Fuzzy Systems and Data Mining VIII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA220370

Community Division Metric Based on Persistent Homology

Hui ZHONG [a,1], Lv Lin QIU [a], Zhi Jian ZHANG [a], Lin JIANG [a], and Xin Yang LI [a]
[a] *Faculty of Science, Kunming University of Science and Technology, China*

Abstract. Community structure is one of the most important structural features of complex networks. However, most of the existing community division metrics only consider the relationship between nodes, and do not consider the overall closeness of the internal and external communities from the perspective of topology. Persistent homology (PH) is a mathematical tool in computational topology, which can capture high-dimensional topological features and is widely used in the analysis of complex networks. In this paper, we define a community partitioning metric based on persistent homology theory, and propose an algorithm of community division based on C_{PH} which provides a new method for community partitioning performance. From the validation experiments, the Louvain algorithm is used to evaluate the community partitioning performance of social networks, and the experimental results show that community division metric based on persistent homology can measure the performance of community partitioning from the perspective of topology, persistent homology can be used as a new way to describe community structure.

Keywords. Persistent homology, Community, Community division metric based on persistent homology, MCPH Algorithm

1. Introduction

With the rapid development of the Internet, social networks have become a good platform for user communication, and while providing people with diverse interaction environments and information resources, complex and diverse networks have been formed. So-called communities are collections of nodes with similar connection structures, with tightly connected nodes within communities and sparsely connected nodes between communities [1]. Community structure is a common and important topological feature of many complex networks in the real world. Detecting the communities in a network helps to gain better insight into the network structure and understand the composition of the network.

Most of the existing community discovery methods are based on the similarity between nodes and the method of probability model, do not consider the topological structure relationship between nodes. Persistent homology can extract important features of long duration in network changes and filter out noise. If the internal nodes of the community are closely connected and the external connections of the community are sparse, then

[1]Corresponding Author: Zhi Jian ZHANG, Faculty of Science, Kunming University of Science and Technology, Kunming; E-mail: zhijian@kust.edu.cn.

the effect of community division should be the best. Therefore, considering the proportion of high-dimensional features within the community and the ratio of low-dimensional features between communities after network division, a new method and metric for the performance of community division is given.

2. Related works

As an important research work in complex networks, community discovery has received extensive attention from domestic and foreign scholars. The Louvain [2] algorithm is a faster community discovery algorithm with optimized modularity, which considers each node as an independent community and computes the modularity increment function continuously to obtain the community structure. The GN algorithm proposed by Girvan [3] and Newman defines the edge betweenness by calculating the frequency of a certain edge appearing on the shortest path between any nodes and continuously removing the edge with the highest edge betweenness to obtain the community structure of the network, but the algorithm has high time complexity and is not suitable for large networks.

Persistent homology as a tool for algebraic topology has started to be used frequently to analyze the topology of networks, and specifically, applications involve various types of networks, including collaboration networks [4, 5], social networks [6, 7, 8], sensor networks [9], brain networks [10, 11], and random networks [12]. The study of social networks through persistent homology has also made great progress, as Carstens and Horadam [4] first used 1-Betti number and 2-Betti number for the analysis of weighted networks, proving that persistent homology corresponds to tangible features of the network and can be used to distinguish collaboration networks from similar random networks. It is shown that the effectiveness and superiority of the method have been demonstrated in solving the problems related to social networks. Therefore, this paper proposes a community division metric based on persistent homology, and gives the corresponding community division evaluation algorithm, which provides new inspiration for community division measure and analysis methods.

3. Community division metric based on persistent homology and traditional community division metrics

At present, domestic and foreign scholars have proposed a series of metrics to evaluate the performance of community division, and the traditional methods to evaluate the performance of community division are modularity and normalized mutual information. Most of the metrics are not considered from the perspective of topology, so persistent homology is used to evaluate and analyze the performance of community division.

3.1. Community division metric based on persistent homology

After the community is divided, the community should have the characteristics that the internal nodes are tightly connected and the inter-community nodes are sparsely connected. In the persistent homology theory, with the persistent change of the filtration value, the persistent generation and disappearance of each-dimensional feature, the 0-dimensional feature represents the change of the connected components,

the 1-dimensional feature represents the change of the 1-dimensional hole, and the 2-dimensional feature represents the 2-dimensional voids. The higher-dimensional features represent more complex structures. Therefore, if the proportion of 0-dimensional features between communities and the proportion of high-dimensional features within communities is higher, the better the effect of community division is. According to this idea, the metric of community division is defined.

