Journal of Applied and Computational Topology
https://doi.org/10.1007/s41468-022-00108-3

First Betti number of the path homology of random directed graphs

Thomas Chaplin[1]

Received: 11 February 2022 / Revised: 3 November 2022 / Accepted: 4 November 2022

Abstract
Path homology is a topological invariant for directed graphs, which is sensitive to their asymmetry and can discern between digraphs which are indistinguishable to the directed flag complex. In Erdős–Rényi directed random graphs, the first Betti number undergoes two distinct transitions, appearing at a low-density boundary and vanishing again at a high-density boundary. Through a novel, combinatorial condition for digraphs we describe both sparse and dense regimes under which the first Betti number of path homology is zero with high probability. We combine results of Grigor'yan et al., regarding generators for chain groups, with methods of Kahle and Meckes in order to determine regimes under which the first Betti number is positive with high probability. Together, these results describe the gradient of the lower boundary and yield bounds for the gradient of the upper boundary. With a view towards hypothesis testing, we obtain tighter bounds on the probability of observing a positive first Betti number in a high-density digraph of finite size. For comparison, we apply these techniques to the directed flag complex and derive analogous results

Keywords Path homology · Directed flag complex · Stochastic topology · Directed graphs · Random graphs

Mathematics Subject Classification 05C20 · 05C80 · 55U15 · 60C05 · 62F03

1 Introduction

In applications, networks often arise with asymmetry and directionality. Chemical synapses in the brain have an intrinsic direction (see (Purves et al. 2018, §5)); gene regulatory networks record the causal effects between genes (e.g. Aalto et al. 2020);

✉ Thomas Chaplin
thomas.chaplin@maths.ox.ac.uk

[1] Mathematical Institute, University of Oxford, Woodstock Rd, Oxford OX2 6GG, Oxfordshire, UK

Published online: 04 December 2022

communications in social networks have a sender and a recipient (e.g. Leskovec and Krevl 2014). A common hypothesis is that the structure of a network determines its function (Ingram et al. 2006; Reimann et al. 2017), at least in part. In order to investigate such a claim, one requires a topological invariant which describes the structure of the network. To obtain such a summary for a digraph, one often symmetrises to obtain an undirected graph, before applying traditional tools from TDA (e.g. Helm et al. 2021). This potentially inhibits the predictive power of the descriptor, since the pipeline becomes blind to the direction of edges. In recent years, particularly in applications related to neuroscience (e.g. Caputi et al. 2021; Reimann et al. 2017), researchers have explored the use of topological methods which are sensitive to the asymmetry of directed graphs.

A much-studied construction, for undirected graphs, is the clique complex (or flag complex)—a simplicial complex in which the k-simplices are the $(k+1)$-cliques in the underlying graph. An obvious extension to the case of directed graphs is the directed flag complex (Lütgehetmann et al. 2020). This is an *ordered* simplicial complex in which the *ordered* k-simplices are the $(k+1)$-directed cliques: $(k+1)$-tuples of distinct vertices $(v_0, \ldots, v_k)$ such that $v_i \to v_j$ whenever $i < j$. An important property of this construction is that is able to distinguish between directed graphs with identical underlying, undirected graphs; it is sensitive to the asymmetry of the digraph.

Path homology (first introduced by Grigor'yan et al. (2012)) provides an alternative construction which, while more computationally expensive, is capable of distinguishing between digraphs which are indistinguishable to the directed flag complex (e.g. Fig. 1, c.f. Chowdhury and Mémoli (2018)). Moreover, the non-regular chain complex, from which path homology is defined, contains the directed flag complex as a subcomplex. Intuitively, the generators of the kth chain group of the directed flag complex are all the directed paths, of length k, such that all shortcut edges are present in the graph. Whereas, the kth chain group of the non-regular chain complex consists of all *linear combinations* of directed paths, of length k, such that any missing shortcuts of length $(k-1)$ are cancelled out.

Other desirable features of path homology include good functorial properties in an appropriate digraph category (Grigor'yan et al. 2014, 2020) and invariance under an appropriate notion of path homotopy (Grigor'yan et al. 2014, Theorem 3.3). Furthermore, path homology is a particularly novel method since it operates directly on directed paths within the digraph, rather than first constructing a simplicial complex. Rather than being freely generated by distinguished motifs, the chain groups for path homology are formed as the pre-images of the boundary maps. As such, finding a basis for the chain groups is often non-trivial, which complicates the understanding

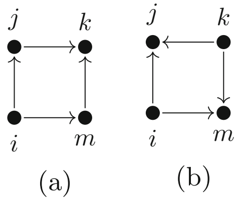


Fig. 1 Two motifs which are indistinguishable to the directed flag complex but have different path homology

of how homology arises in a random digraph. Hence, it is desirable to develop an understanding of the statistical behaviour of path homology, both from an applied perspective and from independent interest.

Key questions include (as discussed for the clique complex in Kahle 2009; Kahle et al. 2014; Kahle and Meckes 2013): when should one expect homology to be trivial or non-trivial; when homology is non-trivial, what are the expected Betti numbers; and how are the Betti numbers distributed?

Notation 1.1 Assume G is disturbed according to a null model depending only on n, e.g. after assuming all other parameters are functions of n.

1. We say a property P holds *with high probability*, if property P holds with probability tending to 1 as $n \to \infty$.
2. Given two random variables X, Y, depending on G, we write $X \sim Y$ *with high probability* if for any $\epsilon > 0$

$$\mathbb{P}\left[1 - \epsilon \leq \frac{X}{Y} \leq 1 + \epsilon\right] \to 1 \quad \text{as } n \to \infty. \tag{1.1}$$

To date, traditional topological invariants enjoy a greater statistical understanding in the context of basic null models. In particular, Kahle showed the following:

Theorem 1.2 (Kahle 2009; Kahle et al. 2014) *For an Erdős–Rényi random undirected graph $G \sim G(n, p)$, denote the kth Betti number (over a field of characteristic 0) of its clique complex $\mathrm{X}(G)$ by β_k. Let f_k denote the number of k-cliques then $\mathbb{E}[f_k] = \binom{n}{k+1} p^{\binom{k+1}{2}}$. Assume $p = n^{\alpha}$, then*

1. *if $-1/k < \alpha < -1/(k+1)$ then $\beta_k \sim \mathbb{E}[\beta_k] \sim \mathbb{E}[f_k] \sim f_k$ with high probability;*
2. *if $-1/k < \alpha < -1/(k+1)$ then $\beta_k > 0$ with high probability;*
3. *if $\alpha < -1/k$ then $\beta_k = 0$ with high probability;*
4. *if $\alpha > -1/(k+1)$ then $\beta_k = 0$ with high probability.*

In essence, this characterises the understanding that, in any given degree, random graphs only have non-trivial, clique complex homology in a ‘goldilocks’ region, wherein graph density is neither too big nor too small. Moreover, the boundaries of this region are dependent on the number of nodes in the graph, scaling as a power law. Our primary contribution is a similar description for two different flavours of path homology, in degree 1.

1.1 Summary of results

In order to derive useful probability bounds, it is often necessary to prescribe a null model which is highly symmetric and depends on few parameters. Therefore, throughout this paper we will be focusing on an Erdős–Rényi random directed graph model, in which the number of nodes is fixed (at n) and each possible directed edge appears independently, with some probability p. Note, this model allows for the existence of a reciprocal pair of directed edges.

Although individual results are potentially stronger, the following theorems characterise the theoretical understanding that we will develop. Denote the kth Betti number of the non-regular path homology of a digraph by $\overrightarrow{\beta}_k$. Firstly, $(\overrightarrow{\beta}_0 + 1)$, is the number of weakly connected components. This coincides with the number of connected components of the flat symmetrisation of the digraph (see Definition 2.10). If $G \sim \overrightarrow{G}(n, p)$ is an Erdős–Rényi random directed graph then its symmetrisation is an Erdős–Rényi random undirected graph, $\bar{G} \sim G(n, \bar{p})$, where $\bar{p} = 1 - (1-p)^2$. Thus, we use a standard result due to Erdős and Rényi (1960), Kahle (2009) to prove the following.

Theorem 1.3 *For an Erdős–Rényi random directed graph $G \sim \overrightarrow{G}(n, p(n))$, let $\overrightarrow{\beta}_0$ denote the $0th$ Betti number of its non-regular path homology over $\mathbb{Z}$. Assume $1 - (1-p(n))^2 = (\log(n) + f(n))/n$, then*

1. *if $\lim_{n\to\infty} f(n) = -\infty$ then $\overrightarrow{\beta}_0 > 0$ with high probability;*
2. *if $\lim_{n\to\infty} f(n) = \infty$ then $\overrightarrow{\beta}_0 = 0$ with high probability.*

The same result holds for regular path homology.

Our primary contribution identifies a similar 'goldilocks' region for the first Betti number of path homology, $\overrightarrow{\beta}_1$.

Theorem 1.4 *For an Erdős–Rényi random directed graph $G \sim \overrightarrow{G}(n, p(n))$, let $\overrightarrow{\beta}_1$ denote the $1st$ Betti number of its non-regular path homology over $\mathbb{Z}$. Let N_1 denote the number of edges, $N_1 = \#E(G)$, then $\mathbb{E}[N_1] = n(n-1)p$. Assume $p(n) = n^\alpha$, then*

1. *if $-1 < \alpha < -2/3$ then $\overrightarrow{\beta}_1 \sim \mathbb{E}[\overrightarrow{\beta}_1] \sim \mathbb{E}[N_1] \sim N_1$ with high probability;*
2. *if $-1 < \alpha < -2/3$ then $\overrightarrow{\beta}_1 > 0$ with high probability;*
3. *if $\alpha < -1$ then $\overrightarrow{\beta}_1 = 0$ with high probability;*
4. *if $\alpha > -1/3$ then $\overrightarrow{\beta}_1 = 0$ with high probability.*

The same result holds for regular path homology.

By way of justifying the assumption $p(n) = n^\alpha$, in Fig. 9a we plot $\mathbb{P}[\overrightarrow{\beta}_1(G) = 0]$, for $G \sim \overrightarrow{G}(n, p)$, in colour against $\log(n)$ and $\log(p)$ along the two spatial axes. We observe two transitions between three distinct regions in parameter space. There is an interim region, in which we observe mostly $\overrightarrow{\beta}_1 > 0$; when p becomes too small we suddenly observe mostly $\overrightarrow{\beta}_1 = 0$, and likewise when p becomes too large. On this plot, the boundaries between the three regions appear as straight lines. Hence a reasonable conjecture is that these boundaries follow a power-law relationship $\log(p) = \alpha \log(n) + c$. Therefore, following power-law trajectories through parameter space will allow us to derive either $\mathbb{P}[\overrightarrow{\beta}_1(G) > 0] \to 1$ or $\mathbb{P}[\overrightarrow{\beta}_1(G) = 0] \to 1$.

Turning our attention to higher degrees, we provide weak guarantees for the asymptotic behaviour of $\overrightarrow{\beta}_k$, for arbitrary $k \geq 1$, at low densities.

Theorem 1.5 *For an Erdős–Rényi random directed graph $G \sim \overrightarrow{G}(n, p(n))$, let $\overrightarrow{\beta}_k$ denote the kth Betti number of its non-regular path homology over $\mathbb{Z}$. Assume $p(n) =$*

n^{α} with $\alpha < -\frac{N+1}{N}$ for some $N \in \mathbb{N}$. Then, $\overrightarrow{\beta}_k = 0$ with high probability for every $k \geq N$. The same result holds for regular path homology.

For comparison, in Sect. 5, we apply the techniques used to prove Theorem 1.4 in order to obtain analogous results for the directed flag complex.

Theorem 1.6 *For an Erdős–Rényi random directed graph $G \sim \overrightarrow{G}(n, p(n))$, let β_k denote the kth Betti number of its directed flag complex homology over $\mathbb{Z}$. Let N_k denote the number of directed k-cliques, $N_k = \operatorname{rank} \overrightarrow{X}_k(G)$, then $\mathbb{E}[N_k] = \binom{n}{k+1}(k+1)!\, p^{\binom{k+1}{2}}$. Assume $p(n) = n^{\alpha}$, then for each $k \geq 0$*

1. *if $-1/k < \alpha < -1/(k+1)$ then $\mathbb{E}[\beta_k] \sim \mathbb{E}[N_k]$;*
2. *if $-1 < \alpha < -1/2$ then $\beta_1 \sim \mathbb{E}[\beta_1] \sim \mathbb{E}[N_1] \sim N_1$ with high probability;*
3. *if $-1 < \alpha < -1/2$ then $\beta_1 > 0$ with high probability;*
4. *if $\alpha < -1$ then $\beta_1 = 0$ with high probability;*
5. *if $\alpha > -1/4$ then $\beta_1 = 0$ with high probability.*

In Sect. 6, we summarise these results and compare path homology and the directed flag complex to more traditional symmetric methods. We provide Table 1 in which we record, for each of the homologies under consideration, the α-region in which we know β_1 is either zero or positive, with high probability (assuming $p = n^{\alpha}$).

In Appendix A, with a view towards hypothesis testing, we derive a tighter explicit bound for $\mathbb{P}(\overrightarrow{\beta}_1(G) > 0)$, which becomes useful when p is large. In order to identify a given Betti number as statistically significant, against a Erdős–Rényi null model, one would usually resort to a Monte Carlo permutation test (e.g. Dwass 1957). This would require the computation of path homology for a large number of random graphs. For large graphs ($n \geq 100$ nodes), this is often infeasible, due to the computational complexity of path homology. However, if graph density falls into one of the regions identified by the results in Appendix A, one can potentially circumvent this costly computation.

2 Background

2.1 Graph theory definitions and assumptions

For clarity, we present a number of standard definitions, and assumptions that we will use throughout this paper. First, we fix our notation for graphs.

Definition 2.1 1. A *(undirected) graph* is a pair $G = (V, E)$, where V is an arbitrary set and E is a set of 2-element subsets of V.
2. A *directed graph* (or *digraph*) is a pair $G = (V, E)$, where V is an arbitrary set and $E \subseteq V \times V$.
3. A *(resp. directed) multigraph* is a (resp. directed) graph $G = (V, E)$ in which E is allowed to be a multiset.
4. In all cases, we call $V(G) := V$ the *set of nodes* or *vertices* and $E(G) := E$ the *set of edges*.

5. A digraph $G = (V, E)$ is *simple* if $E \subseteq (V \times V) \setminus \Delta$, where $\Delta := \{(i, i) \mid i \in V\}$.
6. The *density* of a simple digraph $G = (V, E)$ is the ratio of edges present, relative to the maximum number of possible edges:

$$\text{density}(G) := \frac{\#E}{\#V(\#V - 1)}. \tag{2.1}$$

Assumption 2.2 Throughout this paper, unless stated otherwise, we assume that all digraphs $G = (V, E)$ are simple. This means that they contain no self loops and contain at most one edge between any ordered pair of vertices.

Given a directed graph G, we make the following definitions to refer to subgraphs within G.

Definition 2.3 Given a digraph $G = (V, E)$, we make the following definitions.

1. A *subgraph* is another graph $G' = (V', E')$ such that $V' \subseteq V$ and $E' \subseteq E$; we denote this as $G' \subseteq G$.
2. Given a subgraph $G_1 \subseteq G$ and a subset of edges $E_2 \subseteq E(G)$ we let $G_1 \cup E_2$ denote a new graph with edges

$$E(G_1 \cup E_2) = E(G_1) \cup E_2. \tag{2.2}$$

   and node-set $V(G_1 \cup E_2)$, the smallest superset of $V(G_1)$ that contains all endpoints of edges in E_2.
3. A *(combinatorial) undirected walk* is an alternating sequences of vertices and edges

$$\rho = (v_0, e_1, v_1, e_2, \dots, v_{n-1}, e_n, v_n) \tag{2.3}$$

   such that edges connect adjacent vertices, in either direction. That is, for each i, either $e_i = (v_{i-1}, v_i)$ or $e_i = (v_i, v_{i-1})$.
4. A *(combinatorial) directed walk* is an undirected walk such that all edges are forward edges, that is $e_i = (v_{i-1}, v_i)$ for every i.
5. A *(combinatorial) directed/undirected path* is a directed/undirected walk which never repeats vertices or edges, that is $v_i = v_j$ or $e_i = e_j$ implies $i = j$.
6. A *(combinatorial) directed/undirected cycle* is a directed/undirected walk such that

$$v_i = v_j, i \neq j \iff \{i, j\} = \{0, n\}. \tag{2.4}$$

7. The *length* of a walk is the number of edges it traverses, e.g. the length of ρ in equation (2.3) is n.
8. A *double edge* is an unordered pair of vertices $\{i, j\} \subseteq V$ such that both directed edges are in the graph, i.e. $(i, j), (j, i) \in E$.

Notation 2.4 1. For vertices $i, j \in V$, we write $i \to j$ if $(i, j) \in E$.
2. If $E_2 = \{e\}$ is a singleton then we define $G_1 \cup e := G_1 \cup E_2$.

Remark 2.5 Assumption 2.2 allows for the existence of double edges.

2.2 Analytic and algebraic definitions

Next, we provide definitions of 'Landau symbols', which we use describe the asymptotic behaviour of two functions, relative to one another.

Notation 2.6 Given two functions $f, g : \mathbb{N} \to \mathbb{R}$ we write

1. $f(n) = o(g(n))$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$;
2. $f(n) = \omega(g(n))$ if $\lim_{n\to\infty} \frac{g(n)}{f(n)} = 0$;
3. $f(n) \sim g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$.

Remark 2.7 1. There is an equivalence, $f(x) = \omega(g(x)) \iff g(x) = o(f(x))$.
2. Note that, if X and Y are deterministic random variables then $X \sim Y$ with high probability if and only if $X \sim Y$ in the sense above.

Finally, we make a formal, algebraic definition, which will be required later in order to define path homology.

Definition 2.8 Given a ring R and a set V, we let $R\langle V\rangle$ denote the R-module of formal R-linear combinations of elements of V. That is,

$$R\langle V\rangle := \left\{ \sum_{i=1}^{n} \alpha_i e_{v_i} \mid n \in \mathbb{N},\ \alpha_i \in R,\ v_i \in V \right\} \tag{2.5}$$

where $\{e_v \mid v \in V\}$ are formal symbols which form a basis of the free R-module $R\langle V\rangle$.

2.3 Erdos–Rényi random graphs

Throughout this paper, we will primarily be investigating random directed graphs under an Erdős–Rényi model.

Definition 2.9 1. The *Erdős–Rényi random undirected graph model, $G(n, p)$*, is a probability space of undirected graphs. Each graph has exactly n nodes $\{1, \dots, n\}$ and each directed edge is included, independently, with probability p. A given graph G on n nodes with m edges appears with probability

$$p^m (1-p)^{\binom{n}{2}-m}. \tag{2.6}$$

For a graph drawn from this model, we write $G \sim G(n, p)$.

