An Efficient Algorithm for 1-Dimensional (Persistent) Path Homology

Tamal K. Dey · Tianqi Li · Yusu Wang

Received: 11 August 2020 / Revised: 8 July 2022 / Accepted: 8 July 2022 / Published online: 27 November 2022
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Abstract
This paper focuses on developing an efficient algorithm for analyzing a directed network (graph) from a topological viewpoint. A prevalent technique for such topological analysis involves computation of homology groups and their persistence. These concepts are well suited for spaces that are not directed. As a result, one needs a concept of homology that accommodates orientations in input space. Path-homology developed for directed graphs by Grigoryan et al. (2012) has been effectively adapted for this purpose recently by Chowdhury and Mémoli (2018). They also give an algorithm to compute the path-homology and the persistent path-homology. Our main contribution in this paper is an algorithm that computes this path-homology and its persistence more efficiently for the 1-dimensional ($H_1$) case. In developing such an algorithm, we discover various structures and their efficient computations that aid computing the 1-dimensional path-homology. We implement our algorithm and present some preliminary experimental results.

Keywords Computational topology · Path homology · Directed graph · Topological persistence · Arboricity

Mathematics Subject Classification 65D18 · 05C85 · 68W40 · 57M15 · 46M20

Editor in Charge: Kenneth Clarkson

Tamal K. Dey
tamaldey@purdue.edu

Tianqi Li
li.6108@osu.edu

Yusu Wang
yusuwang@ucsd.edu

1 Department of Computer Science, Purdue University, West Lafayette, IN, USA
2 Department of Computer Science, Ohio State University, Columbus, OH, USA
3 Halıcıoğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA
1 Introduction

When it comes to graphs, traditional topological data analysis has focused mostly on undirected ones. However, applications in social networks [2, 21], brain networks [24], and others require processing directed graphs. Consequently, topological data analysis for these applications needs to be adapted accordingly to account for directedness. Recently, some work [5, 20] have initiated to address this important but so far neglected issue.

Since topological data analysis uses persistent homology as a main tool, one needs a notion of homology for directed graphs. Of course, one can forget the directedness and consider the underlying undirected graph as a simplicial 1-complex and use a standard persistent homology pipeline for the analysis. However, this is less than desirable because the important information involving directions is lost. Currently, there are two main approaches that have been proposed for dealing with directed graphs. One uses directed clique complexes [10, 20] and the other uses the concept of path homology [12]. In the first approach, a $k$-clique in the input directed graph is turned into a $(k - 1)$-simplex if every face in the simplex and the clique has a single source and a single sink. The resulting simplicial complex is subsequently analyzed with the usual persistent homology pipeline. One issue with this approach is that there could be very few cliques with the required condition and thus accommodating only a very few higher dimensional simplices. For example, the induced complex can have only vertices (0-cliques) and edges (1-cliques) if each 2-clique is a directed cycle. The second approach based on path homology alleviates this deficiency. Furthermore, certain natural functorial properties, such as Künneth formula, do not hold for the clique complex [12].

The path homology, originally proposed by Grigoryan, Lin, Muranov, and Yau in 2012 [12] and later studied in [6, 13, 14], has several properties that make it a richer mathematical structure. For example, there is a concept of homotopy under which the path homology is preserved; it accommodates Künneth formula; and the path homology theory is dual to the cohomology theory of digraphs introduced in [14]. Furthermore, persistent path homology developed in [6] is shown to respect a stability property for its persistent diagrams.

To use path homologies effectively in practice, one needs efficient algorithms to compute them. In particular, we are interested in developing efficient algorithms for computing 1-dimensional path homology and its persistent version because even for this case the current state of the art is far from satisfactory: Given a directed graph $G$ with $n$ vertices, the most efficient algorithm proposed in [6] has a time complexity $O(n^9)$ (more precisely, their algorithm takes $O(n^{3+3d})$ to compute the $(d - 1)$-dimensional persistent path-homology).

The main contribution of this paper is stated in Theorem 1.1. The reduced time complexity of our algorithm can be attributed to the fact that we compute the boundary groups more efficiently. In particular, it turns out that for 1-dimensional path homology, the boundary group is determined by bigons, certain triangles, and certain quadrangles in the input directed graph. The bigons and triangles can be determined relatively easily. It is the boundary quadrangles whose computation and size determine the time complexity. The authors in [6] compute a basis of these boundary quadrangles by
constructing a certain generating set for the 2-dimensional chain group by a nice column reduction algorithm (being different from the standard simplicial homology, it is non-trivial to do reduction for path homology). We take advantage of the concept of arboricity and related results in graph theory, together with other efficient strategies, to enumerate a much smaller set generating the boundary quadrangles. Computing the cycle and boundary groups efficiently both for non-persistent and persistent homology groups is the key to our improved time complexity.

**Theorem 1.1** Given a directed graph $G$ with $n$ vertices and $m$ edges, set

$$r = \min \left\{ a(G)m, \sum_{(u,v) \in E} (d_{in}(u) + d_{out}(v)) \right\},$$

where $a(G) = O(n)$ is the so-called arboricity of $G$, and $d_{in}(u)$ and $d_{out}(u)$ are the in-degree and out-degree of $u$, respectively. There is an $O(rm^{\omega - 1} + m\alpha(n))$ time algorithm for computing the 1-dimensional persistent path homology for $G$ where $\omega < 2.373$ is the exponent for matrix multiplication,\(^1\) and $\alpha(\cdot)$ is the inverse Ackermann function. This also gives an $O(rm^{\omega - 1} + m\alpha(n))$ time algorithm for computing the 1-dimensional path homology $H_1$ of $G$. In particular, for a planar graph $G$, $a(G) = O(1)$, and the time complexity becomes $O(n^\omega)$.

The arboricity $a(G)$ of a graph $G$ mentioned in Theorem 1.1 denotes the minimum number of edge-disjoint spanning forests into which $G$ can be decomposed [15]. It is known that in general, $a(G) = O(n)$, but it can be much smaller. For example, $a(G) = O(1)$ for planar graphs and $a(G) = O(g)$ for a graph embedded on a genus-$g$ surface [4]. Hence, for planar graphs, we can compute 1-dimensional persistent path homology in $O(n^\omega)$ time whereas the algorithm in [6] takes $O(n^5)$ time.\(^2\)

**Organization of the paper** After characterizing the 1-dimensional path homology group $H_1$ in Sect. 3, we first propose a simple algorithm to compute it. In Sect. 4, we consider its persistent version and present an improved and more efficient algorithm. In Sect. 5, we develop an algorithm to compute 1-dimensional minimal path homology basis [9, 11], and also show experiments demonstrating the efficiency of our new algorithms.

### 2 Background

We briefly introduce some necessary background for path homology. Interested readers can refer to [12] for more details. The original definition can be applied to structures beyond directed graphs; but for simplicity, we use directed graphs to introduce the notations.

Given a directed graph $G = (V, E)$, we denote $(u, v)$ as the directed edge from $u$ to $v$. A self-loop is defined to be the edge $(u, u)$ from $u$ to itself. Throughout this

---

\(^1\) That is, the fastest algorithm to multiply two $r \times r$ matrices takes time $O(r^{\omega})$.

\(^2\) The original time complexity stated in the paper is $O(n^9)$ for 1-dimensional case. However, one can improve it to $O(n^5)$ by a more refined analysis for planar graphs.

\(\square\) Springer
paper, we assume that $G$ does not have self-loops. We also assume that $G$ does not have multi-edges, i.e., for every ordered pair $u, v$, there is at most one directed edge from $u$ to $v$. For notational simplicity, we sometimes use index $i$ to refer to vertex $v_i \in V = \{v_1, \ldots, v_n\}$.

Let $\mathbb{F}$ be a field with 0 and 1 being the additive and multiplicative identities respectively. We use $-a$ to denote the additive inverse of $a$ in $\mathbb{F}$. An elementary $d$-path on $V$ is simply a sequence $i_0, i_1, \ldots, i_d$ of $d + 1$ vertices in $V$. We denote this path by $e_{i_0, i_1, \ldots, i_d}$. Let $\Lambda_d = \Lambda_d(G, \mathbb{F})$ denote the $\mathbb{F}$-linear space of all linear combinations of elementary $d$-paths with coefficients from $\mathbb{F}$. It is easy to check that the set $\{e_{i_0, \ldots, i_d} : i_0, \ldots, i_d \in V\}$ is a basis for $\Lambda_d$. Each element $p$ of $\Lambda_d$ is called a $d$-path, and it can be written as

$$ p = \sum_{i_0, \ldots, i_d \in V} a_{i_0, \ldots, i_d} e_{i_0, \ldots, i_d}, \quad \text{where } a_{i_0, \ldots, i_d} \in \mathbb{F}. $$

Similar to simplicial complexes, there is a well-defined boundary operator $\partial : \Lambda_d \to \Lambda_{d-1}$:

$$ \partial e_{i_0, \ldots, i_d} = \sum_{i_0, \ldots, i_d \in V} (-1)^j e_{i_0, \hat{i}_j, \ldots, i_d}, $$

where $\hat{i}_j$ means the omission of index $i_j$. The boundary of a path $p = \sum_{i_0, \ldots, i_d \in V} a_{i_0, \ldots, i_d} e_{i_0, \ldots, i_d}$, is thus $\partial p = \sum_{i_0, \ldots, i_d \in V} a_{i_0, \ldots, i_d} \partial e_{i_0, \ldots, i_d}$. We set $\Lambda_{-1} = 0$ and note that $\Lambda_0$ is the set of $\mathbb{F}$-linear combinations of vertices in $V$. [12, Lemma 2.4] shows that $\partial \circ \partial = 0$.

Next, we restrict to real paths in directed graphs. Specifically, given a directed graph $G = (V, E)$, call an elementary $d$-path $e_{i_0, \ldots, i_d}$ allowed if there is an edge from $i_k$ to $i_{k+1}$ for all $k \in [0, d - 1]$. Define $A_d$ as the space of all allowed $d$-paths, that is, $A_d := \text{span} \{e_{i_0, \ldots, i_d} : i_0, \ldots, i_d \text{ is allowed}\}$. An elementary $d$-path $i_0, \ldots, i_d$ is called regular if $i_k \neq i_{k+1}$ for all $k$, and is irregular otherwise. Clearly, every allowed path is regular since there is no self-loop. However, the boundary map $\partial$ on $\Lambda_d$ may create a term resulting into an irregular path. For example, $\partial e_{vu} = e_{vu} - e_{u} + e_{uv}$ is irregular because of the term $e_{uv}$. To deal with this case, the term containing consecutive repeated vertices is identified with 0 [12]. Thus, for the previous example, we get $\partial e_{vu} = e_{vu} - 0 + e_{uv} = e_{vu} + e_{uv}$. The boundary map $\partial$ on $A_d$ is taken to be the boundary map for $\Lambda_d$ restricted on $A_d$ with this modification: where all terms with consecutive repeated vertices created by the boundary map $\partial$ are replaced with 0’s.

