

Fast and Scalable Complex Network Descriptor Using PageRank and Persistent Homology

1st Mustafa Hajij

Department of Mathematics and Computer Science

Santa Clara, California
mhajij@scu.edu

2nd Elizabeth Munch

Department of Computational Mathematics, Science, and Engineering

Michigan State University
Lansing, Michigan
muncheli@msu.edu

3rd Paul Rosen

Department of Computer Science and Engineering

University of South Florida
Tampa, Florida
prosen@usf.edu

Abstract—The PageRank of a graph is a scalar function defined on the node set of the graph which encodes nodes centrality information of the graph. In this article we use the PageRank function along with persistent homology to obtain a scalable graph descriptor and utilize it to compare the similarities between graphs. For a given graph $G(V, E)$, our descriptor can be computed in $O(|E|\alpha(|V|))$, where α is the inverse Ackermann function which makes it scalable and computable on massive graphs. We show the effectiveness of our method by utilizing it on multiple shape mesh datasets.

Index Terms—PageRank, Complex Networks Similarity, Topological Data Analysis, Graph Similarity

I. INTRODUCTION

The problem of studying similarity between graphs has attracted much attention recently in the pattern recognition and machine learning communities. One of the main challenges is to construct an effective similarity measure between graphs that takes into account the complexity of the underlying structure while still being computed efficiently.

In this work, we utilize the *PageRank* vector [2] in conjunction with a tool available in *persistent homology* [10] to define a graph descriptor. More specifically, we view the PageRank as a continuous scalar function [20] defined on the vertices of the graph and utilize this scalar function to induce a *filtration* as defined traditionally in the context of persistent homology. We show that the *persistence diagram* induced by this filtration can be utilized for graph similarity.

Persistent homology provides a robust set of tools for the theoretical and practical capacity to understand the *shape of data* [4] in any number of dimensions and on multiple scales, placing the concept of shape, as applied to data analysis, on a solid mathematical foundation. On the other hand, the PageRank function of a graph stores information regarding the centrality information of the underlying nodes. The filtration induced by the PageRank provides a method to decode the information encoded in this scalar function and stores it in the persistence diagram. The latter, when combined with

bottleneck distance, can then be used for the graph similarity task.

Utilizing the PageRank vector has two main advantages. First, PageRank was originally designed to compute efficiently on very large graphs. The efficiency of the PageRank vector has been studied extensively [14]. The PageRank vector has found many applications, including graph partition [1], image search [16], and citation analysis [17], among others. Second, as we will show here, as a function defined on the nodes of the graph the PageRank vector stores rich structural information about the underlying graph that can be utilized to detect the similarity between different graphs effectively.

Graph similarity lies within the realm of pattern recognition and machine learning [21]. Persistent homology provides unique information about the graphs, discover uncovering insights, and determines which predictors are more related to the outcome. Persistent Homology-based methods have shown excellent performance in several applications including pattern recognition on graphs [5], [8], [18], [19], [25], time-varying data [9], [13], and images [6], [11], [22], among others.

II. BACKGROUND

In this section, we give a brief review of persistent homology and the PageRank vector. While the work here is concerned with graphs, we choose here to introduce persistent homology for simplicial complexes since our work can be generalized easily to more general domains. We assume the reader is familiar with the basics of simplicial homology.

A. Persistent Homology

Let K be a simplicial complex. We will denote the vertices of K by $V(K)$. Let S be an ordered sequence $\sigma_1, \dots, \sigma_n$ of all simplices in K , such that for simplex $\sigma \in K$ every face of σ appear before it σ in S . Then S induces a nested sequence of subcomplexes called a *filtration*: $\phi = K_0 \subset K_1 \subset \dots \subset K_n = K$. A d -homology class $\alpha \in H_d(K_i)$ is said to be *born* at the time i if it appears for the first time as a homology class

in $H_d(K_i)$. A class α *dies* at time j if it is trivial $H_d(K_j)$ but not trivial in $H_d(K_{j-1})$. The *persistence* of α is defined to be $j - i$. Persistent homology captures the birth and death events in a given filtration and summarizes them in a multi-set structure called the *persistence diagram* $P^d(\phi)$. Specifically, the persistence diagram of the a filtration ϕ is a collection of pairs (i, j) in the plane where each (i, j) indicates a d -homology class that is created at time i in the filtration ϕ and killed entering time j .

