Journal of Applied and Computational Topology
https://doi.org/10.1007/s41468-023-00146-5

Multivariate central limit theorems for random clique complexes

Tadas Temčinas[1] · **Vidit Nanda**[2] · **Gesine Reinert**[1,3]

Received: 18 November 2022 / Revised: 1 August 2023 / Accepted: 12 September 2023

Abstract
Motivated by open problems in applied and computational algebraic topology, we establish multivariate normal approximation theorems for three random vectors which arise organically in the study of random clique complexes. These are:

(1) the vector of critical simplex counts attained by a lexicographical Morse matching,
(2) the vector of simplex counts in the link of a fixed simplex, and
(3) the vector of total simplex counts.

The first of these random vectors forms a cornerstone of modern homology algorithms, while the second one provides a natural generalisation for the notion of vertex degree, and the third one may be viewed from the perspective of U-statistics. To obtain distributional approximations for these random vectors, we extend the notion of dissociated sums to a multivariate setting and prove a new central limit theorem for such sums using Stein's method.

Keywords Stein's method · Multivariate normal approximation · Discrete Morse theory · Random graphs · Random Simplicial complexes

Mathematics Subject Classification 60F05 · 60D05 · 05C80

✉ Vidit Nanda
nanda@maths.ox.ac.uk

Tadas Temčinas
tadas.temcinas@keble.ox.ac.uk

Gesine Reinert
reinert@stats.ox.ac.uk

[1] Department of Statistics, University of Oxford, Oxford, UK

[2] Mathematical Institute, University of Oxford, Oxford, UK

[3] The Alan Turing Institute, London, UK

Published online: 21 October 2023

1 Introduction

Methods from applied and computational algebraic topology have recently found substantial applications in the analysis of nonlinear and unstructured datasets (Ghrist 2008; Carlsson 2005). The modus operandi of topological data analysis is to first build a nested family of simplicial complexes around the elements of a dataset, and to then compute the associated persistent homology barcodes (Edelsbrunner and Harer 2010). Of central interest, when testing hypotheses under this paradigm, is the question of what homology groups to expect when the input data are randomly generated. Significant efforts have therefore been devoted to answering this question for various models of noise, giving rise to the field of *stochastic topology* (Kahle 2011; Bobrowski and Kahle 2018; Kahle 2009; Adler et al. 2014; Costa and Farber 2016). Our work here is a contribution to this area at the interface between probability theory and algebraic topology.

Distributional approximations provide a way of understanding random variables in cases where closed-form distributions cannot be easily obtained. This paper establishes the first multivariate normal approximations to three important counting problems in stochastic topology; as these approximations are based on Stein's method, explicit bounds on the approximation errors are provided. Our starting point is the ubiquitous graph model $\mathbf{G}(n, p)$; a graph G chosen from this model has as its vertex set $[n] = \{1, 2, \ldots, n\}$, and each of its possible $\binom{n}{2}$ edges is included independently with probability $p \in [0, 1]$. A natural higher-order generalisation of $\mathbf{G}(n, p)$ is furnished by the random clique complex model $\mathbf{X}(n, p)$, whose constituent complexes $\mathscr{L}$ are constructed as follows. One first selects an underlying graph $G \sim \mathbf{G}(n, p)$, and then deterministically fills out all k-cliques in G with $(k-1)$-dimensional simplices for $k \geq 3$. Higher connectivity is measured by the Betti numbers $\beta_k(\mathscr{L})$, which are ranks of rational homology groups $\mathrm{H}_k(\mathscr{L}; \mathbb{Q})$—in particular, $\beta_0(\mathscr{L})$ equals the number of connected components of the underlying random graph G. In Kahle (2014), Kahle proved the following far-reaching generalisation of the Erdős-Rényi connectivity result: for each $k \geq 1$ and $\epsilon > 0$,

(1) if

$$p \geq \left[\left(\frac{k}{2} + 1 + \epsilon\right) \cdot \frac{\log(n)}{n}\right]^{1/(k+1)},$$

then $\beta_k(\mathscr{L}) = 0$ with high probability; and moreover,

(2) if

$$\left[\frac{k+1+\epsilon}{n}\right]^{1/k} \leq p \leq \left[\left(\frac{k}{2} + 1 - \epsilon\right) \cdot \frac{\log(n)}{n}\right]^{1/(k+1)},$$

then $\beta_k(\mathscr{L}) \neq 0$ with high probability.

