Persistent Homology of Collaboration Networks

C. J. Carstens
School of Mathematical and Geospatial Sciences
RMIT University
Melbourne 3001, Australia
corriejacobien.carstens@rmit.edu.au

K. J. Horadam
School of Mathematical and Geospatial Sciences
RMIT University
Melbourne 3001, Australia
kathy.horadam@rmit.edu.au

Abstract

We apply persistent homology to four collaboration networks. We show that the intervals for the zeroth and first Betti numbers correspond to tangible features of the structure of these networks. Finally, we use persistent homology to distinguish collaboration networks from similar random networks.

1 Introduction

Connections in social networks have different strengths. Some people in a social network are close friends or family, others are colleagues and yet others are merely acquaintances. These differences carry interesting information about the structure of a social network. As argued by Granovetter in his famous paper [1], the ties of different strength also perform different functions. Strong ties are found in clusters or groups, and the weak ties connect these groups.

We may use weighted networks to model a social network with different strengths. Weighted networks are often transformed to unweighted networks by setting a threshold weight w^* and keeping only connections that are stronger than this weight. For instance in [2] a range of different thresholds is scanned to find the optimal threshold for detecting a global community structure in the network. The authors mention that changing the threshold is like changing the resolution at which we inspect the network's structure.

In this paper we discuss a recent technique from the field of computational topology called *persistent homology* that studies all these different levels of resolution *at once*. Instead of finding the optimal threshold weight, which is often only optimal for a specific property, the framework of persistent homology records structural properties and their changes for a whole range of thresholds.

Persistent homology was not developed with networks in mind, but can be applied in this setting very naturally. For instance, it was previously used to compare normal and abnormal brain networks [3–5]. However, in this work only the zeroth Betti numbers were used. To be able to distinguish two types of abnormal brain networks from normal brain networks the authors needed to include geometrical information by constructing the single linkage dendrograms. Using the Gromov-Hausdorff distance between dendrograms resulted in perfect clustering accuracy. We use the same network filtration as Lee et al. but include the second and first Betti numbers as well as the zeroth Betti numbers.

In another paper [6] persistent homology was applied to random, exponential and scale-free networks. There the networks were unweighted and the filtration used is the filtration of i-skeletons. This leads to a different type of barcode, in particular as the j-th homology group is zero for all i-skeletons with $i < j$ and only depends on the j and $j+1$-skeleton of a complex, all intervals in the barcode should either persist for just one step in the filtration or forever.

There are already many invariants in use to study the topology of networks. Examples include average path length, small world property, scale-free degree distribution and clustering coefficients. However we think that the robustness and theoretical grounding of persistent homology can give us new insights into the structure of networks.

2 Persistent homology of weighted networks

In this section we will introduce concepts from computational topology in the setting of networks. For a more elaborate introduction to persistent homology we refer to [7, 8].

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.

$$\emptyset = X_0 \subseteq X_1 \subseteq \cdots X_n = X$$

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 i-th homology group of X.[1] 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 *i-th Betti number of X*. We are mainly interested in computing these Betti numbers.