Persistent homology can be defined given any filtration. For the purposes of this work, the input is a piecewise linear function $f : |K| \rightarrow \mathbb{R}$ defined on the vertices of complex K . Furthermore, we assume the function f has different values on different nodes of K . Any such a function induces the *lower-star* filtration as follows.

Let $V = \{v_1, \dots, v_n\}$ be the set of vertices of K sorted in non-decreasing order of their f -values, and let $K_i := \{\sigma \in K \mid \max_{v \in \sigma} f(v) \leq f(v_i)\}$. The lower-star filtration is defined as:

$$\mathcal{F}_f(K) : \phi = K_0 \subset K_1 \subset \dots \subset K_n = K. \quad (1)$$

The lower-star filtration reflects the topology of the function f in the sense that the persistence homology induced by the filtration 1 is identical to the persistent homology of the sublevel sets of the function f . We denote by $P_f(K)$ to the persistence diagram induced by the lower-star filtration $\mathcal{F}_f(K)$. See Figure 1.

Furthermore, we will denote by $P_f^k(K)$ to the k^{th} persistence diagram induced by the lower-star filtration $\mathcal{F}_f(K)$. In this work, we will only consider the 0-dimensional persistence diagram.

B. Computing the 0-persistence diagram the of a lower-star filtration

For completeness of our treatment we give a brief description for computing the 0-persistence diagram the PageRank defined on the nodes of on a graph G . The computation of the zero persistent diagram $P_f^0(G)$ can actually be done using union-find data structure. We give the details next. If $e = (u, v)$ is an edge of the graph G then we will extend the PageRank vector to e by defining $PR(e) := \max(PR(u), PR(v))$.

Let $V = \{v_1, \dots, v_m\}$ be the node set of G . Let $E = \{e_1, \dots, e_n\}$ be its edge set ordered with respect to their PR -values. The steps of the the algorithm to compute the zero PD associated with the PageRank is given as follows.

The first step in the algorithm creates a connected component C_i for each node v_i in the graph G . Here we assume that the connected components are created using the *disjoint set* data structure.

The second step of the algorithm looks at the edges of G in the ascending order with respect to their PR -values. For each $e = (u, v)$, we check if the nodes u and v of e belong to two different sets. If this is the case, then we merge the two connected components containing u and

Algorithm 1: Computing the Persistence Diagram induced by the PageRank

```

1 Function computePageRankPD( $G, PR : G \rightarrow \mathbb{R}$ )
2   bars = []
3    $U = \emptyset$ 
4   foreach Node  $i$  in  $V(G)$  do
5      $U.\text{make}(i);$ 
6   Sort the edge of the graph  $G$  in ascending order
      using the their  $PR$ -values.
7   foreach Edge  $e = (u, v)$  in  $E(G)$  do
8      $c \leftarrow U.\text{get}(u)$ 
9      $d \leftarrow U.\text{get}(v)$ 
10    if  $c \neq d$  then
11       $U.\text{merge}(c, d)$ 
12    bars.append(( $\max(PR(c), PR(d)), PR(e)$ ))
13  return bars
14 End Function

```

v. Furthermore, we append to the list of *bars* the pair $(\max(PR(c), PR(d)), PR(e))$ where c and d are the roots of the trees that contain the nodes u and v respectively in the disjoint set data structure. The algorithm return the list *bars* representing the birth and death of 0-features of the graph G with respect to PageRank functional values.

The *merge* operation in line 12 in Algorithm 1 assumes the following merge order on the sub-trees in disjoint set data structure. The tree with root c is merged with the tree with root d according to the PR values of c and d . Namely, if $PR(c) > PR(d)$ then we set d to be the parent of c . Otherwise c to be the parent of d .

An illustrative example of running this algorithm on a 1-d function is given in Figure 2.

III. COMPUTING THE DISTANCE BETWEEN THE PERSISTENCE DIAGRAMS

Given two persistence diagrams, we measure the distance between them using the bottleneck distance. Namely, given two persistence diagrams X and Y , let η be a bijection between points in the diagrams. The bottleneck distance is defined as,

$$W_\infty(X, Y) = \inf_{\eta: X \rightarrow Y} \sup_{x \in X} \|x - \eta(x)\|_\infty.$$

For technical reasons we usually add to the persistence diagram infinitely many points on the diagonal and each one of these points with is counted with infinite multiplicity. In our study we utilize the bottleneck distance to quantify the difference between two PR descriptors. Other distances can also be employed such as the Wasserstein distance.