Suppose the $G = (V,E)$ network is divided into communities$G = \{G_1, \cdots, G_K\}$, $p(ex)_{dim0}$ represent the proportion of 0-dimensional features among communities, $p_i(in)_{dim1\,dim2}$ is the proportion of high-dimensional features within each community, and $average(p_i(in)) = \frac{\sum_{i=1}^{K} p_i(in)_{dim1\,dim2}}{K}$ is the average of the proportion of high-dimensional features within communities. In summary, we obtain the definition of the community division performance metric.

$$C_{PH} = p(ex)_{dim0} + \frac{\sum_{i=1}^{K} p_i(in)_{dim1\,dim2}}{K} \tag{3.1}$$

From the equation 3.1, the higher the score of C_{PH} , the closer the connection of the nodes in the community, the better the community division effect. When all nodes in the community are connected to nodes in the community, the value of C_{PH} reaches the maximum value 2, and the community division effect is the best at this time; when all nodes in the community are connected to nodes outside the community, the community division effect is the worst.

Table 1 Comparison of different community classification metrics

Metric	Computational Foundations	Characteristics
Modularity(Q)	Differences between community and random networks after partitioning.	Tight internal community connections and sparse external community connections.
Normalized mutual information(NMI)	Differences between algorithmically divided communities and communities of real network.	Considering the connection relationship between nodes, the similarity of nodes inside the community is high and the similarity of nodes outside the community is low.
Community division metric based on persistent homology(C_{PH})	The proportion of high-dimensional features inside the community, the proportion of low-dimensional feature points outside community.	More high-dimensional features within communities, more low-dimensional features between communities.

3.2. Metric algorithm of community division performance based on persistent homology

Next, combining with the community division measure based on persistent homology, a corresponding community division evaluation algorithm is proposed. The idea of the algorithm is: first, use Louvain algorithm to divide the original network into communities, and using *Rips* complex to calculate the persistent homology among the internal nodes of

the community. Thereby, the proportion of high-dimensional features within each community is obtained, and the average value of the proportion of high-dimensional features in the community is calculated. The aggregated communities are regarded as nodes below, and the ratio of 0-dimensional features between different communities is calculated. Through this strategy, the C_{PH} value of the current community division can be obtained.

Algorithm 1 Metric algorithm of community division performance based on persistent homology

Input:
Distance matrix of original network
Output:
C_{PH} value after community division using Louvain algorithm
1: Louvain algorithm to divide the original network into K communities
2: Using *Rips* complex to compute persistent homology for each divided community
3: **for** each community $G_i \in G$ **do**
4: computing $\sum_{i=1}^{K} p_i(in)_{dim1\,dim2}$
5: **end for**
6: computing $average(p_i(in)) = \frac{\sum_{i=1}^{K} p_i(in)_{dim1\,dim2}}{K}$ within the community
7: Using *Rips* complex to compute persistent homology for inter-community, and computing $p(ex)_{dim0}$
8: **return** C_{PH} using equation 3.1

The time complexity of this algorithm is greatly affected by the Louvain algorithm and the calculation of the persistent homology of the community, and the time complexity of the Louvain algorithm is $O(nlogn)$, the time complexity of computing the network persistent homology by programming is $O(n)$, therefore, the time complexity of the algorithm is $O(nlogn)$.

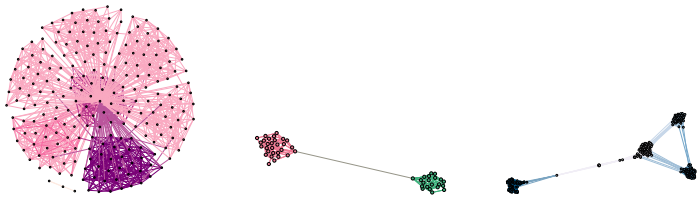
4. Numerical experiments

4.1. Experimental data

In this paper, we evaluate the validity of the persistent homology community segmentation metric by selecting the Enron social network graph from the SNAP [13] database and the artificial network from Network repository [14] and conducting simulation experiments on three different forms of social networks with some parameters of the network shown in Table 2.