Unlike Kahle's result, we study $\mathbf{X}(n, p)$ in the regime where p is a constant. In that case, we may have $\beta_k(\mathbf{X}(n, p)) \neq 0$ for multiple values of k that depend on n. Hence, studying the multivariate distribution of the Betti numbers is of interest in

this regime. Since many results about Betti numbers in the univariate case are based on counting simplices, we hope that understanding the multivariate simplex counts will facilitate multivariate understanding of the Betti numbers. With this result in mind, we motivate and describe three random vectors pertaining to $\mathscr{L} \sim \mathbf{X}(n, p)$; the normal approximation of these three random vectors will be our focus in this paper. All three are denoted $T = (T_1, \ldots, T_d)$ for an integer $d > 0$. For the purposes of this introduction, we add a superscript (1), (2) or (3) to indicate the particular vector.

Random vector 1: critical simplex counts

The computation of Betti numbers $\beta_k(\mathscr{L})$ begins with the chain complex

$$\cdots \xrightarrow{d_{k+1}} \mathrm{Ch}_k \xrightarrow{d_k} \mathrm{Ch}_{k-1} \xrightarrow{d_{k-1}} \cdots \xrightarrow{d_2} \mathrm{Ch}_1 \xrightarrow{d_1} \mathrm{Ch}_0.$$

Here Ch_k is a vector space whose dimension equals the number of k-simplices in $\mathscr{L}$, while $d_k : \mathrm{Ch}_k \to \mathrm{Ch}_{k-1}$ is an incidence matrix encoding which $(k-1)$-simplices lie in the boundary of a given k-simplex. These matrices satisfy the property that every successive composite $d_{k+1} \circ d_k$ equals zero, and $\beta_k(\mathscr{L})$ is the dimension of the quotient vector space $\ker d_k / \mathrm{img}\, d_{k+1}$. In order to calculate $\beta_k(\mathscr{L})$, one is required to put the matrices $\{d_k : \mathrm{Ch}_k \to \mathrm{Ch}_{k-1}\}$ in reduced echelon form, which is a straightforward task in principle. Unfortunately, Gaussian elimination on an $m \times m$ matrix incurs an $O(m^3)$ cost, which becomes prohibitive when facing simplicial complexes built around large data sets (Otter 2017). The standard remedy is to construct a much smaller chain complex which has the same homology groups, and by far the most fruitful mechanism for achieving such homology-preserving reductions is *discrete Morse theory* (Forman 2002; Mischaikow and Nanda 2013; Henselman-Petrusek and Ghrist 2016; Lampret 2019).

The key structure here is that of an *acyclic partial matching*, which pairs together certain adjacent simplices of $\mathscr{L}$; and the homology groups of $\mathscr{L}$ may be recovered from a chain complex whose vector spaces are spanned by unpaired, or *critical*, simplices. One naturally seeks an optimal acyclic partial matching on $\mathscr{L}$ which admits the fewest possible critical simplices. Unfortunately, the optimal matching problem is computationally intractable to solve (Joswig and Pfetsch 2006) even approximately (Bauer and Rathod 2019) for large $\mathscr{L}$. Our first random vector $T^{(1)}$ is obtained by letting $T_k^{(1)}$ equal the number of critical k-simplices for a specific type of acyclic partial matching on $\mathscr{L}$, called the *lexicographical* matching. Knowledge of this random vector serves to simultaneously quantify the benefit of using discrete Morse theoretic reductions on random simplicial complexes and to provide a robust null model by which to measure their efficacy on general (i.e., not necessarily random) simplicial complexes. This is the first time the asymptotic distribution of this random vector is studied.

Random vector 2: link simplex counts

The *link* of a simplex t in $\mathscr{L}$, denoted $\mathbf{lk}(t)$, consists of all simplices s for which the union $s \cup t$ is also a simplex in $\mathscr{L}$ and the intersection $s \cap t$ is empty. The link of t forms a simplicial complex in its own right; and if we restrict attention to the underlying random graph G, then the link of a vertex is precisely the collection of its neighbours. Therefore, the Betti numbers $\beta_k(\mathbf{lk}(t))$ generalise the degree distribution for vertices of random graphs in two different ways—one can study neighbourhoods of higher-dimensional simplices by increasing the dimension of t, and one can examine higher-order connectivity properties by increasing the homological dimension k. The second random vector $T^{(2)}$ of interest to us here is obtained by letting $T_k^{(2)}$ equal the number of k-simplices that would lie in the link of a fixed simplex t in $\mathscr{L}$, if t indeed was a simplex in the random complex. As far as we are aware, ours is the first work that studies this random vector. A different conditional distribution, which follows directly from results on subgraph counts in $\mathbf{G}(n, p)$, has been studied before, see Remark 5.1.

There are compelling reasons to better understand the combinatorics and topology of such links from a probabilistic viewpoint. For instance, the fact that the link of a k-simplex in a triangulated n-manifold is always a triangulated sphere of dimension $(n - k - 1)$ has been exploited to produce canonical stratifications of simplicial complexes into homology manifolds (Asai and Shah 2022; Nanda 2020). Knowledge of simplex counts (and hence, Betti numbers) of links would therefore form an essential first step in any systematic study involving canonical stratifications of random clique complexes.