A. PageRank

This work utilizes the lower-star filtration induced by the PageRank function [2]; more specifically, we consider a

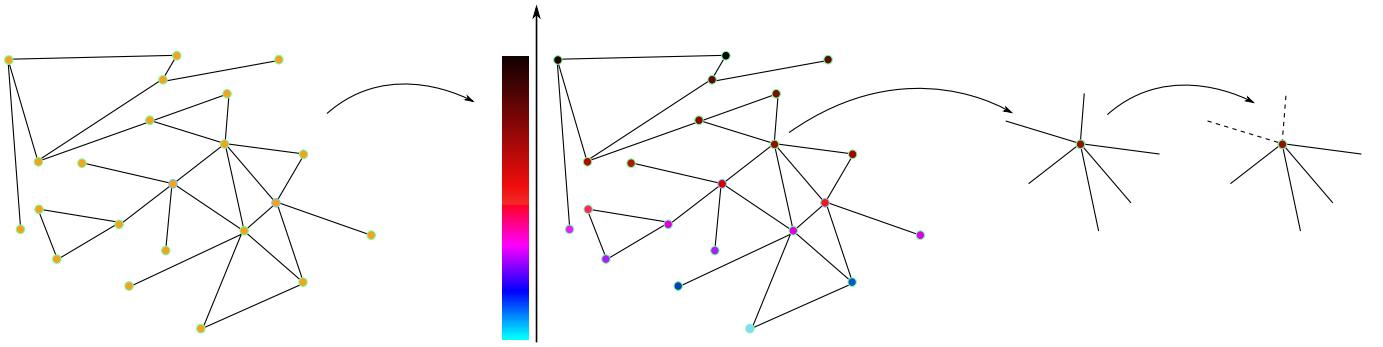


Fig. 1: Left : a graph with a scalar function defined on its nodes. Middle the star of the node v . Right: the lower-star of a vertex v .

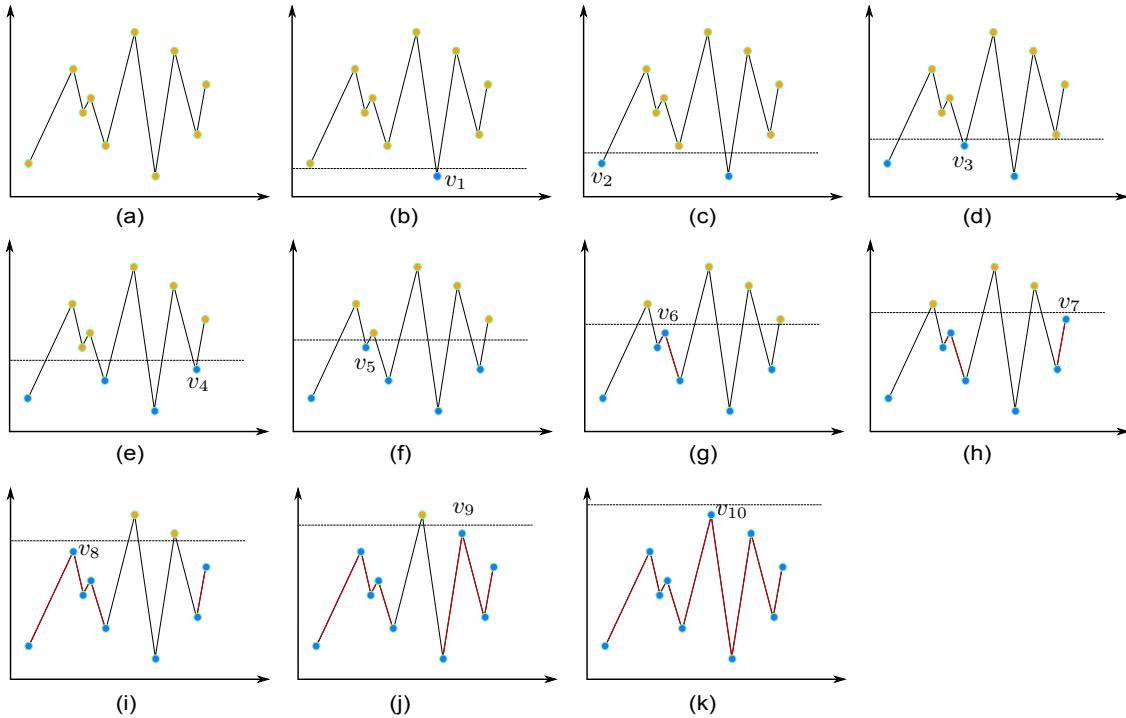


Fig. 2: An example illustrating the computation of the persistence diagram on a scalar function defined on 1-d simplicial complex K . We assume that we have a scalar function $f : V(K) \rightarrow \mathbb{R}$ defined on the vertex set $V(K)$ of K . We order the nodes and the edges in the function using their f values and process them with respect to this order. The values of the function f hence induces a lower star filtration where at every stage in this filtration we introduce a vertex along with the edges that are connected to it and have lower f -values, if it has any.