2. The *Erdős–Rényi random directed graph model, $\overrightarrow{G}(n, p)$*, is a probability space of directed graphs. Each graph has exactly n nodes $\{1, \dots, n\}$ and each directed edge is included, independently, with probability p. A given digraph G on n nodes with m edges appears with probability

$$p^m (1-p)^{n(n-1)-m}. \tag{2.7}$$

For a digraph drawn from this model, we write $G \sim \overrightarrow{G}(n, p)$.

2.4 Symmetrisation

Definition 2.10 Given a directed graph $G = (V, E)$,

1. the *flat symmetrisation* is an undirected graph, $\bar{G} := (V, \bar{E})$, where

$$\{i, j\} \in \bar{E} \text{ with multiplicity } 1 \iff (i, j) \in E \text{ or } (j, i) \in E \text{ or both.} \tag{2.8}$$

2. the *weak symmetrisation* is an undirected multigraph $\mathring{G} := (V, \mathring{E})$, where $\{i, j\}$ appears in $\mathring{E}$ with multiplicity 2 if both $(i, j) \in E$ and $(j, i) \in E$, or with multiplicity if 1 if only one of these edges is present.

Remark 2.11 We can view $\bar{G}$ and $\mathring{G}$ as topological spaces by giving them the natural structure of a simplicial complex and delta complex respectively. Both of these structures have no simplices above dimension 1, so clearly $\beta_k(\bar{G}) = \beta_k(\mathring{G}) = 0$ for all $k > 1$.

Lemma 2.12 *Given a random directed graph $G \sim \overrightarrow{G}(n, p)$, the flat symmetrisation is distributed as $\bar{G} \sim G(n, \bar{p})$ where*

$$\bar{p} := 1 - (1 - p)^2 = 2p - p^2. \tag{2.9}$$

Proof A given undirected edge $\{i, j\}$ appears in $\bar{G}$ if and only if at least one of (i, j) or (j, i) is in G. Therefore

$$\mathbb{P}(\{i, j\} \notin \bar{E}) = \mathbb{P}\big((i, j) \notin E \text{ and } (j, i) \notin E\big) = (1 - p)^2. \tag{2.10}$$

Hence, the undirected edge appears with probability $1 - (1 - p)^2$. The existence of each undirected edge depends on the existence of a distinct pair of directed edges. Hence each undirected edge appears independently. □

Remark 2.13 Since we always assume $p \to 0$, note that $\bar{p} \sim 2p$. This clearly implies that $p = o(n^{-1/k}) \iff \bar{p} = o(n^{-1/k})$.

Definition 2.14 Throughout this paper, we define $\bar{p}$ as in (2.9), whenever the underlying p is clear from context.

Definition 2.15 Given an undirected graph $G = (V, E)$,

1. a *k-clique* is a subset of vertices $V' \subseteq V$, such that $\#V' = k$ and for any two, distinct vertices, $i, j \in V'$, the edge between them is present, i.e. $\{i, j\} \in E$;
2. the *clique complex,* $\mathrm{X}(G)$ is a simplicial complex where the k-simplices are the $(k + 1)$-cliques in G.

We now investigate the behaviour of these ‘symmetric methods’ on random directed graphs. Since the flat symmetrisation of a random digraph $\overrightarrow{G}(n, p)$ is a random graph $G(n, \bar{p})$ and the asymptotics of $\bar{p}$ do not differ greatly from those of p, Theorem 1.2

can be restated immediately for $\beta_k(\mathrm{X}(\bar{G}))$ with the only change being that $\mathbb{E}[f_k] = \binom{n}{k+1}\bar{p}^{\binom{k+1}{2}}$.

Next, we prove that if $p = p(n)$ shrinks too quickly then β_1 will vanish for $\bar{G}$ and $\mathring{G}$, with high probability. This is a special case of the proof given by Kahle (Kahle (2009), Theorem 2.6). We repeat the proof to illustrate that it can be applied to $\beta_1(\bar{G})$, $\beta_1(\mathring{G})$ and, later on, path homology $\overrightarrow{\beta}_1(G)$.

Proposition 2.16 *If $p = p(n) = o(n^{-1})$ then, given a random directed graph $G \sim \overrightarrow{G}(n, p)$, we have*

$$\lim_{n\to\infty} \mathbb{P}(\beta_1(\bar{G}) = 0) = \lim_{n\to\infty} \mathbb{P}(\beta_1(\mathring{G}) = 0) = 1. \tag{2.11}$$

Proof Note that the existence of an undirected cycle in $\bar{G}$ of length $L \in [3, n]$ is a necessary condition for $\beta_1(\bar{G}) > 0$. For each L, by a union bound, the probability of there being an undirected cycle of length L is at most $(n\bar{p})^L$. Hence, the probability that there is an undirected cycle of any length is at most $(n\bar{p})^3/(1-(n\bar{p}))$. The assumption $p = o(n^{-1})$ implies that $\lim_{n\to\infty}(n\bar{p}) = 0$ and hence the bound converges to 0 as $n \to \infty$.

To prove $\overrightarrow{\beta}_1(\mathring{G}) = 0$ with high probability, all that remains is to bound probability of there being an undirected cycle on 2 nodes (i.e. a double edge) in G. The probability that there is some double edge is at most $\binom{n}{2}p^2 \leq n^2p^2$ which tends to 0 since $p = o(n^{-1})$. □

Finally, we investigate conditions under which we expect $\beta_1(\bar{G}) > 0$ and $\beta_1(\mathring{G}) > 0$ with high probability, and determine the growth rate of $\mathbb{E}[\beta_1]$ in each situation. Standard techniques, as employed for the clique complex in Kahle (2009), shows the following.

Proposition 2.17 *If $p = p(n) = \omega(n^{-1})$ then, given a random directed graph $G \sim \overrightarrow{G}(n, p)$,*

$$\mathbb{E}[\beta_1(\bar{G})] \sim \binom{n}{2}\bar{p} \quad \textit{and} \quad \mathbb{E}[\beta_1(\mathring{G})] \sim n(n-1)p. \tag{2.12}$$

Moreover, $\beta_1(\bar{G}) \sim \mathbb{E}[\beta_1(\bar{G})]$ and $\beta_1(\mathring{G}) \sim \mathbb{E}[\beta_1(\mathring{G})]$ with high probability and hence

$$\lim_{n\to\infty} \mathbb{P}(\beta_1(\bar{G}) > 0) = \lim_{n\to\infty} \mathbb{P}(\beta_1(\mathring{G}) > 0) = 1. \tag{2.13}$$

Proof Denoting the original digraph $G = (V, E)$, we deal with the flat symmetrisation first. For convenience, we define $N_1 := \#\bar{E}$ and $N_0 := \#V$. Note that $\mathbb{E}[N_1] = \binom{n}{2}\bar{p}$ and $\mathbb{E}[N_0] = n$. A standard application of the Euler characteristic shows

$$N_1 - N_0 \leq \beta_1 \leq N_1. \tag{2.14}$$

The assumption $p(n) = \omega(n^{-1})$ yields $\mathbb{E}[N_0] = o(\mathbb{E}[N_1])$ and hence $\mathbb{E}[\beta_1] \sim \mathbb{E}[N_1]$.

Now we show $\beta_1 \sim \mathbb{E}[\beta_1]$ with high probability. Since $\mathbb{E}[\beta_1] \sim \mathbb{E}[N_1] \to \infty$, by an application of Chebyshev's inequality Alon and Spencer (2016), it suffices to show that $\mathrm{Var}(\beta_1) = o(\mathbb{E}[\beta_1]^2)$. Using the inequalities (2.14) we can (eventually) bound

$$\frac{\mathrm{Var}(\beta_1)}{\mathbb{E}[\beta_1]^2} = \frac{\mathbb{E}[\beta_1^2] - \mathbb{E}[\beta_1]^2}{\mathbb{E}[\beta_1]^2} \leq \frac{\mathbb{E}\left[N_1^2\right] - \mathbb{E}[N_1 - N_0]^2}{\mathbb{E}[N_1 - N_0]^2}. \tag{2.15}$$

Then since N_1 is a binomial random variable with mean $\mathbb{E}[N_1] \to \infty$ we have $\mathbb{E}\left[N_1^2\right] \sim \mathbb{E}[N_1]^2$. We have already seen that $\mathbb{E}[N_1]^2 \sim \mathbb{E}[N_1 - N_0]^2$ and hence the bound in (2.15) tends to 0.

Now since $\beta_1 \sim \mathbb{E}[\beta_1]$ with high probability and eventually $\mathbb{E}[\beta_1] > 0$, the final conclusion $\beta_1 > 0$ with high probability follows because for any $\epsilon \in (0, 1)$ we can eventually bound

$$\mathbb{P}\left[\beta_1(\bar{G}) > 0\right] \geq \mathbb{P}\left[\beta_1(\bar{G}) \geq (1 - \epsilon)\mathbb{E}[\beta_1]\right] \to 1. \tag{2.16}$$

The case for the weak symmetrisation has an identical proof, except that $\mathbb{E}[\#\mathring{E}] = \mathbb{E}[\#E] = n(n-1)p$. □

3 Path homology of directed graphs

3.1 Definition

Path homology was first introduced by Grigor'yan et al. (2012, 2020). The key concept behind path homology is that, in order to capture the asymmetry of a digraph, we should not construct a simplicial complex, but instead a *path complex*. In a simplicial complex, one can remove any vertex from a simplex and obtain a new simplex in the complex. This property may not hold for directed paths in digraphs; if we bypass a vertex in the middle of a path then we may not obtain a new path. However, we can always remove the initial or final vertex of a path and obtain a new path. This is the defining property of a path complex (Grigor'yan et al. (2012), §1).

Path homology can be defined on any path complex but for this paper we focus on the natural path complex associated to a digraph. Throughout this section we fix a ring R and a simple digraph $G = (V, E)$.

Definition 3.1 We make the following definitions to classify sequences of vertices in V:

1. Any sequence $v_0 \dots v_p$ of $(p+1)$ vertices $v_i \in V$ is an *elementary p-path*.
2. An elementary path is *regular* if no two consecutive vertices are the same, i.e. $v_i \neq v_{i+1}$ for every i. Otherwise, the path is called *non-regular or irregular*.
3. An elementary path is *allowed* if subsequent vertices are joined by a directed edge in the graph, i.e. $(v_i, v_{i+1}) \in E$ for every i.

Remark 3.2 An allowed path coincides with a combinatorial, directed walk.

Definition 3.3 The following R-modules are defined to be freely generated by the generators specified, for $p \geq 0$:

$$\Lambda_p := \Lambda_p(G; R) := R\left\langle \left\{ v_0 \dots v_p \text{ elementary } p\text{-path on } V \right\} \right\rangle \tag{3.1}$$

$$\mathcal{R}_p := \mathcal{R}_p(G; R) := R\left\langle \left\{ v_0 \dots v_p \text{ regular } p\text{-path on } V \right\} \right\rangle \tag{3.2}$$

$$\mathcal{A}_p := \mathcal{A}_p(G; R) := R\left\langle \left\{ v_0 \dots v_p \text{ allowed } p\text{-path in } G \right\} \right\rangle \tag{3.3}$$

For $p = -1$, we let $\Lambda_{-1} := \mathcal{R}_{-1} := \mathcal{A}_{-1} := R$. Given an elementary p-path $v_0 \dots v_p$, the corresponding generator of Λ_p is denoted $e_{v_0 \dots v_p}$. For convenience, given an edge $\tau = (a, b) \in E(G)$ we define $e_\tau := e_{ab}$ as an alias for the basis element of $\mathcal{A}_1$.

We can construct homomorphisms $\Lambda_p \to \Lambda_{p-1}$ for each p.

Definition 3.4 Given $p > 0$, we can define the *non-regular boundary map* $\partial_p : \Lambda_p \to \Lambda_{p-1}$ by setting

$$\partial_p(e_{v_0 \dots v_p}) := \sum_{i=0}^{p} (-1)^i e_{v_0 \dots \hat{v_i} \dots v_p} \tag{3.4}$$

where $v_0 \dots \hat{v_i} \dots v_p$ denotes the elementary $(p-1)$-path $v_0 \dots v_p$ with the vertex v_i omitted. This defines ∂_p on a basis of Λ_p, from which we extend linearly. In the case $p = 0$, we define $\partial_0 : \Lambda_0 \to R$ by

$$\partial_0 \left(\sum_{v \in V} \alpha_v e_v \right) := \sum_{v \in V} \alpha_v \tag{3.5}$$

which yields an element of R.

Remark 3.5 1. A standard check verifies that $\partial_{p-1} \circ \partial_p = 0$ (Grigor'yan et al. (2012), Lemma 2.4) and hence $\left\{ \Lambda_p, \partial_p \right\}$ forms a chain complex.

2. Since we assume all digraphs are simple, there are no self-loops. Therefore, any allowed path must be regular and hence

$$\mathcal{A}_p \subseteq \mathcal{R}_p \subseteq \Lambda_p. \tag{3.6}$$

In order to incorporate information about paths in the graph we would like a boundary operator between the $\mathcal{A}_p$. However, the boundary of an allowed path may not itself be allowed, because it involves removing vertices from the middle of paths. To resolve this, we define a R-module, for each $p \geq 0$, called the *space of* ∂*-invariant p-paths*

$$\Omega_p := \Omega_p(G; R) := \left\{ v \in \mathcal{A}_p \mid \partial_p v \in \mathcal{A}_{p-1} \right\} = \mathcal{A}_p \cap \partial_p^{-1}(\mathcal{A}_{p-1}). \tag{3.7}$$

Since $\partial_{p-1} \circ \partial_p = 0$, we see that $\partial_p(\Omega_p) \subseteq \Omega_{p-1}$. Hence, we can make the following construction.

Definition 3.6 The *non-regular chain complex* is

$$\dots \xrightarrow{\partial_3} \Omega_2 \xrightarrow{\partial_2} \Omega_1 \xrightarrow{\partial_1} \Omega_0 \xrightarrow{\partial_0} R \xrightarrow{\partial_{-1}} 0 \tag{3.8}$$

where each ∂_p is the restriction of the non-regular boundary map to Ω_p.

Definition 3.7 The homology of the non-regular chain complex (3.8) is the *non-regular path homology* of G. The pth homology group is denoted

$$H_p(G; R) := \frac{\ker \partial_p}{\operatorname{im} \partial_{p+1}}. \tag{3.9}$$

The rank of the pth homology group is pth Betti number, denoted $\overrightarrow{\beta}_p(G; R)$.

When computing Ω_p, one often encounters paths $v \in \mathcal{A}_p$ with irregular summands in their boundary. For example,

$$\partial_2(e_{iji}) = e_{ji} - e_{ii} + e_{ij}. \tag{3.10}$$

Since irregular summands are never allowed, these must be cancelled to obtain an element of Ω_p. An alternative construction, which is featured more frequently in the literature, alters the boundary operator to remove these irregularities.

There is a projection map $\pi : \Lambda_p \to \mathcal{R}_p$ which sends every irregular path to 0. This allows us to make the following construction:

Definition 3.8 For each $p \geq 0$, the *regular boundary operator* $\partial_p^{\mathcal{R}} : \mathcal{R}_p \to \mathcal{R}_{p-1}$ is defined by

$$\partial_p^{\mathcal{R}} := \pi \circ \partial_p. \tag{3.11}$$

With this new boundary operator we still have the issue that the boundary of an allowed path may not be allowed. Therefore, we again construct an R-module, for each $p \geq 0$, called the *space of* $\partial^{\mathcal{R}}$*-invariants p-paths*.

$$\Omega_p^{\mathcal{R}} := \Omega_p^{\mathcal{R}}(G; R) := \left\{ v \in \mathcal{A}_p \mid \partial_p^{\mathcal{R}} v \in \mathcal{A}_{p-1} \right\} = \mathcal{A}_p \cap (\partial_p^{\mathcal{R}})^{-1}(\mathcal{A}_{p-1}). \tag{3.12}$$

One can check that, given any irregular path v, either $\partial v = 0$ or ∂v is a sum of irregular paths (Grigor'yan et al. 2012, Lemma 2.9) and hence

$$\partial_{p-1}^{\mathcal{R}} \circ \partial_p^{\mathcal{R}} = \pi \circ \partial_{p-1} \circ \partial_p = \pi \circ 0 = 0. \tag{3.13}$$

Definition 3.9 The *regular chain complex* is

$$\dots \xrightarrow{\partial_3^{\mathcal{R}}} \Omega_2^{\mathcal{R}} \xrightarrow{\partial_2^{\mathcal{R}}} \Omega_1^{\mathcal{R}} \xrightarrow{\partial_1^{\mathcal{R}}} \Omega_0^{\mathcal{R}} \xrightarrow{\partial_0^{\mathcal{R}}} R \xrightarrow{\partial_{-1}^{\mathcal{R}}} 0 \tag{3.14}$$

where each $\partial_p^{\mathcal{R}}$ is the restriction of the non-regular boundary map to $\Omega_p^{\mathcal{R}}$.

Definition 3.10 The homology of the regular chain complex chain complex is the *regular path homology* of G and the kth homology group is denoted

$$H_k^{\mathcal{R}}(G; R) := \frac{\ker \partial_k^{\mathcal{R}}}{\operatorname{im} \partial_{k+1}^{\mathcal{R}}}. \tag{3.15}$$

We denote the *Betti numbers* for these homology groups by $\overrightarrow{\beta}_k^{\mathcal{R}}(G; R)$.

Remark 3.11 1. If R is also a field, then the homology groups H_k and $H_k^{\mathcal{R}}$ are vector spaces and so fully characterised, up to isomorphism, by $\overrightarrow{\beta}_k$ and $\overrightarrow{\beta}_k^{\mathcal{R}}$ respectively.
2. Since we augment the chain complex with R in dimension -1, this is technically a reduced homology, but we omit additional notation for simplicity.
3. As noted in (Grigor'yan et al. (2012), §5.1), given a subgraph $G' \subseteq G$ then, for every $p \geq 0$,

$$\Omega_p(G') \subseteq \Omega_p(G) \quad \text{and} \quad \Omega_p^{\mathcal{R}}(G') \subseteq \Omega_p^{\mathcal{R}}(G). \tag{3.16}$$

Notation 3.12 When G is clear from context, we shall omit it from notation. If the coefficient ring R is omitted from notation, assume that $R = \mathbb{Z}$.

Note that the primary difference between the regular and non-regular chain complex is the boundary operator. The difference between the boundary operators ∂_p and $\partial_p^{\mathcal{R}}$ affects the difference between the R-modules Ω_p and $\Omega_p^{\mathcal{R}}$.

3.2 Proof of Theorem 1.5

As an easy first step, we show that, when graph density is too low, it is very unlikely that there are any long paths within the digraph. Therefore, for large k, $\mathcal{A}_k$ becomes trivial and consequently $\overrightarrow{\beta}_k = 0$.