Random vector 3: total simplex counts

The strategy employed in Kahle's proof of the second assertion above involves first checking that the expected number of k-simplices in $\mathscr{L} \sim \mathbf{X}(n, p)$ is much larger than the expected number of simplices of dimensions $k \pm 1$ whenever p lies in the range indicated by (2). Therefore, one may combine the Morse inequalities with the linearity of expectation in order to guarantee that the expected $\beta_k(\mathscr{L})$ is nonzero—see Kahle (2014, Section 4) for details. To facilitate more refined analysis and estimates of this sort, the third random vector $T^{(3)}$ we study in this paper is obtained by letting $T_k^{(3)}$ equal the total number of k-dimensional simplices in $\mathscr{L}$.

Since $T_k^{(3)}$ is precisely the number of $(k+1)$-cliques in $G \sim \mathbf{G}(n, p)$, this random vector falls within the purview of *generalised U-statistics*. We extend results from Janson and Nowicki (1991) to show not only distributional convergence asymptotically but a stronger result, detailing explicit non-asymptotic bounds on the approximation. Several interesting problems can be seen as special cases—these include classical U-statistics (Lee 1990; Korolyuk and Borovskich 2013), monochromatic subgraph counts of inhomogeneous random graphs with independent random vertex colours, and the number of overlapping patterns in a sequence of independent Bernoulli trials. To the best of our knowledge, this is the first multivariate normal approximation result with explicit bounds where the sizes of the subgraphs are permitted to increase with n.

Main results

The central contributions of this work are multivariate normal approximations for all three random vectors T described above. The approximation error is quantified for finite n in terms of both smooth test functions as well as convex set indicator test functions. As long as the bound with respect to either test function class vanishes asymptotically, it implies asymptotic convergence in distribution. However, the convexset indicator test function result is stronger and can be more useful in statistical applications: for example, when estimating confidence regions, which are usually taken to be convex sets.

We state a simplified version of our normal approximation results here and note that the full statements and proofs have been recorded as Theorems 4.5, 5.3, and Corollary 6.1. Note that the quantities below $B_{5.3}$ and $B_{6.1}$ are explicit, allowing to vary the parameters p and d.

To state the results, for a positive integer d we define a class of test functions $h : \mathbb{R}^d \to \mathbb{R}$, as follows. We say $h \in \mathscr{H}_d$ iff h is three times partially differentiable with third partial derivatives being Lipschitz and bounded. Moreover we denote by $\mathscr{K}$ the class of convex sets in $\mathbb{R}^d$.

Theorem 1.1 *Let $W^{(i)}$ be an appropriately scaled and centered version of random vector $T^{(i)}$ for $i = 1, 2, 3$ as described above. Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$ and Σ_i be the covariance matrix of $W^{(i)}$ for each i. Let $h \in \mathscr{H}_d$.*

(1) There is a constant $B_{1.1.1} > 0$ independent of n and a natural number $N_{1.1.1}$ such that for any $n \geq N_{1.1.1}$ we have

$$\left|\mathbb{E}h(W^{(1)}) - \mathbb{E}h(\Sigma_1^{\frac{1}{2}} Z)\right| \leq B_{1.1.1} \sup_{i,j,k\in[d]} \left\| \frac{\partial^3 h}{\partial x_i \partial x_j \partial x_k} \right\|_\infty n^{-1}.$$

Also, there is a constant $B_{1.1.2} > 0$ independent of n and a natural number $N_{1.1.2}$ such that for any $n \geq N_{1.1.2}$ we have

$$\sup_{A\in\mathscr{K}} |\mathbb{P}(W^{(1)} \in A) - \mathbb{P}(\Sigma_1^{\frac{1}{2}} Z \in A)| \leq B_{1.1.2} n^{-\frac{1}{4}}.$$

(2) There is a quantity $B_{5.3}$ independent of n and defined explicitly such that

$$\left|\mathbb{E}h(W^{(2)}) - \mathbb{E}h(\Sigma_2^{\frac{1}{2}} Z)\right| \leq |h|_3 \, B_{5.3}(n - |t|)^{-\frac{1}{2}};$$

and

$$\sup_{A\in\mathscr{K}} |\mathbb{P}(W^{(2)} \in A) - \mathbb{P}(\Sigma_2^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{5.3}^{\frac{1}{4}} (n - |t|)^{-\frac{1}{8}}.$$

(3) There is a quantity $B_{6.1}$ independent of n and defined explicitly such that

$$\left|\mathbb{E}h(W^{(3)}) - \mathbb{E}h(\Sigma_3^{\frac{1}{2}} Z)\right| \leq |h|_3 \, B_{6.1} n^{-1};$$

and

$$\sup_{A \in \mathscr{K}} |\mathbb{P}(W^{(3)} \in A) - \mathbb{P}(\Sigma_3^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{6.1}^{\frac{1}{4}} n^{-\frac{1}{4}}.$$

En route to proving Theorem 1.1, we also establish the following properties, which are of direct interest in computational topology. Here we assume that $p \in (0, 1)$ and $k \in \{1, 2, \ldots\}$ are constants.

