Applied Network Science

RESEARCH Open Access

Co-occurrence simplicial complexes in mathematics: identifying the holes of knowledge

Vsevolod Salnikov[1], Daniele Cassese[1,2,3*], Renaud Lambiotte[3] and Nick S. Jones[4]

*Correspondence:
daniele.cassese@unamur.be
[1]University of Namur and NaXys, Rempart de la Vierge, 5000 Namur, Belgium
[2]ICTEAM, University of Louvain, Av Georges Lemaître, 1348 Louvain-la-Neuve, Belgium
Full list of author information is available at the end of the article

Abstract

In the last years complex networks tools contributed to provide insights on the structure of research, through the study of collaboration, citation and co-occurrence networks. The network approach focuses on pairwise relationships, often compressing multidimensional data structures and inevitably losing information. In this paper we propose for the first time a simplicial complex approach to word co-occurrences, providing a natural framework for the study of higher-order relations in the space of scientific knowledge. Using topological methods we explore the conceptual landscape of mathematical research, focusing on homological holes, regions with low connectivity in the simplicial structure. We find that homological holes are ubiquitous, which suggests that they capture some essential feature of research practice in mathematics. k-dimensional holes die when every concept in the hole appears in an article together with other $k+1$ concepts in the hole, hence their death may be a sign of the creation of new knowledge, as we show with some examples. We find a positive relation between the size of a hole and the time it takes to be closed: larger holes may represent potential for important advances in the field because they separate conceptually distant areas. We provide further description of the conceptual space by looking for the simplicial analogs of stars and explore the likelihood of edges in a star to be also part of a homological cycle. We also show that authors' conceptual entropy is positively related with their contribution to homological holes, suggesting that *polymaths* tend to be on the frontier of research.

Keywords: Co-occurrence, Topological data analysis, Persistent homology

Introduction

Co-occurrence networks capture relationships between words appearing in the same unit of text: each node is a word, or a group of words, and an edge is defined between two nodes if they appear in the same unit of text. Co-occurrence networks have been used, among other things, to study the structure of human languages (Ferrer-i-Cancho and Solé 2001), to detect influential text segments (Garg and Kumar 2018) and to identify authorship signature in temporal evolving networks (Akimushkin et al. 2017). Other applications include the study of co-citations of patents (Wang et al. 2011), articles (Lazer et al. 2009) and genes (Jenssen et al. 2001; Mullen and et al. 2014). Here we focus on the co-occurrences of concepts (theorems, lemmas, equations) in scientific articles to gain understanding in the structure of knowledge in Mathematics. Similar problems have been

SpringerOpen

considered in scientometrics, even if previous works have limited their analysis to keywords, or words appearing in abstracts (Radhakrishnan et al. 2017; Zhang et al. 2012; Su and Lee 2010), and focused only on binary relations between words, as we clarify below.

The main novelty of our work is to study co-occurrences in a simplicial complex framework, using persistent homology to understand the conceptual landscape of mathematics. The adoption of a simplicial complex framework is motivated by the fact that concepts are inherently hierarchical, so simplicial complexes might seem a natural representation: often elementary conceptual units connect together to form nested sequences of higher-order concepts. A simplicial complex approach to model the semantic space of concepts was already suggested by (Chiang 2007), even if not in a topological data analysis framework (Patania et al. 2017b), while application of topological data analysis tools to visualisation of natural language can be found in (Jo et al. 2011; Wagner and et al. 2012; Sami and Farrahi 2017). Several reasons motivate the use of higher-order methods in this context. First, co-occurrence networks tend to be extremely dense in practice and require additional tools to filter the relations and sparsify the network to extract information (Serrano et al. 2009; Slater 2009). Second, in the original dataset, interactions are not pairwise and it is unclear if the constraints induced by a network framework, in terms of nodes and pairwise edges, do not obscure important structures in the system. By modelling co-occurrence relations as a simplicial complex, we thus go beyond the network description that reduces all the structural properties to pairwise interactions and their combinations, explicitly introducing higher-order relations. Note that this modelling approach, in particular the use of simplicial persistent homology, has found uses when the data is inherently multidimensional (Petri et al. 2013), with applications in neuroscience (Petri et al. 2014; Stolz et al. 2017), biology (Chan et al. 2013; Mamuye et al. 2016) to the study of contagion (Taylor and et al. 2015) and to coauthorship networks (Patania et al. 2017a; Carstens and Horadam 2013).

A second contribution of our work is the analysis of the full text of a large corpus of articles, which allows us to bypass the high-level categorisation provided by keywords but also to identify the use of methodological tools and to gain insight into mathematical praxis. However, the main purpose of this article is to use the resulting dataset of concepts and articles as a testbed in which to apply methods from topological data analysis, and to go beyond a standard network analysis.

Dataset

The dataset analysed has been scraped from arXiv, and includes a total of 54177 articles from 01/1994 to 03/2007, of which 48240 in mathematics (`math`) and 5937 in mathematical physics (`math-ph`). We have limited the timeframe due to naming conventions in arXiv: since 03/2007 subject is not a part of the article identifier, thus if one wants to export it additional queries to metadata are needed. That is easily expandable, but we decided to limit the dataset at this moment for computational speed. The date is extracted from article id, hence it refers to the submission date. Notice that some of the articles in the first years may have been written some years before 1991 (arXiv first article's date). In order to describe the mathematical content of articles from the LaTeX file we look at different concepts occurrences in the text. Clearly the choice of the concepts set can influence the outcome: choosing them manually by a small group of people would result in a strong bias towards the understanding and priorities of the individuals in the group.

Thus we wanted to have something either globally accepted by scientific community or at least created by a sufficiently large group of people. Another point in the selection of a good concepts list is the possibility to make a similar research for other disciplines, thus we chose to get it from some general, easily accessible source. Our strategy consisted in parsing a concepts list from Wikipedia, which includes 1612 equations, theorems, lemmas. Clearly these concepts are not homogeneous, meaning that some of them might represent extremely specific theorems, while others can be very general, like *differential equation*, but similar holds for any text processing with different words having different frequencies. Our position on that is still to minimize the manually introduced bias: we consider that all concepts have similar weight and try to have as complete set as possible. Moreover it is possible that two different names represent for example the same theorem due to historical reasons. For the moment we consider such synonyms as distinct entities as the usage of one of them but not the other may reflect structural properties: for example a lemma might have different names depending on a (sub-)field of mathematics and manually merging them is not correct.

As a next step, we combine both datasets. Among the whole concepts list, 1067 find a match in at least one article. Among the 54177 articles, 35018 contain at least one of the concepts in our list (30369 for mathematics and 4649 for mathematical physics), and we also take the list of authors to analyse their contribution to the conceptual space. We construct the binary (non-weighted) co-occurrence simplicial complex (defined more formally below) over the 1067 nodes by including a $(k-1)$-simplex for each article containing k concepts, provided its concept set is not fully included in the concept set of another article, that is we only keep *facets* of the simplicial complex. Whenever the concept sets of two articles intersect, their corresponding simplices share a face of dimension $(n-1)$, where n is the dimension of the intersection. The corresponding network, namely the 1-skeleton of the co-occurrence simplicial complexes (that is we only look at faces of dimension 0 and 1) has 1067 nodes and 32707 unweighted edges. Figure 1 shows the network and simplicial (concept) degree distribution, where the simplicial degree of a concept is the number of *facets* (articles) it belongs to. The sum of all simplicial degrees is 42009, which means there are 39.37 papers per concept on average.

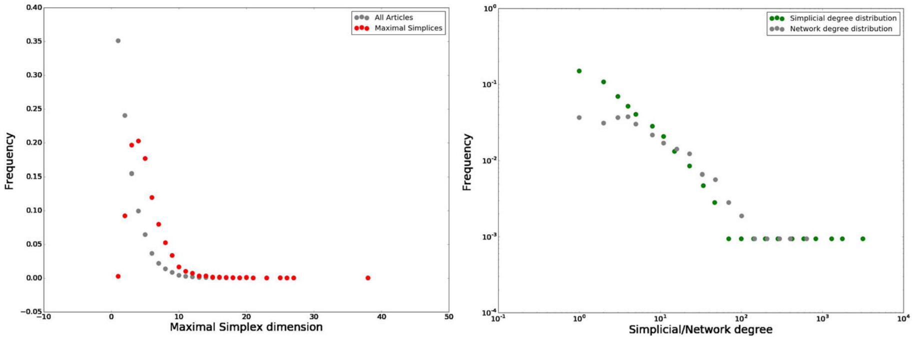

Fig. 1 Dimension of maximal simplices and simplicial degree. On the lefthand side the distribution of number of concepts per article. Here we compare the distribution of the entire dataset (grey) with that of the papers included in the simplicial complex, which shows that most of the articles with few concepts in are not included as they are a subset of other articles. On the righthand side the simplicial and network degree distribution are showed

Simplicial complexes

A simplicial complex is a space obtained as the union of simple elements (nodes, edges, triangles, tetrahedra and higher dimensional polytopes). Its elements are called simplices, where a k-simplex is a set of $k+1$ distinct nodes and its subsets of cardinality $d \leq k$ are called its *d-faces*. We say that two simplices intersect if they share a common face. More formally:

Definition *Let V be a set of vertices, then a n-dimensional simplex is a set of cardinality* $n+1$ *of distinct elements of V,* $\{v_0, v_1, \ldots, v_n\}$*,* $v_i \in V$*. A simplicial complex is a collection K of simplices such that if* $\sigma \in K$ *and* $\tau \subset \sigma$ *then* $\tau \in K$*, so for every simplex in K all its faces are also in K. The k-skeleton of K is the union of all simplices in K up to dimension k.*

Simplicial complexes can be seen as generalisation of a network beyond pairwise interactions, that differ from hypergraphs as all subsets of a simplicial complex must also be simplices. As an illustrative example of how simplicial complexes capture higher-order interactions where networks fail to do so, consider that in a co-occurrence network it is not possible to distinguish between three concepts appearing in the same paper and three concepts appearing in three papers each containing two concepts: in a network both cases are represented by a triangle, while in a simplicial complex the first is a 2-simplex (a filled triangle) and the second is a cycle made of three 1-simplices (an empty triangle).