Proposition 3.13 *Given $N \in \mathbb{N}$, if $p = p(n) = o(n^{-(N+1)/N})$ for some $N \in \mathbb{N}$ then, given a random directed graph $G \sim \overrightarrow{G}(n, p)$, for all $k \geq N$ we have*

$$\lim_{n\to\infty} \mathbb{P}(\overrightarrow{\beta}_k(G) = 0) = 1. \tag{3.17}$$

Proof Note that it suffices to show that $\mathbb{P}(\mathcal{A}_N = \{0\}) \to 1$ as $n \to \infty$ because, if there are no allowed N-paths, then there are certainly no allowed k-paths. If there are no allowed k-paths then $\Omega_k = \{0\}$ and so $\overrightarrow{\beta}_k = 0$.

For $\mathcal{A}_N$ to be non-trivial there must be some combinatorial, directed walk of length N. Equivalently, there must exist a combinatorial, directed cycle or a combinatorial, directed path of length N (or both).

If $p = o(n^{-(N+1)/N})$ then certainly $p = o(n^{-1})$ and hence, following the proof of Proposition 2.16, the probability that there is a directed cycle tends to 0 as $n \to \infty$.

A combinatorial, directed path is a sequence of $N+1$ distinct nodes, each joined by an edge in the forward direction. By a union bound, the probability that there exists such a sequence is at most

$$\binom{n}{N+1}(N+1)!\, p^N \leq n^{N+1} p^N \tag{3.18}$$

which, by the assumption on p, tends to 0 as $n \to \infty$. □

Proof of Theorem 1.5 By Proposition 3.13, it suffices to note that $n^\alpha = o(n^{-(N+1)/N})$ whenever $\alpha < -\frac{N+1}{N}$. □

This theorem is very weak. For example, to obtain $\overrightarrow{\beta}_1 = 0$ with high probability, we require $p = o(n^{-2})$, in which case the expected number of edges in the digraph tends to 0. The weakness of this result stems from its reliance on the chain of inequalities

$$\overrightarrow{\beta}_k \leq \operatorname{rank}(\ker \partial_k) \leq \operatorname{rank} \Omega_k \leq \operatorname{rank} \mathcal{A}_k. \tag{3.19}$$

There is likely a region of graph densities wherein one or more of these inequalities is strict. Hence, in order to obtain stronger results, we require an understanding of Ω_k, at the very least.

3.3 Chain group generators

Proposition 3.14 (Grigor'yan et al. 2012, §3.3) *For any simple digraph $G = (V, E)$,*

$$\Omega_0 = \Omega_0^{\mathcal{R}} = R\langle V\rangle = \mathcal{A}_0 \quad \textit{and} \quad \Omega_1 = \Omega_1^{\mathcal{R}} = R\langle E\rangle = \mathcal{A}_1. \tag{3.20}$$

Proof Certainly $\Omega_0 \subseteq \mathcal{A}_0$ and $\Omega_1 \subseteq \mathcal{A}_1$. Moreover, the boundary of any vertex is just an element of $R = \mathcal{A}_{-1}$ and hence allowed. The boundary of any edge is a sum of vertices and any vertex is an allowed 0-path. Therefore $\mathcal{A}_0 \subseteq \Omega_0$ and $\mathcal{A}_1 \subseteq \Omega_1$. □

We can also see that the non-regular chain complex is a subcomplex of the regular chain complex, which immediately implies an inequality between the Betti numbers. This subcomplex relation was first noted by Grigor'yan et al. (2012, Proposition 3.16).

Proposition 3.15 *For any simple digraph G, the non-regular chain complex is a subcomplex of the regular chain complex. In particular, for each $p \geq 0$, we have*

$$\Omega_p(G) \subseteq \Omega_p^{\mathcal{R}}(G). \tag{3.21}$$

Proof Suppose $v \in \Omega_p$, then $\partial_p(v) \in \mathcal{A}_{p-1}$. We have seen that $\mathcal{A}_{p-1} \subseteq \mathcal{R}_{p-1}$. Hence, if we project $\partial_p(v)$ onto $\mathcal{R}_{p-1}$ via π, we do not remove any summands. Therefore

$$\partial_p^{\mathcal{R}}(v) = \pi\left(\partial_p(v)\right) = \partial_p(v) \in \mathcal{A}_{p-1}. \tag{3.22}$$

Certainly $v \in \mathcal{A}_p$ and hence $v \in \Omega_p^{\mathcal{R}}$. Since the two operators, ∂_p and $\partial_p^{\mathcal{R}}$, agree on Ω_p, the non-regular chain complex is a subcomplex of the regular chain complex. □

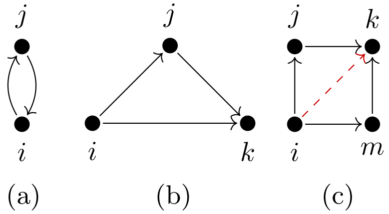


Fig. 2 Generators for $\Omega_2^{\mathcal{R}}(G;\mathbb{Z})$ as described in Proposition 3.17. **a** A double edge. **b** A directed triangle. **c** A long square. The red, dashed line must not be present for the third motif to constitute a long square

Corollary 3.16 *For any simple digraph G, $\vec{\beta}_1^{\mathcal{R}}(G) \leq \vec{\beta}_1(G)$.*

Proof By Proposition 3.14, the two complexes coincide in dimensions 0 and 1 and hence $\operatorname{rank}\ker\partial_1 = \operatorname{rank}\ker\partial_1^{\mathcal{R}}$. By Proposition 3.15, $\operatorname{im}\partial_2 \subseteq \operatorname{im}\partial_2^{\mathcal{R}}$ and hence $\operatorname{rank}\operatorname{im}\partial_2 \leq \operatorname{rank}\operatorname{im}\partial_2^{\mathcal{R}}$. Therefore,

$$\vec{\beta}_1^{\mathcal{R}}(G) = \operatorname{rank}\ker\partial_1^{\mathcal{R}} - \operatorname{rank}\operatorname{im}\partial_2^{\mathcal{R}} \leq \operatorname{rank}\ker\partial_1 - \operatorname{rank}\operatorname{im}\partial_2 = \vec{\beta}_1(G). \tag{3.23}$$

□

Note, given a directed edge $\tau = (i, j)$, $\partial_1(e_{ij}) = \partial_1^{\mathcal{R}}(e_{ij}) = e_j - e_i$. From this, it is easy to obtain the characterisation of the lowest Betti number first stated in Sect. 1.

Proof of Theorem 1.3 A standard argument shows that $\vec{\beta}_0 = \vec{\beta}_0^{\mathcal{R}} = \#C - 1$, where $\#C$ is the number of weakly connected components of the digraph G. Note, $\#C$ coincides with the number of connected components of the symmetrisation $\bar{G}$. The result follows by Lemma 2.12 and a standard result due to Erdős and Rényi (see e.g. Erdős and Rényi 1960; Kahle 2009). □

Unfortunately, higher chain groups do not enjoy such a concise description. However, when working with coefficient over $\mathbb{Z}$, it is possible to write down generators for $\Omega_2^{\mathcal{R}}$, in terms of motifs within the digraph G.

Proposition 3.17 (Grigor'yan et al. (2014), Proposition 2.9) *Let G be any finite digraph. Then any $\omega \in \Omega_2^{\mathcal{R}}(G;\mathbb{Z})$ can be represented as a linear combination of 2-paths of the following three types:*

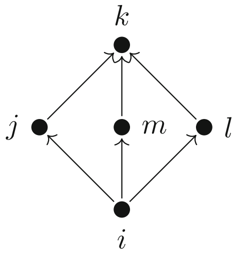
1. *e_{iji} with $i \to j \to i$ (double edges);*
2. *e_{ijk} with $i \to j \to k$ and $i \to k$ (directed triangles);*
3. *$e_{ijk} - e_{imk}$ with $i \to j \to k$, $i \to m \to k$, $i \not\to k$ and $i \neq k$ (long squares).*

Remark 3.18 Note that all vertices i, j, k, m in this theorem are distinct, either due to the existence of an edge (e.g. $i \to j$ implies $i \neq j$) or explicit statement (e.g. $i \neq k$).

The following non-regular corollary follows immediately since, by Proposition 3.15, $\Omega_2 \subseteq \Omega_2^{\mathcal{R}}$.

Corollary 3.19 *Let G be any finite digraph. Then any $\omega \in \Omega_2(G;\mathbb{Z})$ can be represented as a linear combination of 2-paths of the three types enumerated in Proposition 3.17.*

Fig. 3 Linearly dependent long squares with source i and sink k



Note that each of the generators in Proposition 3.17 are elements of $\Omega_2^{\mathcal{R}}$ and hence they form a generating set for $\Omega_2^{\mathcal{R}}(G;\mathbb{Z})$. Note that elements of each type reside in mutually orthogonal components of $\mathcal{A}_2$ because they are supported on distinct basis elements. That is, we can write

$$\Omega_2^{\mathcal{R}} = D \oplus T \oplus S \tag{3.24}$$

where D is freely generated by all double edges e_{iji} in G, and T is freely generated by all directed triangles e_{ijk} in G. The final component, S, is generated by all long squares $e_{ijk} - e_{imk}$ in G. However, they may not be linearly independent, for example, as seen in Fig. 3,

$$(e_{ijk} - e_{imk}) + (e_{imk} - e_{ilk}) + (e_{ilk} - e_{ijk}) = 0. \tag{3.25}$$

Note that double edges are not ∂-invariant paths, i.e. $e_{iji} \notin \Omega_2$. However there are linear combinations of double edges which do belong to Ω_2. For example, suppose $i \to j \to i$ and $i \to k \to i$, then $e_{iji} - e_{iki} \in \Omega_2$. It is possible to state a non-regular version of Proposition 3.17, in which all generators are elements of Ω_2. This can be achieved by replacing double edge generators with such differences of double edges, which share a common base point. However, we omit this result, as it is not necessary for our main contribution.

Remark 3.20 1. An alternative approach to computing rank Ω_2 and rank $\Omega_2^{\mathcal{R}}$ was first seen in (Grigor'yan et al. (2012), Proposition 4.2) and is explored further in Appendix 1.

2. For the interested reader, more results which characterise relations between the Ω_p are available in Grigor'yan et al. (2012, 2020).
3. We can use Proposition 3.17 to obtain an intuition for $H_1^{\mathcal{R}}(G;\mathbb{Z})$. Starting with the cell complex $\mathring{G}$, glue in a 2-cell for each generator identified by Proposition 3.17 by identifying its boundary with the corresponding motif in $\mathring{G}$. Then H_1 of this cell complex coincides with $H_1^{\mathcal{R}}(G;\mathbb{Z})$. Unfortunately, since we do not have generators for $\Omega_p^{\mathcal{R}}$ for $p > 2$, developing intuition in higher degrees is much harder.

Example 3.21 For further intuition, we reproduce the example given in (Grigor'yan et al. (2012), Proposition 4.7) but compute both regular and non-regular path homology. A cycle graph is a weakly connected digraph, on $n \geq 2$, nodes such that each vertex has degree 2. Fix a cycle graph G. Then, $H_1^{\mathcal{R}}(G; \mathbb{Z}) \cong \mathbb{Z}$ unless G is a double edge, directed triangle or long square, in which case $H_1^{\mathcal{R}}(G; \mathbb{Z}) = 0$. Whereas, $H_1(G; \mathbb{Z}) \cong \mathbb{Z}$ unless G is a directed triangle or long square, in which case $H_1(G; \mathbb{Z}) = 0$. For more examples, please consult Grigor'yan et al. (2012).

4 Asymptotic results for path homology

Intuitively, we expect that the two transitions, identified in Fig. 9, correspond to two distinct topological phenomena. When density becomes sufficiently large, cycles start to appear in the graph and $\ker \partial_1$ is non-empty for the first time. Then, when density becomes too large, boundaries enter into Ω_2 which begin to cancel out all of the cycles, removing all homology. In the interim period, we expect that the number of cycles and the number of boundaries is approximately balanced. Therefore, in order to understand the lower boundary we should study $\ker \partial_1$ and in order to understand the upper boundary we should study $\operatorname{im} \partial_2$. In order to show that $\overrightarrow{\beta}_1 > 0$ in the 'goldilocks' region we should compare the growth rates of $\operatorname{rank} \ker \partial_1$ and $\operatorname{rank} \operatorname{im} \partial_2$, or some approximation thereof. Moreover we expect reasonable conditions on $p(n)$ to be of the form $p = o(n^\alpha)$ or $p = \omega(n^\alpha)$ for some α, since conditions of this sort constrain $p(n)$ relative to straight lines through Fig. 9.

4.1 Proof of Theorem 1.4(1)

In order to characterise the behaviour of $\overrightarrow{\beta}_1$ when it is non-trivial, we will follow the approach of Kahle in (2009, §7). The approach is to use the 'Morse inequalities'. In the context of a digraph G, denote the ranks of the chain groups by $N_k := \operatorname{rank} \Omega_k(G; \mathbb{Z})$ and $N_k^{\mathcal{R}} := \operatorname{rank} \Omega_k^{\mathcal{R}}(G; \mathbb{Z})$. Then we have

$$-N_{k-1} + N_k - N_{k+1} \leq \overrightarrow{\beta}_k \leq N_k. \tag{4.1}$$

and a similar set of inequalities between the $N_k^{\mathcal{R}}$ and $\overrightarrow{\beta}_k^{\mathcal{R}}$. It is usually easier to compute the rank of chain groups than the rank homology groups. Hence, we use the limiting behaviour of N_k to investigate the limiting behaviour of $\overrightarrow{\beta}_k$. First we will need estimates for $\mathbb{E}[N_k]$.

Lemma 4.1 *For a random directed graph $G \sim \overrightarrow{G}(n, p)$ we have the following expectations*

$$\mathbb{E}[N_0] = \mathbb{E}\left[N_0^{\mathcal{R}}\right] = n \tag{4.2}$$

$$\mathbb{E}[N_1] = \mathbb{E}\left[N_1^{\mathcal{R}}\right] = n(n-1)p \tag{4.3}$$

$$\mathbb{E}[N_2] \leq \mathbb{E}\left[N_2^{\mathcal{R}}\right] \leq n^2p^2 + n^3p^3 + n^4p^4. \tag{4.4}$$

Proof The first two claims are clear since they count the expected number of nodes and edges in G, respectively. There is no difference between the regular and non-regular chain complex in dimensions 0 and 1.

We use Proposition 3.17 to compute bounds for $\mathbb{E}[N_2^{\mathcal{R}}]$ and then the bound on $\mathbb{E}[N_2]$ follows immediately because $\Omega_2 \subseteq \Omega_2^{\mathcal{R}}$ (by Proposition 3.15). Since both orientations of a double edge constitute a distinct basis element of $\Omega_2^{\mathcal{R}}$, the expected number of double edges is $n(n-1)p^2$, which is bounded above by n^2p^2. The expected number of directed triangles is $6\binom{n}{3}p^3$, because each subset of 3 vertices can support 6 distinct directed triangles.

Counting linearly independent long squares is more involved. For an upper bound, note that any subset of 4 vertices can support 12 long squares (not double counting for the two orientations since they differ by a factor of ±1). Each fixed long square appears with probability $p^4(1-p)$. Therefore an upper bound on the number of *linearly independent* long squares is

$$12\binom{n}{4}p^4(1-p) \leq n^4p^4. \tag{4.5}$$

Combining these counts yields the upper bound on $\mathbb{E}[N_2^{\mathcal{R}}]$. □

Proposition 4.2 *If $G \sim \overrightarrow{G}(n,p)$ where $p = p(n)$, with $p(n) = \omega(n^{-1})$ and $p(n) = o(n^{-2/3})$, then*

$$\mathbb{E}[-N_0 + N_1 - N_2] \sim \mathbb{E}[N_1] \quad \textit{and hence} \quad \mathbb{E}[\overrightarrow{\beta}_1(G)] \sim \mathbb{E}[N_1]\,. \tag{4.6}$$

Moreover, the same relations hold between the $N_k^{\mathcal{R}}$ and $\overrightarrow{\beta}_k^{\mathcal{R}}$.

Proof We prove the non-regular case, but the regular case follows from an identical argument. Using our expectations from Lemma 4.1, we see $\mathbb{E}[N_0] = o(\mathbb{E}[N_1])$ because

$$\lim_{n\to\infty}\frac{\mathbb{E}[N_0]}{\mathbb{E}[N_1]} = \lim_{n\to\infty}\frac{n}{n(n-1)p} = \lim_{n\to\infty}\frac{1}{np} = 0, \tag{4.7}$$

where the final equality follows from the assumption $p = \omega(n^{-1})$. Next, note that

$$0 \leq \lim_{n\to\infty}\frac{\mathbb{E}[N_2]}{\mathbb{E}[N_1]} \leq \lim_{n\to\infty}\frac{n^2p^2 + n^3p^3 + n^4p^4}{n(n-1)p} = \lim_{n\to\infty}\left(p + np^2 + n^2p^3\right). \tag{4.8}$$

The assumption $p = o(n^{-2/3})$ is equivalent to $n^{2/3}p \to 0$ as $n \to \infty$. This is sufficient to ensure $p \to 0$, $np^2 \to 0$ and $n^2p^3 \to 0$ as $n \to \infty$ and so $\mathbb{E}[N_2] = o(\mathbb{E}[N_1])$. Hence the first relation follows and the latter follows immediately from the Morse inequalities (4.1). □

Remark 4.3 If we choose $p = n^\alpha$ to satisfy the hypotheses of Proposition 4.2, then we must have $-1 < \alpha < -2/3$ in which case $\mathbb{E}[N_1]$ is of the order $n^2 p = n^{\alpha+2}$. Then $\alpha + 2 > 1$ so $\mathbb{E}[\overrightarrow{\beta}_1] \to \infty$ at least linearly as $n \to \infty$.

Proposition 4.4 *If $G \sim \overrightarrow{G}(n, p)$ where $p = p(n)$, with $p(n) = \omega(n^{-1})$ and $p(n) = o(n^{-2/3})$, then $\overrightarrow{\beta}_1(G) \sim \mathbb{E}[\overrightarrow{\beta}_1(G)]$ and $\overrightarrow{\beta}_1^{\mathcal{R}}(G) \sim \mathbb{E}[\overrightarrow{\beta}_1^{\mathcal{R}}(G)]$ with high probability.*

Proof We prove the non-regular case but the regular case follows by an identical argument. As in Proposition 2.17, it suffices to show that $\mathrm{Var}(\overrightarrow{\beta}_1) = o(\mathbb{E}[\overrightarrow{\beta}_1]^2)$.

