TDA - filament network classification

November 2019

Abstract

The actin cytoskeleton plays a critical role in plant cells. The filamentous structure of actin proteins can be viewed as a network endowed with a topology. We propose a novel, automated classifier, combining topological data analysis (TDA) with a machine learning framework in order to investigate and leverage the topology of actin networks. Our classifier is non-distance-based, instead using a persistence vectorization. We attain additional power, at a relatively low computational cost, through resampling. We benchmark our classifier against several distance and non-distance based classifiers, using a synthetic dataset to measure accuracy and sensitivity. We succeed in classifying the simulated networks with very high accuracy. Finally, we demonstrate an application with real data from confocal microscopy, classifying myosin-mutant and wildtype *Arabidopsis* root cells.

1 Introduction

The actin cytoskeleton is a complex network of proteins that is present in all eukaryotic cells. In addition to its function as cellular scaffolding, the actin cytoskeleton enables several basic cellular functions including the control of cellular shape and direction of movement [14]. These basic functions are critical to many higher order physiological processes such as cell division, expansion, mobility and motility[5].

Actin filament organization is thought to be largely governed by the interaction of the filaments themselves and by myosin motor proteins. Actin filaments are polar structures, ploymerized by globular actin proteins. Many actin-binding proteins have potential to bind to actin filaments at various sites along the filament. These binding proteins allow actin filaments to spatiotemporally assemble and disassemble. The binding proteins give rise to a dense cross-linking where filaments develop into networks consisting of many filaments and very many binding sites. To understand certain behaviors of cells, it is of tremendous importance to understand the processes that govern actin filament network organization. One key driver of these dynamics may be the relationship between actin-binding proteins, individual filaments and emergent networks.

Our goal in this work is to develop a framework for the classification of actin newtworks. The images of actin networks that are examined in this work are captured via confocal mcroscopy. These images are very high resolution (**TODO**: add typical resolution), but still suffer from several types of noise such as: filaments moving through the focal plain, rounding of the cell at the edges, neighboring cells poluting the image, changes in microscopic conditions/settings, and several more. Therefore, the confocal microsocopy data has the advantage of providing many high quality data at the expense of also including many noisy data. In order to automatically study these images, without myriad interjections by researchers (and thus eliminating potential for an introduction of unforseen bias), an automated tool is called for which is highly robust to these types of noise. Consequently, this work seeks to develop a classifier, robust to the noise of confocal microscopic images of actin networks, and requiring minimal input from researchers.

In the fast developing field of machine learning, topological data analysis (TDA) has become increasingly popular as a tool for noisy network and signal classification. To date, researchers have used TDA to solve many real-world problems including signal identification [9], materials classification [7, 10], shape recognition

[2, 8], histologic image analysis [1, 11, 12], ecology of human mobility [3, 4], and cosmology [13, 15]. A review of TDA and its applications is provided in [16]. A sub-method of TDA, persistence homology, is a popular method used to measure differences in topological features, due to its robustness in the face of purturbation of data. Persistence homology records when homological features (connections and voids) appear and vanish in data. These patterns vary between data. All of the appearances and disappearances of homological features are summarised in persistence barcodes and/or diagrams. In this work, we encode the geometric features of filaments networks into persistence diagrams and show a method of classification on the vectorization of the persistence space (a persistence space is not itself a vector space, so it has no mean for instance **TODO:cite**). We compare this approach to traditional, distance-based classififfications which attempt to summarize the similarities of the actin network topologies in the persistence space.

We are aided in our investigation of classification methods by a high quality dataset of simulated actin networks, which we use to benchmark our candidate methods. We are provided the outputs of simulations which combine theoretical physical properties with experimental stochastic simulations in order to emulate actin network dyanmics. These simulations allow the researcher to control the instantiating factors which will drive the emergent structure of the networks. Varying these initial conditions then enables us to compare the conditional difference in outcomes of the simulated networks. This experimental strategy can provide an opportunity to independently examine the role each factor plays in the process. These factors could include the cross-linker density(number of cross-linkers per certain area), cross-linker stiffness, maximum angle that can exist between two filament segments to be crosslinked, and so on[5, 6]. This control mechanism also allows us to test our methods on a highly controlled and clean dataset, in order to test sensitivity and to compare between methods.