As for networks, also for simplicial complexes we can define simplicial measures that are the higher-order analogs of networks ones, for example (Estrada and Ross 2018) defines several simplicial centrality measures, providing also the characterisation of some families of simplicial complexes. In this paper we use the simplicial analogs of stars to provide a further description of the concept space.

Definition *A simplicial star* S^k_l *consists of a central* $(k-1)$*-simplex that is a face of l k-simplices, and there is no other simplex but their subsimplices.*

S^1_5 for example is the usual star, with a node in the core connected to 5 nodes, while for S^3_5 the core is a triangle. Examples are reported in Fig. 2.

Persistent homology

Persistent homology is a method in topological data analysis (Carlsson 2009; Patania et al. 2017b) based on algebraic topology, that studies the shape of the data by finding holes of different dimensions in the dataset (Table 1). Holes are topological invariants that can be seen as voids bounded by simplices of different dimension: in dimension 0 they are connected components, in dimension 1 loops (voids bounded by edges), in dimension 2 voids bounded by triangles and so on. Here we give a brief and intuitive explanation of the homology of simplicial complexes, for details on how to compute it we refer to (Edelsbrunner and Harer 2008; Horak et al. 2009; Otter et al. 2017).

For every k-simplex of a simplicial complex K, consider the simplicial analog of a path, a k-chain, simply the formal sum of adjacent k-simplices (where by adjacent we mean that they share one $(k-1)$ face) with coefficients in some algebraic ring R uniquely identifying

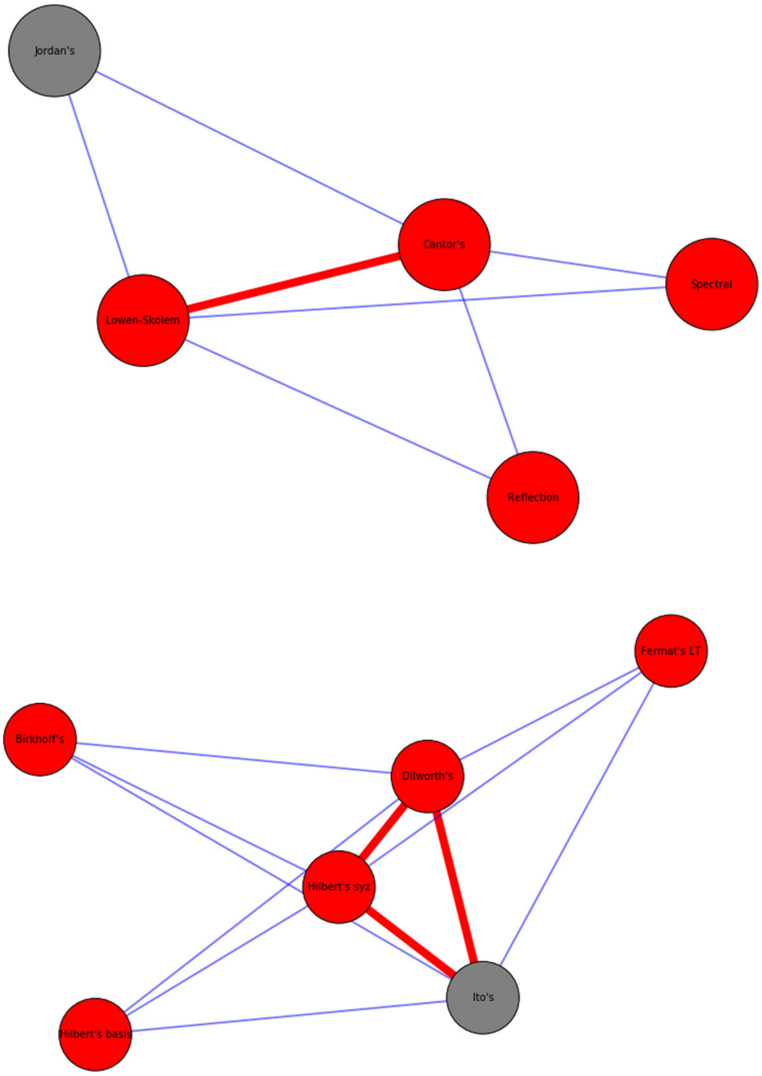

Fig. 2 Simplicial stars. Example of a S_3^2 (top) and a S_3^3 (bottom) star. Red nodes are theorem, grey nodes lemmas, while red edges delimit the simplices in core of the star

the chain (for example a 1-chain is a formal sum of oriented edges). It is a common practice to consider Z or Z_n for R and negative coefficients change the orientation. Without any limitations and for the sake of simplicity one can imagine $R = Z_2$ as it permits us to eliminate questions of orientation: in this case $-1S = S$ for any simplex S. If we consider a k-simplex, it is bounded by its $(k-1)$ faces, and we call the corresponding $(k-1)$-chain, equal to the sum of these faces with coefficients 1 and coherent orientations, the boundary of that simplex. Again to make an illustrative example, the boundary of a 2-simplex (filled triangle) is then a formal sum of its 1-faces (edges). The boundary for a general

Table 1 Dataset

	Years	Papers	Concepts	Authors
Total	1994 - 2007	35018	1612	23471
Included	1994 - 2007	8375	1067	8852

k-chain is defined as the sum of the boundaries of the simplices in the chain taken with corresponding coefficients. Consider the linear map on the space of k-simplices, mapping each k-simplex to its boundary, the *boundary operator*

$$\partial_k : C_k(K) \to C_{k-1}(K)$$

defined on the vector space with basis given by the simplices of K. A k-cycle is defined as a k-chain without a boundary, hence it is an element of the kernel of ∂_k, and a k-boundary is a k-chain which is the boundary of a $(k+1)$-chain, so it is an element of the image of ∂_{k+1}, which is a subset of the kernel of ∂_k as the boundary of a boundary is empty, or $\partial_k \partial_{k+1} = 0$.

So we have defined two interesting subspaces: the collection of k-cycles and the collection of k-boundaries, and we can also take the quotient space as the second is a subset of the first: what is left in the quotient space are those k-cycles that do not bound $(k+1)$-subcomplexes, and these are the k-dimensional voids. More precisely, as there can be more k-cycles around the same hole, the elements of the quotient space can be divided in homological classes, each identifying a hole. This quotient space is the kth homology of the simplicial complex

$$H_k(K) = \frac{\ker(\partial_k)}{\mathrm{Im}(\partial_{k+1})}$$

and its dimension

$$\beta_k(K) = \dim \ker(\partial_k) - \dim \mathrm{Im}(\partial_{k+1})$$

is the number of *homology classes* or k-dimensional voids in the simplicial complex, the k-th *Betti number* of the homology. For example the zero*th* Betti number counts the number of connected components in the graph that constitutes the 1-skeleton of the simplicial complex, the first Betti number the number of loops, the second Betti number counts voids.

To gain some intuition, consider Fig. 3: $\{ab, ac, bc\}$ and $\{ac, ad, cd\}$ are two 1-chains, of which $\{ac, ad, cd\}$ is the boundary of the (filled) triangle acd while $\{ab, ac, bc\}$ is not in the boundary of any 2-chain. So $\{ab, ac, bc\}$ is a 1-dimensional homological cycle H_1. On the other hand the 2-chain $\{bce, bcf, bef, cef\}$ is the boundary of the (filled) tetrahedron $bcef$ hence it is not a 2-dimensional cycle. In Fig. 4 the same 2-chain is a 2-dimensional homological cycle H_2 as there the tetrahedron $bcef$ is not in the simplicial complex anymore (as

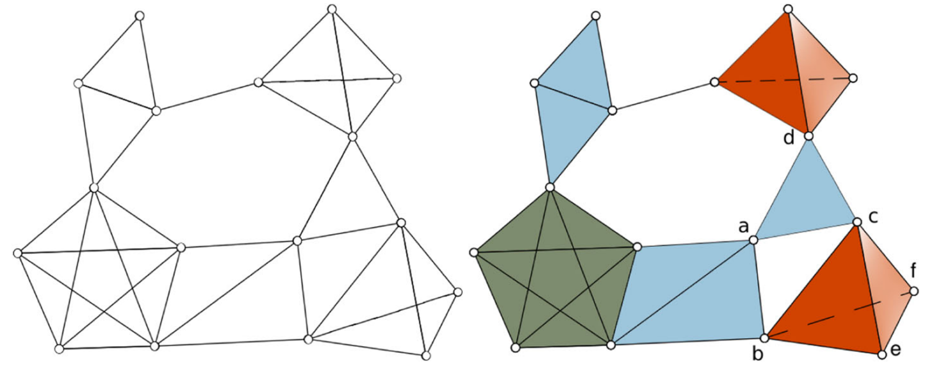

Fig. 3 Network (left) and simplicial complex (right) representation. The (maximal) 2-simplices are in light blue, 3-simplices in red and the 4-simplex is in green. Notice that in the network these are indistinguishable

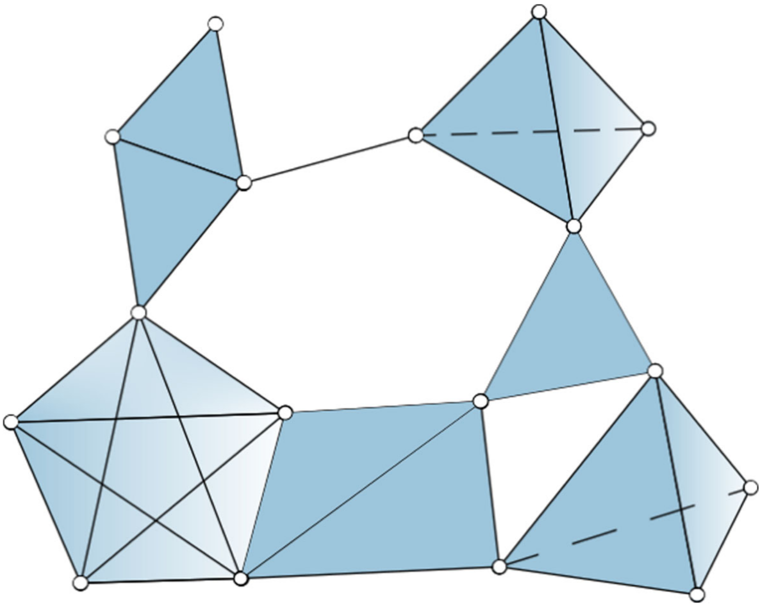

Fig. 4 2-skeleton of a 4-complex in 3. As can be seen 1-dimensional holes are preserved, while 2-dimensional holes are not, in particular there are 3 H_2-holes that were not in the original complex, namely those inside the two 3-simplices, and one inside the 4-simplex

we are now taking the 2-skeleton of the complex in Fig. 3), so $\{bce, bcf, bef, cef\}$ is not the boundary of any higher order chain and its triangular elements bound a void.