We have seen that $\mathbb{E}[N_1] \sim \mathbb{E}[-N_0 + N_1 - N_2]$ and certainly $\mathbb{E}[N_1] \to \infty$ as $n \to \infty$. Therefore, eventually $\mathbb{E}[-N_0 + N_1 - N_2] \geq 0$ so, by the Morse inequalities, eventually we can bound

$$\mathbb{E}[\overrightarrow{\beta}_1]^2 \geq \mathbb{E}[-N_0 + N_1 - N_2]^2 . \tag{4.9}$$

Moreover, we always have $\overrightarrow{\beta}_1^2 \leq N_1^2$ so eventually we can bound

$$\frac{\mathrm{Var}(\overrightarrow{\beta}_1)}{\mathbb{E}[\overrightarrow{\beta}_1]^2} = \frac{\mathbb{E}[\overrightarrow{\beta}_1^2] - \mathbb{E}[\overrightarrow{\beta}_1]^2}{\mathbb{E}[\overrightarrow{\beta}_1]^2} \leq \frac{\mathbb{E}\big[N_1^2\big] - \mathbb{E}[-N_0 + N_1 - N_2]^2}{\mathbb{E}[N_0 + N_1 - N_2]^2} . \tag{4.10}$$

To conclude, it suffices to show $\mathbb{E}\big[N_1^2\big] \sim \mathbb{E}[-N_0 + N_1 - N_2]^2$. By Proposition 4.2, we know $\mathbb{E}[-N_0 + N_1 - N_2]^2 \sim \mathbb{E}[N_1]^2$. Then $\mathbb{E}\big[N_1^2\big] \sim \mathbb{E}[N_1]^2$ because N_1 is a binomial random with mean $\mathbb{E}[N_1] \to \infty$. □

Proof of Theorem 1.4(1) If $p(n) = n^\alpha$ for $-1 < \alpha < -2/3$ then $p = \omega(n^{-1})$ and $p = o(n^{-2/3})$. Moreover, $p = \omega(n^{-1})$ is sufficient to ensure $\mathbb{E}[N_1] \to \infty$ as $n \to \infty$. Since N_1 is a binomial random variable, this implies $N_1 \sim \mathbb{E}[N_1]$ with high probability. Combining Proposition 4.2 and Proposition 4.4 yields the result. □

4.2 Proof of Theorem 1.4(2)

Having done the work of showing $\overrightarrow{\beta}_1 \sim N_1$, showing that $\overrightarrow{\beta}_1 > 0$ is now an easy corollary.

Corollary 4.5 *If $G \sim \overrightarrow{G}(n, p)$ where $p = p(n)$, with $p(n) = \omega(n^{-1})$ and $p(n) = o(n^{-2/3})$, then*

$$\lim_{n\to\infty} \mathbb{P}(\overrightarrow{\beta}_1(G) > 0) = \lim_{n\to\infty} \mathbb{P}(\overrightarrow{\beta}_1^{\mathcal{R}}(G) > 0) = 1. \tag{4.11}$$

Proof As in Proposition 2.17, this follows immediately because $\overrightarrow{\beta}_1 \sim \mathbb{E}[\overrightarrow{\beta}_1]$ and $\overrightarrow{\beta}_1^{\mathcal{R}} \sim \mathbb{E}[\overrightarrow{\beta}_1^{\mathcal{R}}]$ with high probability and both expectations are eventually positive. □

Proof of Theorem 1.4(2) If $p(n) = n^{\alpha}$ for $-1 < \alpha < -2/3$ then $p = \omega(n^{-1})$ and $p = o(n^{-2/3})$. Hence, by Corollary 4.5, $\mathbb{P}[\overrightarrow{\beta}_1(G) > 0] \to 1$ and $\mathbb{P}[\overrightarrow{\beta}_1^{\mathcal{R}}(G) > 0] \to 1$ as $n \to \infty$. □

4.3 Proof of Theorem 1.4(3)

Having understood the behaviour of $\mathbb{E}[\overrightarrow{\beta}_1]$ in the 'goldilocks' region, we turn our attention to the boundaries of this region. As with the symmetric methods, we expect that if p is too small then $\overrightarrow{\beta}_1$ will vanish due to the lack of cycles.

Proposition 4.6 *If $p = p(n) = o(n^{-1})$ then, given directed random graphs $G \sim \overrightarrow{G}(n, p)$, we have*

$$\lim_{n\to\infty} \mathbb{P}(\overrightarrow{\beta}_1(G) = 0) = \lim_{n\to\infty} \mathbb{P}(\overrightarrow{\beta}_1^{\mathcal{R}}(G) = 0) = 1. \tag{4.12}$$

Proof Given a double edge iji, note $\partial_2^{\mathcal{R}}(e_{iji}) = e_{ij} + e_{ji}$. Hence, for the regular case, a necessary condition for $\overrightarrow{\beta}_1^{\mathcal{R}} > 0$ is that there is some undirected cycle, of length at least 3, in the digraph. Whereas, for the non-regular case, a necessary condition is that there is some undirected cycle, of length at least 2, in the digraph. Therefore, the proof of the regular case is identical to the proof that $\overrightarrow{\beta}_1(\bar{G}) = 0$ with high probability and the proof of the non-regular case is identical to the proof that $\overrightarrow{\beta}_1(\mathring{G}) = 0$ with high probability, as seen in Proposition 2.16. □

Proof of Theorem 1.4(3) Assume that $p(n) = n^{\alpha}$. If $\alpha < -1$ then $p = o(n^{-1})$ and hence, by Proposition 4.6, $\mathbb{P}(\overrightarrow{\beta}_1(G) = 0) \to 1$ and $\mathbb{P}(\overrightarrow{\beta}_1^{\mathcal{R}}(G) = 0) \to 1$ as $n \to \infty$. □

4.4 Proof of Theorem 1.4(4)

For the previous subsection we chose p small enough to ensure that it is highly likely that $\ker \partial_1$ is empty. We also observe $\overrightarrow{\beta}_1$ vanishing for larger values of p. In these regimes $\ker \partial_1$ is likely non-empty but all cycles are cancelled out by boundaries. Put another way, we wish to show that, when p is large, every cycle $\omega \in \ker \partial_1$ can be shown to satisfy

$$\omega = 0 \qquad (\text{mod } \operatorname{im} \partial_2). \tag{4.13}$$

The strategy is to find conditions under which cycles supported on many vertices can be reduced down to cycles supported on just 3 vertices, and then show that small cycles can be reduced to 0. For this subsection, we will prove that $\mathbb{P}[\overrightarrow{\beta}_1(G) = 0] \to 1$ which then implies, by Corollary 3.16, that $\mathbb{P}[\overrightarrow{\beta}_1^{\mathcal{R}}(G) = 0] \to 1$, as $n \to \infty$. First, we need to ensure that we can choose a basis for $\ker \partial_1$ which will be amenable to our reduction strategy.

Definition 4.7 Given an element $\omega \in \Omega_1(G)$, we can write ω in terms of the standard basis

$$\omega = \sum_{\tau \in E(G)} \alpha_\tau e_\tau . \tag{4.14}$$

1. We define the *support* of v to be

$$\operatorname{supp}(\omega) := \{\tau \in E \mid \alpha_\tau \neq 0\} . \tag{4.15}$$

2. We call ω a *fundamental cycle* if $\omega \in \ker \partial_1$, $\alpha_\tau \in \{\pm 1\}$ for each $\tau \in E$, and $\operatorname{supp}(\omega)$ forms a combinatorial, undirected cycle in G.

Lemma 4.8 *Given a simple digraph,* $\ker \partial_1$ *has a basis of fundamental cycles in* G.

Proof Take an undirected spanning forest T for G, i.e. a subgraph of T in which every two vertices in the same weakly connected component of G can be joined by a unique undirected path through T. One can check that $\partial_1 : \Omega_1(T; R) \to \Omega_0(T; R)$ has trivial kernel, since there are no undirected cycles in T.

Given an edge outside the forest $\tau = (a, b) \in E(G) \setminus E(T)$, there is a unique undirected path ρ through T which joins the endpoints of τ:

$$\rho = (a = v_0, \tau_1, v_1, \dots, \tau_{k-1}, v_{k-1}, \tau_k, b = v_k). \tag{4.16}$$

for some $v_i \in V(G)$, $\tau_i \in E(G)$. Define

$$\alpha_i := \begin{cases} 1 & \text{if } \tau_i = (v_{i-1}, v_i) \\ -1 & \text{if } \tau_i = (v_i, v_{i-1}) \end{cases} \tag{4.17}$$

and note that

$$\partial_1 \left(e_\tau - \sum_{i=1}^{k} \alpha_i e_{\tau_i} \right) = 0. \tag{4.18}$$

Hence $b_\tau := e_\tau - \sum_{i=1}^{k} \alpha_i e_{\tau_i} \in \ker \partial_1$. Note that b_τ is a fundamental cycle.

The set $B := \{b_\tau \mid \tau \in E(G) \setminus E(T)\}$ is linearly independent because, given $b_\tau \in B$, no other $b_{\tau'} \in B$ involves the basis element e_τ of Ω_1. Note, we can write

$$\Omega_1 = R\langle \{e_\tau \mid \tau \in E(T)\} \rangle \oplus R\langle B \rangle . \tag{4.19}$$

Since there are no cycles in the spanning forest T, the kernel of ∂_1 on the first component is trivial. Therefore, $\operatorname{rank} \ker \partial_1 \leq \# B$ and hence B spans $\ker \partial_1$. □

Now we can describe the strategy by which systematically reduce long fundamental cycles into smaller ones. We design a combinatorial condition on a directed graph which is more likely to occur at higher densities.

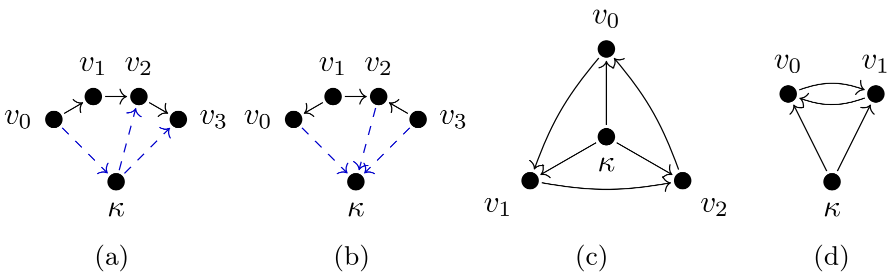


Fig. 4 **a**, **b** Directed centres for undirected paths of length 3. The blue, dashed edges constitute $J_{\sigma,\kappa}$. **c** A cycle centre for a 3-cycle. **d** A cycle centre for a 2-cycle

Definition 4.9 1. An undirected path $\sigma \subseteq G$, on vertices $(v_1, \dots, v_k)$, is said to be *reducible* if there is some *shortcut edge*, $e = (v_i, v_j)$, with $|i - j| > 1$ such that $\overrightarrow{\beta}_1(\sigma \cup e) = 0$. If a path is not reducible then it is called *irreducible*.

2. Given an undirected path $\sigma \subseteq G$ of length 3, on vertices (v_0, v_1, v_2, v_3), and a vertex $\kappa \in V(G) \setminus V(\sigma)$, define the *linking set*

$$J_{\sigma,\kappa} := \{e \in E(G) \mid e = (v_i, \kappa) \text{ or } e = (\kappa, v_i) \text{ for some } i\}\,. \tag{4.20}$$

Such a vertex, κ, is called a *directed centre* for σ if there is some subset of linking edges $J' \subseteq J_{\sigma,\kappa}$ such that $\overrightarrow{\beta}_1(\sigma \cup J') = 0$ and $\sigma \cup J'$ contains an undirected path, of length 2, on the vertices (v_0, κ, v_3).

3. A *cycle centre* for a directed cycle of length k, on vertices $(v_0, \dots, v_{k-1})$, is a vertex $\kappa \in V(G) \setminus \{v_0, \dots, v_{k-1}\}$ such that $(k, v_i) \in E(G)$ for all $i = 0, \dots, k-1$ or $(v_i, k) \in E(G)$ for all i.

In the following examples, we demonstrate the utility of directed centres.

Example 4.10 Figure 5 shows four examples of the reduction strategy described by Lemma 4.11. For illustration, we describe these reductions in more detail below.

1. In Fig. 5a, the initial undirected path of length 3 has a directed centre κ which does not coincide with a vertex in the rest of the cycle. Therefore, we can write

$$\begin{aligned}&[e_{v_0v_1} + e_{v_1v_2} - e_{v_3v_2}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_6v_5} + e_{v_6v_7} + e_{v_7v_0}\\ &\quad = [e_{v_0\kappa} + e_{\kappa v_3}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_6v_5} + e_{v_6v_7} + e_{v_7v_0} \pmod{\operatorname{im}\partial_2}.\end{aligned} \tag{4.21}$$

2. In Fig. 5b, the path has a directed centre $\kappa = v_5$. Replacing the initial path with the smaller path, via the directed centre, yields a sum of two fundamental cycles:

$$\begin{aligned}&[e_{v_0v_1} + e_{v_1v_2} - e_{v_3v_2}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_6v_5} + e_{v_6v_7} + e_{v_7v_0}\\ &\quad = [e_{v_0\kappa} + e_{\kappa v_3}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_6v_5} + e_{v_6v_7} + e_{v_7v_0} \pmod{\operatorname{im}\partial_2}\\ &\quad = [e_{v_0v_5} - e_{v_6v_5} + e_{v_6v_7} + e_{v_7v_0}] + [e_{v_5v_3} + e_{v_3v_4} - e_{v_5v_4}] \pmod{\operatorname{im}\partial_2}.\end{aligned} \tag{4.22}$$

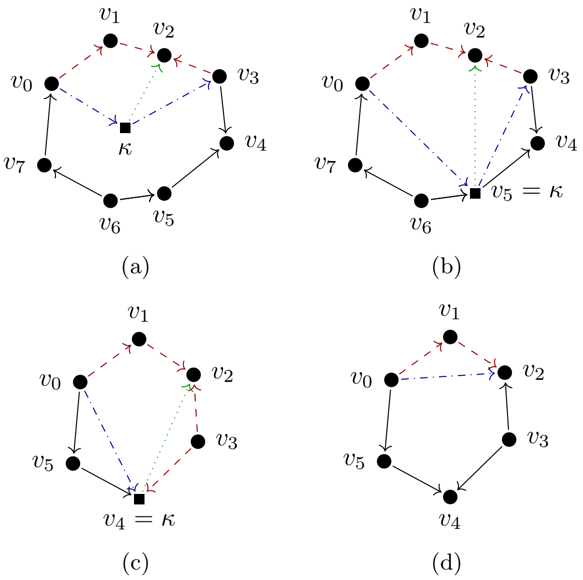


Fig. 5 Examples of the reductions used in Lemma 4.11 which are explained in greater depth in Example 4.10. Black, solid edges indicate the initial cycle. Blue, dash-dotted edges are new edges in the reduced cycle. Red, dashed edges are those removed in the reduced cycle. Green, dotted edges must be present in order to do the illustrated reduction. Square nodes symbolise directed centres for the undirected path (v_0, v_1, v_2, v_3).

3. In Fig. 5c, the path has a directed centre $\kappa = v_4$. Replacing the initial path (v_0, v_1, v_2, v_3) with the smaller path (v_0, κ, v_3) yields a much smaller support since the edge (v_3, v_4) gets cancelled out:

$$\begin{aligned}&[e_{v_0v_1} + e_{v_1v_2} - e_{v_3v_2}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_0v_5}\\ &\quad = [e_{v_0\kappa} - e_{v_3\kappa}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_0v_5} \pmod{\operatorname{im}\partial_2}\\ &\quad = e_{v_0v_4} - e_{v_5v_4} - e_{v_0v_5} \pmod{\operatorname{im}\partial_2}.\end{aligned} \tag{4.23}$$

4. Finally, in Fig. 5d, the initial path is reducible via the shortcut edge (v_0, v_2) and hence

$$\begin{aligned}&[e_{v_0v_1} + e_{v_1v_2} - e_{v_3v_2}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_0v_5}\\ &\quad = [e_{v_0v_2} - e_{v_3v_2}] + e_{v_3v_4} - e_{v_5v_4} - e_{v_0v_5} \pmod{\operatorname{im}\partial_2}.\end{aligned} \tag{4.24}$$

These examples tell the story of each case in the following lemma, in which we confirm that the presence of directed centres allows us to systematically reduce fundamental cycles.

Lemma 4.11 *For any simple digraph G, suppose every irreducible, undirected path of length 3 has a directed centre. Given a fundamental cycle $\omega \in \ker\partial_1$ with $\#\operatorname{supp}(\omega) = k \geq 4$, there exists fundamental cycles $\tilde{\omega}_1, \tilde{\omega}_2 \in \ker\partial_1$ such that*

$$\omega = \tilde{\omega}_1 + \tilde{\omega}_2 \pmod{\operatorname{im}\partial_2} \tag{4.25}$$

with $\#\operatorname{supp}(\tilde{\omega}_1) + \#\operatorname{supp}(\tilde{\omega}_2) \leq k - 1$ and, potentially, one or more $\tilde{\omega}_i = 0$.

Proof Since v is a fundamental cycle, it is supported on some combinatorial, undirected cycle

$$\rho = (v_0, \tau_1, v_1, \ldots, v_{k-1}, \tau_k, v_k = v_0). \tag{4.26}$$

for some $v_i \in V$ and $\tau_i \in E$ ordered such that

$$\omega = \sum_{i=1}^{k} \alpha_i e_{\tau_i} \tag{4.27}$$

where

$$\alpha_i := \begin{cases} 1 & \text{if } \tau_i = (v_{i-1}, v_i) \\ -1 & \text{if } \tau_i = (v_i, v_{i-1}) \end{cases}. \tag{4.28}$$

Since $k \geq 4$, the vertices $(v_0, \ldots, v_3)$ are distinct and, along with the edges τ_1, τ_2, τ_3, form an undirected path of length 3. Either this is reducible via some shortcut edge $\tau \in E$, or there exists a directed centre $\kappa \in V$. In either case, there is some undirected path, from v_0 to v_3, of length at most 2. This path is represented by some $\eta' \in \Omega_1$ with coefficients in $\{\pm 1\}$, such that $\partial_1 \eta' = e_{v_3} - e_{v_1}$ and $\# \operatorname{supp} \eta' \leq 2$.

Since both η' and $\eta := \sum_{i=1}^{3} \alpha_i e_{\tau_i}$ are supported on undirected paths from v_0 to v_3, we have $\partial_1 \left(\eta - \eta'\right) = 0$. Since $\operatorname{supp}(\eta) \subseteq E(G')$ and $\operatorname{supp}(\eta') \subseteq E(G')$ for some subgraph $G' \subseteq G$ with $\overrightarrow{\beta}_1(G') = 0$ (either due to reducibility or a directed centre), there is some $u \in \Omega_2(G)$ such that $\partial_2 u = \eta - \eta'$. Therefore we can replace the initial undirected path of length 3, in v, with an undirected path of length at most 2, i.e

$$\omega = \eta' + \sum_{i=4}^{k} \alpha_i e_{\tau_i} =: \tilde{\omega} \pmod{\operatorname{im} \partial_2}. \tag{4.29}$$

Certainly $\# \operatorname{supp}(\tilde{\omega}) \leq 2 + (k-3) < 3 + (k-3) = \# \operatorname{supp}(\omega)$ so $\tilde{\omega}$ has a strictly smaller support. It remains to prove that $\tilde{\omega}$ can be decomposed into a sum of at most two fundamental cycles. In the case that the path has a directed centre κ, we split into two further sub-cases.

Case 1.1: If $\kappa \neq v_i$ for any i, $\operatorname{supp}(\omega)$ and $\operatorname{supp}(\eta')$ are disjoint so all coefficients of $\tilde{\omega}$ are still ± 1. Moreover, since κ is distinct from the vertices of ρ, replacing $(v_0, \ldots, v_3)$ with (v_0, κ, v_3) certainly yields an undirected cycle in G.