Unfortunately, after restricting to the space of allowed paths $A_d$, the inclusion that $\partial A_d \subset A_{d-1}$ may not hold any more; that is, the boundary of an allowed $d$-path is not necessarily an allowed $(d - 1)$-path. To this end, we adopt a stronger notion of allowed paths: an allowed path $p$ is $\partial$-invariant if $\partial p$ is also allowed. Let $\Omega_d := \{p \in A_d : \partial p \in A_{d-1}\}$ be the space generated by all $\partial$-invariant paths. We then have $\partial \Omega_d \subset \Omega_{d-1}$ (as $\partial^2 = 0$). This gives rise to the following chain complex of $\partial$-invariant allowed paths:

$$ \cdots \Omega_d \xrightarrow{\partial} \Omega_{d-1} \xrightarrow{\partial} \cdots \Omega_d \xrightarrow{\partial} \Omega_0 \xrightarrow{\partial} 0. $$
We can now define the homology groups of this chain complex. The $d$-th cycle group is defined as $Z_d = \text{Ker} \partial |_{\Omega_d}$, and elements in $Z_d$ are called $d$-cycles. The $d$-th boundary group is defined as $B_d = \text{Im} \partial |_{\Omega_{d+1}}$, with elements of $B_d$ being called $d$-boundary cycles (or simply $d$-boundaries). The resulting $d$-dimensional path homology group is $H_d(G, \mathbb{F}) = Z_d / B_d$.

**An example (Fig. 1)** As in any directed graph, the set of edges generate the allowed 1-paths space $A_1$ in this example. There are six allowed elementary 2-paths $e_{cba}, e_{dcb}, e_{fcb}, e_{fga}, e_{fgb}, e_{gba}$ generating $A_2$. Among these six paths, only $e_{gba}$ is $\partial$-invariant, i.e., $e_{gba} \in \Omega_2$. However, note that while neither $e_{fgb}$ nor $e_{fcb}$ is $\partial$-invariant, the 2-path $e_{fgb} - e_{fcb} \in \Omega_2$ because $\partial (e_{fgb} - e_{fcb}) = e_{gb} - e_{fb} + e_{fg} - (e_{cb} - e_{fb} + e_{fc}) = e_{gb} + e_{fg} - e_{cb} - e_{fc} \in A_1$. In fact, $\Omega_2$ is generated by 2-paths $e_{fgb} - e_{fcb}$ and $e_{gba}$.

Consider three cycles: $C_1 = e_{gb} + e_{ba} - e_{ga}$, $C_2 = e_{fc} + e_{cb} - e_{gb} - e_{fg}$ and $C_3 = e_{dc} - e_{fc} + e_{fe} - e_{de}$. It is easy to prove that $C_1$, $C_2$, and $C_3$ are cycles that generate $Z_1$. But, only $C_1$, $C_2$ are 1-boundaries which generate $B_1$. Thus, the 1-dimensional path homology group $H_1$ is generated by the class homologous to $C_3$ and $\dim(H_1) = 1$.

### 2.1 Examples of 1-Boundaries

Below we give three examples of 1-boundaries; see Fig. 2.

**Bigon** A bigon is a 1-cycle $e_{uv} + e_{vu}$ consisting of two edges $(u, v)$ and $(v, u)$ from $E$; see Fig. 2a. Consider the 2-path $e_{uvw}$. We have that its boundary is $\partial e_{uvw} = e_{vu} -$
We denote a boundary quadrangle \( e_{uv} \). Since both \( e_{uv} \) and \( e_{uv} \) are allowed 1-paths, it follows that any bigon \( e_{uv} + e_{uw} \) of \( G \) is necessarily a 1-boundary.

**Boundary triangle** Consider the 1-cycle \( C = e_{uv} - e_{uw} + e_{uv} \) of \( G \) (it is easy to check that \( \partial C = 0 \)). Now consider the 2-path \( e_{uvw} \); its boundary is then \( \partial e_{uvw} = e_{vw} - e_{uw} + e_{uv} = C \). Note that every summand in the boundary is allowed. Thus \( C \) is a 1-boundary. We call any triangle in \( G \) isomorphic to \( C \) a boundary triangle. Note that a boundary triangle always has one sink and one source; see the source is a 1-boundary. We call any triangle in \( G \) a 1-cycle consisting of four edges whose undirected version has the form \((u, v) + (v, w) + (w, z) + (z, u)\). (Note that a quadrangle may not be a boundary quadrangle).

In the remainder of the paper, we use \( R(u, v, w, z) \) to represent a quadrangle; i.e., a 1-cycle consisting of four edges whose undirected version has the form \((u, v) + (v, w) + (w, z) + (z, u)\). (Note that a quadrangle may not be a boundary quadrangle).

**Boundary quadrangle** Consider the 1-cycle \( C = e_{uv} + e_{uv} - e_{uz} - e_{zw} \) from \( G \). It is easy to check that \( C \) is the boundary of the 2-path \( e_{uvw} - e_{uzw} \), as \( \partial (e_{uvw} - e_{uzw}) = e_{uv} - e_{uw} + e_{uv} - (e_{zw} - e_{uw} + e_{uz}) = e_{uv} + e_{uv} - e_{zw} - e_{uz} = C \). We call any quadrangle isomorphic to \( C \) a boundary quadrangle.

In the remainder of the paper, we use \( R(u, v, w, z) \) to represent a quadrangle; i.e., a 1-cycle consisting of four edges whose undirected version has the form \((u, v) + (v, w) + (w, z) + (z, u)\). (Note that a quadrangle may not be a boundary quadrangle).

**3 Computing 1-Dimensional Path Homology \( H_1 \)**

Note that the 1-dimensional \( \partial \)-invariant path space \( \Omega_1 = \Omega_1(G) \) is the space generated by all edges \([12]\) because the boundary of every edge is allowed by definition.

Now consider the 1-cycle group \( Z_1 \subseteq \Omega_1 \); that is, \( Z_1 \) is the kernel of \( \partial \) applied to \( \Omega_1 \). We show below that a basis of \( Z_1 \) can be computed by considering a spanning tree of the undirected version of \( G \), which is denoted by \( G_u \). This is well known when \( \mathbb{F} \) is \( \mathbb{Z}_2 \), see e.g. \([19, \text{Chapter VIII}]\). It is easy to see that this spanning tree based construction also works for arbitrary field \( \mathbb{F} \).

Specifically, let \( T \) be a rooted spanning tree of \( G_u \) with root \( r \), and \( \overline{T} := G_u \setminus T \). For every edge \( e = (v_1, v_2) \in \overline{T} \), let \( c_e \) be the 1-cycle (under \( \mathbb{Z}_2 \)) obtained by summing \( e \) and all edges on the cycles \( \pi_1 \) and \( \pi_2 \) between \( v_1 \) and \( r \), and \( v_2 \) and \( r \) respectively. The cycles \( \{c_e : e \in \overline{T} \} \) form a basis of 1-cycle group of \( G_u \) under \( \mathbb{Z}_2 \) coefficient. Now for every such cycle \( c_e \) in \( G_u \), we also have a cycle in \( \Omega_1(G) \) containing same edges with \( c_e \) which are assigned a coefficient 1 or \(-1\) depending on their orientations in \( G \). We call this 1-cycle in \( \Omega_1(G) \) also \( c_e \). Then we have the following proposition.

**Proposition 3.1** The cycles \( \{c_e : e \in \overline{T} \} \) in \( \Omega_1(G) \) form a basis for \( Z_1 \) under any coefficient field \( \mathbb{F} \).

**Proof** First, it is obvious that the cycles \( \{c_e : e \in \overline{T} \} \) are independent because every cycle contains a unique edge that does not exist in other cycles, which means any cycle cannot written as linear combination of other cycles in the set. Now what remains is to show that the cycles \( \{c_e : e \in \overline{T} \} \) generate \( Z_1 \). Consider any 1-cycle \( c \neq 0 \) in \( Z_1 \) with coefficients in \( \mathbb{F} \). Observe that \( c \) must have at least one edge from \( \overline{T} \) because otherwise it will have non-zero coefficients only on edges in \( T \) whose boundary cannot

\( \square \) Springer
be 0. Consider an edge \( e_0 \in \bar{T} \) with non-zero coefficient \( a_0 \in \mathbb{F} \) in \( c \). The cycle \( c' = -a_0c_0 + c \) has \( e_0 \) with zero coefficient. If \( c' \) is not 0, continue the argument again and we are guaranteed to derive a null cycle ultimately because every time we make the coefficient of an edge belonging to \( \bar{T} \) zero. This means that we have \( c - a_0c_0 - \cdots - a_kc_k = 0 \) for some \( k \). In other words \( c = \sum_{i=0}^{k} a_i c_i \). It immediately follows that \( \{ c_e : e \in \bar{T} \} \) form a basis for \( \mathbb{Z}_1 \).

Now, we show a relation between 1-dimensional homology, cycles, bigons, triangles, and quadrangles. Recall that bigons, boundary triangles, and boundary quadrangles are specific types of 1-dimensional boundaries with two, three or four vertices, respectively; see Sect. 2.1. The following theorem is similar to [13, Proposition 2.9], where the statement there is under coefficient ring \( \mathbb{Z} \).

**Theorem 3.2** Let \( G = (V, E) \) be a directed graph. Let \( Q \) denote the space generated by all boundary triangles, boundary quadrangles, and bigons in \( G \). Then we have \( B_1 = Q \).

**Proof** First, by discussions in Sect. 2.1, it is easy to see that \( Q \subseteq B_1 \subseteq Z_1 \). We now show that \( B_1 \subseteq Q \). To this end, consider any 2-path (2-chain) \( p = \sum a_{uvw}e_{uvw} \in \Omega_2 \), where \( a_{uvw} \in \mathbb{F} \) is the coefficient of the elementary 2-path \( e_{uvw} \). Its boundary is \( \partial p = \sum a_{uvw}(e_{uw} - e_{uw} + e_{vw}) \) which we will argue to be in the space \( Q \). Now set \( C = \partial p \).

Specifically, consider an elementary 2-path \( e_{uvw} \in C \) such that \( e_{uvw} \neq 0 \). We have that \( \partial e_{uvw} = e_{uw} - e_{uw} + e_{vw} \). Note that at this point, we do not yet know whether the allowed 2-path \( e_{uvw} \) is also an \( \partial \)-invariant path yet (i.e., it is not clear whether \( \partial e_{uvw} \in \Omega_1 \)). Nevertheless, as \( e_{uvw} \) is an allowable 2-path, edges \( (u, v), (v, w) \in E \) (that is, both 1-paths \( e_{uv} \) and \( e_{vw} \) are allowed). As for the 1-path \( e_{uw} \), we have three cases:

(i) \( u = w \), we have a bigon.

(ii) \( u \neq w \) and \( (u, w) \) is an edge in \( G \). In this case, \( e_{uv} - e_{uw} + e_{vw} \) is allowed. As it is also a cycle, it follows that \( e_{uv} - e_{uw} + e_{vw} \in \Omega_1 \), and in fact, it forms a boundary triangle.

(iii) \( u \neq w \) and \( (u, w) \notin E \): In this case, \( e_{uw} \) cannot exist in \( \partial C \) because \( C \in \Omega_2 \), meaning that \( \partial C \) must be an allowed 1-path. In other words, \( e_{uw} \) has to be cancelled by the boundary of some other elementary 2-path \( e_{uv'w} \) with \( a_{uv'w} \neq 0 \) in \( C \). Since \( (u, w) \notin E \), the 2-path \( e_{uv'w} \) is not a bigon nor a boundary triangle. That is, there is a 2-chain \( e_{uv'w} \) with \( a_{uv'w} \neq 0 \), whose boundary equals \( e_{uv'} - e_{uw} + e_{v'w} \), and \( (u, v'), (v'w) \in E \). Hence \( \partial(e_{uvw} - e_{uv'w}) = e_{uv} + e_{v'w} - e_{uv'} - e_{v'w} \) forms a boundary quadrangle.