version applicable to undirected graphs [12]. The PageRank function $PR : V \rightarrow \mathbb{R}$ is defined for every vertex $v \in V$ by

$$PR(v) = \frac{(1-d)}{|V|} + d \sum_{u \in N(v)} \frac{PR(u)}{|N(u)|}, \quad (2)$$

where $N(v)$ is the set of neighbors of v ; $0 < d < 1$ is the *damping factor*, typically set at 0.85. Equation (2) can be solved efficiently by the power method [15]. See also [24] for a $\mathcal{O}(\sqrt{\log(n)}/\epsilon)$ distributed algorithm where n is the number of nodes in the graph and ϵ is a fixed constant.

A high PageRank score at v typically means that v is

connected to many nodes, which also have high PageRank scores. For our purpose, it is important to notice that the PageRank is a continuous function [20]. For example, Figure 3 illustrates the continuity of the function on the nodes of the graph on a random geometric graph.

IV. RUNNING TIME

The proposed descriptor can be computed in almost linear time. Once the graph data is loaded, the 0-dimensional persistence diagram can be computed using disjoint sets which take $\mathcal{O}(|E|\alpha(|V|))$, where α is the inverse Ackermann function [7], an extremely slow growing function. The PageRank can

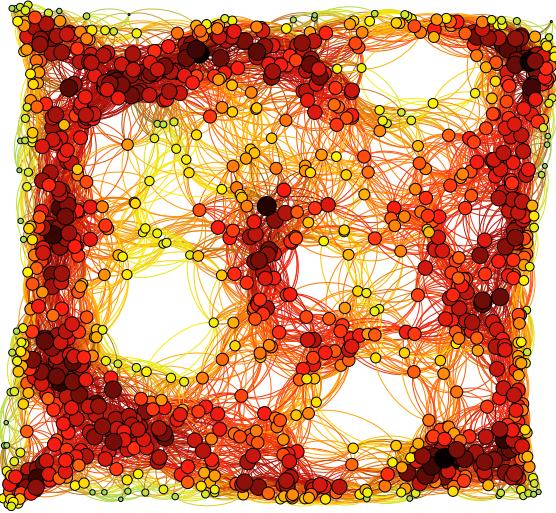


Fig. 3: Example of the PageRank vector computed on a geometric graph. Higher PageRank values indicate higher node centrality. In this figure the PageRank values are indicated by the size of the nodes as well as the the color of the nodes color (nodes with higher PR values have darker colors).

be computed in sub-linear time. For instance see [24] for a $\mathcal{O}(\sqrt{\log(n)}/\epsilon)$ distributed algorithm where n is the number of nodes in the graph and ϵ is fixed constant.

V. RESULTS

To validate the method proposed, we run some experiments on three publicly available datasets. We use mesh datasets to make a visual comparison between similar graphs easier.

In our experiments, we compute the persistence diagram of each mesh obtained from the lower-star filtration induced by the PageRank vector defined on that mesh. The pairwise bottleneck distance is then computed between every pair of persistence diagrams. Finally, the resulting discrete metric space is visualized using a 2d t-SNE projection [27].

The first dataset [26] consists of 60 meshes that are divided into 6 categories: cat, elephant, face, head, horse, and lion. Each category contains ten triangulated meshes. The result is reported in Figure 5 left handside.

The second dataset [23] consists of 30 meshes that are divided into 2 categories: kid A and kid B. The result is reported in Figure 5 right-handside.

The third dataset [3] contains a total of 80 objects, including 11 cats, 9 dogs, 3 wolves, 8 horses, 6 centaurs, 4 gorillas, 12 female figures. The vertex count for each object in this data is about 50K.

In all of our three example datasets, one can clearly observe the effectiveness of the proposed descriptor at capture the geometry of the underlying meshes. In particular, one can easily see that the meshes within the same category are clustered together. We also notice that meshes with similar topology tend to be closer than those with different topology. Observe for instance the clusters of horses and cats in Figure 5.