Case 1.2: If $\kappa = v_i$ for some $i \in \{0, \ldots, k-1\}$ then there are number of possible sub-cases. If $\operatorname{supp}(\omega) \cap \operatorname{supp}(\eta') = \emptyset$ then all coefficients of $\tilde{\omega}$ are still ± 1. However, the replacement procedure has the effect of pinching $\operatorname{supp}(\tilde{\omega})$ into two edge-disjoint, undirected cycles, which share a vertex at κ. Hence, we can easily decompose $\tilde{\omega}$ into a sum of two fundamental cycles $\tilde{\omega}_1$ and $\tilde{\omega}_2$, supported on each of these underlying cycles.

If the intersection is non-empty, then $\operatorname{supp}(\omega) \cap \operatorname{supp}(\eta') \subseteq \{\tau_4, \tau_k\}$, so there are at most two offending edges. Moreover, in order to attain $\partial_1(\eta - \eta') = 0$ these edges must appear with opposite signs in ω and η' respectively. If there are two offending edges then we must have $4 = k - 1$ and the replacements procedure yields $\tilde{\omega} = 0$. If there is only one offending edge then this edge is no longer contained in $\operatorname{supp}(\tilde{\omega})$ and the length of the underlying undirected cycle is further reduced.

Case 2: If the path was reducible, in most cases $\operatorname{supp}(\omega)$ and $\operatorname{supp}(\eta')$ are disjoint and the replacement process simply removes one or two vertices from the undirected cycle. The only remaining case is if $k = 4$ and $\operatorname{supp}(\omega) \cap \operatorname{supp}(\eta') = \{\tau_k\}$, in which case the replacement procedure yields $\tilde{\omega} = 0$. □

Once we have reduced large cycles into smaller ones, we need conditions to ensure that the resulting small cycles are themselves homologous to zero.

Lemma 4.12 *Given a fundamental cycle $\omega \in \ker \partial_1$ such that $\operatorname{supp}(\omega)$ is a directed cycle of length k, if $\operatorname{supp}(\omega)$ has a cycle centre $\kappa \in V$ then*

$$\omega = 0 \pmod{\operatorname{im} \partial_2}. \tag{4.30}$$

Proof For some vertices $v_0, \dots, v_{k-1} \in V$ and edges $\tau_1, \dots, \tau_k \in E$ we can write the underlying cycle as

$$\rho = (v_0, \tau_1, \dots, v_{k-1}, \tau_k, v_0) \tag{4.31}$$

so that

$$\omega = \pm \sum_{i=0}^{k-1} e_{\tau_i}. \tag{4.32}$$

Since κ is a cycle centre, either $\gamma_i := e_{\kappa v_i v_{i+1}} \in \Omega_2$ for every i or $\gamma_i := e_{v_i v_{i+1} \kappa} \in \Omega_2$ for every i (identifying $v_k = v_0$). In either case, by a telescoping sum argument,

$$\partial_1 \left(\sum_{i=0}^{k-1} \gamma_i \right) = \sum_{i=0}^{k-1} e_{\tau_i}. \tag{4.33}$$

After adjusting for a factor of ±1, this concludes the proof. □

Piecing these lemmas together, gives us a topological condition, which implies $\vec{\beta}_1(G) = 0$, and which is likely to occur in high density graphs.

Proposition 4.13 *For any simple digraph G, if every irreducible, undirected path of length 3 has a directed centre, and every directed cycle of length 2 or 3 has a cycle centre, then $\vec{\beta}_1(G) = 0$ and $\vec{\beta}_1^{\mathcal{R}}(G) = 0$.*

Proof We prove the non-regular case from which the regular case immediately follows by Corollary 3.16.

Fix a basis $\{\omega_1, \dots, \omega_k\}$ of fundamental cycles for $\ker \partial_1$, as described by Lemma 4.8. Choose an arbitrary $i \in \{1, \dots, k\}$ It suffices to show that $\omega_i = 0 \pmod{\operatorname{im} \partial_2}$. By Lemma 4.11, we can reduce each ω_i to a sum of fundamental cycles

$$\omega_i = \sum_{j=1}^{k_i} \tilde{\omega}_{i,j} \pmod{\operatorname{im} \partial_2} \tag{4.34}$$

with $\# \operatorname{supp} \tilde{\omega}_{i,j} \leq 3$ for each j.

If $\operatorname{supp}(\tilde{\omega}_{i,j})$ is a directed triangle then $\tilde{\omega}_{i,j}$ is the boundary of the corresponding basis element in Ω_2; hence $\tilde{\omega}_{i,j} = 0 \pmod{\operatorname{im} \partial_2}$. Otherwise, $\operatorname{supp}(\tilde{\omega}_{i,j})$ must be a directed cycle or length 2 or 3. In either case, the support has a cycle centre and hence, by Lemma 4.12, $\tilde{\omega}_{i,j} = 0 \pmod{\operatorname{im} \partial_2}$. Therefore $\omega_i = 0 \pmod{\operatorname{im} \partial_2}$. □

Remark 4.14 Every double edge $(i, j), (j, i) \in E$ appears as an allowed 2-path in $\Omega_2^{\mathcal{R}}$ and $\partial_1^{\mathcal{R}}(e_{iji}) = e_{ji} + e_{ij}$. Therefore, the requirement that every cycle of length 2 has a cycle centre is not strictly necessary to ensure $\overrightarrow{\beta}_1^{\mathcal{R}}(G) = 0$.

Definition 4.15 For each $n \in \mathbb{N}$, the *complete directed graph* on n-nodes, K_n, is defined by

$$V(K_n) := \{1, 2, \dots, n\}, \quad E(K_n) := \{(i, j) \mid i \neq j\}. \tag{4.35}$$

Moreover, we define the following collection of subgraphs contained within K_n:

1. $P_3^n := \{\text{subgraphs } \sigma \subseteq K_n \mid \sigma \text{ is an undirected path of length } 3\}$;
2. For each $k \geq 2$, $C_k^n := \{\text{subgraphs } \sigma \subseteq K_n \mid \sigma \text{ is a directed cycle of length } k\}$.

Given a random graph $G \sim G(n, p)$, we define the following events:

1. for $\sigma \in P_3^n$ or $\sigma \in C_k^n$ for some $k \geq 2$, S_σ is the event that σ is a subgraph of G;
2. for $\sigma \in P_3^n$, I_σ is the event that σ is irreducible in the graph $G \cup \sigma$;
3. for $\sigma \in P_3^n$, $A_{\sigma,\kappa}$ is the event that κ is a directed centre for σ in the graph $G \cup \sigma$;
4. for $\sigma \in C_k^n$ for some $k \geq 2$, $B_{\sigma,\kappa}$ is the event that κ is a cycle centre for σ in the graph $G \cup \sigma$.

Remark 4.16 For a fixed $\sigma \in P_3^n$, the events S_σ, I_σ and $A_{\sigma,\kappa}$ for every $\kappa \in V(G) \backslash V(\sigma)$ are mutually independent. For a fixed $\sigma \in C_k^n$ for some $k \geq 2$, the events S_σ and $B_{\sigma,\kappa}$ for every $\kappa \in V(G) \setminus V(\sigma)$ are mutually independent.

Proposition 4.17 *If* $G \sim \overrightarrow{G}(n, p)$, *where* $p = p(n) = \omega\left((n/\log(n))^{-1/3}\right)$, *then*

$$\lim_{n\to\infty} \mathbb{P}(\overrightarrow{\beta}_1(G) = 0) = \lim_{n\to\infty} \mathbb{P}(\overrightarrow{\beta}_1^{\mathcal{R}}(G) = 0) = 1. \tag{4.36}$$

Proof By Proposition 4.13, it suffices to show that the probability that there exists an irreducible, undirected path of length 3 without directed centre, or a cycle of length

2 or 3 without directed centre, tends to 0 as $n \to \infty$. The probability that there is an irreducible, undirected path of length 3 without a directed centre is at most

$$\sum_{\sigma \in P_3^n} \left(\mathbb{P}[S_\sigma] \cdot \mathbb{P}[I_\sigma] \cdot \mathbb{P}\left[\bigcap_{\kappa \in V(G) \setminus V(\sigma)} A_{\sigma,\kappa}{}^c \right] \right). \tag{4.37}$$

Because the events $\left\{A_{\sigma,\kappa} \mid \kappa \in V(G) \setminus V(\sigma)\right\}$ are independent, $\mathbb{P}\left[\cap_\kappa A_{\sigma,\kappa}{}^c\right] = \prod_\kappa (1 - \mathbb{P}\left[A_{\sigma,\kappa}\right])$. Note that $\# P_3^n = \binom{n}{4} 4!\, 2^2$. This count arises because an undirected path of length 3 is determined by a choice of 4 nodes, an order on the nodes and a choice of orientation on each edge. However, this counts each path twice: once in each direction. Also, each path arises in G with probability $\mathbb{P}[S_\sigma] = p^3$ and clearly $\mathbb{P}[I_\sigma] \leq 1$.

For each $\sigma \in P_3^n$ and $\kappa \in V(G) \setminus V(\sigma)$, there is at least one choice of 3 directed edges, from κ to the vertices of the path, which forms a directed centre. Namely, label the vertices of σ by $(v_0, \ldots, v_3)$. Then we can always choose an edge between κ and v_0 and another edge between κ and v_2 so that there is a long square on $\{\kappa, v_0, v_1, v_2\}$, as illustrated in Fig. 4. The third edge can then be chosen to ensure that there is a directed triangle on $\{\kappa, v_2, v_3\}$. If these three edges are present in G, they constitute $J' \subseteq J_{\sigma,\kappa}$ with the properties required to form a directed centre and hence $\mathbb{P}\left[A_{\sigma,\kappa}\right] \geq p^3$. Therefore we can bound the probability (4.37) further by

$$\binom{n}{4} 4!\, 2^2 p^3 \left[1 - p^3\right]^{n-4} \leq 4n^4 p^3 \exp\left(-p^3(n-4)\right). \tag{4.38}$$

We wish to show that this bound tends to 0 as $n \to \infty$. Since $p \leq 1$, it suffices to show $\lim_{n\to\infty} n^4 \exp(-p^3 n) = 0$. By Lemma 4.18, the condition on p ensures that $\lim_{n\to\infty}(4\log(n) - p^3 n) = -\infty$. By the continuity of the exponential function, $\lim_{n\to\infty} n^4 \exp(-p^3 n) = 0$.

Note $\# C_2^n = \binom{n}{2}$ and $\# C_3^n = 2\binom{n}{3}$. By another union bound, we see that the probability that there is a directed cycle, of length 2, without cycle centre, is at most

$$\sum_{\sigma \in C_2^n} \left(\mathbb{P}[S_\sigma] \cdot \prod_{\kappa \in V(G) \setminus V(\sigma)} \mathbb{P}\left[B_{\sigma,\kappa}{}^c\right] \right) = \binom{n}{2} p^2 [1 - p^2]^{2n-4}. \tag{4.39}$$

Similarly, the probability that there is a directed cycle, of length 3, without cycle centre is at most

$$2\binom{n}{3} p^3 [1 - p^3]^{2n-6}. \tag{4.40}$$

Again, by Lemma 4.18, the condition on p suffices to ensure that these two bounds also tend to 0 as $n \to \infty$. □

Lemma 4.18 *Given $k \in \mathbb{N}_{>0}$ and $A, B > 0$, if $p = \omega\left((n/\log(n))^{-1/k}\right)$ then*

$$\lim_{n\to\infty} (A\log(n) - Bnp^k) = -\infty. \tag{4.41}$$

Proof The condition on p is equivalent to $\lim_{n\to\infty} \frac{np^k}{\log(n)} = \infty$, which implies $(A\frac{\log(n)}{np^k} - B) \to -B$ and $np^k \to \infty$ as $n \to \infty$. □

Proof of Theorem 1.4(4) Assume that $p(n) = n^{\alpha}$. If $\alpha > -1/3$ then an application of L'Hôpital's rule shows $p = \omega\left((n/\log(n))^{-1/3}\right)$ and hence, by Proposition 4.17, $\mathbb{P}(\overrightarrow{\beta}_1(G) = 0) \to 1$ and $\mathbb{P}(\overrightarrow{\beta}_1^{\mathcal{R}}(G) = 0) \to 1$ as $n \to \infty$. □

Remark 4.19 Lemma 4.18 reveals the origin of the ratio 1/3, which appears in Theorem 1.4(4) and Proposition 4.17. In particular, it arises as the ratio between the power of n and the power of p inside the exponential of equation (4.38). The power of n is 1 because there are on the order of n^1 possible directed centres for an undirected path of length 3. The power of p is 3 because we require at least 3 edges from κ to the path, in order for κ to form a directed centre. In Lemma A.7, we will see that this is indeed the minimal number of edges required to form a directed centre.

The bounds used in the proof of Proposition 4.17 are by no means the best possible. Indeed, by splitting P_3^n into four isomorphism classes, it is possible to get exact values for $\mathbb{P}[I_\sigma]$ and $\mathbb{P}\left[A_{\sigma,\kappa}\right]$. We explore this further in Appendix 1 in order to obtain tighter bounds, useful for hypothesis testing.

Moreover, the topological condition for $\overrightarrow{\beta}_1(G) = 0$ presented in Proposition 4.13 was chosen since it is likely to occur at high densities. However, there may (and indeed probably does) exist weaker topological conditions which imply $\overrightarrow{\beta}_1(G) = 0$ and occur at somewhat lower densities. This could potentially allow for a weaker hypothesis on Proposition 4.17. In order to conjecture the weakest possible hypothesis, we conduct a number of experiments in Appendix B.

5 Directed flag complex of random directed graphs

For comparative purposes, we now apply the techniques of Sect. 4 to the directed flag complex, which features more readily in the literature.

Definition 5.1 (Lütgehetmann et al. 2020, Definition 2.2) An *ordered simplicial complex on a vertex set V* is a collection of ordered subsets of V, which is closed under taking non-empty, ordered subsets (with the induced order). A subset in the collection consisting of $(k + 1)$ vertices is called a *k-simplex*.

Definition 5.2 (Lütgehetmann et al. 2020, Definition 2.3) Given a directed graph $G = (V, E)$,

1. a *directed $(k + 1)$-clique* is a $(k + 1)$-tuple of distinct vertices $(v_0, \dots, v_k)$ such that $(v_i, v_j) \in E$ whenever $i < j$;

2. the *directed flag complex,* $\overrightarrow{X}(G)$, (often denoted $\mathrm{dFl}(G)$) is an ordered simplicial complex, whose k-simplicies are the directed $(k+1)$-cliques.

Given a ring R, the *directed flag chain complex* is $\{\overrightarrow{X}_k(G), \partial_k\}_{k\geq -1}$ where, for $k \geq 0$,

$$\overrightarrow{X}_k(G) := \overrightarrow{X}_k(G; R) := R\langle\{(v_0, \dots, v_k) \mid \text{directed } (k+1)\text{-clique in } G\}\rangle\,, \quad (5.1)$$

$$\partial_k(e_{(v_0,\dots,v_k)}) := \sum_{i=0}^{k} (-1)^i e_{(v_0,\dots,\hat{v_i},\dots,v_k)}. \quad (5.2)$$

where $(v_0, \dots, \hat{v_i}, \dots, v_k)$ denotes the directed k-clique $(v_0, \dots, v_k)$ with the vertex v_i removed. This defines ∂_k on a basis of $\overrightarrow{X}_k(G)$, from which we extend linearly. We also define $\overrightarrow{X}_{-1}(G) = R$ and ∂_0 simply sums the coefficients in the standard basis, as in equation (3.5).

The homology of this chain complex is the *directed flag complex homology*. The Betti numbers are denoted $\beta_k(\overrightarrow{X}(G))$. When the coefficient ring R is omitted from notation, assume $R = \mathbb{Z}$.

Firstly, as with path homology, $\beta_0(\overrightarrow{X}(G))$ captures the weak connectivity of a digraph G and hence Theorem 1.3 also holds for the directed flag complex. Next, since we have an explicit list of generators for $\overrightarrow{X}_k(G)$, and they are easy to count, we can calculate the expected rank of the chain groups in every dimension. As before, we denote the ranks of chain groups as $N_k := \operatorname{rank} \overrightarrow{X}_k(G)$ and use these to estimate the Betti numbers.

Lemma 5.3 *For an Erdős–Rényi directed random graph $G \sim \overrightarrow{G}(n, p)$, for any $k \geq 0$ we have*

$$\mathbb{E}[N_k] = \binom{n}{k+1}(k+1)!\, p^{\binom{k+1}{2}}. \quad (5.3)$$

Proof A possible directed clique is uniquely determined by an ordered $(k+1)$-tuple of distinct vertices. Therefore, there are $\binom{n}{k+1}(k+1)!$ possible cliques. For a given clique to be present, one edge must be present in G for every pair of distinct nodes. □

Using the Morse inequalities as before, this allows us to compute the growth rate of the expected Betti numbers, under suitable conditions on $p = p(n)$.

Proposition 5.4 *For $k \geq 0$, if $G \sim \overrightarrow{G}(n, p)$ where $p = p(n)$, with $p(n) = \omega(n^{-1/k})$ and $p(n) = o(n^{-1/(k+1)})$, then*

$$\mathbb{E}\big[-N_{k-1} + N_k - N_{k+1}\big] \sim \mathbb{E}[N_k] \quad \textit{and hence} \quad \mathbb{E}[\beta_k(\overrightarrow{X}(G))] \sim \mathbb{E}[N_k]\,. \quad (5.4)$$

Proof It is easy to check that

$$\frac{\mathbb{E}[N_{k-1}]}{\mathbb{E}[N_k]} = \frac{1}{(n-k)} p^{-k} \sim \frac{1}{np^k} \quad (5.5)$$

which tends to 0 thanks to the condition $p(n) = \omega(n^{-1/k})$. It follows that $\mathbb{E}\big[N_{k+1}\big]/\mathbb{E}[n_k] \sim np^{k+1}$ which tends to 0 thanks to the condition $p(n) = o(n^{-1/(k+1)})$. As in Proposition 4.2, the result then follows by the Morse inequalities. □

Proposition 5.5 *If $G \sim \overrightarrow{G}(n, p)$ where $p = p(n)$, with $p(n) = \omega(n^{-1})$ and $p(n) = o(n^{-1/2})$, then $\beta_1(\overrightarrow{X}(G)) \sim \mathbb{E}[\beta_1(\overrightarrow{X}(G)]$ with high probability.*

Proof The proof is identical to Proposition 4.4 except we only need $p(n) = o(n^{-1/2})$ to ensure $\mathbb{E}[-N_0 + N_1 - N_2] \sim \mathbb{E}[N_1]$. □

Corollary 5.6 *If $G \sim \overrightarrow{G}(n, p)$ where $p = p(n)$, with $p(n) = \omega(n^{-1})$ and $p(n) = o(n^{-1/2})$, then*

$$\lim_{n\to\infty} \mathbb{P}(\beta_1(\overrightarrow{X}(G)) > 0) = 1. \tag{5.6}$$

Proof The proof is identical to Corollary 4.5 expect again we only require $p(n) = o(n^{-1/2})$ to invoke Proposition 5.5. □

As with path homology, degree 1 homology appears in the directed flag complex with the appearance of undirected cycles in the underlying digraph. Therefore, the same conditions show that $\beta_1(\overrightarrow{X}(G)) = 0$ with high probability, when $p = p(n)$ shrinks too quickly.