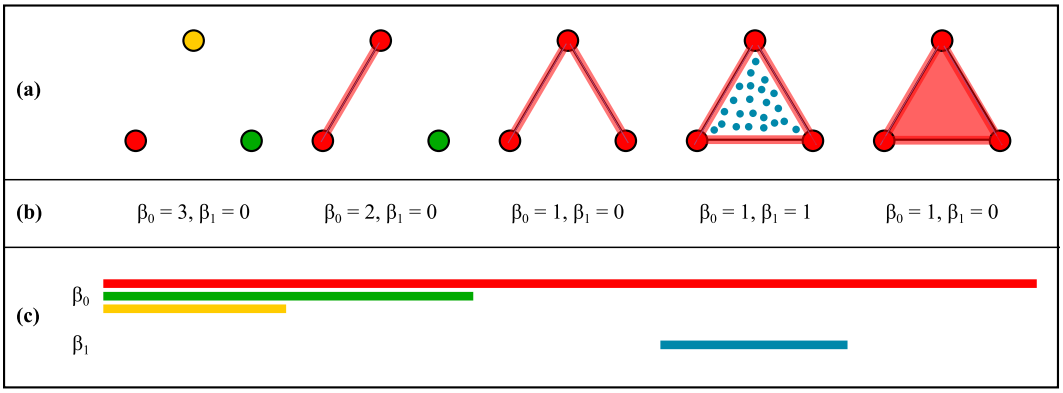


Figure 1: A filtration of a triangle **(a)**. We start with three connected components. The yellow and the green component die in step two and three, but the red component persists the whole filtration. In the fourth step a loop is born, which dies in the final step of the filtration. The zeroth Betti 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)**.

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 [8, 9], where the intervals $[b_k, d_k]$ correspond to filtration values of the birth and death of an element in the i-th homology group. The longer a topological feature is present in the filtration, the longer we say it persists. See Figure 1(c).

[1] We are being sloppy with our notation here for increased readability, but should in fact write $H_i(X; \mathbb{Z}_2)$.

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 [10]. 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.

$$\emptyset \subseteq G(w_0) \subseteq G(w_1) \subseteq \cdots \subseteq G(w_k) \subseteq G$$

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 non-trivial Betti numbers in dimension 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*[2] since it reduces computational efforts. [8, 11].

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 [2]. 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, i.e. 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.

[2] 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.

From the definition of homology we know that the i-th homology groups of a simplicial complex, and thus the i-th 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 can not change the number of connected components. Moreover making use of this fact we can reduce the computational times for computing Betti numbers in dimension 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 [12, 13]. All networks are constructed using a collection of papers. The vertices in the network correspond to the authors of the papers. There is a connection between two scientists if they are co-authors on at least one paper. For each paper that the scientists have collaborated on a weight of $1/(n-1)$ is added to the edge. Here n is the number of authors of the paper. Strong connections correspond to people that collaborate often and in small groups.

Using the above construction we obtain a network that has a very different weight distribution from a more traditional social network as described by Granovetter [1]. In the latter one finds communities of strongly connected individuals and weak ties functioning as local bridges between communities.

Instead, in our collaboration network, 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 scientist be connected by a weak tie with weight 0.125. This implies that they have co-authored a paper with at least seven other authors [3]. 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 co-authors 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.

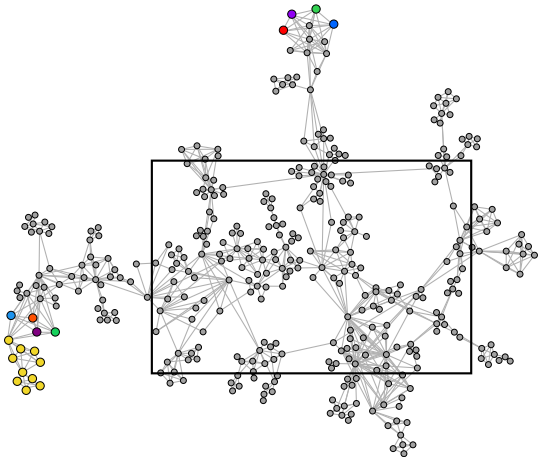


Figure 2: Largest connected component of the network science collaboration network. The enlarged nodes are the nodes that join the largest connected component at the lowest filtration value. Their colours correspond to the component they belonged to before this filtration value is reached.

[3] They could also both have appeared on for instance two papers with 15 authors.

Table 1: Collaboration networks

Network	**#nodes**	**#edges**	w^*	**#edges =** $w*$	**#edges** $< w*$
Network Science	379	914	0.143	47	10
Condensed Matter	36458	171735	0.034	315	0
High-energy Theory	5835	13815	0.056	171	0
Astrophysics	14845	119652	0.018	357	0

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, i.e. 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 notice that the large decreases in number of connected components correspond to values of w^* where many edges are added.