VI. CONCLUSION

In this work, we have illustrated how the PageRank can be utilized in conjunction with persistent homology to study graph similarity and demonstrated our results on small datasets. In future work, we are planning to conduct a more thorough analysis with larger datasets. Moreover, the PageRank is typically defined on directed graphs. This feature of the PageRank vector can be utilized to induce a filtration that is sensitive to the directionality of the edges a directed graph. We are planning to investigate this direction in the future.

VII. ACKNOWLEDGMENT

This work was supported in part by a grant from the National Science Foundation (IIS-1845204).

REFERENCES

- [1] Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486. IEEE, 2006.
- [2] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.
- [3] Alexander M Bronstein, Michael M Bronstein, and Ron Kimmel. *Numerical geometry of non-rigid shapes*. Springer Science & Business Media, 2008.
- [4] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.
- [5] C. J. Carstens and K. J. Horadam. Persistent homology of collaboration networks. *Mathematical Problems in Engineering*, 2013, 2013.
- [6] James R Clough, Ilkay Oksuz, Nicholas Byrne, Veronika A Zimmer, Julia A Schnabel, and Andrew P King. A topological loss function for deep-learning based image segmentation using persistent homology. *arXiv preprint arXiv:1910.01877*, 2019.
- [7] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2009.
- [8] Weinan E, Jianfeng Lu, and Yuan Yao. The landscape of complex networks. *CoRR*, abs/1204.6376, 2012.
- [9] Herbert Edelsbrunner, John Harer, Ajith Mascarenhas, and Valerio Pascucci. Time-varying reeb graphs for continuous space-time data. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 366–372. ACM, 2004.
- [10] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pages 454–463. IEEE, 2000.
- [11] Kathryn Garside, Robin Henderson, Irina Makarenko, and Cristina Massoller. Topological data analysis of high resolution diabetic retinopathy images. *PloS one*, 14(5):e0217413, 2019.
- [12] Vince Grolmusz. A note on the pagerank of undirected graphs. *arXiv preprint arXiv:1205.1960*, 2012.
- [13] Mustafa Hajij, Bei Wang, Carlos Scheidegger, and Paul Rosen. Visual detection of structural changes in time-varying graphs using persistent homology. In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, pages 125–134. IEEE, 2018.
- [14] Taher Haveliwala. Efficient computation of pagerank. Technical report, Stanford, 1999.
- [15] Joe D Hoffman and Steven Frankel. *Numerical methods for engineers and scientists*. CRC press, 2018.
- [16] Yushi Jing and Shumeet Baluja. Pagerank for product image search. In *Proceedings of the 17th international conference on World Wide Web*, pages 307–316, 2008.
- [17] Nan Ma, Jiancheng Guan, and Yi Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*, 44(2):800–810, 2008.
- [18] Giovanni Petri, Martina Scolamiero, Irene Donato, and Francesco Vaccarino. Networks and cycles: A persistent homology approach to complex networks. *Proceedings European Conference on Complex Systems 2012, Springer Proceedings in Complexity*, pages 93–99, 2013.

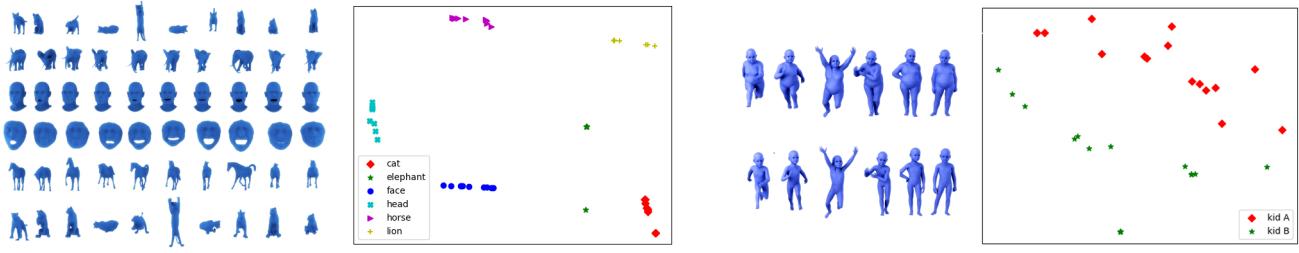


Fig. 4: In both left and right Figure we compute the PageRank’s vector for each mesh in a data set is computed. We then utilize this function to compute 0-persistence diagram associated with the lower-star filtration of PageRank. Then we compute the pairwise bottleneck distance between every pair of of that dataset. The final distance matrix is then visualized using a 2d t-SNE projection. In the left figure, we show the application of our method to a data set consists of 60 triangulated meshes divided into 6 categories [26]. On the other hand the right figure shows the application of this method to kids dataset [23] which consists of 30 meshes, 15 meshes of kid A and 15 meshes of kid B.