Proposition 5.7 *If $p = p(n) = o(n^{-1})$ then, given a random directed graph $G \sim \overrightarrow{G}(n, p)$, we have*

$$\lim_{n\to\infty} \mathbb{P}(\beta_1(\overrightarrow{X}(G)) = 0) = 1. \tag{5.7}$$

Proof The proof is identical to the non-regular case of Proposition 4.6. □

The techniques from Sect. 4.4 can be applied, mutatis mutandis, to show $\beta_1(G) = 0$ with high probability, when $p = p(n)$ shrinks too slowly.

Proposition 5.8 *If $G \sim \overrightarrow{G}(n, p)$, where $p = p(n) = \omega\left((n/\log(n))^{-1/4}\right)$, then*

$$\lim_{n\to\infty} \mathbb{P}(\beta_1(\overrightarrow{X}(G)) = 0) = 1. \tag{5.8}$$

Proof This follows from the same argument as Proposition 4.17. The only difference is that a directed centre for an undirected path of length 3 requires at least 4 edges, in order to form 3 directed cliques. This results in the ratio 1/4 instead of 1/3. □

As in the earlier sections, the results obtain here suffice to prove the summary result, Theorem 1.6, presented in the introduction.

6 Discussion

We have identified asymptotic conditions on $p = p(n)$ which ensure that a random directed graph $G \sim \overrightarrow{G}(n, p)$ has $\overrightarrow{\beta}_1(G) > 0$ with high probability. Moreover, under these conditions we showed that $\mathbb{E}[\overrightarrow{\beta}_1(G)] \sim n(n-1)p$. Beneath the lower boundary of this positive region, we showed that $\overrightarrow{\beta}_1(G) = 0$ with high probability. Immediately after the upper boundary of the positive $\overrightarrow{\beta}_1$ range, our theory is inconclusive, but experimental results (shown in Appendix B) provide evidence that $\overrightarrow{\beta}_1(G) = 0$ with high probability. Further away from the positive region, e.g. when $p = n^\alpha$ for $\alpha > -1/3$, our theory again guarantees that $\overrightarrow{\beta}_1(G) = 0$ with high probability. For comparison, we applied these techniques to the directed flag complex and found similar results, with minor changes to the gradient of the boundary lines. We summarise these results, along with similar results for 'symmetric methods' in Table 1.

These results, along with known results for the clique complex of a random undirected graph, motivate the following research directions:

1. **Tighter upper boundary** – Experimental results, in Appendix 1, indicate that the condition $\alpha > -1/3$ of Theorem 1.4 could be weakened significantly, potentially as far as $\alpha > -2/3$. Indeed this is the best possible result because $\overrightarrow{\beta}_1 > 0$ with high probability for $-1 < \alpha < -2/3$. We saw how the ratio $1/3$ arose from our proof in Remark 4.19, which informs how we might improve this result. Generalising a directed centre to be a set of $k > 1$ nodes, with suitable conditions, might yield better results since there are on the order of n^k set of k nodes. However, this would complicate the probability bound because two directed centres would no longer be

Table 1 Given a random directed graphs $G \sim \overrightarrow{G}(n, p)$, assuming $p = n^\alpha$, we record the known regions of α for which various homologies are either positive, or zero, with high probability

Homology	Expected growth	Positive region (α)	Zero region (α)
Symmetric methods			
$\beta_1(\mathring{G})$	$n(n-1)p$	$(-1, 0]$	$(-\infty, -1)$
$\beta_1(\bar{G})$	$\binom{n}{2}\bar{p}$	$(-1, 0]$	$(-\infty, -1)$
$\beta_1(\mathrm{X}(\bar{G}))$	$\binom{n}{2}\bar{p}$	$(-1, -1/2)$	$(-\infty, -1) \cup (-1/2, 0]$
$\beta_k(\mathrm{X}(\bar{G})), k > 1$	$\binom{n}{k+1}\bar{p}^{\binom{k+1}{2}}$	$\left(-\frac{1}{k}, -\frac{1}{k+1}\right)$	$\left(-\infty, -\frac{1}{k}\right) \cup \left(-\frac{1}{k+1}, 0\right]$
Directed methods			
$\beta_1(\overrightarrow{X}(G))$	$n(n-1)p$	$(-1, -1/2)$	$(-\infty, -1) \cup (-1/4, 0]$
$\overrightarrow{\beta}_1(G)$	$n(n-1)p$	$(-1, -2/3)$	$(-\infty, -1) \cup (-1/3, 0]$
$\overrightarrow{\beta}_1^{\mathcal{R}}(G)$	$n(n-1)p$	$(-1, -2/3)$	$(-\infty, -1) \cup (-1/3, 0]$
$\overrightarrow{\beta}_k(G), k > 1$	**?**	**?**	$\left(-\infty, \frac{k+1}{k}\right)$
$\overrightarrow{\beta}_k^{\mathcal{R}}(G), k > 1$	**?**	**?**	$\left(-\infty, \frac{k+1}{k}\right)$

Moreover, we describe the growth rate of the expected Betti numbers, in the respective positive regions. That is, for Betti number β, we give a function $f(n)$ such that, when α is in the positive region, $\mathbb{E}[\beta(G)] \sim f(n)$

independent. Alternative approaches may include finding a cover of the random graph which can be used to show that the digraph is contractible via path homotopy Grigor'yan et al. (2014).

2. **Higher degrees** – So far we only have weak guarantees for the behaviour of $\overrightarrow{\beta}_k$ for $k > 1$. One potential avenue for improvement is to find conditions under which $\mathbb{P}[\Omega_k = \{0\}] \to 1$ as $n \to \infty$. In order to get better results for vanishing $\overrightarrow{\beta}_k$ at small p, we require a greater understanding of generators of $\ker \partial_k$. In order to get better results at large p, we need a high-density, topological condition which implies that these generators can be reduced to $0 \pmod{\operatorname{im} \partial_{k+1}}$.
3. **Distributional results**—One direction of research is to show that normalised $\overrightarrow{\beta}_1$ converges to a normal distribution as $n \to \infty$. More evidence for this conjecture is given in Appendix 1. This could be done, for example, by Stein's method Chen et al. (2011), as is done by Kahle and Meckes (2013) for $\beta_1(\mathrm{X}(G(n, p)))$.

Acknowledgements The author would like to thank his supervisors Ulrike Tillmann and Heather Harrington for their support and guidance throughout this project. The author would also like to thank Gesine Reinert and Vidit Nanda for their helpful feedback on a prior draft. Finally, the author would like to thank both reviewers for their helpful comments which both helped the presentation and strengthened the result of Theorem 1.41. The author is a member of the Centre for Topological Data Analysis, which is funded by the EPSRC grant 'New Approaches to Data Science: Application Driven Topological Data Analysis' EP/R018472/1.

Data availability The code and data for the experiments of Appendix B, as well as an implementation of the algorithm described in Lemma A.7, are available on both the OSF repository Chaplin (2022) and GitHub Chaplin (2021). All code is written in `MATLAB` and data from the experiments is available in the `.mat` format.

Appendix A: Explicit probability bounds

The results presented in Sect. 4 provide a broad-stroke, qualitative description of the behaviour of $\overrightarrow{\beta}_1$ on random directed graphs. However, they provide no guarantees for a digraph of a fixed size, such as those arising in applications. For hypothesis testing, it is desirable to have explicit bounds on the $\mathbb{P}[\overrightarrow{\beta}_1(G) > 0]$ and $\mathbb{P}[\overrightarrow{\beta}_1(G) = 0]$ for $G \sim \overrightarrow{G}(n, p)$, given n and p. In this section, we describe such bounds arising from the proofs in Sect. 4 and improve on them where possible.

A.1 Positive Betti numbers at low densities

First we refine the bound developed in Proposition 4.6, in order to show that it is unlikely to observe $\vec{\beta}_1(G) > 0$ when graph density is low.

Theorem A.1 *If $G \sim \vec{G}(n, p)$ then*

$$\mathbb{P}\Big[\vec{\beta}_1(G) > 0\Big] \leq \sum_{L=2}^{n} \binom{n}{L} \frac{L!}{2L} (2p)^L. \tag{A1}$$

The same bound holds for $\mathbb{P}[\beta_1(\vec{X}(G)) > 0]$. The same bound holds for $\mathbb{P}[\vec{\beta}_1^{\mathcal{R}}(G) > 0]$ after removing the $L = 2$ term.

Proof We start with the non-regular and directed flag complex case. We follow the same argument as the proof of Proposition 4.6 but make more accurate estimates. A sufficient condition for both $\vec{\beta}_1(G) = 0$ and $\beta_1(\vec{X}(G))$ is that there are no undirected cycles of any length $2 \leq L \leq n$ in G. For each L, there are

$$\binom{n}{L} \frac{L!}{2L} 2^L \tag{A2}$$

possible undirected cycles. This is because an undirected cycle can be determined by a choice of L vertices, an order on those vertices, and a choice of orientation for each edge. However, this over-counts, by a factor of $2L$, since we could traverse the cycle in either direction and start at any vertex. Each cycle of length L appears with probability p^L so a union bound yields the result.

For regular path homology, the only undirected cycles of length 2 are double edges, which are boundaries in the regular path complex. Therefore we can remove the $L = 2$ term from the bound. □

The region of parameter space in which this theorem applies is illustrated in Fig. 6a. For each n, we plot $p_l^t(n)$, the maximum value such that for all $p \leq p_l^t(n)$, Proposition A.1 implies that $\mathbb{P}[\vec{\beta}_1(G) > 0] \leq 0.05$.

A.2 Zero Betti numbers

In order to obtain the best possible estimate, following the second moment method of Corollary 4.5, we need an exact value for $\mathbb{E}[\text{rank}\ \Omega_2]$. We reproduce the approach of (Grigor'yan et al. (2012), Proposition 4.2), making the necessary alterations for the non-regular case. The approach is to determine the number of linearly independent conditions required to describe Ω_2 as a subspace of $\mathcal{A}_2$.

Definition A.2 Given a directed graph $G = (V, E)$,

1. a *semi-edge* is an ordered pair of distinct vertices $(i, k) \in V^2$, $i \neq k$ such that $i \not\to k$ but there is some other vertex $j \in V$, $j \neq i, k$ such that $i \to j \to k$;

2. a *semi-vertex* is a vertex $i \in V$ such that there is some other vertex $j \in V$, $j \neq i$ such that $i \to j \to i$.

We denote the set of all semi-edges by $\mathcal{S}_E$ and all semi-vertices by $\mathcal{S}_V$.

Lemma A.3 *If* $G \sim \overrightarrow{G}(n, p)$ *then*

$$\begin{aligned}\mathbb{E}[\operatorname{rank}\Omega_2(G;\mathbb{Z})] &= n(n-1)^2p^2 - n(n-1)(1-p)\left[1-(1-p^2)^{n-2}\right]\\ &\quad - n\left[1-(1-p^2)^{n-1}\right] \qquad \text{(A3)}\\ \mathbb{E}\left[\operatorname{rank}\Omega_2^{\mathcal{R}}(G;\mathbb{Z})\right] &= n(n-1)^2p^2 - n(n-1)(1-p)\left[1-(1-p^2)^{n-2}\right] \qquad \text{(A4)}\end{aligned}$$

Proof First we deal with the non-regular case. Given $v \in \mathcal{A}_2$, then $v \in \Omega_2$ if and only if $\partial v \in \mathcal{A}_1$. Let A_2 denote the set of all allowed 2-paths in G, then we can write

$$v = \sum_{ijk\in A_2} v^{ijk} e_{ijk} \tag{A5}$$

so that

$$\partial v = \sum_{ijk\in A_2} v^{ijk}(e_{jk} - e_{ik} + e_{ij}). \tag{A6}$$

Since ijk is allowed, so too are ij and jk. Therefore

$$\partial v = -\sum_{ijk\in A_2} v^{ijk} e_{ik} \pmod{\mathcal{A}_1}. \tag{A7}$$

Now we split off terms corresponding to double edges

$$\partial v = -\sum_{\substack{ijk\in A_2\\ i\neq k}} v^{ijk} e_{ik} - \sum_{iji\in A_2} v^{iji} e_{ii} \pmod{\mathcal{A}_1}. \tag{A8}$$

Note ii is never an allowed 1-path. However, for $i \neq k$, ik is an allowed 1-path if $i \to k$, so we can remove these summands

$$\partial v = -\sum_{\substack{ijk\in A_2\\ i\neq k, i\not\to k}} v^{ijk} e_{ik} - \sum_{iji\in A_2} v^{iji} e_{ii} \pmod{\mathcal{A}_1}. \tag{A9}$$

Therefore, $\partial v \in \mathcal{A}_1$ if and only if for each $(i, k) \in V^2$ with $i \neq k$ and $i \not\to k$

$$\sum_{j:\, ijk\in A_2} v^{ijk} = 0 \tag{A10}$$

and for each $i \in V$

$$\sum_{j:\, iji \in A_2} v^{iji} = 0. \tag{A11}$$

Some of the indexing sets of these summations may be empty and hence some of these conditions may be trivial. The remaining conditions are linearly independent and hence it remains to count the number of non-trivial equations. Equation (A10) is non-trivial if and only if (i, k) is a semi-edge and equation (A11) is non-trivial if and only if i is a semi-vertex. Therefore

$$\operatorname{rank} \Omega_2 = \operatorname{rank} \mathcal{A}_2 - \#\mathcal{S}_E - \#\mathcal{S}_V. \tag{A12}$$

Taking expectations

$$\mathbb{E}[\operatorname{rank} \mathcal{A}_2] = n(n-1)^2 p^2, \tag{A13}$$

$$\mathbb{E}[\#\mathcal{S}_E] = n(n-1)(1-p)\left[1 - (1-p^2)^{n-2}\right], \tag{A14}$$

$$\mathbb{E}[\#\mathcal{S}_V] = n\left[1 - (1-p^2)^{n-1}\right] \tag{A15}$$

which concludes the non-regular case.

For the regular case, note that equation (A9) becomes

$$\partial^{\mathcal{R}} v = -\sum_{\substack{ijk \in A_2 \\ i \neq k,\, i \not\to k}} v^{ijk} e_{ik} \quad (\operatorname{mod} \mathcal{A}_1). \tag{A16}$$

since the e_{ii} terms are removed by the projection. Hence all the semi-vertex conditions of equation (A11) are removed. □

Theorem A.4 *If* $G \sim \overrightarrow{G}(n, p)$ *then*

$$\mathbb{P}\left[\overrightarrow{\beta}_1(G) > 0\right] \geq \frac{\max\left(0, -n + n(n-1)p - \mathbb{E}[N_2]\right)^2}{n(n-1)p(1-p) + n^2(n-1)^2 p^2} \tag{A17}$$

where

$$\mathbb{E}[N_2] = n(n-1)^2 p^2 - n(n-1)(1-p)\left[1 - (1-p^2)^{n-2}\right] \\ -n\left[1 - (1-p^2)^{n-1}\right]. \tag{A18}$$

Proof In theory, we could track back the bound from the theorems invoked by Corollary 4.5. Instead we use a more direct bound. Since $\overrightarrow{\beta}_1(G)$ is a non-negative random variable, an application of the Cauchy-Schwarz inequality to $\mathbb{E}[\overrightarrow{\beta}_1 \mathbb{1}_{\overrightarrow{\beta}_1 > 0}]$ gives

$$\mathbb{P}[\overrightarrow{\beta}_1(G) > 0] \geq \frac{\mathbb{E}[\overrightarrow{\beta}_1(G)]^2}{\mathbb{E}[\overrightarrow{\beta}_1(G)^2]}. \tag{A19}$$

Employing the Morse inequalities again, we see

$$\mathbb{P}\left[\overrightarrow{\beta}_1(G) > 0\right] \geq \frac{\max(0, \mathbb{E}[-N_0 + N_1 - N_2])^2}{\mathbb{E}\left[N_1^2\right]} \tag{A20}$$

where $N_k := \operatorname{rank} \Omega_k(G; \mathbb{Z})$. We obtain the numerator using the expectations for N_0 and N_1 from Lemma 4.1 and the expectation of N_2 from Lemma A.3. Then N_1 is a binomial random variable on $n(n-1)$ trials, each with independent probability p, and hence the second moment is

$$\mathbb{E}\left[N_1^2\right] = n(n-1)p(1-p) + n^2(n-1)^2 p^2 \tag{A21}$$

which concludes the proof. □

Remark A.5 Letting N_k denote the rank of the kth chain group in each of the respective chain complexes, it is quick to see that N_1 is the same random variable across all complexes. In order to obtain an analogous theorem to bound $\mathbb{P}[\overrightarrow{\beta}_1^{\mathcal{R}}(G) > 0]$ one need simply replace $\mathbb{E}[N_2]$ with the computation of $\mathbb{E}\left[\operatorname{rank} \Omega_2^{\mathcal{R}}\right]$ from Lemma A.3. To obtain a result for the directed flag complex, one can use the expectations from Lemma 5.3.

Unfortunately, this bound is not useful for practical applications. In Fig. 6b we plot the minimum value of this bound over all $p \in [0, 1]$, for a range of n. Note that the bound does not reach a significance level of 0.1, at any p, until approximately $n = 7.4 \times 10^5$ and does not reach a significance level of 0.05 until approximately $n = 1.2 \times 10^7$. At these large graph sizes, computing $\overrightarrow{\beta}_1(G)$ is infeasible and hence this bound serves no practical use.

A.3 Positive Betti numbers at high densities

Finally, we tackle bounding $\mathbb{P}[\overrightarrow{\beta}_1(G) > 0]$ when graph density is large. In order to improve upon the naive bound available from Proposition 4.17, we provide more avenues for reducing long paths into shorter ones. Specifically, we will obtain better bounds on $\mathbb{P}[I_\sigma]$ and $\mathbb{P}\left[A_{\sigma,\kappa}{}^c\right]$. This will not effect the asymptotic behaviour of the bound, but may provide a substantially lower bound, at a fixed n and p.

In order to achieve this, we must partition P_3^n, the set of all possible undirected paths of length 3, on an n-node graph. Every undirected path of length 3 is uniquely determined by a choice of 4 distinct nodes in the graph, an ordering on these nodes (v_0, v_1, v_2, v_3), and a choice of orientation for the edges joining $\{v_0, v_1\}$ and $\{v_2, v_3\}$.

