Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 815035, 7 pages http://dx.doi.org/10.1155/2013/815035



Research Article

Persistent Homology of Collaboration Networks

C. J. Carstens and K. J. Horadam

School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, VIC 3001, Australia

Correspondence should be addressed to C. J. Carstens; corriejacobien.carstens@rmit.edu.au

Received 29 March 2013; Accepted 19 May 2013

Academic Editor: Tingwen Huang

Copyright © 2013 C. J. Carstens and K. J. Horadam. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Over the past few decades, network science has introduced several statistical measures to determine the topological structure of large networks. Initially, the focus was on binary networks, where edges are either present or not. Thus, many of the earlier measures can only be applied to binary networks and not to weighted networks. More recently, it has been shown that weighted networks have a rich structure, and several generalized measures have been introduced. We use persistent homology, a recent technique from computational topology, to analyse four weighted collaboration networks. We include the first and second Betti numbers for the first time for this type of analysis. We show that persistent homology corresponds to tangible features of the networks. Furthermore, we use it to distinguish the collaboration networks from similar random networks.

1. Introduction

Networks are a useful abstraction for many real-world systems. Some examples are the Internet, communication networks, biological networks, and social networks. Many of these networks are intrinsically weighted [1]. For instance, not all connections in social networks are equal; some people are close friends or family, whereas others are merely acquaintances. By modeling social networks as weighted networks we can analyze the richer structure of links with different strengths [1–3].

There are several approaches to analysing weighted networks. One is to suitably generalize measures for binary networks [2]. Another approach is finding the optimal threshold network for the property of interest. A threshold network is the binary network obtained by setting a threshold weight \boldsymbol{w}^* and keeping only connections with weight higher than \boldsymbol{w}^* . In [4], a range of different thresholds is scanned to find the optimal threshold for detecting a global community structure in a network. The authors regard changing the threshold as changing the resolution at which a network's structure is inspected.

In this paper, we take a different approach. Instead of finding the optimal threshold weight, which is often only

optimal for a specific property, we study all different levels of resolution at once. To do so we use persistent homology, a recent technique from computational topology. The framework of persistent homology records structural properties and their changes for a whole range of thresholds. There are only a few other papers that use persistent homology to analyse networks [5–9] that we are aware of. Both [5, 6] use a different filtration for persistent homology from that of Lee et al.; in particular, they do not use persistent homology to analyse weighted networks. The filtration that we use is the same as in the work by Lee et al., where it was used to compare normal and abnormal brain networks [7–9]. In their work only the zeroth Betti numbers were used.

Here, we include the second and first Betti numbers for the first time. This leads to richer network measures. We show that the first Betti numbers correspond to tangible features of the network and use this richer form of persistent homology to distinguish structured networks from random networks.

2. Persistent Homology of Weighted Networks

In this section we introduce concepts from computational topology in the setting of networks. For a more elaborate introduction to persistent homology we refer to [10, 11].

2.1. Persistent Homology. Persistent homology computes the topological features of a filtration of a space. A filtration of a space can be thought of as the evolution of a space or a growing sequence of spaces. More formally a filtration of a space X is a nested sequence of subspaces beginning with the empty set and ending with X:

$$\emptyset = X_0 \subseteq X_1 \subseteq \cdots X_n = X. \tag{1}$$

See Figure 1(a) for an illustration of a filtration, where X is a triangle. Persistent homology computes the classical homology groups of spaces in such a filtration. In this paper we always use homology with \mathbb{Z}_2 coefficients. We write $H_i(X)$ for the ith homology group of X. (We are being sloppy with our notation here for increased readability but should in fact write $H_i(X; \mathbb{Z}_2)$.) The homology groups with coefficients in \mathbb{Z}_2 will always be of the form $H_i(X) \simeq \mathbb{Z}_2^{\beta_i}$ where β_i is the ith Betti number of X. We are mainly interested in computing these Betti numbers.

Using the inclusion maps $X_j \to X_{j+1}$ we can identify copies of \mathbb{Z}_2 in the homology groups $H_i(X_j)$ and $H_i(X_{j+1})$ of a filtration. This way we can record when a new copy is born, an existing copy persists or dies. The births and deaths correspond to changes in the topology of the filtration. These changes can be depicted as a barcode [11, 12], where the intervals $[b_k, d_k]$ correspond to filtration values of the birth and death of an element in the ith homology group. The longer a topological feature is present in the filtration, the longer we say it persists; see Figure 1(c).