Table 2 Parametric characteristics of the networks

Name of the network	$\|V\|$	$\|E\|$	Average degree
Enron network	221	1209	5.471
Artificial network A	50	301	6.02
Artificial network B	269	1599	5.944



(a) Enron network (b) Artificial network A (c) Artificial network B

Figure 1 Diagram of network topology

4.2. Experimental steps

First, processing the original social network by the random walk algorithm [15], and the distance matrix is calculated by applying the shortest path algorithm. Then, the filtering flow of the complex shape is computed by applying the persistent homology to construct the complex shape, which results in the number of feature points per dimension and the persistence interval, and finally, the barcode and persistence diagram are obtained.

4.3. Community division of Enron network

Persistence diagrams are graphs to display persistence barcodes. Figure 2 shows the persistence diagram and barcode of the original Enron network and artificial network. So,

(a) Enron network (b) Artificial network A

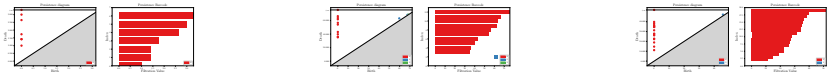
Figure 2 Persistence diagrams and barcodes of networks

simply a persistence diagram depicts the boundary points of barcodes, and they always lie above $y = x$, the dashed diagonal line in a diagram.

We can see the 0-dimensional homology groups and the connected components are all born at t = 0, and new features are continuously generated, some of them die very fast, and the final features are retained indefinitely. Until t=30, 0-dimensional features continue to disappear, and finally the largest connected component is formed, that is, the longest barcode at the top of the barcode (original network). When t=2, 1-dimensional features start to be generated, so some 0-dimensional features disappear at this moment, and more 1-dimensional circles are generated in the process of continuously building complex shapes around t=10. When t=50, 2-dimensional features are generated. The network structure is more complex.

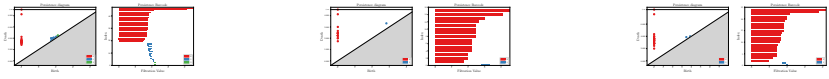
The Louvain algorithm is used to divide the communities of the Enron network, which is divided into 7 communities, and then the network is divided into 10 and 18 communities randomly. Figure 3 shows the result that after using the classical community division algorithm, only 0-dimensional features are generated between the communities, while more feature points are generated between the communities that are divided randomly, and at this time 1 and 2-dimensional feature points reflecting the high-dimensional structure of the network are generated. Because the random division of com-

munities leads to the existence of more edges between communities, which generates some high-dimensional features. The figure 4 shows that the classical community di-



(a) Louvain algorithm (b) Random 10 (c) Random 18

Figure 3 Persistence diagrams and barcodes of inter-rcommunity in Enron network



(a) Louvain algorithm (b) Random 10 (c) Random 18

Figure 4 Persistence diagrams and barcodes of intra-community in Enron network

vision internally generates more high-dimensional persistent features than the random division, because the community is more tightly connected internally at this time, while the edges inside the randomly divided community are connected to the outside of the community and the internal nodes are more loosely connected.

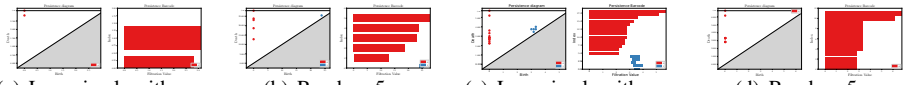
The table 3 is the C_{PH} of the Enron network under different divisions. It can be seen that the community division using the Louvain algorithm has the highest C_{PH} value, which proves that the metrics are effective. And experiments show that the more complex the community structure will be, the more persistent points will be generated, and the greater the number of corresponding high-dimensional features will be.

Table 3 Community division metrics of Enron network

	Classic community division	Dividing into 10 communities	Dividing into 18 communities
$p(ex)_{dim0}$	1	0.769	0.9
$average(p_i(in))$	0.476	0.171	0.184
C_{PH}	1.476	0.940	1.084

4.4. Community division of Artificial network

Since the artificial network A has fewer nodes and the network is not complex enough, persistence diagram and barcode produce fewer high-dimensional features. 0-dimensional features begin to appear at t=0, disappearing around t=3 slowly, and soon form a largest connected component(original network) at t=6. At t=3.5, some 1-dimensional features start to be generated, and a 2-dimensional circle surrounding the 1-dimensional circle is generated. The division metrics of community under different di-



(a) Louvain algorithm (b) Random 5 (c) Louvain algorithm (d) Random 5

Figure 5 Persistence diagrams and barcodes of inter-community and intra-community in artificial network A

visions and shows that the artificial network A using Louvain algorithm has the highest C_{PH} value for community division and the random community division has a lower value, the validity of the C_{PH} measure of community division performance is again confirmed. And figure 5 (a) and (b) are persistence diagrams and barcodes outside the community, figure 5 (c) and (d) are inside community.

