbeke, E. Papastergis, A. A. Ponomareva, S. Rathi, and S. De Rijcke. A new astrophysical solution to the too big to fail problem. *A & A*, 607:A13, 2017.

[9] M. S. Oey and C. J. Clarke. *The Size Distribution of Superbubbles in the Interstellar Medium*, pages 112–119. Cambridge Contemporary Astrophysics. CUP, 1999.

[10] S. Stanimirovic. Shells in the magellanic system. *Proc. of the Int. Astronomical Union*, 2(S237):84–90, Aug 2006.

[11] David Halliday, Robert Resnick, and Jearl Walker. *Fundamentals of Physics*. Wiley, 10 edition edition, August 2013.

[12] GUDHI C++ library,url: gudhi.gforge.inria.fr.