The homological features of complexes are usually studied on a filtration of the complex, that is a sequence of simplicial complexes starting at the empty complex and ending with the full complex, so that the complex at step $n < m$ is embedded in the complex at m for all the steps. In this way it is possible to focus on the persistency of homological features: as the filtration evolves the shape of data changes, so *birth* and *death* of holes can be recorded. A hole is born at step s if it appears for the first time in the corresponding step of the filtration, and dies at t if after step t the hole disappears. The difference between birth and death of homological features is called *persistence*, and can be recorded by a *barcode*, a multiset of intervals bounded below (Carlsson et al. 2005) visualizing the lifetime of the feature and its location across the filtration: the endpoints of each interval are the steps of the filtration where the homological feature is born and dies (Horak et al. 2009). An alternative visualisation is provided by the *persistence diagrams*, which are built by constructing a peak function for each barcode, proportional to its length (Edelsbrunner et al. 2002; Stolz et al. 2017).

The way the filtration is built depends on the analysis that one wants to do on the data, a very common method on a weighted network is the *weighted rank clique filtration* (Petri et al. 2013). This is done by filtering for weights: after listing all edge weights w_t in descending order, at every step t one takes the graph obtained keeping all the edges which weight is greater than or equal to w_t. The simplicial complex at that step of the filtration is built by including all the maximal k-cliques of the graphs as k-simplices. The obtained simplicial complex is called *clique complex*.

In this paper we use a *temporal filtration* instead, as in (Pal et al. 2017). Using article dates we build a temporal filtration

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_T$$

where $(0, T)$ are first and last date in our dataset (time step is one month) and for $i < j$ and each $\mathcal{F}_i$ co-occurrence complex contains simplices of concepts (articles) up to date i. As every article is a simplex we do not need to build a clique complex like in the weighted rank clique filtration.

Reducing the computational burden

Computing persistent homology is very costly if there are large simplices in the simplicial complex, as for each simplex the computation requires to list all the possible subsimplices. For instance in our (rather small) dataset, there are already simplices with 37 vertices and the number of its (k-1)-subsimplices is $\frac{37!}{(37-k!)k!}$, making it impossible to finish the task in reasonable time with standard tools. In order to reduce the computational burden, we put an upper bound on the dimension of simplices, that is we take the subcomplex that only includes simplices up to a maximum dimension $d_M = 5$. In other words, we compute the homology of the d_M*-skeleton* of our simplicial complex K.

In the d_M-skeleton of K all simplices of dimension $d > d_M$ are replaced by collections of their d_M-faces, that is a complex of dimension d_M made by glueing together $\binom{d+1}{d_M+1}$ d_M-simplices along their $d_M - 1$ faces, such that for each $d_M - 1$ face there are $d - d_M$ simplices sharing that face. To make an illustrative example if $d_M = 2$, the 2-skeleton of a 3-simplex is the collection of triangles in the boundary of the tetrahedron.

It is straightforward to show that the d_M-skeleton of K, K^{d_M} is $(d_M - 1)$-homologically equivalent to K, in the sense that they have the same homology groups up to $H_{(d_M-1)}$. Moreover, for $d \leq d_M$, the d-chains group of the d_M-skeleton coincides with the d-chains group of K, as they have the same d_M-simplices, hence it follows that also d-cycles groups coincide (see Fig. 5). This implies that any map ∂_d with $d \leq d_M$ is the same on K^{d_M} and K, hence the set of d-boundaries with $d \leq d_M - 1$ is the same on the two complexes. So $H_d(K) = \frac{\ker(\partial_d)}{\mathrm{Im}(\partial_{d+1})} = H_d(K^{d_M})$ for $d \leq d_M - 1$.

As a trivial example consider the 2-skeleton of the tetrahedron, this contains a homological cycle of dimension 2, as there is a void bounded by triangles inside the tetrahedron. But no homological cycle of dimension 1 nor 0, as all its edges are in the boundary of some 2-simplex. An illustration of the 2-skeleton of a complex can be seen in Fig. 4.

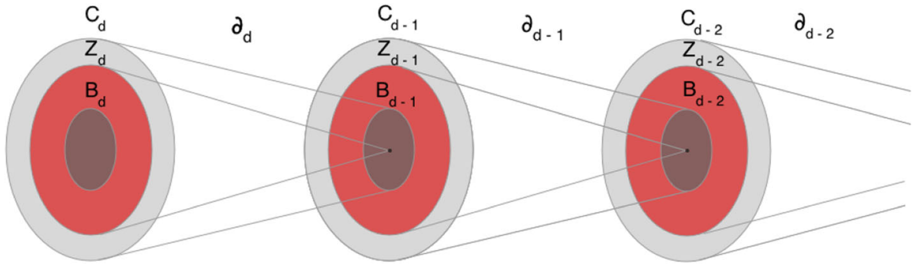

Fig. 5 Chains, cycles and boundaries sets and their maps under the boundary operator for the d-skeleton of a complex K, with $d = d_M$. C_d, Z_d and B_d represent the collections of d chains, cycles and boundaries respectively. Notice that in K we may well have C_{d+1} and bigger dimensional chains with corresponding boundary operators, while in K^{d_M} we start from C_{d_M} as the largest simplices we have are d_M-dimensional

Experimental results

We focus on homological cycles of dimension 1 and 2, respectively two dimensional holes bounded by edges and three dimensional holes bounded by triangles. Persistent homology is computed using javaplex (Adams et al. 2014), and the algorithms implemented here are based on (Zomorodian and Carlsson 2005). Without going into detail, the computation of persistent homology is formulated as a matrix decomposition problem: the boundary operator ∂_k has a standard matrix representation, M_k. The null space of M_k corresponds to cycles Z_k and its range-space to boundaries B_{k-1} so births and deaths of holes are detected when the rank and nullity of M_k change. In other words the algorithm allows to detect births and deaths of holes but not to *localize* them precisely, and for each hole it computes a representative cycle that it is not necessarily the shortest cycle surrounding the hole. We use this cycle representatives for our analysis, so it is necessary to raise some caveats: when we refer to the size of the hole we are referring to the size of its representative, which can be safely considered as a proxy of the size of the hole, as the representative is a random cycle surrounding the hole. The analysis of holes killers is more problematic, as it may be that only some of the concepts in the representative cycle are actually part of the hole. Recall that a homological hole is a k-chain which is not the boundary of any higher-order structure, which means that the concepts in a k-cycle only appear together in sets of at most $k+1$ elements (in the k-simplices making the cycle). A k-dimensional hole dies when all of its concepts appear in an article together with other $k+1$ concepts in the hole. Consider for example the H_1 cycle on the left of Fig. 6, its concepts appear in the same paper in couples and at most in couples. Hence they are related but at the same time they are conceptually distant. This hole could be killed by an article including all of its concepts, or by a collection of articles that include at least 3 of its concepts, covering all concepts in the hole. Clearly these articles can appear at different steps of the filtration, so that the hole progressively "shrinks" until all of its concepts are covered, and that step of the filtration is registered as the death of the hole. We use this information to detect potential hole killers: we check all articles appearing in the filtration when the hole dies, and we select those having an intersection with the cycle representative which is at least $k+2$. If there is more than one, we take the simplex with largest intersection with simplices in the cycle. So when we refer to a hole killer, we mean

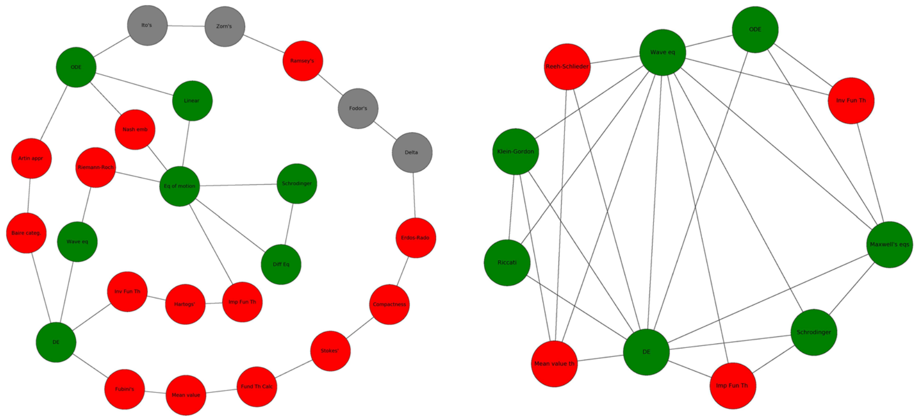

Fig. 6 Example of a H_1 cycle, a hole bounded by edges (left) and a H_2 cycle, hole bounded by triangles (right). Nodes color: red for theorems, grey for lemmas and green for equations

the last simplex that closes the representative cycle. By using this approach we are able to find hole killers only for a subset of representative cycles, and these representative cycles are those more likely to have a large intersection with the shortest cycle surrounding the hole, so we can use these representative cycles and their killers to illustrate some examples.