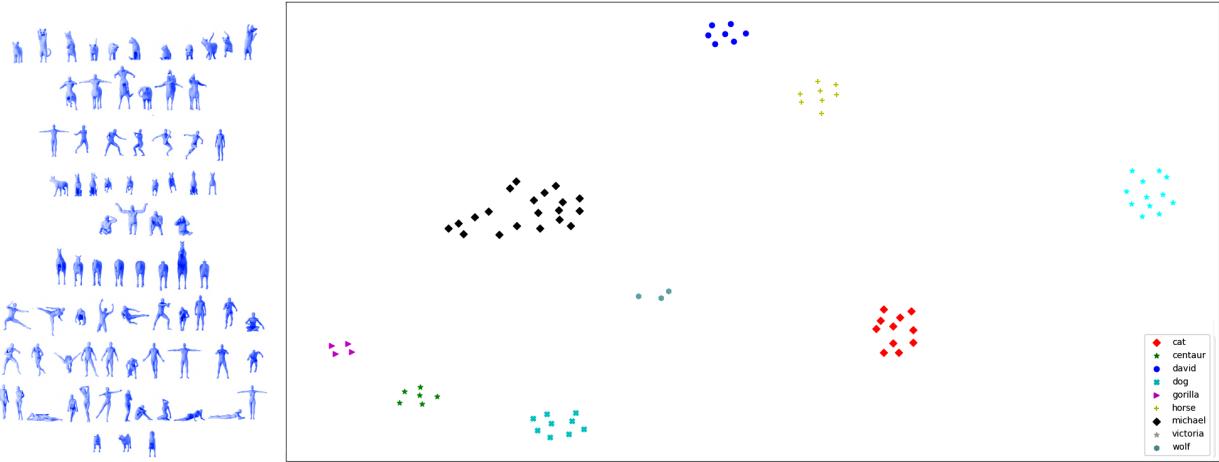


Fig. 5: On the left the dataset [3] which consists a total of 80 objects, including 11 cats, 9 dogs, 3 wolves, 8 horses, 6 centaurs, 4 gorillas, 12 female figures. The vertex count for this dataset is about 50,000. On the right the t-SNE projection obtained from the distance matrix of the pairwise bottleneck distance between the persistence diagrams associated with the lower-star filtration of the PageRank vectors.

- [19] Giovanni Petri, Martina Scolamiero, Irene Donato, and Francesco Vaccarino. Topological strata of weighted complex networks. *PLoS ONE*, 8(6):e66506, 2013.
- [20] Luca Pretto. Analysis of web link analysis algorithms: The mathematics of ranking. In Maristella Agosti, editor, *Information Access through Search Engines and Digital Libraries*, pages 97–111. Springer, 2008.
- [21] Saif Ur Rehman, Asmat Ullah Khan, and Simon Fong. Graph mining: A survey of graph mining techniques. In *Seventh International Conference on Digital Information Management (ICDIM 2012)*, pages 88–92. IEEE, 2012.
- [22] Alejandro Robles, Mustafa Hajij, and Paul Rosen. The shape of an image: A study of mapper on images. *arXiv preprint arXiv:1710.09008*, 2017.
- [23] Emanuele Rodolà, Samuel Rota Bulo, Thomas Windheuser, Matthias Vestner, and Daniel Cremers. Dense non-rigid shape correspondence using random forests. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4184, 2014.
- [24] Atish Das Sarma, Anisur Rahaman Molla, Gopal Pandurangan, and Eli Upfal. Fast distributed pagerank computation. In *International Conference on Distributed Computing and Networking*, pages 11–26. Springer, 2013.
- [25] Ashley Suh, Mustafa Hajij, Bei Wang, Carlos Scheidegger, and Paul Rosen. Persistent homology guided force-directed graph layouts. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):697–707, 2019.
- [26] Robert W Sumner and Jovan Popović. Deformation transfer for triangle meshes. *ACM Transactions on graphics (TOG)*, 23(3):399–405, 2004.
- [27] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.