After testing several methods of actin network classification on the simulated data, we choose the top performing method and addapt it to the microscopy images. We perform a classification between myosin-mutant and wildtype *Arabidopsis* root cells.

Move to discussion?: **In this work, we propose a machine learning approach to classify filament networks generated with varied cross-linker density. Our method leverages the topology of the actin networks through Topological Data Analysis(TDA). Our exploratory work is the first time filament networks have been studied by direct topological classification. This work could serve as a pilot for future research in actin cytoskeleton organization. In the future, this work should be useful in the course of research on cytoplasmic streaming to be able to classify real cells based on images of their actin networks. This could provide biologists a method of disentangling the interaction of myosin motor proteins, the actin network, and streaming, i.e. by imaging the actin structure and clustering cells based on their actin network topology, the researcher may be able to fix a network structure while varying parameters specific to myosin.**

The structure of this work is as follows: In section 2, we describe the data and introduce the background of persistence homology. Section 3 demonstrates two algorithms for classifying filament networks. Section 4 exhibits the numerical results. Section 5 will give conclusions and a discussion of future directions.

2 Persistence Homology and Filament Networks

To quantify the differences in filament networks, we need to tranform our data in a manner that reveals its hidden geometric features. We perform this transform with simplicial complexes in a manner typical of persistence homology. We use the 2-dimensional coordinates of sampled points along the filaments as initial nodes. Simplcial complexes provide a bridge between the data space and a topological space in which computation of distances between sets of data points can be realized. A simplicial complex is a finite collection of simplices of different dimensions such that faces of simplices are also simplices, and intersections of the simplices are either empty or a face of both [**?**]. In particular, higher dimensional simplices are

constructed from lower dimensional simplices. Vertices are 0-dim simplices. A 1-dim simplex is called an edge and is created by its two vertices as faces (note that a higher dimensional edge is constructed from lower dimensional points). A 2-dim simplex or a triangle has three edges as faces. Further more, a 3-dim simplex or a tetrahedron has four triangles as faces, another nesting of several lower dimensional features to build one of higher dimension.

2.1 Data

We are provided data in two spaces. The focus of this work is to develop a method for the classification of actin networks in confocal microscopic imagery. This data comes in the form of one 2-dimensional image per cell. Ou benchmarking data come from simulations, of which the output is simply a point-cloud per cell.

Microscopy data: The microscopy data were provided in the form of one grayscale image per cell. The actin filaments fluoresce in the images and so the intensity of each pixel of an image can be thought of as indicating the likely presense of a filament in that region of the cell. In order to study the homology of an actin network, we must perform a topological tranformation on the data. Since the images contain hundreds of thousands of pixels, we chose to sample from the images a set of points, where the probability of choosing a pixel is propotionate to the pixel's intensity. We make a choice of a number of points that we think is likely to sufficiently summarize a network. We then can perform our topological transformation on these new point clouds.

TODO: Add WT and mutant cell images here.

Simulated data: Our synthetic data come from simulations with varried numbers of crosslinking protiens. As discussed, actin filaments are thought to be organized by cross-linking on actin-binding proteins. Filaments and inter-filament structure can then be simulated by a physical model [5, 6]. The change of initial conditions in a eukaryotic cell will cause variation in later measurement of filament networks. Our network data is simulated by three different cross-linker densities. Higher cross-linker density means more opportunities for filaments to be cross-linked, i.e. the binding and unbinding processes can be more active. As shown in Figure 1(a), three kinds of filaments networks were simulated with different numbers of cross-linkers: 825, 1650 and 3300. All simulated cells were bound by a $20\,\mu m \times 20\,\mu m$ square. Therefore, the cross-linker density of each network is 2.06, 4.13 and 8.25 per μm^2, respectively. In each network, there are a total of 100 filaments with average length $10\,\mu m$, where filaments are modeled as polar worm-like chains in red and blue dots represent barbed ends of these filaments. We also record the locations of the actin beads that make up the filaments, which are shown as small black circles in Fig. 1(b). Each actin bead is of radius $0.5\,\mu m$. The actin beads of the simulated networks will act as our point clouds in the topological transormations of these synthetic data.