Figure 7 reports the barcodes for H_1 and H_2, where holes are ordered by their death time (that is the step of the filtration at which they disappear), and Fig. 8 holes persistence. The first thing to notice is that most holes persist up to the end of the filtration, that is up to 03/2007, meaning that there are several areas of low connectivity (both in H_1 and in H_2) in the conceptual space, while this was not emerging from the network analysis of our data. Moreover new holes are born continuously, along all the steps of the filtration. The fact that holes continue to be born at every point in time, and most of them don't die finds a possible interpretation in that the evolution of research in mathematics proceeds by connecting new conceptual areas in a cyclic way, and rarely these concepts contribute all together to the production of scientific advances (that would kill the hole). So this suggests that the death of conceptual holes may be a sign of important advances in mathematics as the emergence of a new subfield.

We investigate which are the most important concepts in H_1 and H_2 by counting the number of times each concept appears in a cycle, divided by the number of times it appears in different articles, to correct for the fact that very common concepts are also more likely to appear in cycles. The results are reported in Fig. 9, notice that most of the concepts are theorems there are 5 concepts in common between the most important 20 of H_1 and H_2, even if their ordering is not preserved.

A clear feature emerging from both H_1 and H_2 is that longer cycles are more difficult to break. Considering only those cycles that die before the end of the filtration (otherwise they have infinite persistence) we compute the average persistence among all cycles of given length. Figure 10 shows the plot of cycles length versus average persistence of cycles of the corresponding length: despite some noise, a positive trend appears clearly. This suggests that concepts appearing in a long cycle that are not successive elements of the k-chain (so that don't appear in the same article) have a great distance among them in the conceptual space of mathematics compared to those appearing in shorter cycles. Notice that this relation is somehow natural, as the more concepts there are in a k-cycle, we need more articles (each including at least $k+2$ of the concepts in the cycle) in order to kill the hole.

Looking at the distribution of killers' size, the largest simplex in H_0 is made of 27 concepts, while both in H_1 and H_2 has 38 concepts (it is actually the same article in both). Interestingly and not surprisingly, these two articles are both surveys, the largest for H_0 regards open questions in number theory (Waldschmidt 2004) and the largest for H_1 is a survey on differential geometry (Yau 2006). The most frequent killers' sizes for H_0, H_1 and H_2 are respectively 4,7,11.

To give a better idea of the meaning of holes death, let us consider some examples. The smallest cycle in H_1 is an empty triangle, one of such cycles is given by the three simplices {(*Schur's lemma*, *Stone-Von Neumann Theorem*), (*Schur's lemma*, *Spectral Theorem*), (*Spectral Theorem*, *Stone-Von Neumann Theorem*)} which is killed by a 2-simplex when the three concepts appear together in the paper (Mantoiu et al. 2004). Another interesting example is a 5-step long cycle in H_1, {(*Boltzmann equation*, *Alternate Interior*

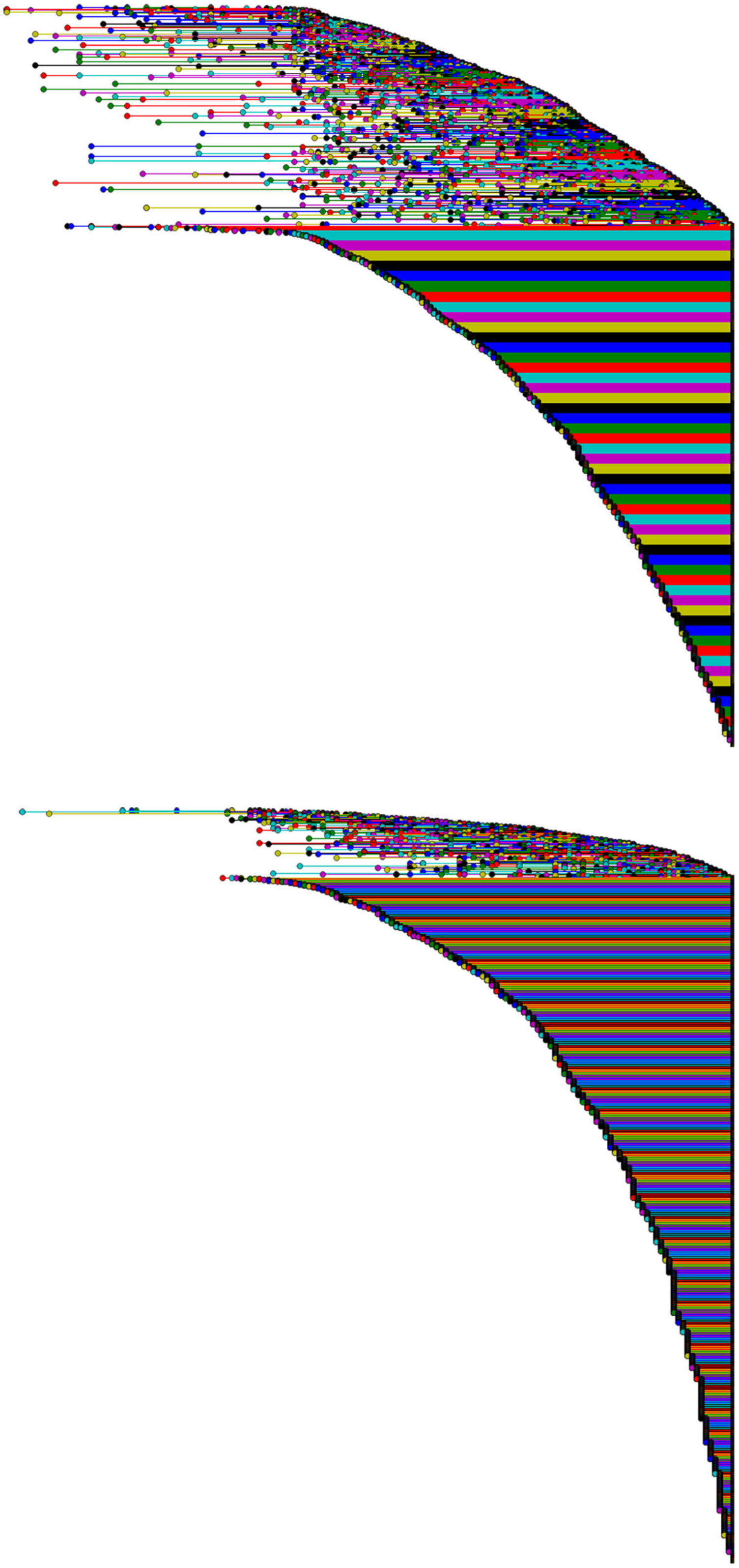

Fig. 7 Barcodes for H_1 (top) and H_2 (bottom), ordered for death time. The left and right endpoints of each segment represent the first and last step in the filtration where the hole appears

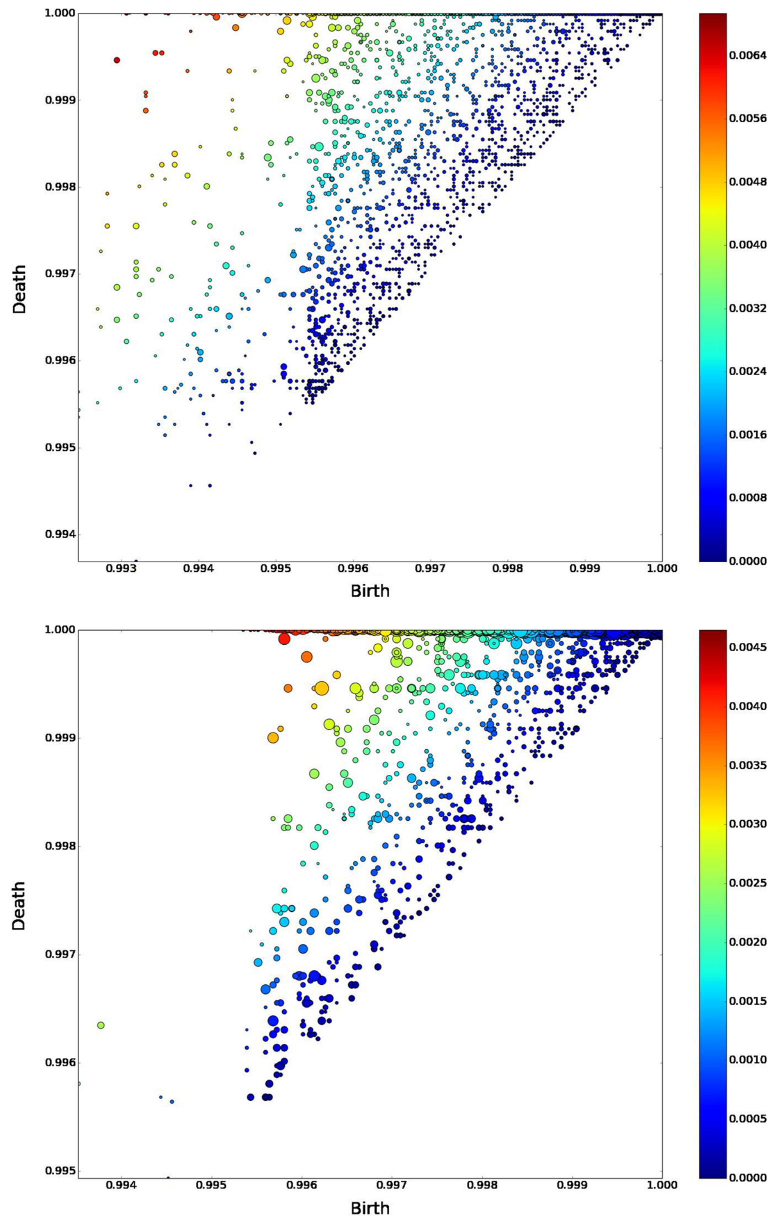

Fig. 8 Persistence diagrams for H_1 and H_2. Persistence diagrams report the same information of barcodes, where each barcode is mapped to a peak function that is proportional to the life of the homological feature. The heatmap reports information on the length of the cycle and the coordinates are birth and death times