Next, we try to conduct experiments on a larger artificial dataset. The nodes of artificial network B are more densely connected to generate more high-dimensional features.

The intra-community persistence diagram and barcode of artificial network B using classical division and random division, respectively. After calculation, artificial network B can be divided into four communities by Louvain algorithm, the highest C_{PH} value obtained is 1.154, and the C_{PH} of random division as 6 and 8 communities are 0.875 and 0.763, respectively, which are lower than the classical division algorithm.

In summary, the community division using Louvain discovery algorithm is tighter within the community and less connected between communities, and the persistent homology community division metric can measure the performance of community division. Meanwhile, persistent homology can describe the density inside and outside, capturing the community with different characteristics and shape. Furthermore the complexity of the community is proportional to the number of persistence points.

5. Conclusion and Future work

At present, there is no uniform evaluation criterion for community discovery algorithms, and for the same network, different community discovery algorithms may obtain different community structures, and different evaluation criteria may also obtain different optimal community structures. In this paper, we found that after using the classic community discovery algorithm to divide the community, the persistent homology between communities only produces 0-dimensional features, and basically no high-dimensional features are generated. However, more high-dimensional features are generated within the community, because the community is relatively close at this time. When dividing nodes into the corresponding community randomly, more high-dimensional features will be generated between them, and the internal division of the community is not so close, so there are few high-dimensional features generated. The C_{PH} can measure the community division based on the degree of inter-community connectivity, intra-community closeness and the proportion of each dimensional feature point, which can be used as a means to evaluate the stability of community division, and provides an effective new method for describing the community division performance in social network graphs.

When using persistent homology to analyze the network, as the network continues to increase the number of nodes, the number of simplex will increase sharply, and high-dimensional simplex complex will continue to be constructed. We only conduct experiments on small networks, therefore, combining machine learning methods with persistent homology on a larger network for experiment is future work.

References

[1] Shen H W. Community structure of complex networks. Springer Science and Business Media, 2013.

[2] GUILLAUME L. Fast unfolding of communities in large networks. Journal Statistical Mechanics: Theory and Experiment, 2008, 10: P1008.
[3] Girvan M, Newman M E J. Community structure in social and biological networks. Proceedings of the national academy of sciences, 2002, 99(12): 7821-7826.
[4] Carstens C J, Horadam K J. Persistent homology of collaboration networks. Mathematical problems in engineering, 2013, 2013.
[5] Wilkerson A C, Moore T J, Swami A, et al. Simplifying the homology of networks via strong collapses//2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 5258-5262.
[6] Kee K F, Sparks L, Struppa D C, et al. Social groups, social media, and higher dimensional social structures: A simplicial model of social aggregation for computational communication research. Communication Quarterly, 2013, 61(1): 35-58.
[7] Fellegara R, Fugacci U, Iuricich F, et al. Analysis of geolocalized social networks based on simplicial complexes//Proceedings of the 9th ACM SIGSPATIAL Workshop on Location-based Social Networks. 2016: 1-8.
[8] Rieck B, Leitte H. 'Shall I compare thee to a network?': Visualizing the Topological Structure of Shakespeares Plays. 2016.
[9] De Silva V, Ghrist R. Homological sensor networks. Notices of the American mathematical society, 2007, 54(1).
[10] Lee H, Chung M K, Kang H, et al. Discriminative persistent homology of brain networks//2011 IEEE international symposium on biomedical imaging: from nano to macro. IEEE, 2011: 841-844.
[11] Lee H, Kang H, Chung M K, et al. Persistent brain network homology from the perspective of dendrogram. IEEE transactions on medical imaging, 2012, 31(12): 2267-2277.
[12] Horak D, Maleti S, Rajkovi M. Persistent homology of complex networks. Journal of Statistical Mechanics: Theory and Experiment, 2008, 2009(03): P03034.
[13] Jure Leskovec and Andrej Krevl. SNAP Graph Library: Stanford Large Network Dataset Collection. http://snap.stanford.edu/data. June 2014.
[14] Rossi R, Ahmed N. The network data repository with interactive graph analytics and visualization//Twenty-ninth AAAI conference on artificial intelligence. 2015.
[15] Doyle P G, Snell J L. Random walks and electric networks. American Mathematical Soc. 1984.