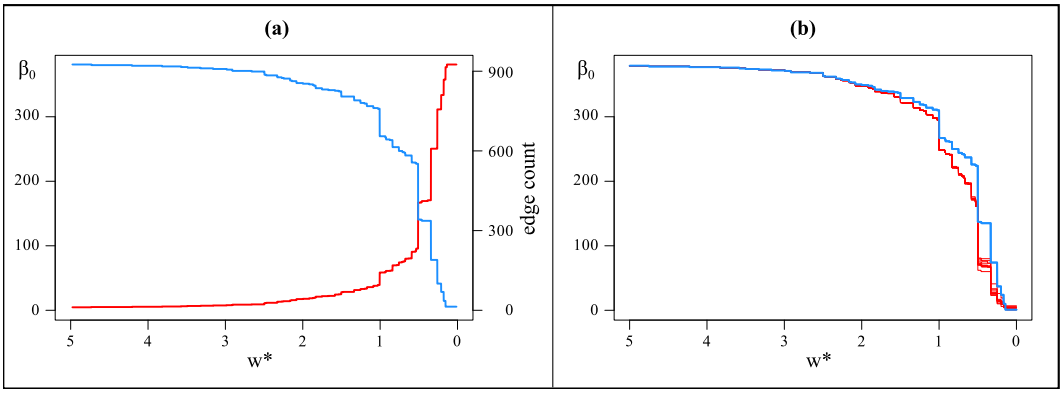
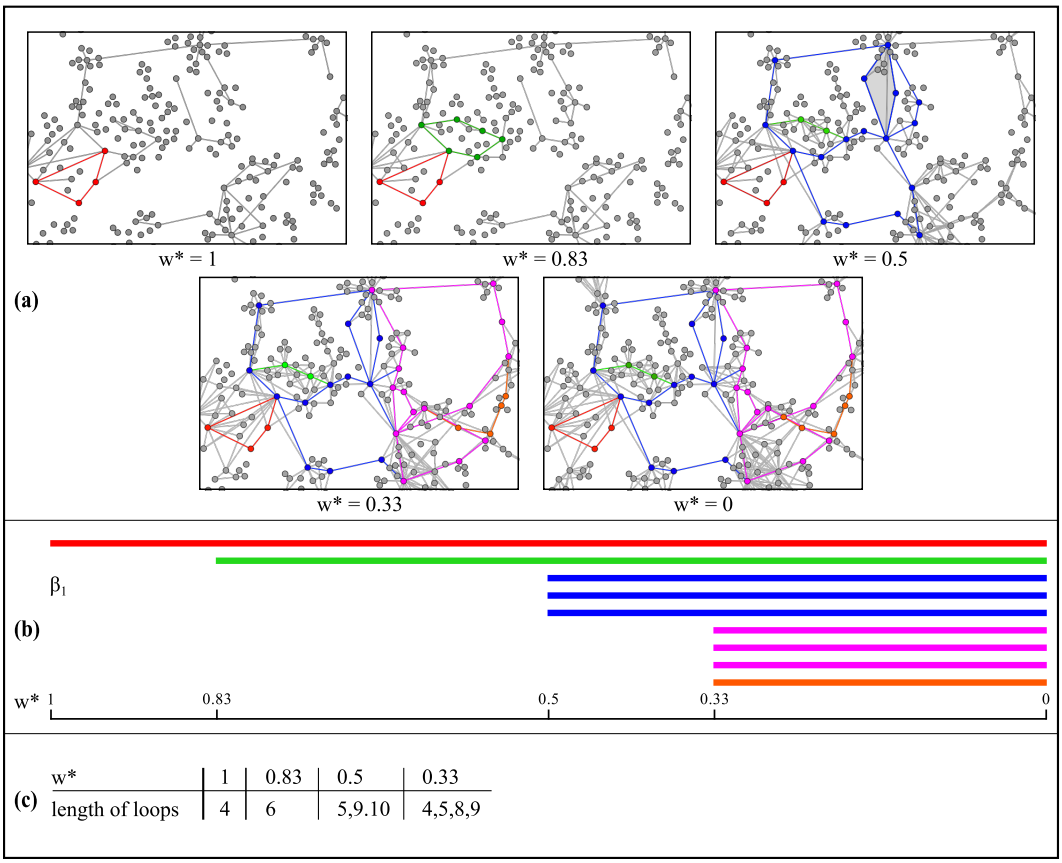