Here, we will restrict our attention to the zero-, one- and two-dimensional homology of spaces. This will reduce our computations significantly, since we do not need to include parts of our space that are higher dimensional than two-dimensional. We will make this statement more precise in the following section.

It is well known that β_0 equals the number of connected components of a space [13]. β_1 and β_2 roughly count the number of loops and voids in a space. We will restrict our results to these dimensions, but there are Betti numbers for all positive $n \in \mathbb{N}$, corresponding to higher dimensional holes in a space. However, for finite spaces most of these will be zero since homology groups are zero in dimensions larger than the dimension of the space itself.

2.2. Weighted Network. A weighted graph is a graph G = (V, E) together with a weight function $w : E \to \mathbb{R}$. As mentioned in the introduction, a weighted graph can be converted to an unweighted graph by keeping only the edges stronger than a certain threshold weight w^* . For a weighted graph G we will denote this threshold subgraph by $G(w^*)$. In every threshold subgraph all of the vertices of G are present.

Note that for two different thresholds $w_i^* > w_j^*$, we obtain an inclusion $G(w_i^*) \subseteq G(w_j^*)$. All edges that are present in $G(w_i^*)$ have weight larger than w_i^* , so larger than w_j^* , and thus they are included in $G(w_j^*)$. For a sequence of weights $w_0 > w_1 > \cdots > w_k$ we obtain a series of graphs and inclusions as follows:

$$\emptyset \subseteq G(w_0) \subseteq G(w_1) \subseteq \dots \subseteq G(w_k) \subseteq G.$$
 (2)

Such a sequence of graph inclusions is called a graph filtration.

Since a graph can be equipped with a topology to turn it into a a one-dimensional space, we can directly apply persistent homology to a graph filtration. We will then obtain nontrivial Betti numbers in dimensions zero and one only.

We can encode more of the topological information of the graph into a higher dimensional space, a simplicial complex. There are many different ways to construct a filtration of simplicial complexes from a graph filtration. A common choice is the clique complex since it reduces computational efforts [11, 14]. (The clique complex is also known as the Vietoris Rips complex and the flag complex. We use the term clique complex as it has more meaning in terms of social networks.)

We obtain the clique complex of a graph by "filling in" all cliques, that is, all complete subgraphs. A 3-clique will turn into a filled triangle and a 4-clique into a solid tetrahedron and similarly for higher dimensional cliques. A nice property of the clique complex is that cliques correspond to highly connected groups of nodes that may represent communities [4]. When computing the first Betti numbers of such a clique complex we count the number of loops in the complex. In the original graph a triangle is a loop and increases the first Betti number by one. In the clique complex all triangles are filled, and the loop is no longer there; see Figure 1(a). This means that all loops that we detect in the clique complex have four or more vertices. The simplest possible loop is formed by four vertices connected as a square with no diagonal connections.

A vertex is also known as a 0-simplex, an edge as a 1-simplex, a triangle as a 2-simplex, and a tetrahedron as a 3-simplex. A face of a simplex σ is a subsimplex of σ . For instance, a triangle has six faces, the three edges and three points in its boundary. A simplicial complex is a set of simplices such that any face of a simplex is also in the simplicial complex and such that the intersection of any two simplices is a face of both.

Let K be the clique complex of a graph G. The 0-skeleton of K is the simplicial complex consisting of just the vertices of G. The 1-skeleton of K is the set of all vertices and edges of G, that is, the graph itself. The 2-skeleton is the set of all vertices, edges, and triangles. In general the i-skeleton of a simplicial complex K is the subcomplex consisting of all j-simplices with $j \leq i$. We denote the i-skeleton by $K^{(i)}$. Notice that for G' a subgraph of G, the i-skeleton of the clique complex $K'^{(i)}$ is a subcomplex of $K^{(i)}$. This means we obtain a filtration of $K^{(i)}$ from a graph filtration of G. And in particular since all our clique complexes are finite dimensional, we obtain a filtration of K.