We now update \( C' = C - a_{uvw}e_{uvw} \) for (i) and (ii), or \( C' = C - a_{uvw}(e_{uvw} - e_{uv'w}) \) for case (iii). It is easy to see that \( C' \) contains fewer terms with non-zero coefficients than \( C \). We repeat the above argument to \( C' \) till it becomes the zero. As a result, \( C = \partial p \) is decomposed to be a combination of bigons, boundary triangles and boundary quadrangles. It follows that for any 2-chain \( p \in \Omega_2 \), we have its boundary \( \partial p \in Q \), which proves \( B_1 \subseteq Q \). Putting everything together, we have that \( B_1 = Q \).

**Corollary 3.3** The 1-dimensional path homology group satisfies that \( H_1 = Z_1/Q \).
Fig. 3 There are \( n \) vertices but \( l_s \cdot l_t = \Theta(n^2) \) quadrangles, \( l_s = \lfloor (n-2)/2 \rfloor \) and \( l_t = \lceil (n-2)/2 \rceil \)

3.1 A Simple Algorithm

Theorem 3.2 and Corollary 3.3 provide us a simple framework to compute \( H_1 \). Below we only focus on the computation of the rank of \( H_1 \); but the algorithm can easily be modified to output a basis for \( H_1 \) as well. Later in Sect. 4, we will develop a more efficient and sophisticated algorithm for the 1-dimensional persistent path homology \( H_1 \), which as a by-product, also gives a more efficient algorithm to compute \( H_1 \).

In the remaining of this paper, we represent each cycle in \( Z_1 \) with a vector. Assume all edges are indexed from 1 to \( m \) as \( e_1, \ldots, e_m \) where \( m \) is the number of edges. Then, each 1-cycle \( C \) is an \( m \)-dimensional vector, where \( C[i] \in \mathbb{F} \) records the coefficient for edge \( e_i \) in \( C \).

**Algorithm 1** A simple first algorithm to compute rank of \( H_1 \)

1: procedure \textsc{CompH1-Simple}(\( G, t \))
2: \hspace{1em} (Step 1): Compute rank of 1-cycle group \( Z_1 \)
3: \hspace{1em} (Step 2): Compute rank of 1-boundary group \( B_1 \)
4: \hspace{2em} (Step 2.a) Compute a generating set \( C \) of 1-boundary cycles that generates \( B_1 \)
5: \hspace{2em} (Step 2.b) From \( C \) compute a basis for \( B_1 \)
6: \hspace{1em} Return rank(\( H_1 \)) = rank(\( Z_1 \)) − rank(\( B_1 \)).
7: end procedure

(Step 1): cycle group \( Z_1 \) By Proposition 3.1, \( \text{rank}(Z_1) = |E| - |V| + 1 \) for directed graph \( G = (V, E) \). The computation of the rank takes \( O(1) \) time (or \( O(|V| \times |E|) \) time if we need to output a basis of it explicitly).

(Step 2): boundary group \( B_1 \) Note that by Theorem 3.2, we can compute the set of all bigons, boundary triangles and boundary quadrangles as a generating set \( C \) of 1-boundary cycles (meaning that it generates the boundary group \( B_1 \)) for (Step 2.a). However, such a set \( C \) could have size \( \Omega(n^2) \) even for a planar graph, where \( n = |V| \); see Fig. 3. (For a general graph, the number of boundary quadrangles could be \( \Theta(n^4) \).)³

³ See [8] for the definition of \( \Omega \) and \( \Theta \).
To make (Step 2.b) efficient, we wish to have a generating set $C$ of 1-boundary cycles with small cardinality. To this end, we leverage a classical result of [4] to reduce the size of $C$.

Given an undirected graph $G$ where the number of multi-edges between any two vertices is constant, its arboricity $\alpha(G)$ is the minimum number of edge-disjoint spanning forests which $G$ can be decomposed into [15]. An alternative definition is

$$\alpha(G) = \max_{H \text{ is a subgraph of } G} \frac{|E(H)|}{|V(H)| - 1}.$$  

From this definition, it is easy to see (and well-known, see e.g. [4]) that:

**Observation 3.4** Given an undirected graph $G$ where the number of multi-edges between any two vertices is constant:

(i) If $G$ is a planar graph, or a graph with bounded vertex degrees, then $\alpha(G) = O(1)$.
(ii) If $G$ is a graph embedded on a genus $g$ surface, then $\alpha(G) = O(g)$.
(iii) In general, if $G$ does not contain self-loops, then $\alpha(G) = O(n)$.

We will leverage some classical results from [4]. First, to represent quadrangles, we use the following triple-list representation [15]: a triple-list $(u, v, \{w_1, w_2, \ldots, w_l\})$ means that for each $i$, $w_i$ is adjacent to both $u$ and $v$, where we say $u'$ and $v'$ are adjacent if either $(u', v')$ or $(v', u')$ are in $E$ (i.e., $u'$ and $v'$ are adjacent when disregarding directions). Given such a triple-list $\xi = (u, v, \{w_1, w_2, \ldots, w_l\})$, it is easy to see that $u, w_i, v, w_j$ form the consecutive vertices of a quadrangle in the undirected version of graph $G$; and we also say that the undirected quadrangle $R(u, w_i, v, w_j)$ is covered by this triple-list. We say that a vertex $z$ is in a triple-list $(u, v, \{w_1, w_2, \ldots, w_l\})$ if it is in the set $\{w_1, w_2, \ldots, w_l\}$.

The size of a triple-list is the total number of vertices contained in it. This triple-list $\xi$ thus represents $\Theta(l^2)$ number of undirected quadrangles in $G$ succinctly with $\Theta(l)$ size.

**Proposition 3.5** [4]

(i) Let $G$ be a connected undirected graph with $n$ vertices and $m$ edges. There is an algorithm listing all the triangles in $G$ in $O(\alpha(G)m)$ time.
(ii) There is an algorithm to compute a set of triple-lists which covers all quadrangles in a connected graph $G$ in $O(\alpha(G)m)$ time. The total size complexity of all triple-lists is $O(\alpha(G)m)$.

Using the above result, we can have the following theorem.

**Theorem 3.6** Let $G = (V, E)$ be a directed graph with $n$ vertices and $m$ edges. We can compute a generating set $C$ of 1-boundary cycles for $B_1$ with cardinality $O(\alpha(G)m)$ in time $O(\alpha(G)m)$.

**Proof** We will show that we can generate all bigons, boundary triangles and boundary quadrangles, via a generating set $C$ satisfying the requirement in the theorem.

First, note that the total number of bigons is $O(m)$ since there are $m$ directed edges and for each edge, there is at most one bigon created as there is no multi-edge in $G$. Next, by Proposition 3.5, we have that the total number of undirected
triangles is $O(a(G)m)$, which further bounds the total number of directed triangles, as well as that of the boundary triangles by $O(a(G)m)$. In particular, we enumerate all $O(a(G)m)$ undirected triangles in $G$ in $O(a(G)m)$ time, and if the vertices form a boundary triangle, we add it to the generating set $C$. This adds all $O(a(G)m)$ number of boundary triangles to $C$ in $O(a(G)m)$ time.

The case for quadrangles is more involved. The total number of quadrangles could be $\Theta(n^4)$ and we aim to find a subset of $O(a(G)m)$ that generate all boundary quadrangles to add to $C$.

By Proposition 3.5, in $O(a(G)m)$ time, we can compute a list $L$ of triple-lists with $O(a(G)m)$ total size complexity that generate all undirected quadrangles. Note that this also implies that $|L| = O(a(G)m)$.

Now for a triple-list $\xi_{uv} = (u, v, \{w_1, w_2, \ldots, w_l\})$, we will generate three lists as follows (see Fig. 4):

**Type-1 list** a triple-list $\xi_{uv}^{(1)} = (u, v, \{w_1^1, w_2^1, \ldots, w_l^1\})$ where $w_i^1$ is in $\xi_{uv}$ if and only if $w_i^1 \in \xi_{uv}$, and both edges $(u, w_i^1)$ and $(w_i^1, v)$ are in $E$. We say that a boundary quadrangle can be generated by the type-1 list $\xi_{uv}^{(1)}$ if it is of the form $(u, v | w_i^1, w_j^1)$, with $i \neq j \in [1, l_1]$.

**Type-2 list** a triple-list $\xi_{uv}^{(2)} = (u, v, \{w_1^2, w_2^2, \ldots, w_l^2\})$ where $w_i^2$ is in $\xi_{uv}$ if any only if $w_i^2 \in \xi_{uv}$, and the edges $(v, w_i^2)$ and $(w_i^2, u)$ are in $E$. We say that a boundary quadrangle can be generated by the type-2 list $\xi_{uv}^{(2)}$ if it is of the form $(v, u | w_i^2, w_j^2)$, with $i \neq j \in [1, l_2]$.

**Type-3 list** a so-called quadruple-list $\xi_{uv}^{(3)} = (u, v, \{w_1^3, w_2^3, \ldots, w_l^3\})$, where each $w_i^3$ and $w_j^3$ are in $\xi_{uv}$, and edges $(w_i^3, u), (w_i^3, v), (u, w_i^3)$ and $(v, w_i^3)$ are in $E$, for $1 \leq i \leq l_3$, $1 \leq j \leq l_4$. We say that a boundary quadrangle can be generated by the type-3 list $\xi_{uv}^{(3)}$ if it is of the form $(w_i^1, w_j^1 | u, v)$, with $i \in [1, l_3]$ and $j \in [1, l_4]$.

It is easy to see that $l_1 + l_2 + l_3 + l_4 = O(l)$. Let $\tilde{L}$ denote all lists generated by triples in $L$. Note that total size complexity of $\tilde{L}$ (which is the sum of the size of each triple-list or quadruple-list in $\tilde{L}$) is still $O(a(G)m)$.

---

**Fig. 4** Three lists extracted from $\xi = (u, v, \{w_1, w_2, \ldots, w_l\})$
Denote as $\mathcal{R}$ the set of all boundary quadrangles in $G$; note that by definition, each element in $\mathcal{R}$ is a 1-boundary. On the other hand, each boundary quadrangle in $\mathcal{R}$ is generated by some list in $\hat{\mathcal{L}}$. Indeed, given an arbitrary boundary quadrangle $(x, y | z, s)$, by Proposition 3.5, we know that its undirected version $R(x, z, y, s)$ must be covered by some triple-list $\xi_{uv} \in \mathcal{L}$. If $x = u$ and $y = v$, then the boundary quadrangle is then generated by the type-1 list $\xi_{uv}^{(1)}$. If $x = v$ and $y = u$, then it is generated by the type-2 list $\xi_{uv}^{(2)}$. Otherwise, by the definition of $\xi_{uv}$ covering the undirected quadrangle $R(x, z, y, s)$, it must mean that $u = z$, $v = s$ or $u = s$, $v = z$.

Thus this boundary quadrangle is generated by the type-3 list $\xi_{uv}^{(3)}$. This proves that all boundary quadrangles in $\mathcal{R}$ are generated by the lists in the set $\hat{\mathcal{L}}$.

Finally, we will add boundary quadrangles to $\mathcal{C}$ as follows: We inspect each list in $\hat{\mathcal{L}}$.

