16 September 2019

Misha,

Thank you very much for the referee reports of 19 August. We are pleased that both referees have very positive views of our manuscript, which we have revised thoroughly (to address the Minor Revisions) based on their helpful comments. We have also gone through the paper from start to finish to polish our presentation further. Below we respond to all of their points, and we look forward to receiving a decision on the revised manuscript.

Thanks,

Mason (on behalf of both authors)

———

Referee #1 (Remarks to the Author):

This is a referee report for "Persistent Homology of Geospatial Data: A Case Study with Voting" by Michelle Feng and Mason Porter, submitted to the SIAM Review, Research Spotlights.

Main Contribution: This paper discusses persistent homology as applied to geospatial data. In particular, various methods of constructing a simplicial complex, including two proposed by the authors, are discussed. These are the distance-based simplicial complexes Vietoris-Rips and alpha complex as well as two methods that incorporate adjacency information. These methods are applied in the context of California precinct-level voting data of the 2016 presidential election with the goal of being able to distinguish spatial features (precincts that voted differently than neighboring precincts).

Assessment: I found the paper to be well-written and an interesting read. With some small modifications, I feel that it would be appropriate for SIREV.

Response: Thank you for the kind comments.

Comments:

There were a few spots in the paper where the exposition was not clear. The most notable is in section 3.3 where the construction of level-set complexes was discussed. From my understanding, a plane is triangulated. There are many ways to do this, and the details were not discussed (e.g. what is the

grid size used in this paper? How was that choice made? Did it differ by precinct? Buried in the appendix, which should be referred to in this section, it states that triangles were formed by adding diagonals from NW to SE on each grid square). It might be useful to include the triangulation used in Fig. 6. Also, I believe the 0-level set of M is not just the boundary of M but all precincts (including the interior) that voted in the same way. Movement of the boundary appears to be moving parallel to the surface of the manifold (i.e. along the surface of the Earth not perpendicular to it); however, the discussion of normal direction and moving vertically through space was not clear. The velocity v and time T are also not explicitly discussed. In discussion of the persistence of this complex, it is discussed that short-persistence true features and short-persistence noise are distinguished because the latter will appear at later time steps in the evolution. A specific example of how and why this appears in the simplicial complex would be useful.

Response: This is a very good point, and we agree that it is important to improve our exposition here. We have reorganized and expanded Section 3.3 to explain further details of the level-set method for front propagation. We have also added a new figure to illustrate a particular example. Additionally, we have added a remark to the final paragraph of Section 3.3 to describe an example of short-persistence noise that can be seen in Figure 6.

In section 3.2 (bottom of page 7), we learn how the filtration of points is formed but need to emphasize again that the edges are created by adjacency.

Response: We have made this change. See Section 3.2, in the paragraph following equation 3.1.

Also, as mentioned at top of page 16, generators of a feature are not necessarily unique. How were generators selected? What does this nonuniqueness say about the results?

Response: We have added a note to the first paragraph of Section 4.2 explaining that while the generators are non-unique, in our case, that any choice of generators surrounds the relevant voting island if there is one.

I wondered about using minimum distances between points in precincts rather than between centroids (as discussed on page 11) as a means to construct a simplicial complex. I can see why this might be computationally difficult, but it would certainly be more interesting than the centroid distance measure.

Response: This is a good point, and we initially considered trying this instead. We chose not to do this because of the computational difficulty of doing so, though we agree that in principle it is

otherwise likely a better choice than centroid distance. We have added a sentence at the end of the second paragraph of Section 3.4.2

Along these lines, instead of specifically constructing a simplicial complex, I wondered why the authors did not use something like sublevel set persistent homology on the "surface" formed by the precincts. For instance, the function value could be the different levels of voter preferences and a cubical complex could be constructed to sweep up through pixel values on a discretized bounding region, including pixels whenever the voter preference is below some threshold, edges when to adjacent pixels are present, and squares when all four vertices are present. I believe this would incorporate both scaling and contiguity information. This method may prove more amenable to detecting holes than those presented in the paper.

Response: This is a good point, and we expect that this would incorporate scaling and contiguity information in a similar way to the level-set filtration, while keeping the voting strength scaling of the adjacency one. We added a sentence to the penultimate paragraph of the Conclusion to indicate that we think this would be interesting future work.

I had some confusion regarding the loops in the discussion of specific examples. Page 15, "The darkness of the loop"...all of the loops look the same color, so this was not at all clear. Also, showing a blow up of specifically discussed loops may help the explainability. Page 16, discussion of eight holes was not clear, there appear to only be 6 in figure but there are 9 intervals in the barcode. Please rework this discussion. Also, why are there so many loops in the adjacency complex in Fig. 13; this caption could be improved.

Response: We have updated all of the figures in the paper with a new criterion to identify features with significant persistence (based on a suggestion by Referee 2). Features that are persistent enough to qualify are now highlighted with darker, thicker loops. We have also rewritten captions and discussion to match the modified figures.

Minor Corrections:

The first sentence in the abstract says that "A crucial step in the analysis of persistent homology...into a simplicial complex." This is not quite accurate, for instance the sub-level set filtration does not build a simplicial complex at all.

Response: We have adjusted this sentence to reflect that the filtration can be any appropriate topological object, which in our case is a simplicial complex.

On page 1, "If X is a network H_0(X) records the number of connected

components" The number is found by the dimension of H_0(X)

Response: We have made this change; see the second paragraph of Section 1.

Page 1, it is mentioned that "other topological invariants are less computationally tractable", such as...?

Response: We have added a reference to homotopy; see the end of the second paragraph of Section 1.