From the definition of homology we know that the ith homology groups of a simplicial complex and thus the ith Betti numbers are completely determined by the (i+1)-skeleton. In particular this means that to compute the zero-dimensional persistent homology of the clique complex of a graph, we only need the original graph filtration. This is not surprising; the graph contains all connectivity information. Filling in triangles cannot change the number of connected components. Moreover, making use of this fact we can reduce

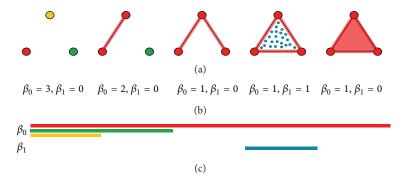


FIGURE 1: A filtration of a triangle (a). We start with three connected components. The yellow and the green components die in step two and three, but the red component persists the whole filtration. In the fourth step a loop is born, which dies in the final step of the filtration. The zeroth Betti number equals the number of connected components. The first Betti number equals the number of loops (b). We use a barcode to visualise the birth and death of the Betti numbers (c).

the computational times for computing Betti numbers in dimensions one and two, by only constructing the clique complex up to the 3-skeleton.

3. Collaboration Networks

We have applied persistent homology to four collaboration networks of scientists [15, 16]. These networks were obtained from http://www-personal.umich.edu/~mejn/netdata/. The four networks were constructed using four collections of papers. The vertices in the network correspond to the authors of the papers. There is a connection between two scientists if they are coauthors on at least one paper. These connections are assigned weights by taking into account how often scientist collaborate and how closely they collaborate. A paper contributes a weight to the connections between all of its authors, however the more authors a paper has the smaller the contributed weight. To be precise, a paper that has n authors contributes a weight of 1/(n-1). Strong connections correspond to people that collaborate often and in small groups.

Through this construction we obtain a network that has a very different weight distribution from a more traditional social network as described by Granovetter [3]. In the latter, one finds communities of strongly connected individuals and weak ties functioning as local bridges between communities.

Instead, in these collaboration networks, weak ties are necessarily part of communities. And in fact, the weaker the tie, the larger the community that it is part of. For example, let two scientists be connected by a weak tie with weight 0.125. This implies that they have coauthored a paper with at least seven other authors (they could also both have appeared on, e.g., two papers with 15 authors). Let us for simplicity assume this is the case. This paper with nine authors corresponds to a 9-clique in our network. All edges in this clique have weight larger or equal to 0.125. If we inspect edges with lower weight than 0.125 we find even more coauthors and larger cliques.

3.1. Collaboration Network of Network Scientists. We will use the network scientists data to explore β_0 and β_1 in detail. We have restricted our persistence computation to

the clique complex of the largest connected component of this collaboration network. This component consists of 379 vertices and 914 edges. The weights in this component range from 0.125 to 4.75.

We will first discuss the zeroth Betti numbers of the clique complex filtration. As discussed in the previous section, we may restrict to the 1-skeleton of the complex for this computation, that is, the graph itself. We start our filtration with $w^*=5$; all vertices of the graph are present but none of the edges are since none of the edges have weight larger than or equal to 5. We immediately find that $\beta_0=379$ since there are 379 connected components, all the individual vertices.

As we lower w^* in our filtration, more and more edges are added to the graph, and β_0 will decrease as the graph becomes more connected. Finally G(0) is connected, so we will end with $\beta_0 = 1$. In Figure 3(a) we can see how β_0 responds to lowering the threshold weight w^* . We have also plotted the number of edges in the graph and noticed that the large decreases in number of connected components correspond to values of w^* where many edges are added.

The network is not connected while $w^*>0.143$. Only after adding the 47 edges with weight equal to 0.143, the network becomes connected; see Table 1. Before adding these edges there are ten components, eight of these consisting of single nodes. We find that these nodes are all part of two 8-cliques; see Figure 2. These authors are only very loosely connected to the rest of the network. We expect that the largest connected component grows rapidly in the filtration and that further lowering the threshold corresponds to adding nodes that are in the periphery of the network. This requires further investigation.