**Case 1** If it is a type-1 list, say of the form $\xi_{uv}^{(1)} = (u, v, \{w_{1}^{1}, w_{2}^{1}, \ldots, w_{l_{1}}^{1}\})$, we add $l_{1} - 1$ number of boundary quadrangles of the form $(u, v | w_{1}^{1}, w_{1}^{1})$ to $\mathcal{C}$ for each $i \in [2, l_{1}]$.

**Case 2** If it is a type-2 list of the form $\xi_{uv}^{(2)} = (u, v, \{w_{1}^{2}, w_{2}^{2}, \ldots, w_{l_{2}}^{2}\})$, we then add $l_{2} - 1$ number of boundary quadrangles of the form $(v, u | w_{1}^{2}, w_{1}^{2})$ to $\mathcal{C}$ for each $i \in [2, l_{2}]$.

**Case 3** If it is a type-3 list of the form $\xi_{uv}^{(3)} = (u, v, \{w_{1}^{3}, w_{2}^{3}, \ldots, w_{l_{3}}^{3}\}, \{w_{1}^{4}, w_{2}^{4}, \ldots, w_{l_{4}}^{4}\})$, then we add $l_{3} + l_{4} - 1$ number of boundary quadrangles of the form $(w_{i}^{3} | u, v)$ or $(w_{j}^{4} | u, v)$ into $\mathcal{C}$, for each $i \in [1, l_{4}]$ and $j \in [2, l_{3}]$.

Note that for each case above, the subset of quadrangles we add to $\mathcal{C}$ can generate all the boundary quadrangles generated by the corresponding list $\xi_{uv}^{(k)}$ for $k = 1, 2$, or 3. Indeed, for (Case 1): each boundary quadrangle $(u, v | w_{1}^{1}, w_{1}^{1})$ generated by $\xi_{uv}^{(1)}$ can be written as the linear combination of $(u, v | w_{1}^{1}, w_{j}^{1})$ and $(u, v | w_{1}^{1}, w_{j}^{1})$, both of which are added to $\mathcal{C}$. (Case 2) can be argued in a symmetric manner. For (Case 3), consider a boundary quadrangle $(w_{j}^{3}, w_{1}^{4} | u, v)$ generated by $\xi_{uv}^{(3)}$. It can be written as the combination of $(w_{3}^{3}, w_{1}^{4} | u, v)$, $(w_{3}^{3}, w_{1}^{4} | u, v)$, and $(w_{3}^{3}, w_{1}^{4} | u, v)$, all of which on the righthand side are added to $\mathcal{C}$.

Hence in summary, the set of quadrangles we add to $\mathcal{C}$ will generate all boundary quadrangles $\mathcal{R}$. Furthermore, note that in each case, the number of quadrangles we add to $\mathcal{C}$ is linear in the size of the list from $\hat{\mathcal{L}}$ being considered. Hence the total number of boundary quadrangles ever added to $\mathcal{C}$ is bounded by the total size of $\hat{\mathcal{L}}$ which is further bounded by $O(a(G)m)$. Putting everything together, the theorem then follows.

It then follows from Theorem 3.2 that (Step 2.a) can be implemented in $O(a(G)m)$ time, producing a generating set of cardinality $O(a(G)m)$. Finally, representing each boundary cycle in $\mathcal{C}$ as a vector of dimension $m = |E|$, we can then compute the rank of cycles in $\mathcal{C}$ in $O(|\mathcal{C}|m^{\omega-1}) = O(a(G)m^{\omega})$, where $\omega < 2.373$ is the exponent for matrix multiplication [3]. Putting everything together, we have that

**Theorem 3.7** Given a directed graph $G = (V, E)$ with $n = |V|$ and $m = |E|$, Algorithm 1 computes the rank of the 1-dimensional path homology group $H_{1}$ in

Springer
Fig. 5 A dense graph with no boundary quadrangle

\( O(\alpha(G)m^\omega) \) time. The algorithm can be extended to compute a basis for \( H_1 \) with the same time complexity.

For example, by Observation 3.4, if \( G \) is a planar graph, then we can compute \( H_1 \) in \( O(n^\omega) \). For a graph \( G \) embedded on a genus \( g \) surface, \( H_1 \) can be computed in \( O(gn^\omega) \) time. In contrast, we note that the algorithm of [6] takes \( O(n^5) \) time for planar graphs.

### 4 Computing Persistent Path Homology \( H_1 \)

The concept of arboricity used in the previous section does not consider edge directions. Indeed, our algorithm to compute a generating set \( C \) as given in the proof of Theorem 3.6 first computes a (succinct) representation of all quadrangles, whether they contribute to boundary quadrangles or not. On the other hand, as Fig. 5 illustrates, a graph \( G \) can have no boundary quadrangle despite the fact that the graph is dense (with \( \Theta(n^2) \) edges and thus \( \alpha(G) = \Theta(n) \) arboricity). Another way to view this is that the example has no allowed 2-path, and thus no \( \partial \)-invariant 2-paths and consequently no 1-boundary cycles. Our algorithm will be more efficient if it can also respect the number of allowed elementary 2-paths.

In fact, a more standard and natural way to compute a basis for the 1-boundary group proceeds by taking the boundary of \( \partial \)-invariant 2-paths. The complication is that unlike in the simplicial homology case, it is not immediately evident how to compute a basis for \( \Omega_2 \) (the space of \( \partial \)-invariant 2-paths). Nevertheless, Chowdhury and Mémoli presented an elegant algorithm to show that a basis for \( B_1 \) (and \( H_1 \)) can still be computed using careful column-based matrix reductions [6]. The time complexity of their algorithm is \( O(\left(\sum_{(u,v) \in E} (d_{\text{in}}(u) + d_{\text{out}}(v) ) \right) mn^2) \) which depends on the number of elementary 2-paths.\(^4\)

In this section, we present an algorithm that can take advantage of both of the previous approaches (the algorithm of [6] and Algorithm 1). Similarly to [6], we

\(^4\) The time complexity given in the paper [6] assumes that the input directed graph is complete, and takes \( O(n^9) \) to compute \( H_1 \). However, a more refined analysis of their time complexity shows that it can be improved to \( O(\left(\sum_{(u,v) \in E} d_{\text{in}}(u) + d_{\text{out}}(v) \right) mn^2) \).
will now consider the persistent path homology setting, where we will add directed edges in $G$ one by one incrementally by considering them in non-decreasing order of their weights. Hence our algorithm can compute the persistent $H_1$ w.r.t. a filtration. However different from [6], instead of reducing a matrix with columns corresponding to all elementary allowed 2-paths, we will follow a similar idea as in Algorithm 1 and add a generating set of boundary cycles each time we consider a new directed edge.

4.1 Persistent Path Homology

We now introduce the definition of the persistent path homology [6]. The persistent vector space is a family of vector spaces together with linear maps $\{ U^\delta \to U^{\delta'} \}_{\delta \leq \delta' \in \mathbb{R}}$ so that: (1) $\mu_{\delta,\delta}$ is the identity for every $\delta \in \mathbb{R}$, and (2) $\mu_{\delta,\delta'} \circ \mu_{\delta',\delta''} = \mu_{\delta,\delta''}$ for each $\delta \leq \delta' \leq \delta'' \in \mathbb{R}$.

Let $G = (V, E, w)$ be a weighted directed graph where $V$ is the vertex set, $E$ is the edge set, and $w$ is the weight function $w : E \to \mathbb{R}^+$. For every $\delta \in \mathbb{R}^+$, a directed graph $G^\delta$ can be constructed as $G^\delta = (V^\delta = V, E^\delta = \{ e \in E : w(e) \leq \delta \})$. This gives rise to a filtration of graphs $\{ G^\delta \hookrightarrow G^{\delta'} \}_{\delta \leq \delta' \in \mathbb{R}}$ using the natural inclusion map $i_{\delta,\delta'} : G^\delta \hookrightarrow G^{\delta'}$.

**Definition 4.1** [6] The 1-dimensional persistent path homology of a weighted directed graph $G = (V, E, w)$ is defined as the persistent vector space $H_1 := \{ H_1(G^\delta) \to H_1(G^{\delta'}) \}_{\delta \leq \delta' \in \mathbb{R}}$. The 1-dimensional path persistence diagram $Dg(G)$ of $G$ is the persistence diagram of $H_1$.

To compute the path homology $H_1(G)$ of an unweighted directed graph $G = (V, E)$, we can order edges in $E$ arbitrarily with the index of an edge in this order being its weight. The rank of $H_1(G)$ can then be retrieved from the 1-dimensional persistent homology group induced by this filtration by considering only those homology classes that “never die”.

4.2 A More Efficient Algorithm

In what follows, to simplify presentation, we assume that we are given a directed graph $G = (V, E)$, where edges are already sorted $e_1, \ldots, e_m$ in increasing order of their weights. Let $G^{(i)} = (V, E^{(i)} = \{ e_1, \ldots, e_i \})$ denote the subgraph of $G$ spanned by the edges $e_1, \ldots, e_i$; and set $G^{(0)} = (V, \emptyset)$. We now present an algorithm to compute the 1-dimensional persistent path homology induced by the nesting sequence $G^{(0)} \subseteq G^{(1)} \subseteq \ldots \subseteq G^{(m)}$. In particular, in Algorithm 2, as we insert each new edge $e_s$, moving from $G^{(s-1)}$ to $G^{(s)}$, we maintain a basis for $Z_1(s) := Z_1(G^{(s)})$ and for $B_1(s) := B_1(G^{(s)})$, updated from $Z_1(s-1)$ and $B_1(s-1)$ and output new persistent pairs. On a high level, this algorithm follows the standard procedure in [7].
**Algorithm 2** Compute 1-D persistent path homology for a directed graph $G = (V, E)$

1: procedure PERSISTENCE($G$)  
2: Order the edges in non-decreasing order of their weights: $e_1, \ldots, e_m$.  
3: Set $G^{(0)} = (V, \emptyset)$, current basis for 1-boundary group is $B = \emptyset$.  
4: for $s = 1$ to $m$ do  
5: Call GENSET($s$) to compute a generating set $C_s$ containing a basis for newly generated 1-boundary cycles moving from $G^{(s-1)}$ to $G^{(s)}$.  
6: Call FINDPAIRS($s$) to output new persistent pairs, and update the boundary basis $B$ for $G^{(s)}$.  
7: end for  
8: end procedure

4.2.1 Procedure GENSET.

Note that $G^{(s)}$ is obtained from $G^{(s-1)}$ by inserting a new edge $e_s = (u, v)$ to $G^{(s-1)}$. At this point, we have already maintained a basis $B$ for $B_1(G^{(s-1)})$. Our goal is to compute a set of generating boundary cycles $C_s$ such that $B \cup C_s$ contains a basis for $B_1(G^{(s)})$.

We first inspect the effect of adding edge $e_s = (u, v)$ to $G^{(s)}$. Two cases can happen:

(Case-A): The endpoints $u$ and $v$ are in different connected components in (the undirected version of) $G^{(s-1)}$, and after adding $e_s$, those two components are merged into a single one in $G^{(s)}$. In this case, no cycle is created, nor does the boundary group change. Thus $Z_1(G^{(s-1)}) = Z_1(G^{(s)})$ and $B_1(G^{(s-1)}) = B_1(G^{(s)})$. We say that edge $e_s$ is negative in this case (as it destroys a class in $H_0$).