Top of page 2, need to emphasize that the homology is tracked for each Scale.

Response: We have adjusted the third paragraph of Section 1.

In Section 2.2, should include general references for the reader and refer to Appendix A.

Response: We have added references; see the first paragraph after Definition 2.1.

Page 4, "m less than the dimension of X". This does not necessarily have to be true. For instance, a point cloud sampled from a circle can have infinitely high homology (see work by Adamasek and Adams)

Response: We have added a footnote to make this point.

Page 4, "Each homology group H_m(X_i) is a vector space" this is only true if field coefficients are chosen

Response: We have made this change; see the first paragraph after Definition 2.1.

Page 4, "However, by making the 'right' choice of construction for the persistence complex" seems a bold statement

Response: We have reworded this statement; see the second paragraph after Definition 2.1.

Page 7, Fig. 2. When reading, I wondered why this was only the red precincts. At this point in the paper there had been no discussion regarding how red/blue connect (or rather don't) in VR or any simplicial complex.

Response: We have added a note to the last paragraph of Section 3.1 that we consider only the red precincts.

Page 11, references [1,7,8,...] suggest that "we evaluate the features that

result from PH based on some criteria other than persistence" This statement is misleading, though, as these approaches really transform persistence into functional, kernel, or vector representations. Please rework this.

Response: We have reworked this statement to clarify that these methods do not rely solely on looking at the most persistent features, but instead at the entire persistence. See the second paragraph of Section 3.4.1.

Typos:

Page 2 paragraph starting with "Data sets" uses should be use and missing a period.
Page 4 Some subscripts appear to be incorrect f_{b_x-1},b_x
Page 5 "The geographic data comes in the from..." should be form
Page 5, the word compute is used a lot in the Vietoris-Rips description.

Response: Thank you for pointing out these typos. We have corrected them and reworded the Vietoris–Rips description.

Referee #2 (Remarks to the Author):

The paper focuses on the comparison of a set of techniques to build filtrations, i.e. sequences of simplicial complexes, in terms of the representations that they provide for datasets that live in two-dimensional space, mostly geospatial datasets. In particular they compare the standard Rips-Vietoris and alpha complex filtrations with two new proposals, named adjacency filtration and level-set filtration.

The authors then construct these filtrations, compute their persistent homology in low dimension (0,1) and proceed to illustrate the results on the basis of the issues (contiguity, scaling) that they highlighted previously, finding indeed that two new filtrations appear to mitigate the problems induced by the standard ones in the case of low-dimensional datasets with spatially extended nodes.

The paper is overall well written and quite detailed in the explanation of the problems, the filtrations and the background material. Overall I think the paper is appropriate for SIREV, although I have some concerns, which I list below, that I think should be addressed before publication:

Response: Thank you for these kind comments and for your suggestions to improve our paper.

- My main concern is that there appears to be a certain mismatch of depth across various parts of the analysis. The description of the filtrations is quite detailed, and so is the discussion of the computational complexity, computation time, etc (as shown by all the tables included). On the contrary, the results of the analysis on the voting maps and their discussion are to a large degree qualitative. In multiple instances, the authors refer to intervals in the barcode as long/short/medium, without providing a quantitative reference to measure what is to be considered short. I fully agree with the authors that low persistence intervals can in the right situations carry a lot of information, and to a large degree I can follow the argument they make in describing the results, but at the same time I think it would be important to provide some kind of reference (for example by perturbing the maps) to characterize what is significant and what is not.

Response: Thank you for this comment. We have added a simple measure to quantify barcodes as either “long persistence” or not. See Equation 4.1. We have also reworked the figures to highlight long-persistence bars in the barcodes and features on the maps to help evaluate whether these long-persistence features are useful.

- along a similar line, when discussing the results of the various filtrations, especially the novel ones, the authors often describe how successful they are in identifying correctly blue/red islands in the regions of red/blue sea. I am convinced that the filtrations are indeed working, but this is only by visual inspection. It would be important to establish some ground truth (in terms of where the homology cycles should be) and then quantify how well the various filtrations perform in recovering it.

Response: We have added an evaluation of how well various filtrations perform when we look only at the long-persistence features. See Table 4. Because the biggest difference in the methods is the amount of noise (all of the methods are mostly able to recover the most obvious holes, but the distance-based methods also pick up a lot of extraneous “features” and do a poor job of distinguishing them), we have chosen to quantify the percentage of detected long-persistence features which correspond to a hole in the ground truth (found by human eye).

- if that's possible, it would also be interesting to show the performance results on all the counties (even just for the adjacency and level-set cases, where the computation is feasible) included in the dataset, whereas now only a handful appear as illustrative examples.

Response: We have added a table showing the performance results of long-persistence features in every county and every simplicial complex for which we computed/identified features (see Table 4).

- finally, from a purely explanatory perspective, the description of the Level-set filtration is quite short in comparison to the space given to more intuitive filtrations, like the VR, alpha and adjacency ones. For example, I did not understand how the final time T is chosen, does it stabilize on its own? How does the \delta_{b,r} value enters in this case? As a fixed threshold at the beginning or is it related to the evolution of the level set? I understand the article is not meant to be a technical review, but this filtration is decidedly the least common one and I think it warrants some more detail.

Response: We have added more description and an additional illustrative example to Section 3.1 to help explain the level-set filtration in more detail. We comment on the final time *T* in the fourth paragraph after Equation (3.2). The \delta_{b,r}, which we defined for the adjacency complex, does not enter into the level-set complex; instead, we consider all precincts that have the same voting preference. We added a brief parenthetical note to clarify this point (in the second paragraph of Section 3.3).