2.2 Persistence Homology

In order to build simplicial complexes, we adopt the procedure of forming Vietoris-Rips complexes on each dataset (actin network) by introducing a sequence of ϵ-balls with increasing radius ϵ centered at each data point (a sampled pixel for image data or an actin bead in the synthetic data). Simplicial complexes are constructed based on intersections of these ϵ-balls and each value of ϵ corresponds to an unordered group of homological features, which is called a homology group. Considering values of ϵ as a timeline, we only record when a homological feature appears and disappears. These indexes are called the birth times and death times of a particular homological feature. Moreover, the lifespan (death minus birth) of a homological feature is referred to as the feature's persistence. A set of homological features gives rise to a set of persistence measures. At the end of this procedure, when radius ϵ is sufficiently large so that the homology group remains unchanged by any further increase to the radius, information of a filament network's persistence homology (the set of persistpent homology measurements) is summarized in a persistence diagram.

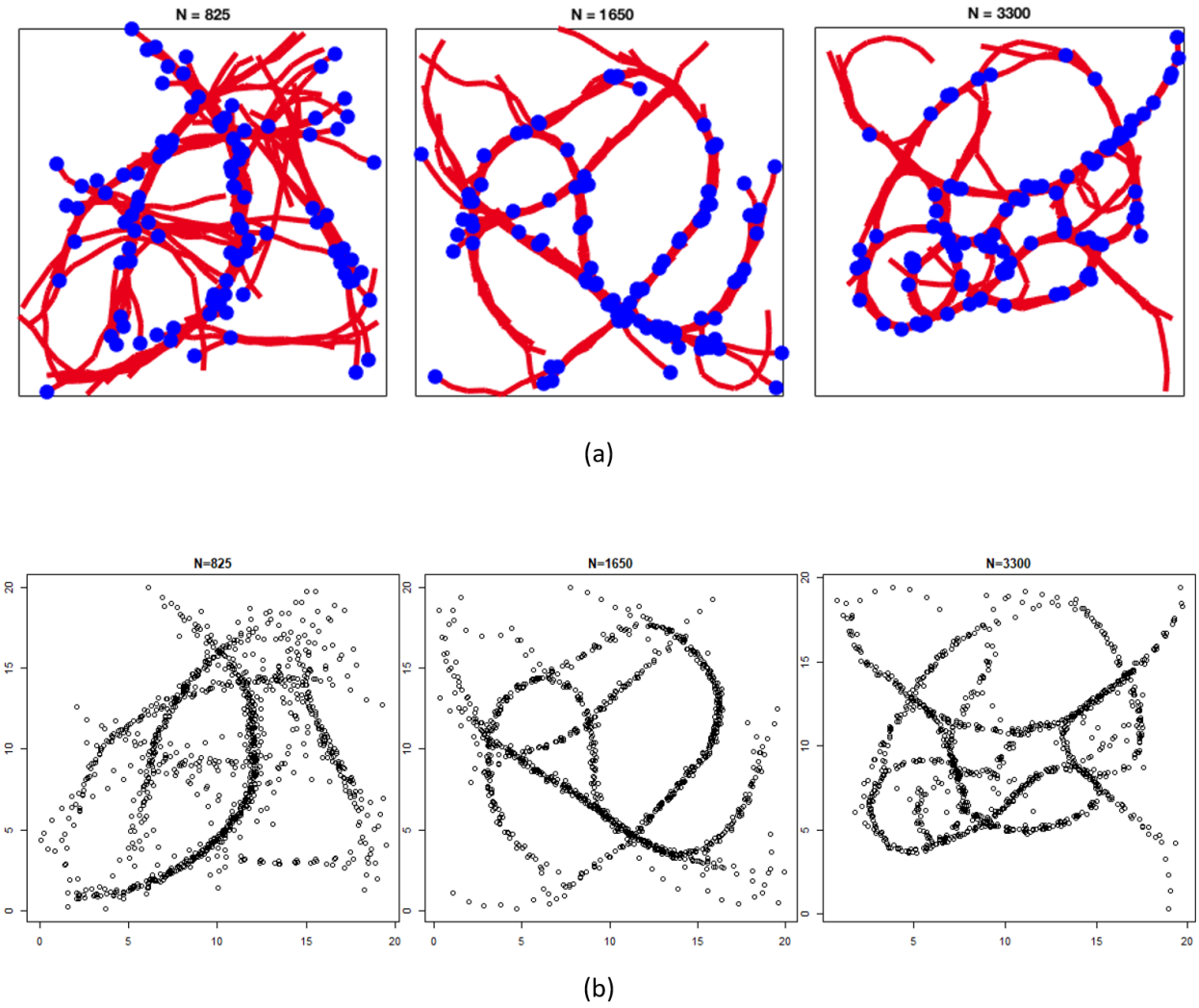

Figure 1: Filament networks. Panel (a) shows three filament networks generated by 825, 1650 and 3300 cross-linkers, respectively, in a 20 µm × 20 µm area. Each network contains 100 filaments which are represented as red lines. The blue dots are the barbed ends of these filaments. Panel (b) shows the locations of the actin beads that make up the filaments exhibited in Panel (a).

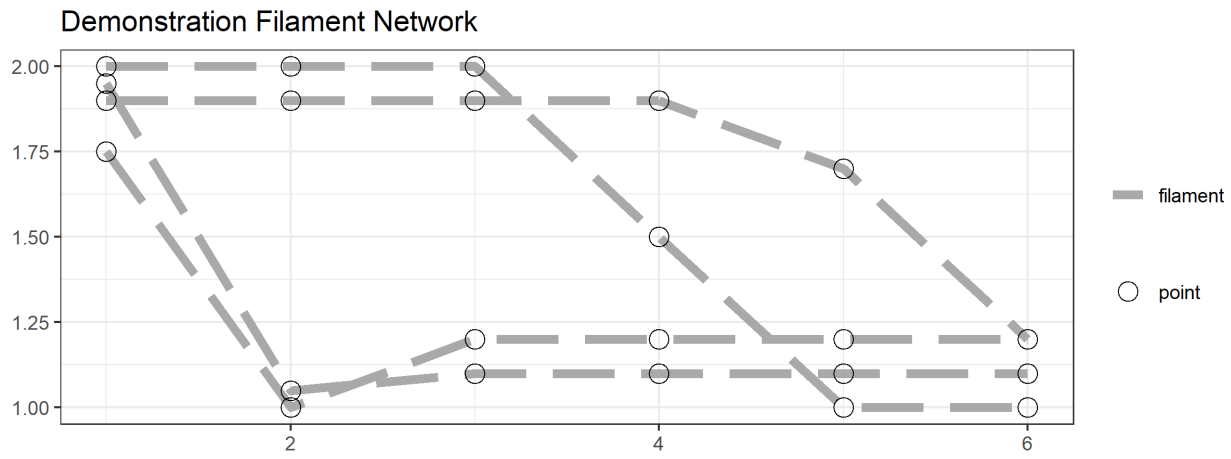

Figure 2: Demonstration fillament network. This network contains 3 filaments. Points are sampled along the filaments, in order to produce a point cloud from which persistence homology can be studied.

For clarity, we demonstrate the formation of the persistence on a filament network. We use a simplified filament network which is shown in detail in Figure 2. Figure 3 depicts the process of discovering and summarizing the persistence homology of the simple network shown in Figure 2. These illustrations should make clear the connection between filament networks, the Vietoris-Rips Complex and the encoding of the persistence information.