The algorithm maintains the set of negative edges seen so far, which is known to form a spanning forest $T_s$ of $V$. (Here, we abuse the notation slightly and say that a set of directed edges span a tree for a set of vertices if they do so when directions are ignored.) The algorithm maintains $T_s$ via a union-find data structure.

(Case-B): The endpoints $u$ and $v$ are already in the same connected component in $G^{(s-1)}$. After adding this edge $e_s$, new cycles are created in $G^{(s)}$. Hence $e_s$ is positive in this case (as it creates an element in $Z_1$; although different from the standard simplicial homology, it may not necessarily create an element in $H_1$ as we will see later).

Whether $e_s$ is positive or negative can be easily determined by performing two Find operations in the union-find data structure representing $T_{s-1}$. A Union($u, v$) operation is performed to update $T_{s-1}$ to $T_s$ if $e_s$ is negative.

We now describe how to handle (Case-B). After adding edge $e_s$, multiple cycles containing $e_s$ can be created in $G^{(s)}$. Nevertheless, by Proposition 3.1, the dimension of $Z_1$ increases only by 1. On the other hand, the addition of $e_s$ may create new boundary cycles. Interestingly, it could increase the rank of $B_1$ by more than 1. See Fig. 6 for an example where rank($B_1$) increases by 3; and note that this number can be made arbitrarily large.

As mentioned earlier, in this case, we wish to compute a set of generating boundary cycles $C_s$ such that $B \cup C_s$ contains a basis for $B_1(G^{(s)})$.

Similarly to Algorithm 1, using Theorem 3.2, we choose some bigons, boundary triangles and boundary quadrangles and add them to $C_s$. In particular, since $C_s$ only
Fig. 6 The insertion of edge \((u, v)\) increases the rank of the boundary group by 3 accounts for the newly created boundary cycles, we only need to consider bigons, boundary triangles and boundary quadrangles that contain \(e_s\). We now describe the construction of \(C_s\), which is initialized to be \(\emptyset\).

(i) Bigons. At most one bigon can be created after adding \(e_s\) (namely, the one that contains \(e_s\)). We add it to \(C_s\) if this bigon exists.

(ii) Boundary triangles. There could be three types of newly created boundary triangles containing \(e_s = (u, v)\). The first case is when \(u\) is the source and \(v\) is the sink; see Fig. 7a. In this case multiple 2-paths may exist from \(u\) to \(v\), \(e_{uw_1}v, e_{uw_2}v, \ldots, e_{uw_p}v\), forming multiple boundary triangles of this type containing \(e_s\). However, we only need to add one triangle of them into \(C_s\), say \((u, v \mid w_1)\) since every other triangle \((u, v \mid w_j)\) can be written as a linear combination of \((u, v \mid w_1)\) and an existing boundary quadrangle \((u, v \mid w_1, w_j)\) in \(G^{(s-1)}\).

Fig. 7 Three types of boundary triangles incident to \(e_s = (u, v)\)
For the second case (see Fig. 7b) where \( u \) is the source but \( v \) is not the sink, we include all such boundary triangles to \( C_s \), i.e., \((u, w_1 \mid v)\), \( \ldots \), \((u, w_p \mid v)\). We also add all boundary triangles of the last type in which \( v \) is the sink but \( u \) is not the source to \( C_s \); see Fig. 7c. It is easy to see that \( C_s \cup B \) can generate all new boundary triangles containing \( e_s = (u, v) \).

(iii) **Boundary quadrangles.** Given an edge \( e_s = (u, v) \), there are two types of the boundary quadrangles incident to it: one has \( u \) as the source; the other has \( v \) as the sink. We focus on the first case; see Fig. 8a. The second case can be handled symmetrically. In particular, we will first compute a set \( M \) and then select a subset of quadrangles from \( M \) for adding to \( C_s \).

Specifically, take any successor \( w \) of \( v \), that is, there is an edge \((v, w) \in G^{(s−1)}\) forming an allowed 2-path \( e_{uvw} \) in \( G^{(s)} \). Before introducing the edge \( e_s \), there may be multiple allowed 2-paths \( e_{uv_1w}, e_{uv_2w}, \ldots, e_{uv lw} \) in \( G^{(s−1)} \); see Fig. 8a. For each such 2-path \( e_{uv_kw}, 1 \leq k \leq l \), a new boundary quadrangle \((u, w \mid v, v_k)\) containing \( e_s = (u, v) \) will be created. However, among all such 2-paths \( e_{uv_1w}, \ldots, e_{uv lw} \), we will pick just one 2-path, say \( e_{uvjw} \) and only add the quadrangle \((u, w \mid v, v_j)\) formed by \( e_{uvw} \) and \( e_{uvjw} \) to \( M \). Observe that any other boundary quadrangle containing 2-path \( e_{uvw} \), say \((u, w \mid v, v_k)\), can be written as a linear combination of the quadrangle \((u, w \mid v, v_j)\) and boundary quadrangle \((u, w \mid v_k, v_j)\) which is already in \( G^{(s−1)} \) (and in the span of \( B \) which is a basis for \( B_1(G^{(s−1)}) \)). In other words, \((u, w \mid v, v_j) \cup B \) generates any other boundary quadrangle containing 2-path \( e_{uvw} \). We perform this for each successor \( w \) of \( v \). Hence this step adds at most \( d_{out}^{G^{(s−1)}}(v) \) number of boundary quadrangles to the set \( M \).

Not all quadrangles in \( M \) will be added to \( C_s \). In particular, suppose we have \( p \) quadrangles \( A = \{(u, w_j \mid v, z) : 1 \leq j \leq p\} \subseteq M \) incident to the newly inserted edge \( e_s = (u, v) \) as well as another vertex \( z \), i.e., there are edges \((u, z), (w_j, z), \) and \((v, w_j)\), \( 1 \leq j \leq p \); see Fig. 8b. If there does not exist any other vertex \( u' \) such that edges \((u', z), (u', v) \in G^{(s−1)}\), then we add all quadrangles in \( A \) to \( C_s \). If this is not the case, let \( u' \) be another vertex such that \((u', z)\) and \((u', v)\) are already in \( G^{(s−1)} \); see
Fig. 8b. In this case, we only add one quadrangle from set A, say, \((u, w_1 | v, z)\) to the generating set \(C_s\).

It is easy to check that any other quadrangle \((u, w_j | v, z), 1 < j \leq p\), can be written as the combination of \((u, w_1 | v, z)\), \((u', w_1 | v, z)\), and \((u', w_j | v, z)\). As the latter two quadrangles are boundary quadrangles from \(G^{(s-1)}\), they can already be generated by \(B\). The entire process takes time \(O(|M|) = O(d_{out}^{G^{(s-1)}}(v))\). It is also easy to see that \(B \cup C_s\) can generate any boundary quadrangle containing \(e_s = (u, v)\) and with \(u\) being its source.

The case when \(v\) is the sink of a boundary quadrangle is handled symmetrically in time \(O(d_{in}^{G^{(s-1)}}(u))\). Hence the total time to compute a generating set \(C_s\) is \(O(d_{in}^{G^{(s-1)}}(u) + d_{out}^{G^{(s)}}(v))\) when inserting a single edge \(e_s = (u, v)\).

### 4.2.2 Procedure FindPairs

Given the generating set \(C_s\) and the previous basis \(B\) for \(B_1(G^{(s-1)})\), we know that \(C_s \cup B\) generate the new boundary group \(B_1(G^{(s)})\). We now extract a basis \(B_{new}\) for \(B_1(G^{(s)})\) from \(B \cup C_s\).

We represent each 1-cycle \(\gamma\) by an \(m\)-dimensional vector, also denoted by \(\gamma\), so that \(\gamma = \sum_{i=1}^{m} \gamma[i]e_i\). (Note that \(e_1, \ldots, e_m\) are sorted according to their filtration order.) A set of \(k\) cycles can now be viewed as a \(m \times k\) matrix, where the \(i\)-th column corresponds to the vector representation of the \(i\)-th cycle.

Thus columns in the matrix \(B\) correspond to cycles in an existing basis for \(B_1(G^{(s-1)})\), and are already linearly independent. Our goal is now to compute a basis of the form \([B | B']\) for the matrix \([B | C_s]\), and the columns in \(B\) and \(B'\) form a new basis \(B_{new}\) for \(B_1(G^{(s)})\).

To do so, we follow the standard persistence algorithm which would also output persistence-pairings for \(H_1\). Specifically, Let low(\(j\)) be the row index of the last non-zero entry in column \(j\) in a matrix \(A\). A matrix \(A\) is in reduced form if the low(\(j\)) in each column \(j\) is unique. We compute persistent pairs by always maintaining the basis in reduced form [7]. Here assume that \(B\) is already in reduced form. We then perform standard column reduction to convert \([B | C_s]\) into reduced form \([B | R]\). For each non-zero column \(R[j]\) in \(R\), let \(k = \text{low}(j)\) be the index of its lowest non-zero entry. Let \(e_k\) be the edge corresponding to this entry in the cycle corresponding to column \(R[j]\). This means that the cycle corresponding to \(R[j]\) is created when \(e_k\) introduced, but killed when introducing \(e_s\), since currently it is a boundary. Then we add the persistence pairing \((w(e_k), w(e_s))\) to the output persistence diagram \(Dg_1G\) for the 1-dimensional path homology.

The collection of non-zero columns in \(R\) gives rise to \(B'\). Afterwards, we update \(B\) to be \(B \cup B'\), and proceed to process the next edge \(e_{s+1}\).

### 4.3 Analysis of Algorithm 2

**Correctness** Notice that the invariant that \(B\) is a basis for \(G^{(s)}\) at the end of the for-loop (line 7 of Algorithm 2) is maintained. Furthermore, \(B\) is always in reduced form which

\[ Springer\]
is maintained via left-to-right column additions only. Hence the algorithm computes the 1-dimensional persistent path homology correctly [7].

**Time complexity analysis** The remainder of this section is devoted to determining the time complexity of Algorithm 2. Specifically, we first show the following theorem.

**Theorem 4.2** Across all stages \( s \in [1, m] \), the total cardinality of the generating set \( C = \bigcup_s C_s \) is \( O(\min\{a(G)m, \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\}) \). The total time taken by procedure NEWBASIS\((s)\) for all \( s \in [1, m] \) is \( O(m + \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))) \).

**Proof** We will count separately the number of bigons, boundary triangles, and boundary quadrangles added to any \( C_s \). Set \( r = \min\{a(G)m, \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\} \).

(i) **Bigons:** First, it is easy to see that for each edge \( e_s = (u, v) \) with \( s \in [1, m] \), at most one bigon (incident to \( e_s \)) is added. Besides, if \( d_{out}(v) = 0 \), there is no bigon incident to \( e_s \). Hence the total number ever added to \( C \) is \( O(\min\{m, \sum_{(u,v)\in E}(d_{out}(v))\}) = O(r) \) and it takes \( O(m) \) time to compute them.

(ii) **Boundary triangles:** For boundary triangles, we know from Proposition 3.5 that there are altogether \( O(a(G)m) \) triangles (thus at most \( O(a(G)m) \) boundary triangles) in a graph \( G \) and they can all be enumerated in \( O(a(G)m) \) time. Obviously, the number of boundary triangles ever added to \( C \) is at most \( O(a(G)m) \).