(1) The expected number of critical k-simplices is one order of n smaller that the expected total number of k-simplices; see Lemma 4.2.
(2) The variance of the number of critical k-simplices is at least of the order n^{2k}, as shown in Lemma 4.3. An upper bound of the same order can be proved similarly. The variance of the total number of k-simplices is also of the same order.
(3) Knowing the expected value and the variance one can prove concentration results using different concentration inequalities, for example, Chebyshev's inequality. This would show that not only the expected value of the number of critical simplices is smaller compared to all simplices but also that large deviations from the mean are unlikely, hence implying that the substantial improvement of one order of n is not only expected but also likely.
(4) For counting critical simplices to high accuracy in probability, it is not necessary to check every simplex. Certain simplices have a very small chance of being critical, and can be safely ignored. The probability of this omission causing an error is vanishingly small asymptotically; see Proposition 4.4.

The main ingredient in establishing such results is often an abstract approximation theorem that can be applied to the random variables of interest. While there is no shortage of multivariate normal approximation theorems (Fang 2016; Raiíc 2004; Meckes 2009; Chen 2011), the existing ones are not sufficiently fine-grained for proving multivariate normal approximations to the random vectors studied here. We therefore return to the pioneering work of Barbour et al. (1989), who proved a univariate central limit theorem (CLT) for a decomposable sum of random variables using Stein's method, treating the case of dissociated sums as a special case. Our approximation result (Theorem 3.2) forms an extension of their ideas to the multivariate setting.

Related work

There are different versions of distributional approximation results for subgraph counts in $\mathbf{G}(n, p)$ that can be interpreted as simplex counts in $\mathbf{X}(n, p)$. For example, a multivariate central limit theorem for *centered* subgraph counts in the more general setting of a random graph associated to a graphon can be found in Kaur and Röollin (2021). That proof is based on Stein's method via a Stein coupling. Translating this result for

uncentered subgraph counts would yield an approximation by a function of a multivariate normal. In Reinert and Röllin (2010), an exchangeable pair coupling led to Reinert and Rollin (2010, Proposition 2) which can be specialised to joint counts of edges and triangles; our approximation significantly generalises this result beyond the case where $k \in \{1, 2\}$. Several univariate normal approximation theorems for subgraph counts are available; recent developments in this area include Privault and Serafin (2020), which uses Malliavin calculus together with Stein's method, and Eichelsbacher and Rednoß (2023), which uses the Stein-Tikhomirov method. Stein's method is used in Kahle and Meckes (2013) to show a CLT for the Betti numbers of $\mathbf{X}(n, p)$ in a sparse regime; in Owada et al. (2021) limit theorems for Betti numbers and Euler characteristic are proven in a dynamical random simplicial complex model, also using Stein's method.

Theorem 3.2 is not the first generalisation of the results in Barbour et al. (1989) to a multivariate setting, see for example (Fang 2016; Raiíc 2004). The key advantage of our approach is that it allows for bounds which are non-uniform in each component of the vector W. This is useful when, for example, the number of summands in each component are of different order or when the sizes of dependency neighbourhoods in each component are of different order. The applications considered here are precisely of this type, where the non-uniformity of the bounds is crucial. Moreover, we do not require the covariance matrix Σ to be invertible, and can therefore accommodate degenerate multivariate normal distributions.

Organisation

In Sect. 2 we recall concepts from the theory of simplicial complexes, which we later use. In Sect. 3 we state the theorems that serve as main tools in proving the CLTs. In order to maintain focus on the main results, we defer the proofs of this section until the end of the paper. In Sect. 4 we prove an approximation theorem for critical simplex counts of lexicographical matchings. Two technical computations required in this section have been consigned to the Appendix. In Sect. 5 we prove an approximation theorem for count variables of simplices that are in the link of a fixed simplex. In Sect. 6 we study simplex counts in the random clique complex and prove a CLT for this random variable. This CLT is a corollary of a multivariate normal approximation of generalised U-statistics, which might be of independent interest. Finally, in Sect. 7 we prove our main tools: the abstract approximation theorem (Theorem 3.2) as well as the approximation theorem for U-statistics (Theorem 3.9). We first use smooth test functions and then extend the results to convex set indicators using a smoothing technique from Gan et al. (2017).