Figure 3: On the left **(a)** we plot the zeroth Betti number against the threshold w^* in blue. The total number of edges present at each stage of the filtration is plotted in red. On the right **(b)** we again plot the zeroth Betti number in blue. There are ten red plots, each corresponding to the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

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-Réyni 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 [3, 14] 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 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 scientist 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 non-zero. 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.



(c)

w*	1	0.83	0.5	0.33
length of loops	4	6	5,9,10	4,5,8,9

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.

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 difference 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,

Table 2: Collaboration networks

Network	**# intervals β_1**	**# intervals β_2**	**p**	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

A. Barrat, 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; $\beta_0, \beta_1, \beta_2$. 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 to the network scientist collaboration network.

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

Let $G(n,p)$ be an Erdös-Réyni graph with p the probability of an edge being present, i.e. $p \sim \frac{2m}{n(n-1)}$. Erdös and Réyni showed that if $p \gg \log n/n$ then $G(n,p)$ is almost always connected [15]. In [16], 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 k-th homology group of a clique complex of a random graph is almost always zero if α is outside the interval $(\frac{-1}{k}, \frac{-1}{2k+1})$. 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 $\alpha < -0.5$. This is the case for all values of α in our filtration. 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 [17] for some basic graph manipulations and for graph visualisations. For the persistence computations we used javaPlex [18]. 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.

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; 267s (on a MacBook Pro 2.4 GHz Intel Core 2 Duo with 4GB 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 [19]. 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 RandomUtility 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 the 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 were also 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.

Acknowledgments

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

References

[1] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78:1360–80, 1973.

[2] G. Palla, I. Derényi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435:814–818, 2005.

[3] H. Lee, H. Kang, M. Chung, B. Kim, and D. Lee. Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging*, 2012. To appear, available online at `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6307875`.

[4] H. Lee, M.K. Chung, H. Kang, B.N. Kim, and D.S. Lee. Discriminative persistent homology of brain networks. In *Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on*, pp. 841–844, 2011.

[5] H. Lee, M. Chung, H. Kang, B. Kim, and D. Lee. Computing the shape of brain networks using graph filtration and Gromov-Hausdorff metric. In *Medical Image Computing and Computer Assisted Intervention (MICCAI), 14th International Conference on*, pp. 289–296, 2011.

[6] D. Horak, S. Maletić, and M. Rajković. Persistent homology of complex networks. *Journal of Statistical Mechanics: Theory and Experiment*, 03:P03034, 2009.

[7] H. Edelsbrunner and J. L. Harer. Persistent homology - a survey. In *Surveys on Discrete and Computational Geometry. Twenty Years Later*, pp. 257–282. American Mathematical Society, 2008.

[8] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–306, 2009.

[9] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin American Mathematical Society*, 45(1):61–75, 2008.

[10] A. Hatcher. *Algebraic Topology*. Cambridge University Press, New York, NY, 2002.

[11] A. J. Zomorodian. Fast construction of the Vietoris-Rips complex. *Computers & Graphics*, 34(3):263–271, 2010.

[12] M.E.J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[13] M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74(3):036104, 2006.

[14] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

[15] P. Erdös and A. Rényi. On random graphs i. *Publ. Math. Debrecen*, 6:290–297, 1959.

[16] M. Kahle. Topology of random clique complexes. *Discrete Mathematics*, 309(6):1658–1671, 2009.

[17] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. Software available at http://gephi.org, 2009.

[18] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. Javaplex: A research software package for persistent (co)homology. Software available at http://code.google.com/javaplex, 2011.

[19] A. J. Zomorodian. *Topology for Computing*. Cambridge University Press, England, 2005.