We now argue that the number of boundary triangles added to \( C \) is also bounded by \( O\left(\sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\right) \). Note that, for every 2-path, at most one boundary triangle is added to the set. Since the number of 2-paths is indeed \( \Theta\left(\sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\right) \), the number of triangles we add is \( O\left(\sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\right) \). Recall that there are three cases for boundary triangles added; see Fig. 7. The time spent for the first case for every \( s \) is \( O(1) \) by recording any 2-path \( e_s(u,w,v) \), and \( O(d_{in}(u) + d_{out}(v)) \) for the last two cases. Thus the total time spent at adding boundary triangles incident to \( e_s \) and identifying triangles to be added to \( C_s \) for all \( s \in [1, m] \) takes \( O(m + \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))) \) time.

(iii) **Boundary quadrangles:** The situation here is somewhat opposite to that of the boundary triangles: Specifically, it is easy to see that this step accesses at most \( O(d_{in}(u) + d_{out}(v)) \) boundary quadrangles when handling edge \( e_s = (u, v) \). Hence the number of boundary quadrangles it can add to \( C_s \) is at most \( O(d_{in}(u) + d_{out}(v)) \). The total number of boundary quadrangles ever added to \( C \) is thus bounded by \( O\left(\sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\right) \).

We now prove that the number of boundary quadrangles ever added to \( C \) is also bounded by \( O(a(G)m) \). We use the existence of a succinct representation of all quadrangles as specified in Proposition 3.5 to help us argue this upper bound. Notice that our algorithm does not compute this representation. It is only used to provide this complexity analysis.

Specifically, by Proposition 3.5, we can compute a list \( L \) of triple-lists with \( O(a(G)m) \) total size complexity, which generates all undirected quadrangles. Following the proof of Theorem 3.6, we can further refine this list, where each triple-list \( \xi \in L \) further gives rise to three lists that are of type 1, 2, or 3. Let \( \tilde{L} \) denote this refinement of \( L \), consisting of lists of type 1, 2, or 3. From the proof of Theorem 3.6,
we know that the total size complexity for all lists in \( \hat{L} \) is still \( O(a(G)m) \). This also implies that the cardinality of \( \hat{L} \) is bounded by \( |\hat{L}| = O(a(G)m) \).

We now denote by \( \mathcal{R} \) the set of all boundary quadrangles ever added to \( \mathcal{C} = \bigcup_s \mathcal{C}_s \) by Algorithm 2. Furthermore, let

\[
P := \{ (\xi, w) : \xi \in \hat{L}, w \in \xi \}.
\]

Below we show that we can find an injective map \( \pi : \mathcal{R} \to P \). But first, note that \(|P|\) is proportional to the total size complexity of \( \hat{L} \) and thus is bounded by \( O(a(G)m) \).

We now establish the injective map \( \pi : \mathcal{R} \to P \). Specifically, we process each boundary quadrangle in the order that they are added to \( \mathcal{C} \). Consider a boundary quadrangle \( R = R(u, v, w, z) \) added to \( \mathcal{C}_s \) while processing edge \( e_s = (u, v) \). There are two cases: The first is that \( R \) is of the form \( (u, w \mid v, z) \) in which \( u \) is the source of this quadrangle. The second is that it has the form \( (w, v \mid u, z) \) in which \( v \) is the sink. We describe the map \( \pi(R) \) for the first case, and the second one can be analyzed symmetrically.

By construction, there is at least one triple-list \( \xi \in L \) covering \( R = (u, w \mid v, z) \). There are three possibilities:

1. \textbf{(Case-a):} The triple-list \( \xi \) is of the form \( \xi = \xi_{uw} = (u, w, \ldots) \). In this case, the boundary quadrangle \( R = (u, w \mid v, z) \) is in a type-1 list \( \hat{L} = \{u, w, S\} \in \hat{L} \), and both \( v, z \in S \). We now claim that the pair \((\xi_{uw}^{(1)}, v) \in P \) has not yet been mapped (i.e., there is no \( R' \in C \) with \( \pi(R') = (\xi_{uw}^{(1)}, v) \) yet), and we can thus set \( \pi(R) = (\xi_{uw}^{(1)}, v) \in P \). Suppose on the contrary there already exists \( R' \in C \) that we processed earlier than \( R \) with \( \pi(R') = (\xi_{uw}^{(1)}, v) \). In that case, \( R' = (u, w \mid v, z') \) must contain the 2-path \( e_{uw} \) as well. Since \( R' \) is processed earlier than \( R \), and edge \( e_s = (u, v) \) is the most recent edge added, \( R' \) must be added when we process \( e_s \) as well (as \( R' \) contains \( e_s \)).

2. \textbf{(Case-b):} The triple-list \( \xi \) is of the form \( \xi = \xi_{wu} = (w, u, \ldots) \). In this case, this quadrangle is covered by the type-2 list \( \xi_{wu}^{(2)} \in \hat{L} \). We handle this in a manner symmetric to (Case-a) and map \( \pi(R) = (\xi_{wu}^{(2)}, v) \).

3. \textbf{(Case-c):} The last case is that \( R \) is generated by triple-list \( \xi \) of the form \( \xi_{vz} = (v, z, \ldots) \). In this case, the quadrangle \( R = (u, w \mid v, z) \) will be covered by the type-3 list \( \xi_{vz}^{(3)} = (v, z, S_1, S_2) \) with \( u \in S_1 \) and \( w \in S_2 \); see Fig. 9. We now argue that at least one of \((\xi_{vz}^{(3)}, u)\) and \((\xi_{vz}^{(3)}, w)\) has not been mapped under \( \pi \) yet.

Suppose this is not the case and we already have both \( \pi(Q_1) = (\xi_{vz}^{(3)}, u) \) and \( \pi(Q_2) = (\xi_{vz}^{(3)}, w) \). Then \( Q_1 \) is necessarily of the form \( (u, w' \mid v, z) \) and \( Q_2 \) is of the form \( (u', w \mid v, z) \); and both \( Q_1 \) and \( Q_2 \) are processed before \( R \). See Fig. 9. Furthermore, \( Q_1 \) is only added when we process edge \( e_s \). However, in this case, once \( Q_1 \) is added, Algorithm 2 will not add further quadrangle containing edges \( (u, v) \) and \( (u, z) \) (recall the handling of Fig. 8b). Hence \( R \) cannot be added to \( \mathcal{C}_s \) in this case. In other words, it cannot be that both \( Q_1 \) and \( Q_2 \) already exist, and hence we can set \( \pi(R) \) to
be one of \((\xi^{(3)}_{uv}, u)\) and \((\xi^{(3)}_{uv}, w)\) that is not yet mapped. Consequently, the map \(\pi\) we construct remains injective.

We process all quadrangles in \(C\) in order. The final \(\pi: \mathcal{R} \to \mathcal{P}\) is injective, meaning that \(|\mathcal{R}| \leq |\mathcal{P}|\) and thus \(|\mathcal{R}| = O(\hat{a}(G)m)\). Putting everything together, we have that the total number of boundary quadrangles added to \(C\) is bounded by \(O\left(\min\{\hat{a}(G)m, \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\}\right)\). Finally, Algorithm 2 spends \(O(m + \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)))\) time to handle both cases in Fig. 8. The theorem then follows.

We now prove our main result, Theorem 1.1. In particular, below we first show that Algorithm 2 takes \(O(rm^2)\) time to compute the 1-dimensional persistent path homology, where

\[
 r = \min \left\{ \hat{a}(G)m, \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)) \right\}.
\]

We then describe how to improve the time complexity to \(O(rm^{\omega-1} + ma(n))\) to establish Theorem 1.1.

Specifically, by using a union-find data structure to maintain a spanning forest for \(G^{(s)}\), it takes \(O(m\alpha(n))\) time, where \(\alpha(\cdot)\) is the inverse Ackermann function, to check whether each edge \(e_s\) is negative or positive (forming a cycle or not). As explained in the algorithm, if it is negative, then nothing needs to be done. If it is positive, then we need to compute a generating set as well as find new basis for the boundary group and compute persistence.

By Theorem 4.2, the total time complexity spent on all executions of line 5 of Algorithm 2 is \(O(m + \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)))\). We now argue that this value is bounded by \(O(rm + ma(n))\). This clearly holds if \(r = \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\). If on the other hand \(r = \hat{a}(G)m\), then \(r = \Omega(m)\) and since \(\sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)) = \)

![Fig. 9 At least one of \((\xi^{(3)}_{uv}, u)\) and \((\xi^{(3)}_{uv}, w)\) has not been mapped yet](image-url)
\[ O(mn), \text{ we have that} \]

\[
m + \sum_{(u,v)\in E} (d_{in}(u) + d_{out}(v)) = O(r) + O(mn) = O(rm + m\alpha(n)).
\]

Next we bound the total time taken by all executions of line 6 of Algorithm 2 (i.e., on calling procedure FINDPAIRS(s) for all \( s \in \{1, m\} \)). Note that at the \( s \)-th stage, we need to reduce \( |C_s| \) number of columns in a matrix of the form \([B|C_s]\), each of which is of length \( m \). Furthermore, since when reducing a specific column in a left-to-right manner, the number of non-zero columns to its left is bounded by the size of rank \( (B_1(G^{(s)})) = O(m) \), each column-reduction takes \( O(m^2) \) time. As \( |C| = \sum_s |C_s| = O(r) \) by Theorem 4.2, procedure FINDPAIRS(s) for all \( s \in \{1, m\} \) takes \( O(m^2) \) time. It follows that Algorithm 2 takes time \( O(rm + rm^2 + m\alpha(n)) = O(rm^2 + m\alpha(n)) \).

**Improving the time complexity** We see that the dominating term for the time complexity of Algorithm 2 is the time spent on procedure FINDPAIRS(s) for all \( s \in \{1, m\} \). To this end, instead of reducing the columns for each \( s = 1, \ldots, m \), we compute the earliest basis from the matrix \([C_1|C_2|\ldots|C_m]\) by the algorithm in [17]. Here the earliest basis of a matrix \( A \) with rank \( r \) means the set of columns \( B_{opt} = \{a_i_1, \ldots, a_i_r\} \) if the column indices \( \{i_1, \ldots, i_r\} \) are the lexicographically smallest index set such that the corresponding columns of \( A \) have full rank. Since \( \bigcup_s C_s = O(r) \) and the number of independent columns is \( O(m) \), this can be done in \( O((r/m)m^{\omega}) = O(rm^{\omega-1}) \) time. This finishes the proof of Theorem 1.1.

**Remark 1** We note that neither term in \( r = \min \{a(G)m, \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\} \) always dominates. In particular, it is easy to find examples where one term is significantly smaller (asymptotically) than the other. For example, for any planar graph \( G \), \( a(G)m = O(n) \). However, it is easy to have a planar graph where the second term \( \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)) = \Omega(n^2) \); see e.g., Fig. 3. On the other hand, it is also easy to have a graph \( G \) where \( \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)) = O(1) \) yet \( a(G)m = \Theta(n^3) \). Indeed, consider the bipartite graph in Fig. 5, where for each edge \( (u,v) \in E, d_{in}(u) + d_{out}(v) = 0 \). However, this graph has \( a(G) = \Theta(n), m = \Theta(n^2) \), and thus \( a(G)m = \Theta(n^3) \).