Notation

Throughout this paper we use the following notation. Given positive integers n, m we write $[m, n]$ for the set $\{m, m+1, \ldots, n\}$ and $[n]$ for the set $[1, n]$. Given a set X we write $|X|$ for its cardinality, $\mathscr{P}(X)$ for its powerset, and given a positive integer k we write $C_k = \{\, t \in \mathscr{P}([n]) \mid |t| = k \,\}$ for the collection of subsets of $[n]$ which are of size k. For a function $f : \mathbb{R}^d \to \mathbb{R}$ we write $\partial_{ij} f = \frac{\partial^2 f}{\partial x_i \partial x_j}$ and $\partial_{ijk} f =$

$\frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}$. Also, we write $|f|_k = \sup_{i_1,i_2,\ldots,i_k \in [d]} \left\| \partial_{i_1 i_2 \ldots i_k} f \right\|_\infty$ for any integer $k \geq 1$, as long as the quantities exist. Here $||\cdot||_\infty$ denotes the supremum norm while $||\cdot||_2$ denotes the Euclidean norm. The notation ∇ denotes the gradient operator in $\mathbb{R}^d$. The notation $\mathrm{Id}_{d\times d}$ denotes the $d \times d$ identity matrix. The vertex set of all graphs and simplicial complexes is assumed to be $[n]$. We also use Bachmann-Landau asymptotic notation: we say $f(n) = O(g(n))$ iff $\limsup_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty$ and $f(n) = \Omega(g(n))$ iff $\liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0$. The notation that $f(n) = \omega(g(n))$ indicates that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$.

2 Simplicial complex preliminaries

2.1 First definitions

Firstly, we recall the notion of a simplicial complex (Spanier 1966, Ch 3.1); these provide higher-dimensional generalisations of a graph and constitute data structures of interest across algebraic topology in general as well as applied and computational topology in particular.

A simplicial complex $\mathscr{L}$ on a vertex set V is a set of nonempty subsets of V (i.e. $\varnothing \notin \mathscr{L} \subseteq \mathscr{P}(V)$) such that the following properties are satisfied:

(1) for each $v \in V$ the singleton $\{v\}$ lies in $\mathscr{L}$, and
(2) if $t \in \mathscr{L}$ and $s \subset t$ then $s \in \mathscr{L}$.

The *dimension* of a simplicial complex $\mathscr{L}$ is $\max_{s\in\mathscr{L}} |s| - 1$. Elements of a simplicial complex are called *simplices*. If s is a simplex, then its dimension is $|s| - 1$. A simplex of dimension k can be called a k-simplex. Note that the notion of one-dimensional simplicial complex is equivalent to the notion of a graph, with the vertex set V and edges as subsets.

Given a graph $G = (V, E)$ the clique complex $\mathscr{X}$ of G is a simplicial complex on V such that

$$t \in \mathscr{X} \iff \forall u, v \in t,\ \{u, v\} \in E.$$

Recall that $\mathbf{G}(n, p)$ is a random graph on n vertices where each pair of vertices is connected with probability p, independently of any other pair. The $\mathbf{X}(n, p)$ random simplicial complex is the clique complex of the $\mathbf{G}(n, p)$ random graph, which is a random model studied in stochastic topology (Kahle 2009, 2014). Note that $t \in \mathscr{X}$ if and only if the vertices of t span a clique in G. Thus, elements in $\mathbf{X}(n, p)$ are cliques in $\mathbf{G}(n, p)$.

2.2 Links

The link of a simplex t in a simplicial complex $\mathscr{L}$ is the subcomplex

$$\mathbf{lk}(t) = \{\, s \in \mathscr{L} \mid s \cup t \in \mathscr{L} \text{ and } t \cap s = \varnothing \,\}.$$

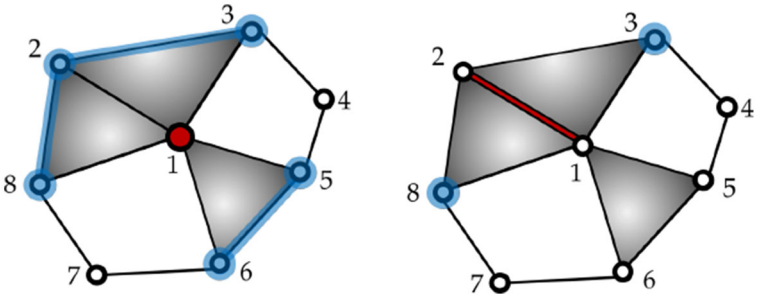

Fig. 1 Left: the link (highlighted in blue) of the vertex 1 (highlighted in red). Right: the link (highlighted in blue) of the edge {1, 2} (highlighted in red). The two-dimensional simplices are shaded in grey

Example 2.1 If we look at a graph as a one dimensional simplicial complex, then the vertices are sets of the form $\{i\}$ and edges are sets of the form $\{i, j\}$. For a vertex $t = \{v\}$, the edges of the form $s = \{v, u\}$ will not be in the link of t because $t \cap s = \varnothing$ is not satisfied. If we pick $s = \{i, j\}$ and $v \notin s$, then $s \cup t \in \mathscr{L}$ is not satisfied. So there will be no edges in the link. However, if $s = \{u\}$ and u is a neighbour of v, then $s \cup t \in \mathscr{L}$ and $s \cap t = \varnothing$. Hence the link of a vertex will be precisely the other vertices that the vertex is connected to; the notion of the link generalises the idea of a neighbourhood in a graph.