We were curious to see if the zeroth Betti numbers could distinguish this collaboration network from random Erdös-Rényi graphs with the same number of nodes and edges and with the same weights assigned to the edges. We generated 1000 random graphs and used the Bottleneck distance [7, 17] between barcodes to compare the networks. We found that for the random graphs the average pairwise distance was 0.157 (s.d. 0.019) whereas the average distance from the collaboration network to the random graphs was 0.332 (s.d. 0.003). We can definitely use this measure to

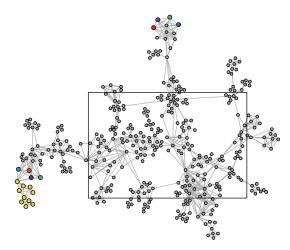


FIGURE 2: Largest connected component of the network science collaboration network. The enlarged nodes are the nodes that join the largest connected component at the lowest filtration value. Their colours correspond to the component they belonged to before this filtration value is reached.

distinguish the two network topologies. Even though at the start and end of the filtration these networks are very similar we can still detect a structural difference by looking at connected components during the filtration of the networks. In Figure 3(b) we have plotted the zeroth Betti numbers of ten random graphs and the zeroth Betti numbers of the networks scientist collaboration network.

Next we inspect the first Betti numbers of the clique complex associated to our network. To do so we built the 2-skeleton, which includes all vertices, edges, and triangles. Note that a filled triangle is added whenever three scientists are pairwise connected. As mentioned in the previous section, without filling in these triangles, each triple of pairwise collaborating scientists would be a loop and increase the first Betti number by one. However, we are interested in the loops in the network on a larger scale. In Figure 4 we illustrate the final stages of the graph filtration where the first Betti number is nonzero. We show the correspondence between the loops in the complex and the barcode we have computed. We find the largest loops for the lowest threshold weights, Figure 3(c). However, many of the edges that are part of these loops have high weights.

We investigated if the first Betti numbers give us further power to distinguish between the collaboration network and the random networks. We found that for random networks we obtain much higher first Betti numbers. For 1000 randomly generated networks we found an average of 520.65 (s.d. 4.39) intervals, while our structured network only has 9 intervals. The reason that this number is so much higher for random networks is that there is less clustering and thus fewer triangles that are filled in and more loops with more than three edges.

Using the first Betti numbers it is enough to only compare the final networks to distinguish between random and structured networks. We hope that the persistent homology of the whole filtration will be able to detect more subtle structural differences to distinguish networks that are more similar in structure. Notice how all of the loops that were born persisted to the end of the filtration. It would have been possible for a loop to die. For instance, if the four scientists (A. Vazquez, A. Vespignani, A. Barrat, and M. Weigt) appearing in the red loop found at $w^*=1$ would have collaborated on a paper, there would be diagonal edges appearing at $w^*=0.33$ which would kill the loop.

For this network all higher Betti numbers are trivial.

3.2. Physics Collaboration Networks. In this section we perform analysis on three larger collaboration networks. Again we restrict our attention to the largest connected component of each network. In Table 1 the number of nodes and edges for all of these networks are given. We computed the barcodes for the first three Betti numbers; β_0 , β_1 , and β_2 . We found that β_0 stayed high for the largest part of the filtration and then quickly decreased to 1 at the end of the filtration in all three cases. In all cases, the smallest weight was needed to create the connected component; see Table 1. This is slightly different behaviour from the network scientist collaboration network.

We investigated if we can distinguish these collaboration networks from random networks using the persistence barcodes. We noticed that all three networks have several intervals corresponding to second Betti numbers.

Let G(n,p) be an Erdös-Rényi graph with p the probability of an edge being present, that is, $p \sim 2m/n(n-1)$. Erdös and Rényi showed that if $p \gg \log n/n$ then G(n,p) is almost always connected [18]. In [19], Kahle shows that there are analogous results for higher dimensional connectivity of the clique complexes of random graphs. In particular, if we define α by $p = n^{\alpha}$, Kahle shows that the kth homology group of a clique complex of a random graph is almost always zero if α is outside the interval (-1/k, -1/(2k+1)). In Table 2 the final values α_f for the three collaboration networks can be found.

A filtration of a random network corresponds to increasing p over time, or increasing α from $-\infty$ to α_f . For the clique complex of a random network G(n,p) we expect the second Betti number to be zero for $\alpha<-0.5$. For all three networks the value of α satisfies this condition. However, we find a large number of intervals for both the condensed matter network and the astrophysics network. This clearly distinguishes these networks from random networks. Inspection of the zeroth and first Betti numbers is ongoing research.

4. Software

We used Gephi [20] for some basic graph manipulations and for graph visualisations. For the persistence computations we used javaPlex [21]. This package was developed to compute the persistent homology of point cloud data. In these computations one starts with a collection of points embedded in Euclidean space, then associates a graph filtration to these points, and finally builds a filtration of simplicial complexes of which the persistent homology is computed.