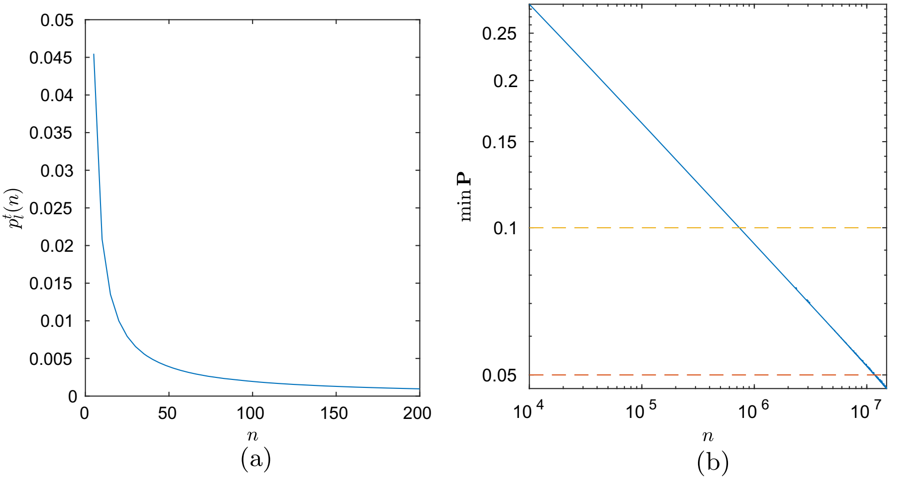


Fig. 6 **a** If (n, p) falls beneath the line $p_l^t(n)$ then Theorem A.1 implies $\mathbb{P}(\overrightarrow{\beta}_1(\overrightarrow{G}(n, p)) > 0) \leq 0.05$. Both axes are linearly scaled. **b** For $n \in [10^4, 1.5 \times 10^7]$, we plot the minimum value of the bound of Theorem A.4 for $p \in [0, 1]$. Both axes are logarithmically scaled

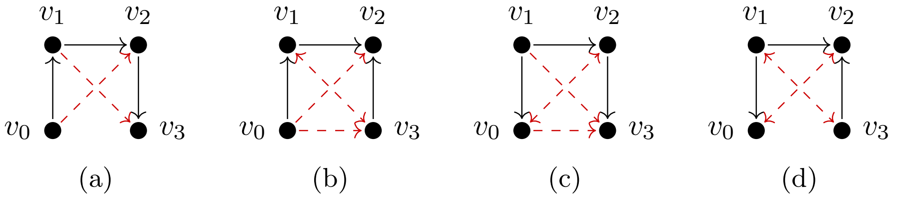


Fig. 7 The four possible '3-path motifs' which partition P_3^n, shown with black edges. From left to right we see the edge orientations for σ belonging to $P_{3,0}^n$, $P_{3,1}^n$, $P_{3,2}^n$ and $P_{3,3}^n$ respectively. Dashed, red edges indicate edges which must not be present for the path to be considered irreducible

The edge joining the middle two nodes is assumed to be 'forward', i.e. via (v_1, v_2), so as to avoid double counting each path. Therefore, we can partition P_3^n, based on the choices of edge orientations, into one of four classes, $P_{3,m}^n$ for $m \in \{0, 1, 2, 3\}$. These classes are visualised in Fig. 7.

Lemma A.6 *If* $\sigma \in P_{3,m}^n$ *then* $\mathbb{P}[I_\sigma] = (1 - p)^{c_m}$*, where* $\mathbf{c} = (c_m) = (2, 4, 4, 4)^\mathsf{T}$.

Proof The red, dashed edges shown in Fig. 7 identify 'shortcut' edges; if any one of these edges is present then σ is reducible. Moreover, these are the only shortcut edges, since the addition of any other edge would create a subgraph with $\overrightarrow{\beta}_1 > 0$. Hence, if $\sigma \in P_{3,m}^n$, it is irreducible if and only if none of these c_m edges are present. Since the existence of each edge is independent, the result follows. □

Lemma A.7 *If* $\sigma \in P_{3,m}^n$ *then*

$$\mathbb{P}[A_{\sigma,\kappa}] = \sum_{l=0}^{8} q_{m,l}\, p^l (1 - p)^{8-l}, \tag{A22}$$

where

$$Q = (q_{m,l}) = \begin{pmatrix} 0 & 0 & 0 & 2 & 11 & 22 & 23 & 8 & 1 \\ 0 & 0 & 0 & 4 & 19 & 33 & 25 & 8 & 1 \\ 0 & 0 & 0 & 4 & 19 & 33 & 25 & 8 & 1 \\ 0 & 0 & 0 & 2 & 16 & 34 & 26 & 8 & 1 \end{pmatrix}. \tag{A23}$$

Proof Assume that $\sigma \in P^n_{3,m}$ and $\kappa \in V(G) \setminus V(\sigma)$. Let $q_{m,l}$ denote the number of l-element subsets $J \subseteq \{(v,\kappa), (\kappa, v) \mid v \in V(\sigma)\}$ such that κ is a generic directed centre for σ in the graph $\sigma \cup J$. Note, $q_{m,l}$ is well-defined because, for a fixed m, all $\sigma \cup J$ are isomorphic for $\sigma \in P_{3,m}$ and $\kappa \in V(G) \setminus V(\sigma)$.

Then, conditioning on the cardinality of $J_{\sigma,k}$, we can write

$$\mathbb{P}[A_{\sigma,\kappa}] = \sum_{l=0}^{8} \mathbb{P}[A_{\sigma,\kappa} \mid \#J_{\sigma,\kappa} = l] \cdot \mathbb{P}(\#J_{\sigma,\kappa} = l) \tag{A24}$$

$$= \sum_{l=0}^{8} \frac{q_{m,l}}{\binom{8}{l}} \cdot \binom{8}{l} p^l (1-p)^{8-l}. \tag{A25}$$

Since $q_{m,l}$ depends only on m and l, we can compute an exact value which does not depend on σ or κ. This is done as follows:

1. For each $m = 0, 1, 2, 3$, initialise a graph G_m with nodes $V(G_m) = \{v_0, \ldots, v_3, \kappa\}$. The edge set $E(G_m)$ consists solely of an undirected path σ, of length 3, on the vertices $(v_0, \ldots, v_3)$, such that $\sigma \in P_{3,m}$ (the black, solid edges of Fig. 7).
2. Define the set of all possible linking edges

$$L := \{(v_i, \kappa), (\kappa, v_i) \mid i = 0, 1, 2, 3\}.$$

3. For each subset $J \subseteq L$, construct $G_{m,J} := (V(G_m), E(G_m) \cup J)$.
4. Set $\alpha_{m,J} := 1$ if $G_{m,J}$ contains an undirected path of length 2 from v_0 to v_3 and $\overrightarrow{\beta}_1(G_{m,J}) = 0$. Otherwise set $\alpha_{m,J} := 0$.
5. For each subset $J \subseteq L$, set $\gamma_{m,J} := 1$ if $\alpha_{m,J'} = 1$ for any $J' \subseteq J$. Otherwise set $\gamma_{m,J} = 0$.
6. Then

$$q_{m,l} = \sum_{\substack{J \subseteq L: \\ \#J = l}} \gamma_{m,J}.$$

This algorithm was implemented as a `MATLAB` script and was subsequently used to compute the matrix Q given in the statement of the theorem. In step 4, Betti numbers are computed via the `pathhomology` package Yutin (2022), using the `symbolic` option. This uses `MATLAB`'s symbolic computational toolbox in order to avoid any numerical errors. Details on how to access the codebase are available in Sect. 1.3. □

Theorem A.8 *If* $G \sim \overrightarrow{G}(n, p)$ *then*

$$\mathbb{P}\left[\overrightarrow{\beta}_1(G) > 0\right] \leq \binom{n}{4} 4! \sum_{m=0}^{3} p^3 (1-p)^{c_m} \left(1 - \sum_{l=0}^{8} q_{m,l}\, p^l (1-p)^{8-l}\right)^{n-4} + \binom{n}{2} p^2 \left[1 - p^2\right]^{2n-4} + 2\binom{n}{3} p^3 \left[1 - p^3\right]^{2n-6}, \tag{A26}$$

where $\mathbf{c} = (c_m) = (2, 4, 4, 4)^{\mathsf{T}}$ *and*

$$Q = (q_{m,l}) = \begin{pmatrix} 0 & 0 & 0 & 2 & 11 & 22 & 23 & 8 & 1 \\ 0 & 0 & 0 & 4 & 19 & 33 & 25 & 8 & 1 \\ 0 & 0 & 0 & 4 & 19 & 33 & 25 & 8 & 1 \\ 0 & 0 & 0 & 2 & 16 & 34 & 26 & 8 & 1 \end{pmatrix}. \tag{A27}$$

Proof We follow the same approach as Proposition 4.17, but with tighter bounds. One can bound the probability that there is an irreducible, undirected 3-path, without a directed centre by

$$\sum_{m=0}^{3} \sum_{\sigma \in P^n_{3,m}} \left(\mathbb{P}[S_\sigma] \cdot \mathbb{P}[I_\sigma] \cdot \prod_{\kappa \in V(G) \setminus V(\sigma)} [1 - \mathbb{P}(A_{\sigma,\kappa})] \right). \tag{A28}$$

By Lemma A.6 and Lemma A.7 we can bound this further by

$$\binom{n}{4} 4! \sum_{m=0}^{3} p^3 (1-p)^{c_m} \left(1 - \sum_{l=0}^{8} q_{m,l}\, p^l (1-p)^{8-l}\right)^{n-4} \tag{A29}$$

since $\# P^n_{3,m} = \binom{n}{4} 4!$ for each m. The same bounds, from Proposition 4.17, apply to the probability that there is a directed cycle, of length 2 or 3, without cycle centre. Combining these bounds concludes the proof. □

Remark A.9 By Corollary 3.16 the bound identified in Theorem A.8 also holds for $\overrightarrow{\beta}^{\mathcal{R}}_1(G)$. However, as noted in Remark 4.14, we can obtain a stronger bound by removing the term

$$\binom{n}{2} \left[1 - p^2\right]^{2n-4} \tag{A30}$$

which corresponds to undirected cycles of length 2.

To demonstrate the utility of this theorem, in Fig. 8, at each n, we plot the minimum $p^t_u(n)$ such that the bounds from Proposition 4.17 (respectively Theorem A.8) imply $\mathbb{P}(\overrightarrow{\beta}_1(\overrightarrow{G}(n, p)) > 0) \leq 0.05$ for all $p \geq p^t_u(n)$. We compute $p^t_u(n)$ via MATLAB's

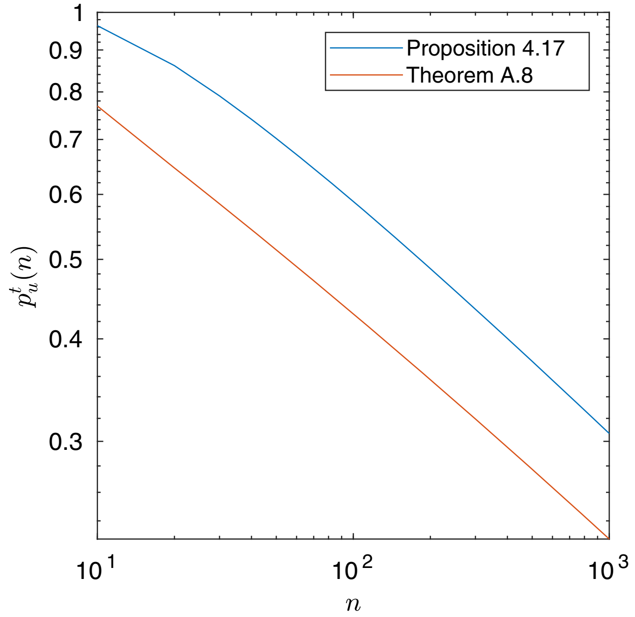


Fig. 8 Boundaries of the parameter regions in which Proposition 4.17 and Theorem A.8 apply. If (n, p) falls above a given line then the corresponding theorem implies $\mathbb{P}(\overrightarrow{\beta}_1(\overrightarrow{G}(n, p)) > 0) \leq 0.05$

`fzero` root-finding algorithm, with an initial interval of $[0.1, 1]$. For graphs on $n \leq 470$ nodes, the region of p in which Theorem A.8 applies is at least 0.1 larger. With $p_u(t)$ plotted on logarithmic axes, the boundaries both appear to be straight lines with the same gradient. This demonstrates that the bound derived in Theorem A.8 would not allow for weaker asymptotic conditions on $p(n)$ in Proposition 4.17.

Finally, these same techniques can be applied to the directed flag complex to get a similar explicit bound.

Theorem A.10 *If $G \sim \overrightarrow{G}(n, p)$ then*

$$\mathbb{P}\Big[\beta_1(\overrightarrow{X}(G)) > 0\Big] \leq \binom{n}{4} 4! \sum_{m=0}^{3} p^3 (1-p)^{c_m} \left(1 - \sum_{l=0}^{8} q_{m,l}\, p^l (1-p)^{8-l}\right)^{n-4} + \binom{n}{2} p^2 \Big[1 - p^2\Big]^{2n-4} + 2\binom{n}{3} p^3 \Big[1 - p^3\Big]^{2n-6}, \tag{A31}$$

where $\mathbf{c} = (c_m) = (2, 3, 3, 4)^\intercal$ *and*

$$Q = (q_{m,l}) = \begin{pmatrix} 0 & 0 & 0 & 0 & 5 & 16 & 19 & 8 & 1 \\ 0 & 0 & 0 & 0 & 7 & 20 & 20 & 8 & 1 \\ 0 & 0 & 0 & 0 & 7 & 20 & 20 & 8 & 1 \\ 0 & 0 & 0 & 0 & 8 & 22 & 21 & 8 & 1 \end{pmatrix}. \tag{A32}$$

Table 2 Scope of the four data collection experiments

Exp. #	Homology	Samples	n-range	p-range
1	$\overrightarrow{\beta}_1(G)$	100	[20, 150]	$[10^{-4}, 0.15]$
2	$\overrightarrow{\beta}_1(G)$	100	[20, 100]	[0.05, 0.35]
3	$\beta_1(\overrightarrow{X}(G))$	200	[20, 200]	$[5 \times 10^{-4}, 0.3]$
4	$\beta_1(\overrightarrow{X}(G))$	200	[20, 200]	[0.1, 0.45]

Proof This follows from the same argument as Theorem A.8. The only changes are the counts of possible shortcut edges for each isomorphism class in Lemma A.6, and the output of the algorithm in the proof of Lemma A.7. □

Appendix B: Experimental results

B.1 Data collection

In order to further investigate the behaviour of path homology on random directed graphs, we sample empirical distributions of Betti numbers. Table 2 records the four experiments that were conducted. In each experiment, a number of random graphs were sampled from $G \sim \overrightarrow{G}(n, p)$, for n evenly spaced in intervals of 5 in the noted n-range, and 50 values of $p$logarithmically spaced in the noted p-range. Then, we compute the first Betti number of either non-regular path homology or directed flag complex, as noted in the table.

By logarithmically spaced in the range $[a, b]$ we mean that values are chosen evenly spaced between $\log(a)$ and $\log(b)$ and then we apply the exponential function. We discuss the reason for this logarithmic spacing in Appendix 1. Non-regular path homology is computed with the `pathhomology` package Yutin (2022). Unlike in Appendix A, we do not use the `symbolic` option and hence Betti numbers are subject to numerical errors, due to error in rank computations. Directed flag complex homology is computed with the `flagser` package Lütgehetmann et al. (2020), without approximation turned on. Also note that, due to computational restrictions, Experiments 1 and 2 were occasionally stopped and restarted. These experiments were run before `rng` persistence was implemented and hence reproduction attempts may yield slightly different results.

B.2 Illustrations

In Fig. 9, we merge the samples from Experiments 1 and 2 for $n \in [20, 100]$. We then use the colour axis to plot statistics for each of these samples, against the logarithm of each parameter on the two spatial axes. In Fig. 9a we record the empirical probability that $\overrightarrow{\beta}_1 = 0$ for each of the samples. In Fig. 9b, we record $\langle \overrightarrow{\beta}_1(G) \rangle / n(n-1)p$, where $\langle \overrightarrow{\beta}_1(G) \rangle$ is the mean $\overrightarrow{\beta}_1$ for each random sample of graphs. In the following

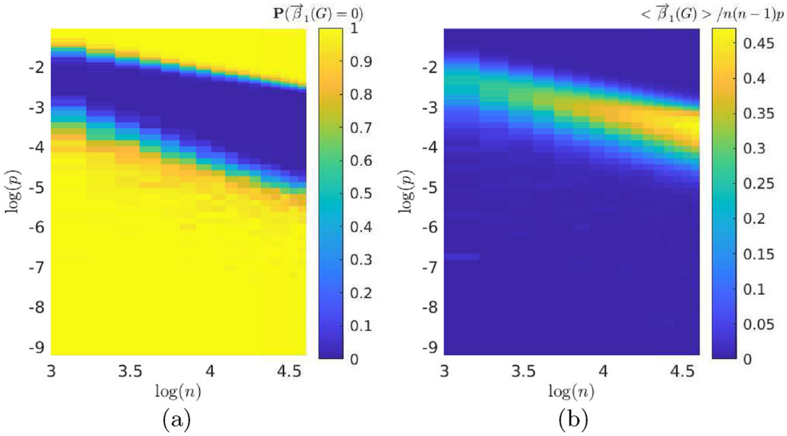


Fig. 9 Statistics for the path homology of samples of 100 random directed graphs $G \sim \vec{G}(n, p)$, sampled at a range of parameter values, plotted against the natural logarithm of the parameters. Our primary contribution is descriptions of the boundaries of the darker, blue region in (a) and a limiting value for the lighter, yellow region of (b) as $n \to \infty$

informal discussion, we refer to these figures as proxies for the exact probabilities and expectations of the distributions from which we sample.

We observe two distinct transitions between three distinct regions in parameter space. Namely, when graph density is relatively low $\mathbb{P}[\vec{\beta}_1 = 0]$ is almost 1. Next, there is a ‘goldilocks’ region in which suddenly $\mathbb{P}[\vec{\beta}_1 = 0]$ is close to 0 and $\mathbb{E}[\vec{\beta}_1]$ appears to be growing. Finally, when density is too large, we transition back to a regime in which $\mathbb{P}[\vec{\beta}_1 = 0]$ is almost 1. Thanks to the logarithmic scaling of the parameter axes, the boundaries between these regions appear to be straight lines.

Theorem 1.4 describes the fate of straight line trajectories through this diagram. Theorem 1.4(3-4) says that a straight line with gradient $m < -1$ (resp. $m > -1/3$) will eventually cross into and remain in the lower (resp. upper) yellow region of Fig. 9a, where $\mathbb{P}(\vec{\beta}_1 = 0)$ is close to 1. Theorem 1.4(2) says that a straight line with gradient $-1 < m < -2/3$ will eventually reach the blue region of Fig. 9a, where $\mathbb{P}(\vec{\beta}_1 = 0)$ is close to 0. In particular, this implies that the gradient of the lower boundary region tends towards -1 and the gradient of the upper boundary is eventually in the region $[-2/3, -1/3]$. Finally, Theorem 1.4(1) says that a straight line with gradient $-1 < m < -2/3$ will eventually reach the yellow region of Fig. 9b and the colour will approach 1.