Example 2.2 Now consider the simplicial complex depicted in Fig. 1: it has 8 vertices, 12 edges and 3 two-dimensional simplices that are shaded in grey. On the left hand side of the figure we see highlighted in blue the link of the vertex 1, which is highlighted in red. So $\mathbf{lk}(\{1\}) = \{\{2\}, \{3\}, \{5\}, \{6\}, \{8\}, \{2, 3\}, \{2, 8\}, \{5, 6\}\}$. On the right hand side of the figure we see highlighted in blue the link of the edge $\{1, 2\}$, which is highlighted in red. That is, $\mathbf{lk}(\{1, 2\}) = \{\{3\}, \{8\}\}$.

2.3 Discrete Morse theory

A *partial matching* on a simplicial complex $\mathscr{L}$ is a collection

$$\Sigma = \{\, (s, t) \mid s \subseteq t \in \mathscr{L} \text{ and } |t| - |s| = 1 \,\}$$

such that every simplex appears in at most one pair of Σ. A Σ-*path* (of length $k \geq 1$) is a sequence of distinct simplices of $\mathscr{L}$ of the following form:

$$(s_1 \subseteq t_1 \supseteq s_2 \subseteq t_2 \supseteq \ldots \supseteq s_k \subseteq t_k)$$

such that $(s_i, t_i) \in \Sigma$ and $|t_i| - |s_{i+1}| = 1$ for all $i \in [k]$. A Σ-path is called a *gradient path* if $k = 1$ or s_1 is not a subset of t_k. A partial matching Σ on $\mathscr{L}$ is called *acyclic* iff every Σ-path is a gradient path. Given a partial matching Σ on $\mathscr{L}$, we say that a simplex $t \in \mathscr{L}$ is *critical* iff t does not appear in any pair of Σ.

For a one-dimensional simplicial complex, viewed as a graph, a partial matching Σ is comprised of elements $(v; \{u, v\})$ with v a vertex and $\{u, v\}$ an edge. A $\Sigma-$path

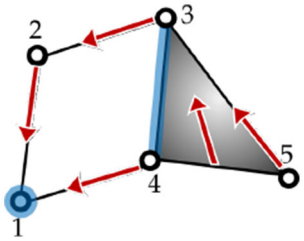

Fig. 2 Lexicographical matching given by the red arrows. Critical simplices are highlighted in blue

is then a sequence of distinct vertices and edges

$$v_1, \{v_1, v_2\}, v_2, \{v_2, v_3\}, \ldots, v_k, \{v_k, v_{k+1}\}$$

where each consecutive pair of the form $(v_i, \{v_i, v_{i+1}\})$ is constrained to lie in Σ.

We refer the interested reader to Forman (2002) for an introduction to discrete Morse theory and to Mischaikow and Nanda (2013) for seeing how it is used to reduce computations in the persistent homology algorithm. In this work we aim to understand how much improvement one would likely get on a random input when using a specific type of acyclic partial matching, defined below.

Definition 2.3 Let $\mathscr{L}$ be a simplicial complex and assume that the vertices are ordered by $[n] = \{1, \ldots, n\}$. For each simplex $s \in \mathscr{L}$ define

$$I_{\mathscr{L}}(s) := \{j \in [n] \mid j < \min(s) \text{ and } s \cup \{j\} \in \mathscr{L}\}.$$

Now consider the pairings

$$s \leftrightarrow s \cup \{i\},$$

where $i = \min I_{\mathscr{L}}(s)$ is the smallest element in the set $I_{\mathscr{L}}(s)$, defined whenever $I_{\mathscr{L}}(s) \neq \varnothing$. We call this the *lexicographical matching*.

Due to the $\min I_{\mathscr{L}}(s)$ construction in the lexicographical matching, the indices are decreasing along any path and hence it will be a gradient path, showing that the lexicographical matching is indeed an acyclic partial matching on $\mathscr{L}$.