**Remark 2** We note that the time complexity of the algorithm proposed by Chowdhury and Mémoli in [6] to compute the \( (d - 1) \)-dimensional persistence path homology takes \( O(n^{3+3d}) \) time. However, for the case \( d = 2 \), a more refined analysis shows that in fact, their algorithm takes only \( O(mn + (\sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)))mn^2) \) time. Compared with our algorithm, which takes time \( O(rm^\omega + \alpha(n)m) \) with \( r = \min \{a(G)m, \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v))\} \) and \( \omega < 2.373 \), observe that our algorithm can be significantly faster (when \( a(G)m \) is much smaller than \( \sum_{(u,v)\in E}(d_{in}(u) + d_{out}(v)) \)). For example, for planar graphs, our algorithm takes \( O(n^{3d}) \) time, whereas the algorithm of [6] takes \( O(n^5) \) time.
5 Applications

We first show in Sect. 5.1 that our new algorithm can be extended to compute a minimal 1-dimensional path homology basis. We then show in Sect. 5.2 some preliminary experimental results for our algorithms, including showing the efficiency of our algorithm compared to the previous best algorithm over several datasets.

5.1 Minimal 1-Dimensional Path Homology Basis

We also design an algorithm for computing a minimally weighted basis for (1-dimensional) path homology. In particular, let $g = \text{rank}(H_1)$, and assume that $G = (V, E)$ is equipped with non-negative edge weights $w : E \to \mathbb{R}^+$. Given any 1-cycle $\gamma = \sum c_i e_i$, for each $i \in [1, m]$ set $\hat{c_i} = 1$ if $c_i \neq 0$; and $\hat{c_i} = 0$ otherwise. Then the length of $\gamma$, $\mu(\gamma)$, equals $\mu(\gamma) = \sum \hat{c_i} w(e_i)$.

Now abusing the notations slightly, we say that a set of $g$ 1-cycles $\{\gamma_1, \ldots, \gamma_g\}$ forms a 1-dimensional homology basis if the homology classes they represent, $\{[\gamma_1], \ldots, [\gamma_g]\}$, forms a basis for $H_1(G)$. The total length of this homology basis equals the total lengths of all cycles involved, i.e., $\sum_{i=1}^g \mu(\gamma_i)$.

Definition 5.1 Given a weighted directed graph $G = (V, E)$, let $g$ be the rank of the 1-dimensional homology group $H_1(G)$. Let $\mu : Z_1 \to \mathbb{R}^+ \cup \{0\}$ be the length of each cycle $C \in Z_1$. A homology basis $\gamma_1, \ldots, \gamma_g$ is called minimal if $\sum_{i=1}^g \mu(\gamma_i)$ is minimal among all bases of $H_1$.

A minimal homology basis of the example in Fig. 1 Let the weight of every edge be 1, and the length of a cycle be the total weights of all its edges. As we analysed in Sect. 2, the path homology group is generated by the class that is homologous to $C_3 = e_{dc} - e_{fc} + e_{fe} - e_{de}$. In other words, the path homology basis is any cycle homologous to $C_3$. There are three other cycles $C_2 + C_3 = e_{dc} + e_{fe} - e_{de} + e_{eb} - e_{gb} - e_{fg}$, $C_1 + C_3 = e_{dc} - e_{fc} + e_{fe} - e_{de} + e_{gb} + e_{ba} - e_{ga}$ and $C_1 + C_2 + C_3 = e_{dc} + e_{fe} - e_{de} + e_{eb} - e_{fg} + e_{ba} - e_{ga}$ homologous to $C_3$ because $C_1$ and $C_2$ are 1-boundaries. By Definition 5.1, the minimal homology basis is $\{C_3\}$ since the lengths of all $C_2 + C_3$, $C_1 + C_3$, and $C_1 + C_2 + C_3$ are greater than $C_3$.

Algorithm 3 Compute 1-dimensional minimal homology basis of $G = (V, E)$

1: procedure MINHOMOBASIS($G$)  
2: Compute annotations of all edges.  
3: Compute Horton set: For every vertex $v$, compute a shortest path tree $T$ rooted at $v$. Take all cycles generated by a non-tree edge and $T$.  
4: Sort the cycles in non-decreasing order of their lengths.  
5: Use matrix reduction to compute the minimal homology basis.  
6: end procedure

Algorithm We now describe our algorithm (Algorithm 3). First, we compute the so-called annotation of a cycle which is a binary vector that represents its homology class as explained below [1]. Then we use this annotation to compute a minimal
homology basis. In particular, for every 1-chain $\gamma \in \Omega_1$, we assign $\gamma$ a $g$-bit vector $\alpha(\gamma)$, called the annotation of $\gamma$, such that any two cycles $C$ and $C'$ are homologous if and only if their annotations $\alpha(C)$, $\alpha(C')$ are identical. As before, set $r = \min \left\{ a(G)m, \sum_{(u,v) \in E} (d_{in}(u) + d_{out}(v)) \right\}$.

We compute the annotation according to [1]. First we compute the annotation for every edge. We construct a cycle basis $Z$ in which any cycle can be expressed in simple and efficient terms. Note that as in the previous section, Algorithm 2 can not only compute a basis for $H_1(G)$, it can in fact partition edges into negative edge set and positive edge set, in which all negative edges $E(T)$ form a spanning tree $T$, while every positive edge $e \in \hat{E} = E \setminus E(T)$ can create a new cycle together with $T$, denoted as $\gamma(T, e)$. Note that $\gamma(T, e)$ can be written as a $m$-bit vector. According to Proposition 3.1, all such cycle form a cycle basis $Z$. For every tree edge $e \in E(T)$, let $\gamma(T, e)$ be a $m$-bit vector where every element is 0. It has been proved in [1] that for any cycle $C = \sum_i c_i e_i$ where $c_i$ is the coefficient of $e_i$ in $C$, it holds that $C = \sum_i c_i \gamma(T, e_i)$. Note that Algorithm 2 indeed computes a boundary basis $B$ for $B_1(G)$ and a homology basis $H$ for $H_1(G)$, such that $B \cup H$ forms another cycle basis $\hat{Z}$ for the cycle-group $Z_1(G)$ where rank($\hat{Z}$) = rank($Z$) = $m - n + 1$. Thus for every cycle $\gamma(T, e)$, we can solve a linear system $\hat{Z}x = \gamma(T, e)$. As a result, $x$ will be a $(m - n + 1) \times 1$ vector. We assign last $g = \text{rank}(H_1)$ bits as the annotation $\alpha(e)$ of $e$. We can solve for the annotation for all edges in $O(m^2)$ time where $m$ is the number of edges. Combined with time needed to compute the homology basis $H$, we conclude that computing the annotations of all edges costs $O(\min \left\{ a(G)m, \sum_{(u,v) \in E} (d_{in}(u) + d_{out}(v)) \right\} m^{\omega-1} + m^\omega) = O(r m^{\omega-1} + m^\omega)$ time. After computing the annotation $\alpha(e)$ for every edge $e$, for every 1-cycle $C = \sum_i c_i e_i$ where $c_i$ is the coefficient for the edge $e_i$, we compute its annotation $\alpha(C) = \sum_i c_i \alpha(e_i)$.

We now compute a minimal homology basis as follows. Using the results in [18], we observe that there is a collection of cycles, called Horton family, that includes a minimal homology basis. Consider the undirected version $G_u$ of the graph $G$. Now for every vertex $r$, we compute a minimum spanning tree $T$ rooted at $r$ and let $\bar{T} = G_u \setminus T$. Then, for every non-tree edge $e = (u, v) \in \bar{T}$, let $c_r$ be the 1-cycle (under $\mathbb{Z}_2$) obtained by summing $e$ and all edges on the paths $\pi_1$ and $\pi_2$ between $u$ and $r$, and $v$ and $r$ respectively. For every such cycle $c_r$ in $G_u$, we also have a cycle in $\Omega_1(G)$ containing the same edges, which are assigned a coefficient 1 or $-1$ depending on their orientations in $G$. Collecting all these cycles in $\Omega_1(G)$ for all minimal spanning trees rooted at every vertex in $G$, we obtain the Horton family. Clearly, the cardinality of the Horton family is $O(nm)$. Once we get a list of the Horton family, we sort all cycles in non-decreasing order of their lengths. This creates an annotation matrix with $O(nm)$ columns, each corresponding to the annotation of a cycle in the sorted list. We observe that a column basis of this matrix whose indices are lexicographically smallest corresponds to a minimal path homology basis of the graph $G$. Accordingly, we compute a basis $\{a_1, \ldots, a_r\}$ satisfying the following: the column indices $\{i_1, \ldots, i_r\}$ are the lexicographically smallest index set such that the corresponding columns have full rank. The result $\{a_{i_1}, \ldots, a_{i_r}\}$ is a minimal homology basis. This can be done by using same steps as in [18], replacing vectors of cycles with their annotations.
According to [1], we have the following: (1) given annotations for edges, the time to compute the annotations of Horton cycles, is $O(mng + mn \log n)$; (2) given $O(nm)$ Horton cycles, extracting a minimal homology basis needs $O(nmg^{\omega-1})$ time. Thus, the total time to compute minimal homology basis is $O(rm^{\omega-1} + m^\omega + nmg^{\omega-1})$, where $r = \min\{a(G)m, \sum_{(u,v) \in E}(d_{in}(u) + d_{out}(v))\}$.

5.2 Experimental Results

We implemented our algorithms for computing the 1-dimensional path homology and the 1-dimensional persistent path homology. The implementation is in Python 3 using the coefficient field $\mathbb{F} = \mathbb{R}$.

5.2.1 Comparisons with Previous Results

We test both of our software\(^5\) as well as the software\(^6\) in [6] on some datasets. All experiments are done on a PC Macbook Pro with Processor 2.9 GHz Intel Core i5, Memory 8GB 1867 MHz DDR3. We first introduce the datasets, and then analyze the output.

**U.S. economic sector data** This dataset is released by the U.S. Bureau of Economic Analysis(https://www.bea.gov/) involves the “make” and “use” tables of the production of commodities by industries. We obtained “use” table data for 15 industries across the year range 1997–2015; we have 19 tables each showing the yearly asymmetric flow of commodities across industries. We used same preprocessing used by [6] after which we get 19 complete directed graphs.

**Asian migration and remittance** This dataset is the Asian net migration and remittance networks in 2015 including 50 countries and regions, obtained from UN Global Migration Database and on bilateral remittances from the World Bank database as reported respectively in [26] and [25]. In [16] this data is analyzed via directed clique complex. Here we use the same preprocessing as [16] and work on persistent path homology.

**Directed cycle graph with 1000 vertices** This dataset is a directed cycle graph with 1000 vertices. Every vertex has both indegree and outdegree to be exactly 1.

**C. elegans** This dataset is a chemical synapse network of C. elegans [24]. It is a directed graph with 279 vertices and 2194 edges. Each vertex represents a neuron, and every edge reflects the synaptic contact between corresponding neurons; the source is the sender, and the sink is the receiver. Every edge has a weight representing the number of chemical synapses the receiver gets from the sender.

**Citation networks–Citeseer and Cora** [23] These two datasets are both citation benchmarks. Both are directed networks indicating the citation between papers. There are 2708 vertices and 5429 edges in Cora, and 3279 vertices and 4608 edges in Citeseer.

\(^5\) Code is available in https://github.com/tianqicoding/1dPPH.

\(^6\) Code can be found in https://github.com/samirchowdhury/pph-matlab.
The datasets show the migration and remittance between countries and regions in Asia. We list all countries and regions in Table 2.