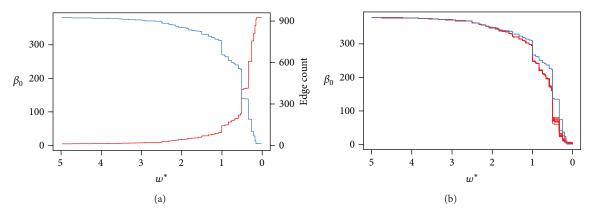


FIGURE 3: On the left (a) we plot the zeroth Betti number against the threshold w^* in blue. The total number of edges present at each stage of the filtration is plotted in red. On the right (b) we again plot the zeroth Betti number in blue. There are ten red plots, each corresponding to the sequence of zeroth Betti numbers of random graphs with the same number of vertices, edges and the same weights.

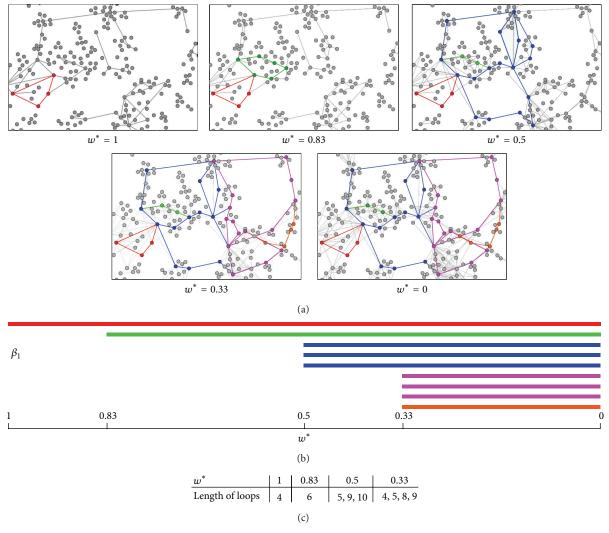


FIGURE 4: We only show the central part, see Figure 2, of the network of 379 network scientists since all loops occur here (a). We find the first relatively small loop between four scientists for appearing for threshold weight 1. As we decrease the threshold weight, more loops appear. Notice how we have shaded two triangles for $w^* = 0.5$; this is to indicate that there is no loop there; there are three new blue loops added at this stage. We notice that for smaller threshold values we find larger loops. In (b) we show the barcode for the first Betti numbers of this filtration. In (c) we enlist the length of the loops that appear at each filtration value.

Network No. of nodes No. of edges w^* No. of edges = w^* No. of edges $< w^*$ Network science 379 914 0.143 47 10 Condensed matter 36458 171735 0.034 315 0 171 0 High-energy theory 5835 13815 0.056 Astrophysics 14845 119652 0.018 357 0

TABLE 1: Collaboration networks.

TABLE 2: Collaboration networks.

Network	No. of intervals β_1	No. of intervals β_2	Р	α
Condensed matter	11361	274	0.00026	-0.79
High-energy theory	1389	2	0.00081	-0.82
Astrophysics	4879	222	0.0011	-0.71

We wrote code in JAVA that imports a weighted edge list and converts it to a graph filtration. Subsequently we used javaPlex to build the clique complex filtration and compute the persistence intervals. The computation of the persistence intervals is the bottleneck in this computation. This took longest for the astrophysics network; 267 s (on a MacBook Pro 2.4 GHz Intel Core 2 Duo with 4 GB RAM), presumably since it is the densest network. For our current purposes these computation times are sufficient; however, if we want to apply the same computations to larger networks we need faster algorithms. This should be possible as described in Chapter 12 of [22]. To generate the random networks with *n* vertices and *m* edges we wrote code that randomly picks endpoints for *m* edges, avoiding double edges and loops. We used the Random Utility class from javaPlex to pick these endpoints.

5. Conclusions

By applying persistent homology to four collaboration networks of scientists we have shown that it gives us interesting information about the structure of weighted networks. We found that due to the construction of collaboration networks, weak ties form cliques and strong ties act as local bridges between those cliques. This is contrary to what has been described in other social networks. We would like to investigate this in greater detail in future work.