Example 2.4 Consider the simplicial complex $\mathscr{L}$ depicted in Fig. 2. The complex has 5 vertices, 6 edges and one two-dimensional simplex that is shaded in grey. The red arrows show the lexicographical matching on this simplicial complex: there is an arrow from a simplex s to t iff the pair (s, t) is part of the matching. More explicitly, the lexicographical matching on $\mathscr{L}$ is

$$\Sigma = \{(\{2\}, \{1, 2\}), (\{3\}, \{2, 3\}), (\{4\}, \{1, 4\}), (\{5\}, \{3, 5\}), (\{4, 5\}, \{3, 4, 5\})\}.$$

Note that $\{3, 4\}$ cannot be matched because the set $I_{\mathscr{L}}(\{3, 4\})$ is empty. Also, in any lexicographical matching $\{1\}$ is always critical as there are no vertices with a smaller label and hence the set $I_{\mathscr{L}}(\{1\})$ is empty. So under this matching there are two

critical simplices: {1} and {3, 4}, highlighted in blue in the figure. Hence, if we were computing the homology of this complex, considering only two simplices would be sufficient instead of all 12 which are in $\mathscr{L}$—a significant improvement.

3 Probabilistic tools

In this section we introduce the approximation theorems that are used to study the random variables of interest. In order not to obscure our main results, the proofs are deferred to Sect. 7.

3.1 CLT for dissociated sums

Let n and d be positive integers. For each $i \in [d] =: \{1, 2, \ldots, d\}$, we fix an index set $\mathbb{I}_i \subset [n] \times \{i\}$ and consider the union of disjoint sets $\mathbb{I} := \bigcup_{i\in[d]} \mathbb{I}_i$. Associate to each such $s = (k, i) \in \mathbb{I}$ a real centered random variable X_s and form for each $i \in [d]$ the sum

$$W_i := \sum_{s\in\mathbb{I}_i} X_s.$$

Consider the resulting random vector $W = (W_1, \ldots, W_d) \in \mathbb{R}^d$. The following notion is a natural multivariate generalisation of the dissociated sum from McGinley and Sibson (1975); see also Barbour et al. (1989).

Definition 3.1 We call W a *vector of dissociated sums* if for each $s \in \mathbb{I}$ and $j \in [d]$ there exists a *dependency neighbourhood* $\mathbb{D}_j(s) \subset \mathbb{I}_j$ satisfying three criteria:

(1) the difference $\left(W_j - \sum_{u\in\mathbb{D}_j(s)} X_u\right)$ is independent of X_s;
(2) for each $t \in \mathbb{I}$, the quantity $\left(W_j - \sum_{u\in\mathbb{D}_j(s)} X_u - \sum_{v\in\mathbb{D}_j(t)\setminus\mathbb{D}_j(s)} X_v\right)$ is independent of the pair (X_s, X_t); and finally,
(3) X_s and X_t are independent if $t \notin \bigcup_j \mathbb{D}_j(s)$.

Let W be a vector of dissociated sums as defined above. For each $s \in \mathbb{I}$, by construction, the sets $\mathbb{D}_j(s)$, $j \in [d]$ are disjoint (although for $s \neq t$, the sets $\mathbb{D}_j(s)$ and $\mathbb{D}_j(t)$ may not be disjoint). We write $\mathbb{D}(s) = \bigcup_{j\in[d]} \mathbb{D}_j(s)$ for the disjoint union of these dependency neighbourhoods. With this preamble in place, we state the abstract approximation theorem that is the main ingredient in the proofs of our normal approximation results.

Theorem 3.2 *Let $h \in \mathscr{H}_d$. Consider a standard d-dimensional Gaussian vector $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$. Assume that for all $s \in \mathbb{I}$, we have $\mathbb{E}\{X_s\} = 0$ and $\mathbb{E}\left|X_s^3\right| < \infty$. Then, for any vector of dissociated sums $W \in \mathbb{R}^d$ with a positive semi-definite covariance matrix Σ,*

$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z)\right| \leq B_{3.2}\, |h|_3\,,$$

where $B_{3.2} = B_{3.2.1} + B_{3.2.2}$ *is the sum given by*

$$B_{3.2.1} := \frac{1}{3} \sum_{s \in \mathbb{I}} \sum_{t,u \in \mathbb{D}(s)} \left(\frac{1}{2} \mathbb{E}\,|X_s X_t X_u| + \mathbb{E}\,|X_s X_t|\,\mathbb{E}\,|X_u| \right)$$

$$B_{3.2.2} := \frac{1}{3} \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{v \in \mathbb{D}(t) \setminus \mathbb{D}(s)} \left(\mathbb{E}\,|X_s X_t X_v| + \mathbb{E}\,|X_s X_t|\,\mathbb{E}\,|X_v| \right).$$

The theorem above together with a smoothing technique will be used to prove the following approximation theorem in terms of convex set indicators.

Theorem 3.3 *Consider a standard d-dimensional Gaussian vector* $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$. *For any centered vector of dissociated sums* $W \in \mathbb{R}^d$ *with a positive semi-definite covariance matrix* Σ *and finite third absolute moments we have*

$$\sup_{A \in \mathscr{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{3.2}^{\frac{1}{4}},$$

where the quantity $B_{3.2}$ *as in Theorem* 3.2.