Angles Theorem), (*Boltzmann equation, Vlasov equation*), (*Inverse function theorem, Vlasov equation*), (*Arzelá - Ascoli theorem, Alternate Interior Angles Theorem*), (*Arzelá - Ascoli theorem, Inverse function theorem*)}, that is killed by the 10-simplex which nodes are: { *Blum's speedup theorem, Boltzmann equation, Alternate Interior Angles Theorem* , *Kramers theorem, Perpendicular axis theorem, Ordinary differential equation, Kronecker's theorem, Arzelá - Ascoli theorem, Navier - Stokes equations, Vlasov equation, Moreau's theorem*} (Gottlieb 2000). The article, classified in arXiv as *Probability*, establishes the conditions for a family of n-particles Markov processes to propagate chaos, and shows its application to kinetic theory. We think this is a another possible interpretation of killing

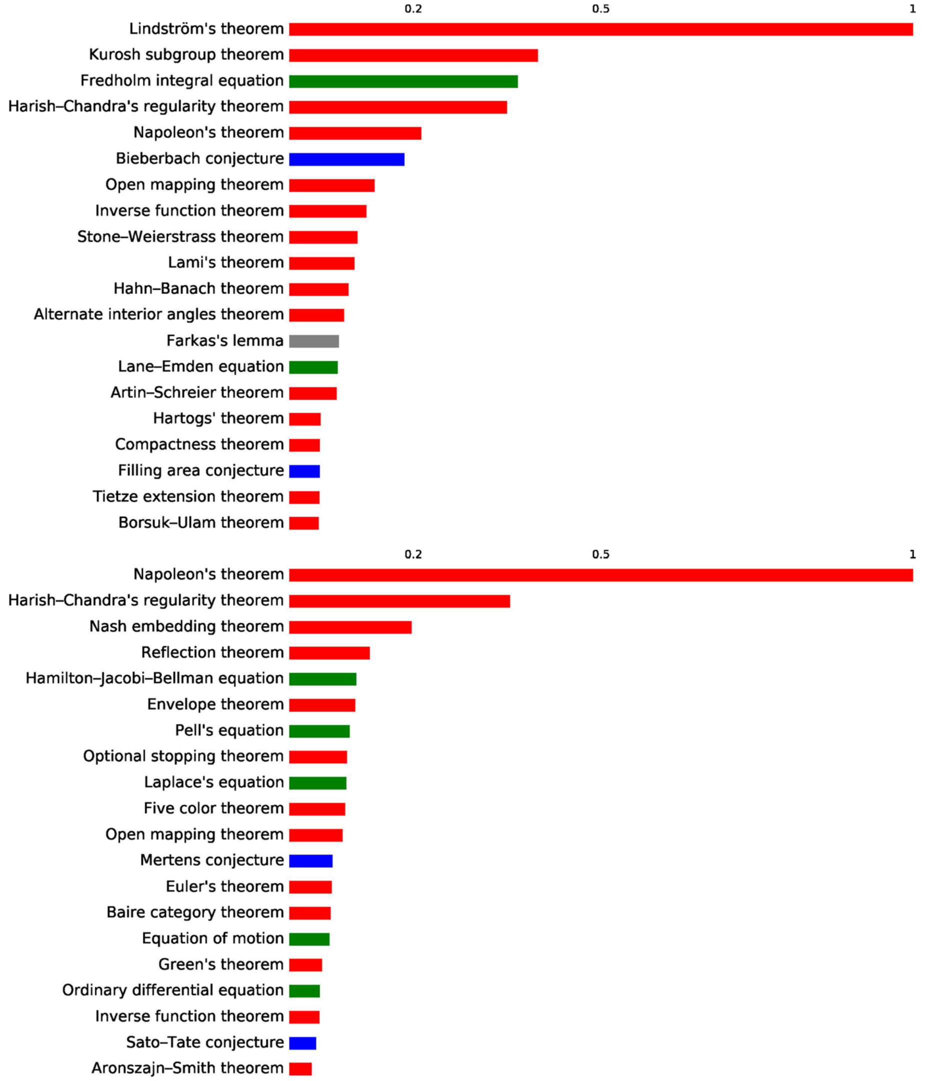

Fig. 9 Concept importance in cycles 20 most important concepts appearing in H_1 cycles (top) and H_2 cycles (bottom). The bars length corresponds to the number of appearances in cycles divided by the number of appearance in edges and triangles respectively

holes: a theoretical result that has several applications, hence bridges related areas and closes a homological cycle.

Simplicial stars represent potentially interesting structures in the conceptual space, which can be visualised as small substructures attached to the 'surface' of a densely connected cluster, like receptors on the membrane of a cell. To grasp the intuition consider that if we partition the concepts in a S^k star between those in the core (the 0-faces of the core k-simplex) and those in the periphery (the zero faces of the $(k+1)$ simplices that have the core as common face, without the core faces) by definition there can't be any edge (or higher-order simplex) between any of the concepts in the periphery. This means that periphery nodes 'touch' the surface of a densely connected area, and each of them belongs to a different simplex lying on the surface, while nodes in the core are one step

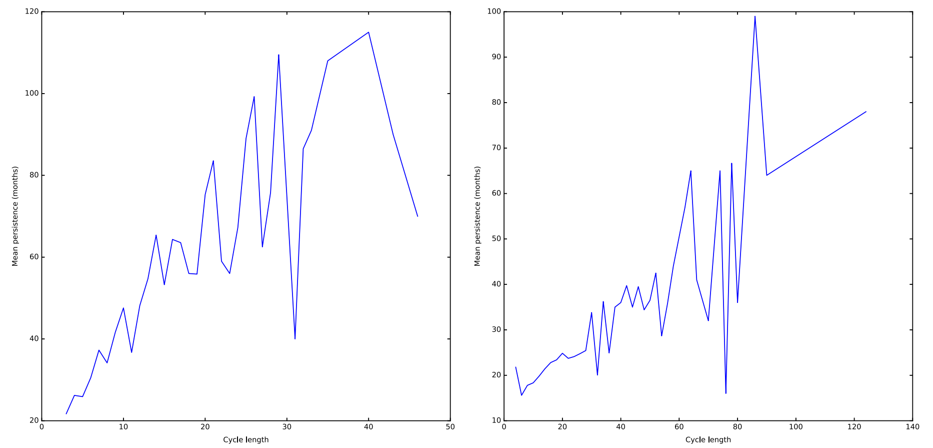

Fig. 10 Longer cycles live longer. The plots report cycle size on the *x*-axis and cycle mean persistence on the *y*-axis for H_1 (left) and H_2 (right). In both cases the trend shows that longer cycles are more difficult to break

far from the surface. Considering only those with at least two simplices we count 567 S^2 stars and 284 S^3 stars. We do not check for higher-order stars for computational reasons.

Figures 11 and 12 report the 20 most important concepts in the cores and peripheries of S^3 and S^2 respectively, adjusted for the number of times concepts appear in triangles (for cores) and tetrahedra (for peripheries). Notice that in both cases all except one concept are theorems/conjectures. Figure 9 reports the ranking of the first 20 concepts in H_1 and H_2, where here we adjust for the frequency of appearance of a concept in the simplex that constitutes the cycles, hence edges and triangles respectively. Even in these two rankings the large majority of the concepts are theorems.

In order to check if edges that appear in cycles are also likely to appear in stars, we find the intersection between the set of concepts in stars (differentiating between edges in the core and edges in the periphery) and cycles and divide by the total numbers of edges in the corresponding cycle. As appearing in a star is a Bernoulli variable, we can easily compute the standard deviations for our estimated probabilities. Table 2 reports the results: it is interesting that edges in the cores of both S^2 and S^3 are more likely to appear in cycles of all dimensions than edges in the peripheries. It is particularly striking that edges in the peripheries of S^3 never appear in any cycle. Moreover edges in stars (both cores and peripheries) are more likely to appear in cycles than a random edge, except for one case: a random edge is more likely to be in a H_3 cycle than an edge in the periphery of a S^2 star.

Authors analysis

We use the conceptual content of articles to classify the activity of researchers, by constructing for each author an activity vector:

$$\mathbf{a}_i = \left(a_i^1, \ldots, a_i^N\right)$$

where a_i^c represents the relative importance of concept c in the activity of author i, given by the number of articles, of which a is one of the authors, containing concept c, divided by the total number of concepts the author used in different articles, so that the sum of the entries of $\mathbf{a}_i$ is one for all authors. As in (Gurciullo and et al. 2015) we can use this vector