The persistent path homologies of migration and remittance networks are motivated by the flow structure. We are concerned with flows that contain a single sink and a single source. Unlike [16], we regard cycles not only with boundary triangle structure but also with boundary quadrangle structure as trivial: two flows with one or two edges from the source to the sink are equivalent. With this interpretation, we capture those non-trivial cycles, as well as compute the persistence.

We first do some preprocess on the dataset as in [16]. For remittance data, let $r_{ab}$ be the remittance from $a$ to $b$. We create an edge $(a, b)$ if $r((a, b)) = r_{ab} - r_{ba} > 0$; there are no bigons in the graph. We set the edge weight $w(e)$ to $w(e) = \max_e r(e) + 1 - r(e)$, inducing a filtration for path homologies. The edge weights transform largest remittance to the smallest weight. As a consequence, cycles with large weights are born early, and cycles that are born early but killed off by edges with smaller weights have large persistence. Same preprocessing is employed on migration dataset.

In migration network, there are 44 generating cycles while there are 61 in [16]. This means that there are 17 generating homology classes which are non-trivial in [16] that become trivial when they are born because of the boundary quadrangles: there are two
2-path flows from the source to the sink. As in [16], all cycles are killed at the end, indicating small migration flow across countries and regions within cycles. We list all generating cycles in Figs. 10 and 11. Here all red nodes denote sources, green ones denote sinks and blue ones denote other vertices. Figures 10 and 11 reflect that neighbor countries tend to attract immigrants from the same countries, e.g. there are big migrations from India to both Kuwait and United Arab Emirates. On the other hand, people in neighbor countries migrate similarly, e.g. considerable number of people in both China (mainland) and Philippines move to Hong Kong and Japan. Besides, we capture a directed cycle Kazakhstan → Kyrgyzstan → Tajikistan → Kazakhstan, indicating a directed flow.

The generating cycles for remittance network are listed in Figs. 10 and 11. Most generating cycles are consistent with cycles in migration generating cycles: the generating cycle for the remittance network is a generating cycle in migration network with all arrows reversed. For example, there is a generating cycle in Fig. 10: India → United

Table 2  Asian migration and remittance

| Order | Country                      | Abbrev. | Order | Country           | Abbrev. |
|-------|------------------------------|---------|-------|-------------------|---------|
| 1     | Afghanistan                  | AF      | 26    | Lebanon           | LB      |
| 2     | Armenia                      | AM      | 27    | Malaysia          | MY      |
| 3     | Azerbaijan                   | AZ      | 28    | Maldives          | MV      |
| 4     | Bahrain                      | BA      | 29    | Mongolia          | MN      |
| 5     | Bangladesh                   | BD      | 30    | Myanmar           | MM      |
| 6     | Bhutan                       | BT      | 31    | Nepal             | NP      |
| 7     | Brunei Darussalam            | BN      | 32    | Oman              | OM      |
| 8     | Cambodia                     | KH      | 33    | Pakistan          | PK      |
| 9     | China (mainland)             | CN      | 34    | Philippines       | PH      |
| 10    | Hong Kong                    | HK      | 35    | Qatar             | QA      |
| 11    | Macau                        | MO      | 36    | Republic of Korea | KR      |
| 12    | Cyprus                       | CY      | 37    | Saudi Arabia      | SA      |
| 13    | Dem. People’s Rep. of Korea | KP      | 38    | Singapore         | SG      |
| 14    | Georgia                      | GE      | 39    | Sri Lanka         | LK      |
| 15    | India                        | IN      | 40    | State of Palestine | PS      |
| 16    | Indonesia                    | ID      | 41    | Syria             | SY      |
| 17    | Iran                         | IR      | 42    | Tajikistan        | TJ      |
| 18    | Iraq                         | IQ      | 43    | Thailand          | TH      |
| 19    | Israel                       | IL      | 44    | Timor-Leste       | TI      |
| 20    | Japan                        | JP      | 45    | Turkey            | TR      |
| 21    | Jordan                       | JO      | 46    | Turkmenistan      | TM      |
| 22    | Kazakhstan                   | KZ      | 47    | United Arab Emirates | AE  |
| 23    | Kuwait                       | KW      | 48    | Uzbekistan        | UZ      |
| 24    | Kyrgyzstan                   | KG      | 49    | Vietnam           | VN      |
| 25    | Laos                         | LA      | 50    | Yemen             | YE      |
Arab Emirates $\leftrightarrow$ Indonesia $\rightarrow$ Saudi Arabia $\leftrightarrow$ India. It corresponds to a generating cycle in Fig. 12: India $\leftrightarrow$ United Arab Emirates $\rightarrow$ Indonesia $\leftrightarrow$ Saudi Arabia $\rightarrow$ India. The reason for that migration could be that people tend to send money back home.

### 5.4 C. elegans

**Path homology** We first compute the 1-dimensional path homology of this network. The rank of this path homology is 17, whereas it is reported in [22] that the rank of 1-dimensional directed clique homology is 183. This means that boundary quadrangles have a great effect on the synaptic contact structure; there are a number of cycles that can be written as linear combination of boundary triangles and quadrangles, but cannot be represented only by boundary triangles.

We also compare with the 1-dimensional path homology on a set of 1000 directed Erdős–Rényi random graphs with 279 vertices and the connection probability 0.028 (thus with 2194 expected connections). The average first Betti number of these ER random graphs is 114.91, while the first Betti number for path homology in C. elegans is 17. It supports the claim from [24] that the C. elegans graph is significantly different from Erdős–Rényi random graph with the same number of vertices and similar number of edges.

**Minimal homology basis** We also computed the minimum homology basis (with respect to lengths of cycles) for the C. elegans network. Interestingly, all the 17 generating cycles in the computed minimal homology basis are (non-boundary) quadrangles. Furthermore, there are no triangles or quadrangles which are directed cycles.
Fig. 11 Generating cycles for persistent path homology on migration network (cont.)

Fig. 12 Generating cycles for persistent path homology on remittance network
ever, when we tested on 1000 Erdős–Rényi random graphs, we found that each of them has directed triangles or quadrangles in its respective minimal homology basis. Our results provide some information on four neuron subnetworks for the chemical synapse network for \textit{C. elegans} (the analysis of four neuron subnetworks is of interest and was previously carried out for the gap-junction network of \textit{C. elegans} [24]).

6 Concluding Remarks

A natural question is whether it is possible to have a more efficient algorithm for computing (persistent) path homology of higher dimensions improving the work of [6]. Another question is whether we can compute a minimal path homology basis faster improving our current time bound $O(rm^{\omega - 1} + m^\omega + nm^g^{\omega - 1})$.

References

1. Busaryev, O., Cabello, S., Chen, C., Dey, T.K., Wang, Y.: Annotating simplices with a homology basis and its applications. In: 13th Scandinavian Symposium and Workshop on Algorithm Theory (Helsinki 2012). Lecture Notes in Computer Science, vol. 7357, pp. 189–200. Springer, Heidelberg (2012)
2. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Paris 2009), pp. 199–208. ACM, New York (2009)
3. Cheung, H.Y., Kwok, T.C., Lau, L.C.: Fast matrix rank algorithms and applications. J. ACM 60(5), # 31 (2013)
4. Chiba, N., Nishizeki, T.: Arboricity and subgraph listing algorithms. SIAM J. Comput. 14(1), 210–223 (1985)
5. Chowdhury, S., Mémoli, F.: A functorial Dowker theorem and persistent homology of asymmetric networks. J. Appl. Comput. Topol. 2(1–2), 115–175 (2018)
6. Chowdhury, S., Mémoli, F.: Persistent path homology of directed networks. In: 29th Annual ACM-SIAM Symposium on Discrete Algorithms (New Orleans 2018), pp. 1152–1169. SIAM, Philadelphia (2018)
7. Cohen-Steiner, D., Edelsbrunner, H., Morozov, D.: Vines and vineyards by updating persistence in linear time. In: 22nd International Symposium on Computational Geometry (Sedona 2006), pp. 119–126. ACM, New York (2006)
8. Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C.: Introduction to Algorithms. MIT Press, Cambridge (2009)
9. Dey, T.K., Li, T., Wang, Y.: Efficient algorithms for computing a minimal homology basis. In: 13th Latin American Symposium on Theoretical Informatics (Buenos Aires 2018). Lecture Notes in Computer Science, vol. 10807, pp. 376–398. Springer, Cham (2018)
10. Dłotko, P., Hess, K., Levi, R., Nolte, M., Reimann, M., Scolamiero, M., Turner, K., Müller, E., Markram, H.: Topological analysis of the connectome of digital reconstructions of neural microcircuits (2016). arXiv:1601.01580
11. Erickson, J., Whittlesey, K.: Greedy optimal homotopy and homology generators. In: 16th Annual ACM-SIAM Symposium on Discrete Algorithms (Vancouver 2005), pp. 1038–1046. ACM, New York (2005)
12. Grigor’yan, A., Lin, Y., Muranov, Yu., Yau, S.-T.: Homologies of path complexes and digraphs (2012). arXiv:1207.2834
13. Grigor’yan, A., Lin, Y., Muranov, Yu., Yau, S.-T.: Homotopy theory for digraphs (2014). arXiv:1407.0234
14. Grigor’yan, A., Lin, Y., Muranov, Yu., Yau, S.-T.: Cohomology of digraphs and (undirected) graphs. Asian J. Math. 19(5), 887–931 (2015)
15. Harary, F.: Graph Theory. Addison-Wesley Series in Mathematics, Addison-Wesley, Boston (1971)
16. Ignacio, P.S.P., Darcy, I.K.: Tracing patterns and shapes in remittance and migration networks via persistent homology. EPJ Data Sci. 8(1), # 1 (2019)
17. Jeanenne, C.: LSP matrix decomposition revisited (2006). http://www.ens-lyon.fr/LIP/Pub/Rapports/RR/RR2006/RR2006-28.pdf
18. Liebchen, C., Rizzi, R.: A greedy approach to compute a minimum cycle basis of a directed graph. Inform. Process. Lett. 94(3), 107–112 (2005)
19. Massey, W.S.: A Basic Course in Algebraic Topology. Graduate Texts in Mathematics, vol. 127. Springer, New York (1991)
20. Masulli, P., Villa, A.E.P.: The topology of the directed clique complex as a network invariant. Springer-Plus 5(1), # 388 (2016)
21. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network motifs: simple building blocks of complex networks. Science 298(5594), 824–827 (2002)
22. Reimann, M.W., Nolte, M., Scolamiero, M., Turner, K., Perin, R., Chindemi, G., Dłotko, P., Levi, R., Hess, K., Markram, H.: Cliques of neurons bound into cavities provide a missing link between structure and function. Front. Comput. Neurosci. 11, # 48 (2017)
23. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., Eliassi-Rad, T.: Collective classification in network data. AI Mag. 29(3), 93–106 (2008)
24. Varshney, L.R., Chen, B.L., Paniagua, E., Hall, D.H., Chklovskii, D.B.: Structural properties of the Caeorhabditis elegans neuronal network. PLoS Comput. Biol. 7(2), # e1001066 (2011)
25. The International Development Association, the World Bank Group: Bilateral remittance estimates for 2015 using migrant stocks, host country incomes, and origin country incomes (millions of US$) (2016). https://www.worldbank.org/en/topic/migrationremittancesdiasporaissues/brief/migration-remittances-data
26. United Nations Department of Economic & Social Affairs: PD trends in international migrant stock: migrants by destination and origin (United Nations database, pop/db/mig/stock/rev.2015) (2015). https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates15.asp

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.