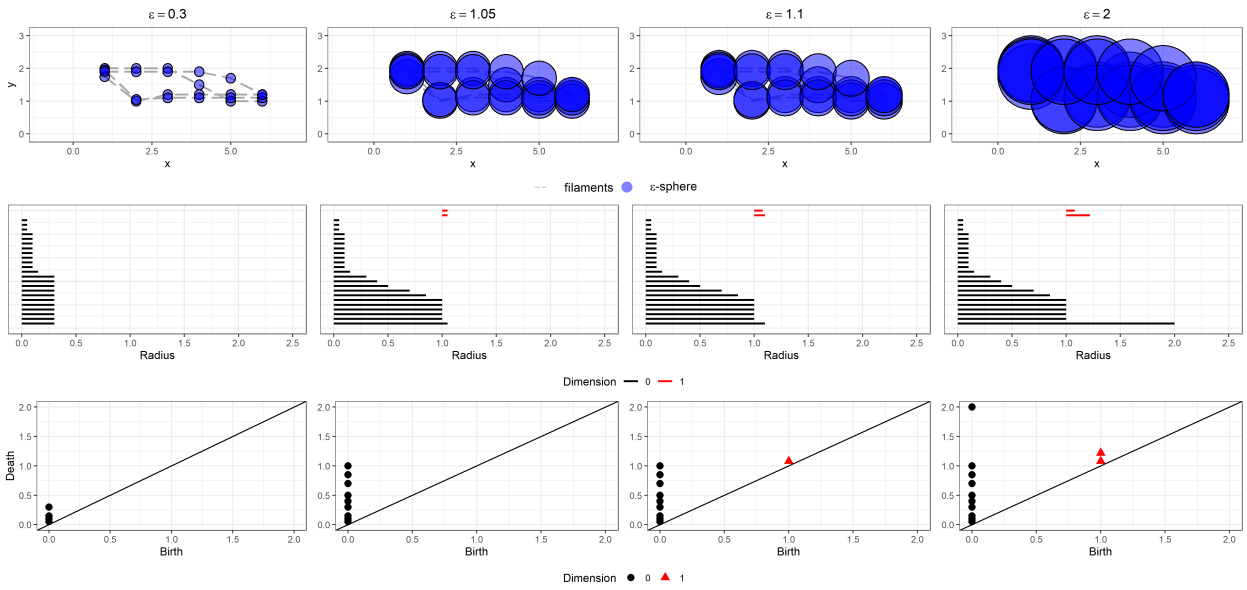

Figure 3: Investigation of the persistence homology of the demonstration filament network in Figure 2. The first row of figures, showes the growing ϵ-spheres about the sampled points of the fillaments. The second row shows the corresponding persistence barcode. The third row shows the corresponding persistence diagram. The columns progress with the algorithm, right-to-left.

When $\epsilon = 0$, the sampled points of the filament network are each their own connected component. As the ϵ-spheres grow, connecetd components begin to merge. In the first column of te figure, when $\epsilon = 0.3$, several

of the sampled points have already connected. We denote the points which have died (having merged into a larger connected component), by ending their bar in the presistence barcode below, and plotting a point at 0 on the x-axis and at their precise time of death on the y-axis in the persistence diagram below that. When $\epsilon \approx 1$, two holes form. The holes are evident in the top row, second column, of the figure, where one can see one larger and one smaller hole. We begin plotting bars for the holes in the corresponding persistence barcode below. Note that there are not yet reccords of the holes in the persistence diagram, because we require the death time of the holes in order to plot them in the 2-dimensional diagram.

In the third column of Figure 3, when $\epsilon = 1.1$, the smaller hole has closed, and become part of a larger connected component. Now a single point is plotted in the persistence diagram below and the bar is terminated.