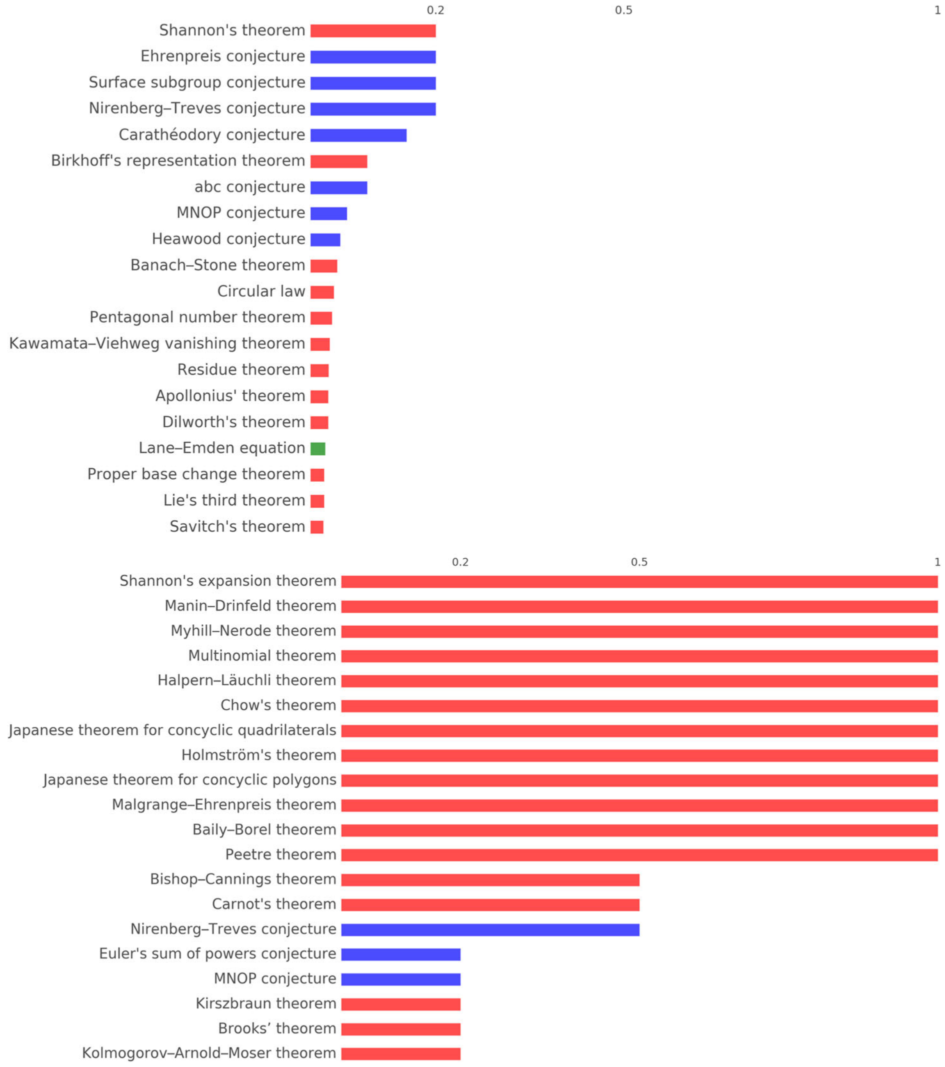

Fig. 11 Concept importance in S^3 stars 20 most important concepts appearing in the core (top) and the periphery (bottom) of S^3 stars. The bars reports the number of appearances in cores or peripheries divided by the number of appearance in triangles and tetrahedra respectively

to map authors' research activity on the basis of the broadness of their contribution to the concept space, as captured by the entropy:

$$\lambda_i = -\sum_c a_i^c \ln a_i^c$$

$\lambda_i \geq 0$ and is 0 only for those authors who only do research about one concept. We suggest a classification of authors based on their entropy level, we define author i as *specialist* if $\lambda_i < 1$, *polymath* if $\lambda_i > 2$, and *mixed* if $1 \leq \lambda_i \leq 2$. The choice of the thresholds is arbitrary, and it is made just for exposition's sake, hence this classification is not to be intended in a true sense of the words: a little caveat here is that this estimator of authors' activity is not very informative for those authors who published only one article. Also, an author with high entropy may be a specialist in one specific topic that has application across disciplines, hence she has a diverse range of collaborations more than

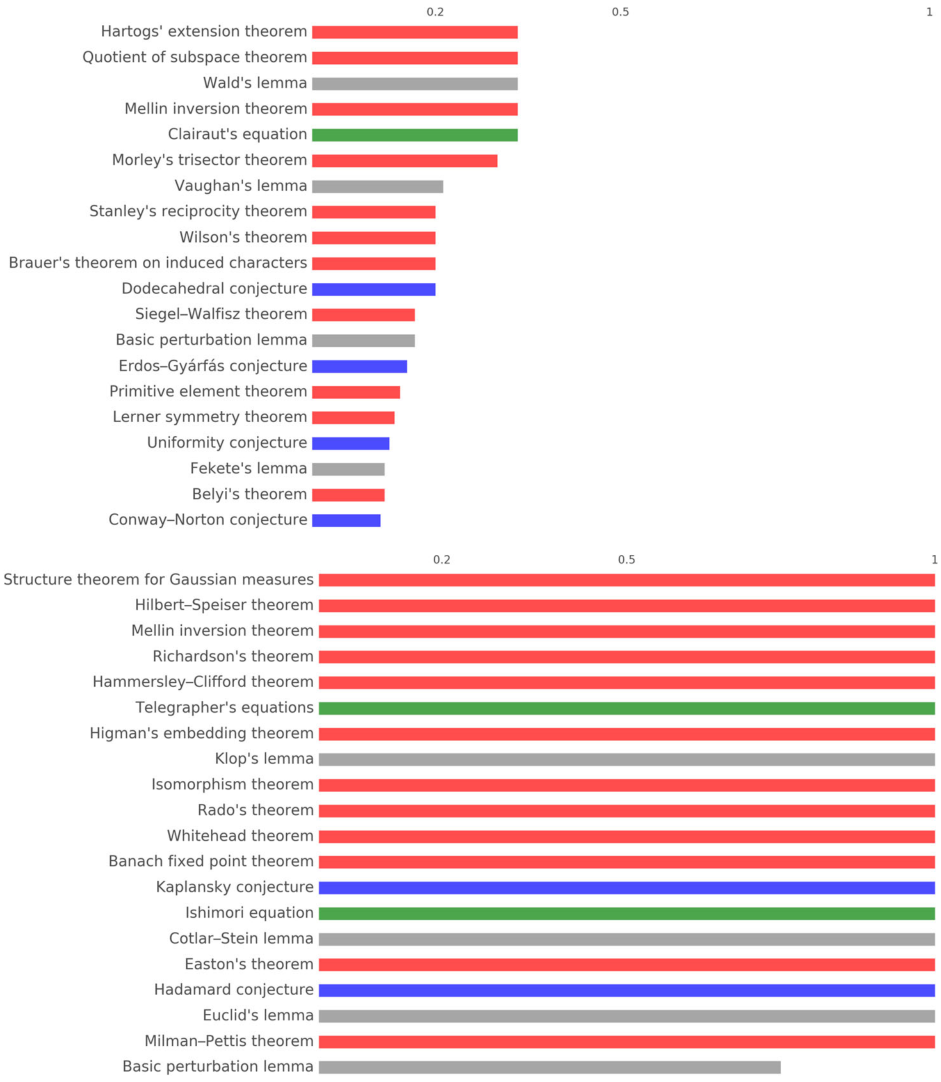

Fig. 12 Concept importance in S^2 stars 20 most important concepts appearing in the core (top) and the periphery (bottom) of S^2 stars. The bars reports the number of appearances in cores or peripheries divided by the number of appearance in edges and triangles respectively

being a *polymath* in a strict sense. Of course we cannot disentangle which is each author's contribution to a paper, but, provided we are clear on the information conveyed by this measure of specialisation, we can make use of it to capture the relation between conceptual breadth of research (being it made by a single authors who really is a *polymath* or by a research group) and homological cycles.

While it is expected that there is a positive relation between number of concepts in the activity vector and its entropy, we want to see how this relation compares with a null model. The null model is constructed as follows: for each concept in our list we add to an urn as many of its copies as the number of times it appears in different articles. Then for each author we count how many concepts she used across all her publications, or equivalently we sum the number of concepts used in each paper she authored (for example if an author published three papers using concept c_1, one using concept c_2 and one using concept c_3 her total count will be 5). Call k^M the max of these counts across all authors, then for $k \in \left[1, k^M\right]$ (k integer), we extract k concepts at random from the urn, and we compute the entropy of the extracted k-tuple, repeating the operation 1000 times for each k and computing the average entropy over the 1000 extractions. To compare with authors, we group them according to the number of concepts used, and compute the average entropy for each k. We estimate the relation $\hat{e}$ between number of concepts k and average entropy, using least squares, finding logarithmic relation $\hat{e}_{null} = A + B\log(1+k)$ for the null model, while for the data $\hat{e}_{data} = A + B\log(1+k)/k$ as can be seen in Fig. 13. For the data (A, B) is $(3.459, -5.257)$ and $(0.170, 0.863)$ for the null model, showing how, for large enough values of k, the fit for the data always lies below the null model, and more importantly, that the relation emerging from the data has a horizontal asymptote, while the null model does not. This is telling us that there is an upper bound on the conceptual entropy, and even very prolific *polymaths* show some degree of specialisation, in the sense that as the number of articles increases, they eventually stop broadening their research including new concepts, and tend to publish new research regarding concepts they already explored.

To understand if there is any relation between authors' profile in terms of their specialisation and their contribution to homological cycles, we compute for author i a measure of his *homological importance*

$$h_i = \sum_{c \in C_i, k} \mathbf{1}_{H_k}(c)$$

where C_i is the set of concepts used by author i, and $\mathbf{1}_{H_k}(c)$ is the indicator function, giving 1 if concept c is in the homological cycle H_k.

To correct for the fact that most frequent concepts tend to appear in cycles more often, we exclude the first 100 most frequent concepts from the computation of the homological importance. After removing the 100 most frequent concepts still the 86.5% of authors contribute at least once to a homological cycle. This confirms that homological cycles are ubiquitous, and really constitute a feature of mathematical research. Figure 14 shows the scatterplot of authors' conceptual entropy and homological importance. It reveals a non-linear positive relation between the two, so more interdisciplinary authors contribute more to homological cycles, thus confirming the intuition that cycles are made by concepts that belong to different areas of mathematics, which are mostly unconnected among them. *Polymaths* are often found on the boundary of these voids surrounded by concepts belonging to different conceptual areas. In Fig. 15 we show the relation $\hat{c}$ between average entropy and average importance in cycles (least squares fit), and how it compares with the null model. The relation between average entropy $\hat{\lambda}$ and average contribution to cycles is exponential both for the null model and for the data, with $\hat{c} = A + B^{\hat{\lambda}}$, and the null model

Table 2 Probabilities and standard errors for edges in the cores and peripheries of stars and of a random edge to be in H_k

	H_1	H_2	H_3
S^2 (cores)	0.794 (0.005)	0.806 (0.004)	0.785 (0.003)
S^2 (peripheries)	0.571 (0.006)	0.51 (0.005)	0.547 (0.003)
S^3 (cores)	0.743 (0.005)	0.905 (0.003)	0.91 (0.002)
S^3 (peripheries)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Random Edge	0.240 (0.002)	0.300 (0.002)	0.641 (0.003)

always lies above the data, with $(A, B) = (3.04, 3.92)$ for the null model and $(-4.50, 3.05)$ for the data.