We used persistent homology to analyse the structure of weighted networks. The inclusion of the first and second Betti numbers gave us a richer measure to work with than in the existing literature. We were able to use persistent homology to distinguish these collaboration networks from random networks. Using the one- and two-dimensional Betti numbers of the network we did not need to take the weights into account. We are hoping that using the weights will give us the ability to distinguish networks that are more similar in structure. This is left as future work.

Acknowledgment

Research by both authors was partially supported by the Australian Department of Defence under Research Agreement 4500743680.

References

- [1] M. E. J. Newman, "Analysis of weighted networks," *Physical Review E*, vol. 70, no. 5, Article ID 056131, 9 pages, 2004.
- [2] T. Opsahl and P. Panzarasa, "Clustering in weighted networks," *Social Networks*, vol. 31, no. 2, pp. 155–163, 2009.
- [3] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1983.
- [4] G. Palla, I. Derényi, I. Farkas, and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, vol. 435, no. 7043, pp. 814–818, 2005.
- [5] V. de Silva and R. Ghrist, "Coverage in sensor networks via persistent homology," *Algebraic & Geometric Topology*, vol. 7, pp. 339–358, 2007.
- [6] D. Horak, S. Maletić, and M. Rajković, "Persistent homology of complex networks," *Journal of Statistical Mechanics*, vol. 2009, no. 3, Article ID P03034, 2009.
- [7] H. Lee, H. Kang, M. K. Chung, B. N. Kim, and D. S. Lee, "Persistent brain network homology from the perspective of dendrogram," *IEEE Transactions on Medical Imaging*, vol. 31, no. 12, pp. 2267–2277, 2012.
- [8] H. Lee, M. K. Chung, H. Kang, B.-N. Kim, and D. S. Lee, "Discriminative persistent homology of brain networks," in Proceedings of the 8th IEEE International Symposium on Biomedical Imaging: from Nano to Macro (ISBI '11), pp. 841–844, April 2011.
- [9] H. Lee, M. K. Chung, H. Kang, B. N. Kim, and D. S. Lee, "Computing the shape of brain networks using graph filtration and Gromov-Hausdorff metric," in *Proceedings of the 14th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'11)*, vol. 6891, pp. 289–296, 2011.
- [10] H. Edelsbrunner and J. Harer, "Persistent homology—a survey," in Surveys on Discrete and Computational Geometry. Twenty Years Later, vol. 453 of Contemporary Mathematics, pp. 257–282, American Mathematical Society, Providence, RI, USA, 2008.
- [11] G. Carlsson, "Topology and data," *Bulletin American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009.
- [12] R. Ghrist, "Barcodes: the persistent topology of data," *Bulletin American Mathematical Society*, vol. 45, no. 1, pp. 61–75, 2008.
- [13] A. Hatcher, Algebraic Topology, Cambridge University Press, Cambridge, Mass, USA, 2002.
- [14] A. Zomorodian, "Fast construction of the Vietoris-Rips complex," *Computers & Graphics*, vol. 34, no. 3, pp. 263–271, 2010.

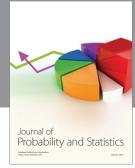
- [15] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 2, pp. 404–409, 2001.
- [16] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, Article ID 036104, 19 pages, 2006.
- [17] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, "Stability of persistence diagrams," *Discrete & Computational Geometry*, vol. 37, no. 1, pp. 103–120, 2007.
- [18] P. Erdős and A. Rényi, "On random graphs. I," *Publicationes Mathematicae Debrecen*, vol. 6, pp. 290–297, 1959.
- [19] M. Kahle, "Topology of random clique complexes," *Discrete Mathematics*, vol. 309, no. 6, pp. 1658–1671, 2009.
- [20] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," 2009, http://gephi.org.
- [21] A. Tausz, M. Vejdemo-Johansson, and H. Adams, "Javaplex: a research software package for persistent (co)homology," 2011, http://code.google.com/p/javaplex/.
- [22] A. J. Zomorodian, Topology for Computing, vol. 16 of Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, New York, NY, USA, 2005.



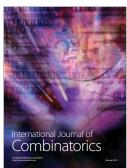










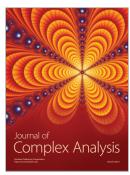


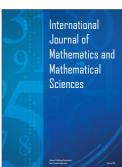


Submit your manuscripts at http://www.hindawi.com











Journal of Discrete Mathematics