As the algorithm progresses, the larger hole eventually dies and its corresponding bar in the persistence barcode is terminated. A point is added at the corresponding birth and death in the persistence diagram (a seconf red triangle appears in the final column). The algorithm could continue to $\epsilon = \inf$, but it is evident in this example that no more homologic information will be discovered as the spheres continue to grow. We choose to arbitrarily terminate the algorithm at $\epsilon = 2$, terminate the final bar in the final barcode, and plot at (0,2) in the persistence diagram to denote final death time.

Overall, persistence homology indirectly summarizes the hidden shape of the data and transcribes this shape to the persistence diagram. With the persistent homology of each point cloud, a classifier can be generated either from the distance [**?**] between persistence diagrams or by alternative vectorizations of the diagrams [**?**, **?**, **?**].

References

[1] Francisco Belchi, Mariam Pirashvili, Joy Conway, Michael Bennett, Ratko Djukanovic, and Jacek Brodzki. Lung topology characteristics in patients with chronic obstructive pulmonary disease. *Scientific reports*, 8(1):5341, 2018.

[2] Thomas Bonis, Maks Ovsjanikov, Steve Oudot, and Frédéric Chazal. Persistence-based pooling for shape pose recognition. In *International Workshop on Computational Topology in Image Context*, pages 19–29. Springer, 2016.

[3] Yen-Chi Chen and Adrian Dobra. Measuring human activity spaces from gps data with density ranking and summary curves. *arXiv preprint arXiv:1708.05017*, 2017.

[4] Yen-Chi Chen et al. Generalized cluster trees and singular measures. *The Annals of Statistics*, 47(4):2174–2203, 2019.

[5] Simon L Freedman, Shiladitya Banerjee, Glen M Hocky, and Aaron R Dinner. A versatile framework for simulating the dynamic mechanical structure of cytoskeletal networks. *Biophysical journal*, 113(2):448–460, 2017.

[6] Simon L Freedman, Glen M Hocky, Shiladitya Banerjee, and Aaron R Dinner. Nonequilibrium phase diagrams for actomyosin networks. *Soft matter*, 14(37):7740–7747, 2018.

[7] Yasuaki Hiraoka, Takenobu Nakamura, Akihiko Hirata, Emerson G Escolar, Kaname Matsue, and Yasumasa Nishiura. Hierarchical structures of amorphous solids characterized by persistent homology. *Proceedings of the National Academy of Sciences*, 113(26):7035–7040, 2016.

[8] Chunyuan Li, Maks Ovsjanikov, and Frederic Chazal. Persistence-based structural recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1995–2002, 2014.

[9] Andrew Marchese and Vasileios Maroulas. Topological learning for acoustic signal identification. In *2016 19th International Conference on Information Fusion (FUSION)*, pages 1377–1381. IEEE, 2016.

[10] Vasileios Maroulas, Farzana Nasrin, and Christopher Oballe. Bayesian inference for persistent homology. *arXiv preprint arXiv:1901.02034*, 2019.

[11] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[12] Nikhil Singh, Heather D Couture, JS Marron, Charles Perou, and Marc Niethammer. Topological descriptors of histology images. In *International Workshop on Machine Learning in Medical Imaging*, pages 231–239. Springer, 2014.

[13] Thierry Sousbie, Christophe Pichon, and Hajime Kawahara. The persistent cosmic web and its filamentary structure–ii. illustrations. *Monthly Notices of the Royal Astronomical Society*, 414(1):384–403, 2011.

[14] Clément Thomas, Stéphane Tholl, Danièle Moes, Monika Dieterle, Jessica Papuga, Flora Moreau, and André Steinmetz. Actin bundling in plants. *Cell motility and the cytoskeleton*, 66(11):940–957, 2009.

[15] Rien van de Weygaert, Erwin Platen, Gert Vegter, Bob Eldering, and Nico Kruithof. Alpha shape topology of the cosmic web. In *2010 International Symposium on Voronoi Diagrams in Science and Engineering*, pages 224–234. IEEE, 2010.

[16] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018.