Discussion

This work is a first attempt to explore the importance of homological holes in mathematics, and it is important to issue certain caveats here. It seems clear that our observations would not have been possible within a classical network analysis of co-occurrences, and we also think that holes deaths can be informative in capturing advances in the discipline. However, we do not claim that we are describing the essence of mathematical practice here. If on one hand, by extracting concepts from the whole text of the article instead of just focusing on the keywords we avoid the bias of authors' own classification of their work, on the other hand it is possible that for some articles the conceptual content is not well captured by our approach: we cannot exclude that the set of concepts that better identify the content of the article do not find an exact match in our list, while those finding a match are poorly representative of the article. We believe that such cases are a minority as our list is very comprehensive, still this is an aspect to take in consideration when analysing our results, even if this is an issue that is potentially arising every time one does text analysis, irrespective of the tools used to explore the data. In this direction, an

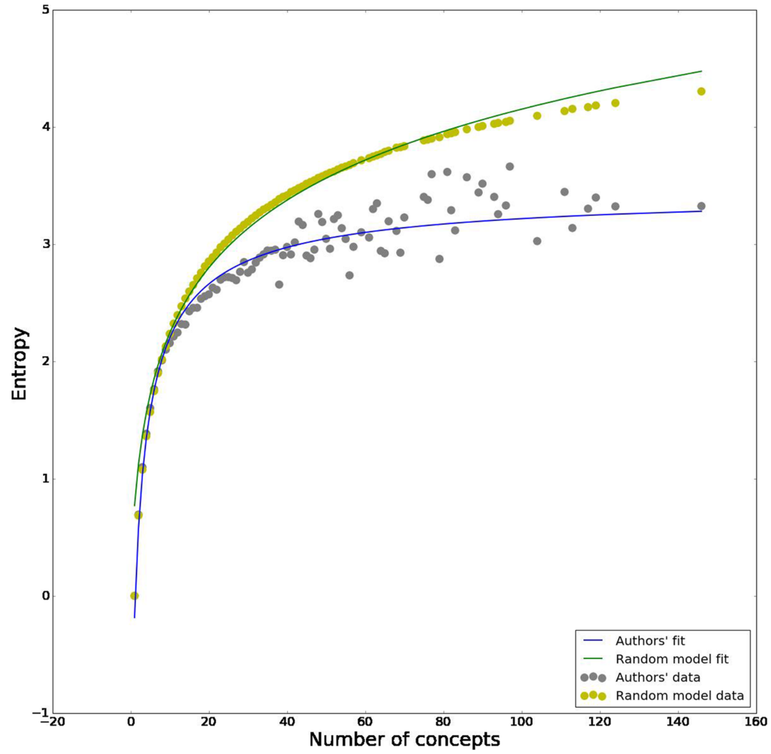

Fig. 13 Authors' entropy versus null model. Here we compare the entropy for authors in our dataset with average entropy over a null model. Entropy for authors is computed as the average over all authors with *k* concepts, while entropy for the null model is the average over 1000 random extractions of *k* concepts from our list. Average entropy for authors is always below the null model, and it is bounded above by a horizontal asymptote

important step would be to assess the robustness of our observations by purposely introducing errors in our data analysis, for instance by focusing only on a fraction of our list of identified concepts, and keeping the other unknown.

Overall, this paper suggests new directions for the study of co-occurrences by focusing on their higher-order properties and there is substantial space for further development. The first and most urgent point, in our opinion, regards the necessity to validate the findings from the data by having well-founded null models for comparison. Recently some null models for simplicial complexes have been developed (Young et al. 2017; Courtney and Bianconi 2016), which we did not use here for computational reasons, as producing samples from our relatively large dataset proved too challenging to manage with our resources. Another very important point is to devise methods and find algorithms to localise homological holes with precision, as this could be very useful for some specific applications, and it can still be considered an open problem. In this regard we think that it would be pivotal to unify the study of the geometry and the topology of the structure. A geometrical approach has been used in (Eckmann and Moses 2002) where they find that curvature is a good measure of thematic cohesion in the WWW. More recently (Wu et al. 2015; Bianconi and Rahmede 2017) studied the emergence of geometry in growing simplicial complexes, without a pre-existing embedding space and metric. This approach could be naturally extended to our case, as we have no natural embedding space for concepts, and also because the co-occurrence simplicial complex that we study is dynamic by nature, as new simplices appear at each time step eventually glueing to existing simplices. Embedding concepts in a space could facilitate the localisation of holes, in this regard (Bianconi and Rahmede 2017) find that the natural embedding for their growing simplicial complexes is hyperbolic, and in their case the position of the incoming nodes depends on the position of the nodes of the face of the simplex to

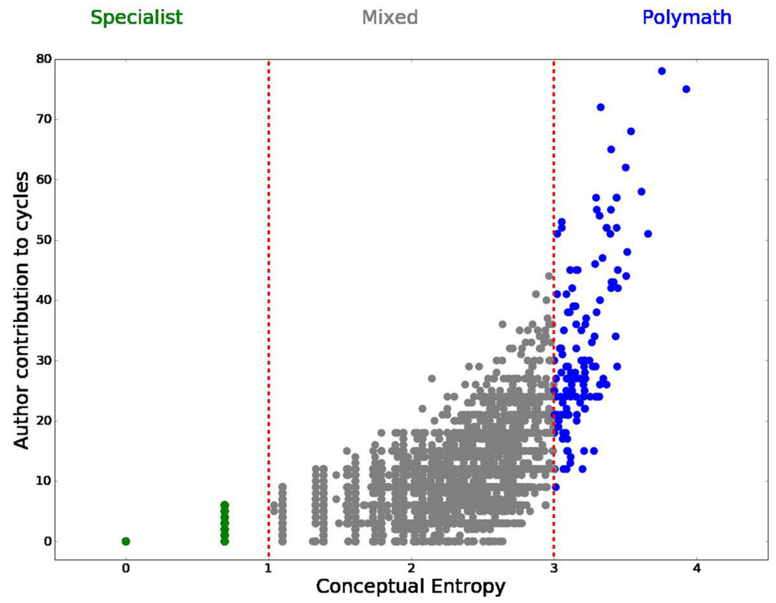

Fig. 14 Author contribution to cycles. The plot shows that contribution to homological cycles in H_1, H_2 and H_3 is positively correlated to the conceptual entropy

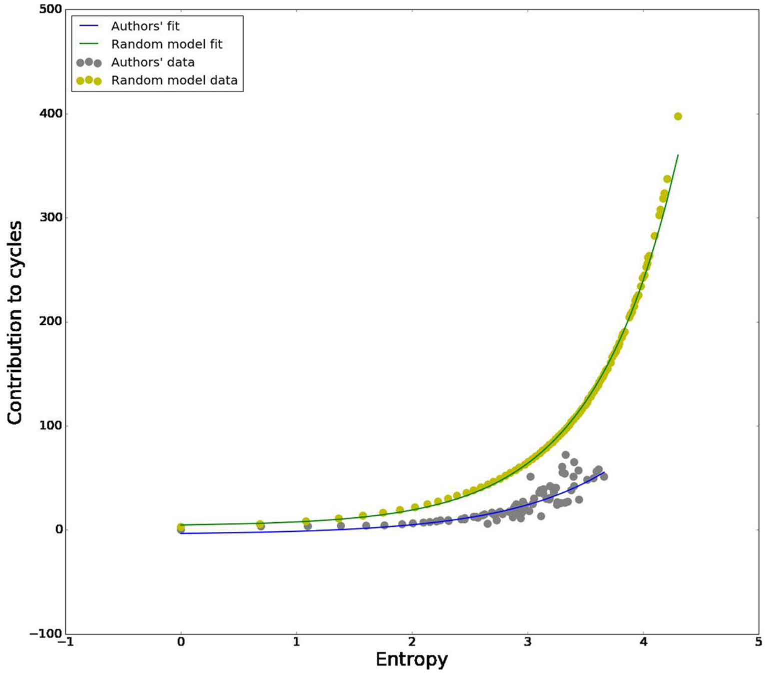

Fig. 15 Authors' contribution to cycles versus null model. Here we compare the contribution to homological cycles as a function of authors' entropy in our dataset with average contribution over the null model, the latter being always above the fit for authors' data

which the new simplex glues. Such a framework could be ideal to the purpose of precise holes identification in our case. Moreover adopting a growing simplicial complex model (generalized to allow for simplices of different size at each time step) and fitting it to our data we could be able to predict the future evolution of the conceptual space of mathematics.

Conclusions

In this paper, we have studied the topological structure of conceptual co-occurrences in mathematics articles, using data from arXiv. We modelled co-occurrences in a simplicial framework, focusing on higher-order relations between concepts and applying topological data analysis tools to explore the evolution of research in mathematics. We find that homological holes are ubiquitous in mathematics, appearing to show an intrinsic characteristic of how research evolves in the field: holes are likely to represent groups of concepts that are closely related but do not belong to a unitary subfield, and the death of a hole is either a sign that anticipates a potential advance in that conceptual area (for example a review trying to bridge the concepts and suggesting research lines), or an actual advance, that is an article (or a set of articles) that unifies a subgroup of concepts in the cycle, for example a theoretical result with application to different areas. Less interesting, but we cannot exclude it as we have no other way to verify than reading each of the papers killing a hole, a hole-killer (especially if of very large size) could be a scarcely relevant article mentioning many concepts without providing any true contribution.

We also find that the higher the number of concepts in a hole, the longer it takes to die, hence the length of a hole is a good proxy of how distant these concepts are, in terms of their likelihood to appear together in an article. So in this sense large holes could be seen as potential spaces for important advances in mathematics. Moreover we further explore the structure of co-occurrences by looking at the simplicial analogs of stars in higher dimension, which represent groups of concept (those in the core of the star) that supports and connects many otherwise unrelated concepts, and we find that concepts appearing in stars tend also to appear in holes more often than they would do at random, suggesting that both structures lie at the frontier of mathematical research.