In Fig. 10, we merge the samples from Experiments 3 and 4 for $n \in [20, 200]$. This figure provides Theorem 1.6 with a similar interpretation, as above, except that the upper boundary is eventually in the region $[-1/2, -1/4]$.

It is worth reiterating that these interpretations and results all hold *in the limit*. That is, Theorem 1.4 provides no guarantees for a finite line segment, regardless of its gradients or length. However, we do observe that the boundaries between the three regions converge onto straight lines, of the correct gradient, relatively quickly

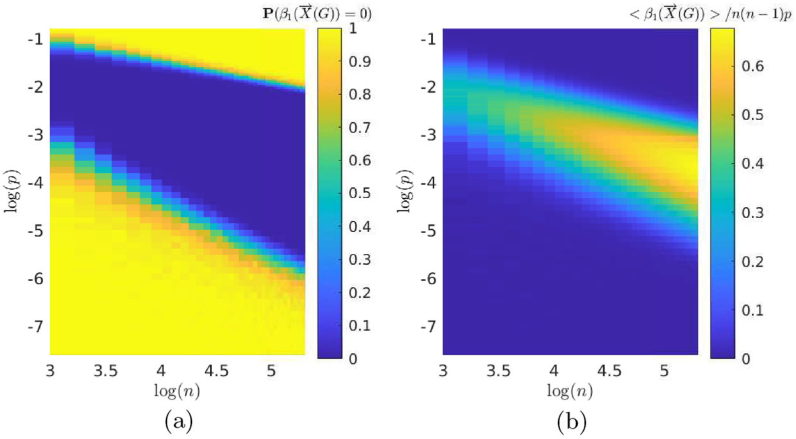


Fig. 10 Statistics for the directed flag complex homology of samples of 200 random graphs $G \sim \overrightarrow{G}(n, p)$ sampled at a range of parameter values

(i.e. within $n \leq 50$ nodes). We will see empirical evidence for this in the following section.

B.3 Finding boundaries

Note, Theorem 1.4 says nothing of the region $-2/3 < \alpha < -1/3$. In the following discussion we attempt to determine, empirically, the equations of the boundaries between the positive region and the two zero regions identified in Fig. 9. This provides evidence to support the conjecture that the zero region for path homology can be expanded to $(-\infty, -1) \cup (-2/3, 0]$.

Conjecture B.1 *For an Erdős–Rényi random directed graph* $G \sim \overrightarrow{G}(n, p(n))$, *let* $\overrightarrow{\beta}_1$ *denote the* 1*st Betti number of its non-regular path homology. Assume* $p(n) = n^{\alpha}$, *if* $\alpha > -2/3$ *then* $\overrightarrow{\beta}_1(G) = 0$ *with high probability. The same holds for regular path homology.*

Using the samples from Experiment 1, for each n, we determine the maximum $p_l(n)$ such that for all $p \leq p_l(n)$ we observe $\hat{\mathbb{P}}[\overrightarrow{\beta}_1(G) = 0] \geq 0.95$ for $G \sim \overrightarrow{G}(n, p)$, where $\hat{\mathbb{P}}$ denotes the empirical probability, derived from our sampled distribution. Similarly, using the samples from Experiment 2, we determine the minimum $p_u(n)$ such that for all $p \geq p_u(n)$ we observe $\hat{\mathbb{P}}[\overrightarrow{\beta}_1(G) = 0] \geq 0.95$ for $G \sim \overrightarrow{G}(n, p)$. Since we anticipate a power-law relationship, the logarithmic spacing of p allows us to achieve greater precision as n increases, since the values of $\log(p)$ are evenly spaced. Moreover, we chose the boundaries of the p-region so that precision is greater near the lower boundary in Experiment 1 and near the upper boundary in Experiment 2.

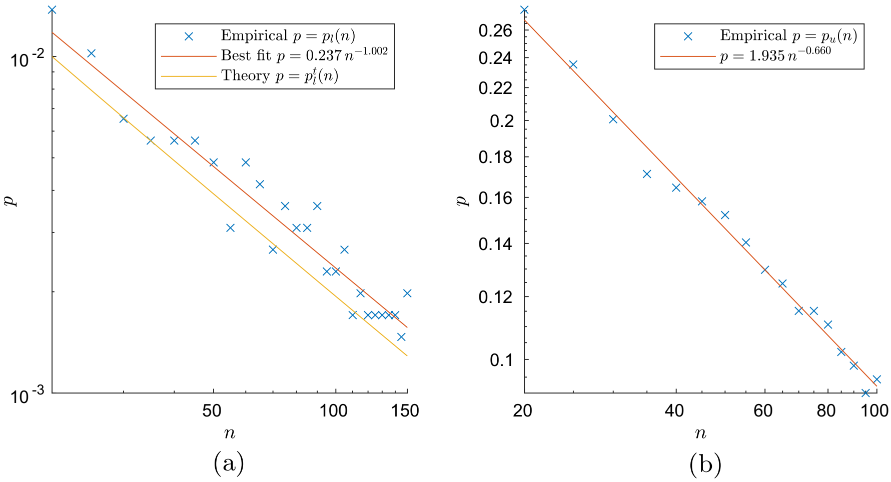


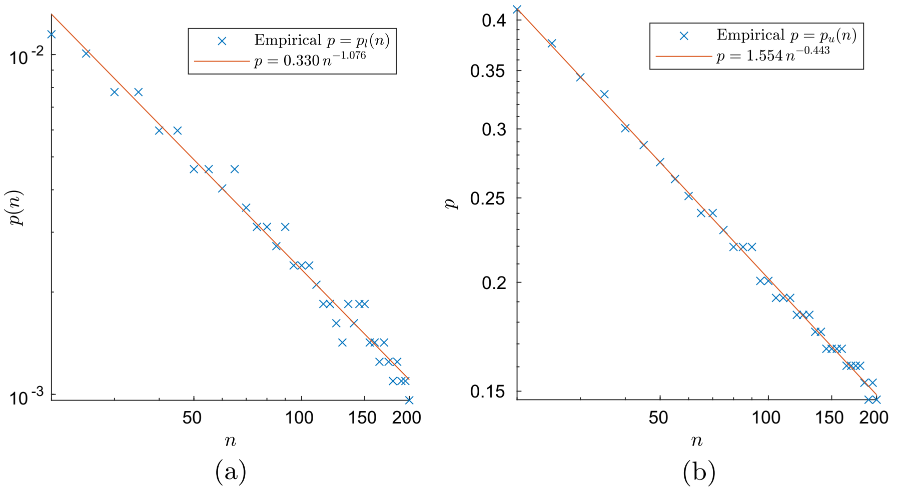
Fig. 11 Over a range of parameters (n, p) we measure the first Betti number, $\overrightarrow{\beta}_1(G)$, for 100 sampled random graphs $G \sim \overrightarrow{G}(n, p)$. Then, at each n, we determine maximum $p_l(n)$ (and minimum $p_u(n)$) such that for all $p \leq p_l(n)$ (and all $p \geq p_u(n)$) at most 5% of graphs sampled from $\overrightarrow{G}(n, p)$ have $\overrightarrow{\beta}_1(G) > 0$. The figures show log–log plots of these two functions. In both cases, we fit a line of best fit in order to obtain an approximate power-law relationship

We then compute a least-squares, line of best fit between $\log(p_l(n))$ and $\log(n)$, to obtain a power-law relationship of the form $p_l(n) = An^{\gamma}$. We repeat this for the upper boundary to obtain a similar relationship for $p_u(n)$. The results of this experiment are shown in Fig. 11.

Figure 11a shows that the empirical lower boundary has a similar dependency on n to that predicted by Theorem 1.4(2, 3), i.e. $p_l(n) \sim n^{-1}$. Moreover, Fig. 11a contains a plot of $p_l^t(n)$, as defined in Appendix A.1. For a parameter pair (n, p) falling below this line, Theorem A.1 implies $\mathbb{P}[\overrightarrow{\beta}_1(G) > 0] \leq 0.05$. We observe that this theoretical boundary of significance lies very close to the observed, experimental boundary, indicating that Theorem A.1 is close to the best possible bound.

Conversely, Fig. 11b predicts an upper boundary of $p_u(n) \sim n^{-0.660}$. This is consistent with Theorem 1.44 since $-0.660 < -1/3$, but indicates that there is significant room for improvement. This suggests that the hypothesis of Theorem 1.44 can be weakened to $\alpha > -2/3$. However, short of a stronger theoretical result, we require more experiments with graphs on $n > 150$ nodes to confirm this; computational complexity is currently the limiting factor.

In Fig. 12 we repeat the same analysis with Experiments 3 and 4 respectively, in order to discern the boundaries of the positive region for directed flag complex homology. Again, Fig. 12a shows that the empirical lower boundary has a similar dependency on n to that predicted by Theorem 1.6(3, 4), i.e. $p_l(n) \sim n^{-1}$. Fig. 11b shows an upper boundary of $p_u(n) \sim n^{-0.443}$. This is also consistent with Theorem 1.65 since $-0.437 < -1/4$. This provides evidence that the zero region for directed flag complex can be expanded as far as $(-\infty, -1) \cup (-1/2, 0]$.



(a) (b)

Fig. 12 As for Fig. 11 but for 200 samples of $\beta_1(\overrightarrow{X}(G))$

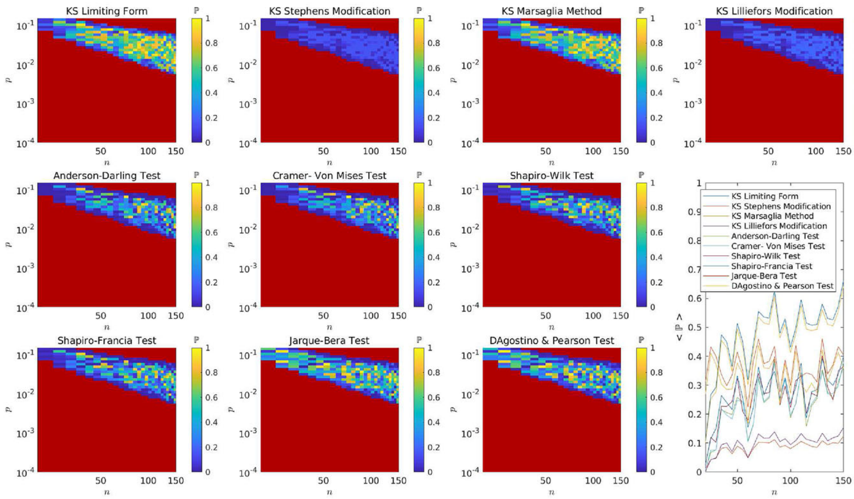


Fig. 13 Results for 10 normality tests on samples of $\overrightarrow{\beta}_1(G)$ for $G \sim \overrightarrow{G}(n, p)$ at a range of n and p. Red rectangles indicate that at least 5% of samples were zero and hence are excluded from the experiment. Colour indicates the $\mathbb{P}$-value for the hypothesis test in question. Finally, for each test and at each n, we average the $\mathbb{P}$-value of the range of relevant p, which is recorded on the line graph. Note adjacent densities, p, on the horizontal axis are shown with equal width, despite being logarithmically spaced

B.4 Testing for normal distribution

In analogy to known results for the clique complex (Kahle and Meckes 2013, Theorem 2.4), one conjecture is that, in the known positive region, the normalised Betti number $\overrightarrow{\beta}_1$ approaches a normal distribution.

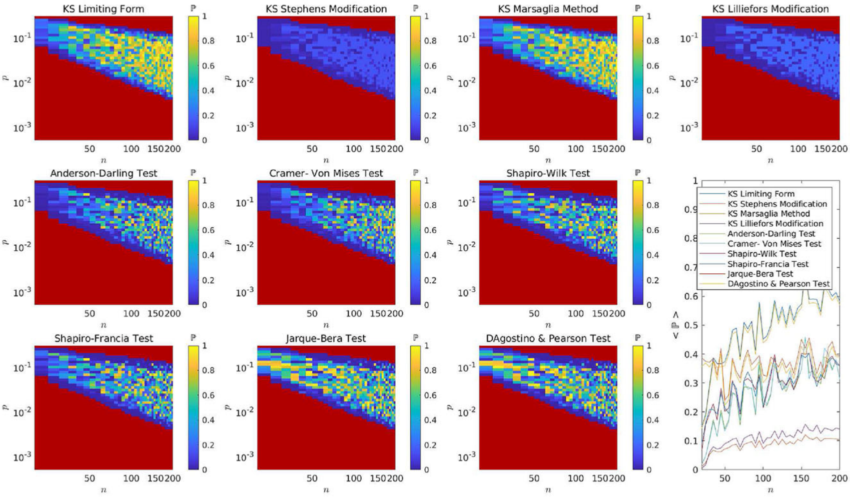


Fig. 14 As Fig. 13 but for $\beta_1(\overrightarrow{X}(G))$

Conjecture B.2 *If $G \sim G(n, p)$, where $p = p(n)$ with $p(n) = \omega(n^{-1})$ and $p(n) = o(n^{-2/3})$, then*

$$\frac{\overrightarrow{\beta}_1(G) - \mathbb{E}[\overrightarrow{\beta}_1(G)]}{\sqrt{\mathrm{Var}\left(\overrightarrow{\beta}_1(G)\right)}} \Longrightarrow \mathcal{N}(0, 1) \text{ as } n \to \infty \tag{B.1}$$

where $\mathcal{N}(0, 1)$ is the normal distribution with mean 0 *and variance* 1.

To provide some empirical evidence towards this conjecture, we perform 10 normality tests on the distributions of $\overrightarrow{\beta}_1$, obtained in Experiment 1. We restrict our focus to the samples in which at most 5% of samples were zero, so that we are in a parameter region where we hope our conjecture would apply.

We normalise each of the remaining samples and perform 10 hypothesis tests under the null hypothesis that the samples come from normal distributions. To avoid confusion with the null model parameter, we refer to the significance of these hypothesis tests as $\mathbb{P}$-values. These tests are computed with the `MATLAB` package `normalitytest` Öner et al. (2017). The $\mathbb{P}$-values (and names of the tests) are recorded in Fig. 13, along with the average $\mathbb{P}$-value against edge-inclusion probability p. In all tests, we see a noisy but consistent trend: there is a decreasing amount of evidence for discarding the null hypothesis as $n \to \infty$.

In Fig. 14, we repeat this analysis with the distributions of $\beta_1(\overrightarrow{X}(G))$ collected in Experiment 3. Again we observe a similar but stronger trend: there is a decreasing amount of evidence for discarding the null hypothesis as $n \to \infty$.

While no individual test is sufficient to conclude that $\overrightarrow{\beta}_1$ tends towards a normal distribution, the ensemble of tests provide good evidence towards this claim. Larger sample sizes, as well as samples at larger n, are required for more convincing evidence.

References

Aalto, A., Viitasaari, L., Ilmonen, P., Mombaerts, L., Gonçalves, J.: Gene regulatory network inference from sparsely sampled noisy data. Nature Commun. **11**(1), 3493 (2020). https://doi.org/10.1038/s41467-020-17217-1

Alon, N., Spencer, J.H.: The Probabilistic Method, 4th edn. Wiley series in discrete mathematics and optimization. Wiley, Hoboken, New Jersey (2016)

Caputi, L., Pidnebesna, A., Hlinka, J.: Promises and pitfalls of topological data analysis for brain connectivity analysis. NeuroImage **238**, 118245 (2021). https://doi.org/10.1016/j.neuroimage.2021.118245

Chaplin, T.: First Betti number of the path homology of random directed graphs - Code and Data Repository. https://github.com/tomchaplin/phrg-code

Chaplin, T.: First Betti number of the path homology of random directed graphs - Code and Data Repository. OSF (2022). https://doi.org/10.17605OSF.IO/ZVUMB. https://osf.io/zvumb/

Chen, L.H.Y., Goldstein, L., Shao, Q.-M.: Fundamentals of Stein's Method, pp. 13–44. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-15007-4_2

Chowdhury, S., Mémoli, F.: Persistent Path Homology of Directed Networks, pp. 1152–1169 (2018). https://doi.org/10.1137/1.9781611975031.75

Dwass, M.: Modified randomization tests for nonparametric hypotheses. Ann. Math. Stat. **28**(1), 181–187 (1957). https://doi.org/10.1214/aoms/1177707045

Erdős, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci **5**(1), 17–60 (1960)

Grigor'yan, A., Lin, Y., Muranov, Y., Yau, S.-T.: Homologies of path complexes and digraphs (2012). arXiv:1207.2834 [math.CO]

Grigor'yan, A., Lin, Y., Muranov, Y., Yau, S.-T.: Homotopy theory for digraphs. Pure Appl. Math. Q. **10**(4), 619–674 (2014). https://doi.org/10.4310/PAMQ.2014.v10.n4.a2

Grigor'yan, A.A., Lin, Y., Muranov, Y.V., Yau, S.-T.: Path complexes and their homologies. Journal of Mathematical Sciences **248**(5), 564–599 (2020). https://doi.org/10.1007/s10958-020-04897-9

Helm, A., Blevins, A.S., Bassett, D.S.: The growing topology of the *C. elegans* connectome. bioRxiv (2021). https://doi.org/10.1101/2020.12.31.424985

Ingram, P.J., Stumpf, M.P., Stark, J.: Network motifs: structure does not determine function. BMC Genomics **7**(1), 1–12 (2006). https://doi.org/10.1186/1471-2164-7-108

Kahle, M.: Topology of random clique complexes. Discrete Math. **309**(6), 1658–1671 (2009). https://doi.org/10.1016/j.disc.2008.02.037

Kahle, M.: Sharp vanishing thresholds for cohomology of random flag complexes. Ann. Math. (2014). https://doi.org/10.4007/annals.2014.179.3.5

Kahle, M., Meckes, E.: Limit theorems for Betti numbers of random simplicial complexes. Homol. Homotopy Appl. **15**(1), 343–374 (2013). https://doi.org/10.4310/HHA.2013.v15.n1.a17

Leskovec, J., Krevl, A.: SNAP Datasets: Stanford Large Network Dataset Collection (2014). http://snap.stanford.edu/data

Lütgehetmann, D., Govc, D., Smith, J.P., Levi, R.: Computing persistent homology of directed flag complexes. Algorithms (2020). https://doi.org/10.3390/a13010019

Öner, M., Deveci Kocakoç, İ.: Jmasm 49: A compilation of some popular goodness of fit tests for normal distribution: Their algorithms and matlab codes (matlab). Journal of Modern Applied Statistical Methods 16(2), 30 (2017). https://doi.org/10.22237/jmasm/1509496200

Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W., LaMantia, A., McNamara, J., White, L. (eds.): Neuroscience, 6th edn. Sinauer Associates, New York (2018)

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). https://doi.org/10.3389/fncom.2017.00048

Yutin, M.: Performant Path Homology. https://github.com/SteveHuntsmanBAESystems/PerformantPathHomology

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