We also explore authors' conceptual profile by ordering them on the basis of their conceptual entropy, so that we can differentiate between those authors who tend to specialise and publish mostly about few concepts, and others that do research on a broad range of topics, that we call *polymaths*. Comparing authors' profiles with a random model, we find that authors' entropy as a function of how may concepts authors use across different publications, is bounded above, while in the null model, entropy is always increasing for larger set of random concepts. This is reasonable, and means that even the more prolific *polymaths*, even if they publish a large number of articles, will still tend to specialise to some extent, instead of doing research always on new topics. Moreover we find that *polymaths* contribute to homological holes more than specialists, so *polymaths* are often at the frontier of research.

Further work could be done by using larger datasets, as it would be very interesting to explore the birth and death of holes in a larger time-span, and to study simplicial co-occurrences in other disciplines, in order to see if any difference appears in the way research evolves in different fields. Furthermore, conceptual spaces emerging from co-occurrences relations could be explored adding a further dimension to the filtration: in our case we focus on a temporal filtration, disregarding the weights of simplices, this could be extended by filtering along time and weight using multidimensional persistence (Carlsson and Zomorodian 2009).

Acknowledgements
The authors thank two anonymous reviewers for their constructive comments and Henry Adams, Oliver Vipond and Alexey Medvedev for useful discussions and suggestions.

Funding
Daniele Cassese wishes to thank support from FNRS (Belgium).

Availability of data and materials
Data available upon request.

Authors' contributions
All authors conceived the study; VS and DC performed the numerical simulations and created the Figures; All authors wrote and reviewed the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1] University of Namur and NaXys, Rempart de la Vierge, 5000 Namur, Belgium. [2] ICTEAM, University of Louvain, Av Georges Lemaître, 1348 Louvain-la-Neuve, Belgium. [3] Mathematical Institute, University of Oxford, Woodstock Road, OX2 6GG Oxford, UK. [4] Department of Mathematics, Imperial College, South Kensington Campus, SW7 2AZ London, UK.

Received: 19 March 2018 Accepted: 18 June 2018
Published online: 28 August 2018

References

Adams H, Tausz A, Vejdemo-Johansson M (2014) javaplex: A research software package for persistent (co)homology. In: Hong H, Yap C (eds). Mathematical Software – ICMS 2014. Springer, Berlin, Heidelberg. pp 129–136

Akimushkin C, Amancio DR, Oliveira ONJ (2017) Text authorship identified using the dynamics of word co-occurrence networks. PLos ONE 12(1):1-101

Bianconi G, Rahmede C (2017) Emergent hyperbolic network geometry. Sci Rep 7(41974). https://doi.org/10.1038/srep41974

Carlsson G (2009) Topology and data. Bullettin AMS 46(2):255–308

Carlsson G, Zomorodian A (2009) The theory of multidimensional persistence. Discrete Comput Geom 42:71–93

Carlsson G, Zomorodian A, Collins A, Guibas LJ (2005) Persistence barcodes for shapes. Int J Shape Model 11:149–188

Carstens CJ, Horadam KJ (2013) Persistent homology of collaboration networks. Math Probl Eng 2013:85035. http://dx.doi.org/10.1155/2013/815035

Chan JM, Carlsson G, Rabadan R (2013) Topology of viral evolution. PNAS 110(46):18566–18571

Chiang IJ (2007) Discover the semantic topology in high-dimensional data. Expert Syst Appl 33:256–262

Courtney O, Bianconi G (2016) Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. Phys Rev E 93:062311. https://doi.org/10.1103/PhysRevE.93.062311

Eckmann J-P, Moses E (2002) Curvature of co-links uncovers hidden thematic layers in the world wide web. PNAS 99(9):5825–5829

Edelsbrunner H, Harer J (2008) Persistent homology - a survey. Contemp Math 453(2):255–308

Edelsbrunner H, Letscher D, Zomorodian A (2002) Topological persistence and simplification. Discret Comput Geom 28(4):511–533

Estrada E, Ross G (2018) Centralities in simplicial complexes. applications to protein interaction networks. J Theor Biol 438:46–60

Ferrer-i-Cancho R, Solé RV (2001) The small world of human language. Proc R Soc Lond B 268:2261–2265. https://doi.org/10.1098/rspb.2001.1800

Garg M, Kumar M (2018) Identifying influential segments from word co-occurrence networks using ahp. Cogn Syst Res 47:23–41

Gottlieb AD (2000) Markov transitions and the propagation of chaos. ArXiv Math E-prints. math/0001076

Gurciullo S, et al. (2015) Complex politics: A quantitative semantic and topological analysis of uk house of commons debates. ArXiv E-prints. 1510.03797

Horak D, Maletic S, Rajkovic M (2009) Persistent homology of complex networks. J Stat Mech Theory Exp 3:3–34

Jenssen T-K, Laegreid A, Komorowski J, Hovig E (2001) A literature network of human genes for high-throughput analysis of gene expression. Nat Genet 28:21–28

Jo Y, Hopcroft JE, Lagoze C (2011) The web of topics: Discovering the topology of topic evolution in a corpus. In: Proceedings of the 20th International Conference on World Wide Web, March 28 - April 01, 2011, Hyberabad, India. ACM, New York. pp 256–266

Lazer D, Mergel I, Friedman A (2009) Co-citation of prominent social network articles in sociology journals: The evolving canon. Connections 29(1)

Mamuye AL, Rucco M, Tesei L, Merelli E (2016) Persistent homology analysis of rna. Mol Based Math Biol 4:14–25

Mantoiu M, Purice R, Richard S (2004) Twisted crossed products and magnetic pseudodifferential operators. ArXiv Math Phys E-prints. math-ph/0403016

Mullen EK, et al. (2014) Gene co-citation networks associated with worker sterility in honey bees. BMC Syst Biol 8(38). https://doi.org/10.1186/1752-0509-8-38

Otter N, Porter MA, Tillmann U, Grindrod P, Harrington H (2017) A roadmap for the computation of persistent homology. EPJ Data Sci 6(17). https://doi.org/10.1140/epjds/s13688-017-0109-5

Pal S, Moore TJ, Ramanathan R, Swami A (2017) Comparative topological signatures of growing collaboration networks. In: Gonçalves B, Menezes RSR, Zlatic V (eds). Proceedings of the 8th Conference on Complex Networks CompleNet 2017. Stoneham: Butterworth-Heinemann, Cham. pp 16–27

Patania A, Petri G, Vaccarino F (2017a) The shape of collaborations. EPJ Data Sci 6(18). https://doi.org/10.1140/epjds/s13688-017-0114-8

Patania A, Petri G, Vaccarino F (2017b) Topological analysis of data. EPJ Data Sci 6(7). https://doi.org/10.1140/epjds/s13688-017-0104-x

Petri G, Expert P, Turkheimer F, Carhart-Harris R, Nutt D, Hellyer PJ, Vaccarino F (2014) Homological scaffolds of brain functional networks. J R Soc Interface 11(20140873). http://dx.doi.org/10.1098/rsif.2014.0873

Petri G, Scolamiero M, I D, F V (2013) Topological strata of weighted complex networks. PLoS ONE 8(6). https://doi.org/10.1371/journal.pone.0066506

Radhakrishnan S, Erbis S, Isaacs JA, Kamarthi S (2017) Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. PLos ONE 12(3):23–41. e0172778. https://doi.org/10.1371/journal.pone.0172778

Sami IR, Farrahi K (2017) A simplified topological representation of th text for local and global context. In: Proceedings of the 2017 ACM on Multimedia Conference, Mountain View, California, USA. ACM, New York. pp 1451–1456

Serrano MA, Boguña M, Vespignani A (2009) Extracting the multiscale backbone of complex weighted networks. PNAS 106:6483–6488

Slater PB (2009) A two-stage algorithm for extracting the multiscale backbone of complex weighted networks. PNAS 106(26). https://doi.org/10.1073/pnas.0904725106

Stolz BJ, Harrington HA, Porter MA (2017) Persistent homology of time-dependent functional networks constructed from coupled time series. Chaos 27(047410). https://doi.org/10.1063/1.4978997

Su H-N, Lee P-C (2010) Mapping knowledge structure by keyword co-occurrence: a first look at journal papers in technology foresight. Scientometrics 85:65–79. https://doi.org/10.1007/s11192-010-0259-8

Taylor D, et al. (2015) Topological data analysis of contagion maps for examining spreading processes on networks. Nat Commun 6(7723). https://doi.org/10.1038/ncomms8723

Wang X, Zhang X, Xu S (2011) Patent co-citation networks of fortune 500 companies. Scientometrics 88(3):761–770

Yau S-T (2006) Perspectives on geometric analysis. ArXiv Math e-prints. math/0602363

Young J-C, Petri G, Vaccarino F, Patania A (2017) Construction of and efficient sampling from the simplicial configuration model. Phys Rev E 96(3):032312. https://doi.org/10.1103/PhysRevE.96.032312

Wagner H, et al. (2012) Computational topology in text mining. In: Ferri M, Frosini P, Landi C, Cerri A, Di Fabio B (eds). Computational Topology in Image Context. Springer, Berlin, Heidelberg. pp 68–78

Waldschmidt M (2004) Open diophantine problems. Mosc Math J 4(1):245–305

Wu Z, Menichetti G, Rahmede C, Bianconi G (2015) Emergent complex network geometry. Sci Rep 5(10073). https://doi.org/10.1038/srep10073

Zhang J, Xie J, Hou W, Tu X, Xu J, et al (2012) Mapping the knowledge structure of research on patient ahderence: knowledge domain visualization based co-word analysis and social network analysis. PLoS ONE 7(4):34497. https://doi.org/10.1371/journal.pone.0034497

Zomorodian A, Carlsson G (2005) Computing persistent homology. Discret Comput Geom 33:249–274