The next result provides a simplification of Theorems 3.2 and 3.3 under the assumption that one uses bounds that are uniform in $s, t, u \in \mathbb{I}$. Its proof follows immediately from writing the sum over $\sum_{s\in\mathbb{I}} \sum_{t,u\in\mathbb{D}(s)}$ as the sum over $\sum_{i\in[d]} \sum_{j\in[d]} \sum_{k\in[d]} \sum_{s\in\mathbb{I}_i} \sum_{t\in\mathbb{D}_j(s)} \sum_{u\in\mathbb{D}_k(s)}$.

Corollary 3.4 *We have the following two bounds:*

(1) Under the assumptions of Theorem 3.2,

$$\left| \mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z) \right| \leq B_{3.4}\,|h|_3\,.$$

(2) Assuming the hypotheses of Theorem 3.3,

$$\sup_{A \in \mathscr{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{3.4}^{\frac{1}{4}}.$$

Here $B_{3.4}$ *is a sum over* $(i, j, k) \in [d]^3$ *of the form*

$$B_{3.4} := \frac{1}{3} \sum_{(i,j,k)} |\mathbb{I}_i|\,\alpha_{ij} \left(\frac{3\alpha_{ik}}{2} + 2\alpha_{jk} \right) \beta_{ijk};$$

and α_{ij} *is the largest value attained by* $\left|\mathbb{D}_j(s)\right|$ *over* $s \in \mathbb{I}_i$, *and*

$$\beta_{ijk} = \max_{s,t,u} \left(\mathbb{E}\,|X_s X_t X_u|\,, \mathbb{E}\,|X_s X_t|\,\mathbb{E}\,|X_u| \right)$$

as (s, t, u) *range over* $\mathbb{I}_i \times \mathbb{I}_j \times \mathbb{I}_k$.

In most of our applications, the variables X_s are centered and rescaled Bernoulli random variables. Hence, the following lemma is useful.

Lemma 3.5 *Let ξ_1, ξ_2, ξ_3 be Bernoulli random variables with expected values μ_1, μ_2, μ_3 respectively. Let $c_1, c_2, c_3 > 0$ be any constants. Consider variables $X_i := c_i(\xi_i - \mu_i)$ for $i = 1, 2, 3$. Then we have*

$$\mathbb{E}\,|X_1X_2X_3| \leq c_1c_2c_3\,\{\mu_1\mu_2(1-\mu_1)(1-\mu_2)\}^{\frac{1}{2}};$$
$$\mathbb{E}\,|X_1X_2|\,\mathbb{E}\,|X_3| \leq c_1c_2c_3\,\{\mu_1\mu_2(1-\mu_1)(1-\mu_2)\}^{\frac{1}{2}}.$$

Proof Note that X_3 can take two values: $-c_3\mu_3$ or $c_3(1-\mu_3)$. As $0 \leq \mu_3 \leq 1$, we have

$$\mathbb{E}\,|X_1X_2|\,\mathbb{E}\,|X_3| \leq c_3\mathbb{E}\,|X_1X_2|;$$

$$\mathbb{E}\,|X_1X_2X_3| \leq c_3\mathbb{E}\,|X_1X_2|.$$

Applying the Cauchy-Schwarz inequality and direct calculation of the second moments gives

$$\mathbb{E}\,|X_1X_2| \leq \left\{\mathbb{E}\left\{X_1^2\right\}\mathbb{E}\left\{X_2^2\right\}\right\}^{\frac{1}{2}} = c_1c_2\,\{\mu_1\mu_2(1-\mu_1)(1-\mu_2)\}^{\frac{1}{2}},$$

which finishes the proof. □

3.2 CLT for *U*-statistics

Here we consider generalised U-statistics, which were first introduced in Janson and Nowicki (1991). The result we derive could be of independent interest but, most importantly, the approximation theorem for simplex counts follows as a consequence. Let $\{\xi_i\}_{1\leq i\leq n}$ be a sequence of of independent random variables taking values in a measurable subset $\mathscr{X} \subseteq \mathscr{U}$ and let $\{Y_{i,j}\}_{1\leq i<j\leq n}$ be an array of of independent random variables taking values in a measurable subset $\mathscr{Y} \subseteq \mathscr{U}$ which is independent of $\{\xi_i\}_{1\leq i\leq n}$. We use the convention that $Y_{i,j} = Y_{j,i}$ for any $i < j$. For example, one can think of X_i as a random label of a vertex i in a random graph where $Y_{i,j}$ is the indicator for the edge connecting i and j. Given a subset $s \subseteq [n]$ of size m, write $s = \{s_1, s_2, \ldots, s_m\}$ such that $s_1 < s_2 < \ldots < s_m$ and set $\mathscr{X}_s = (\xi_{s_1}, \xi_{s_2}, \ldots, \xi_{s_m})$ and $\mathscr{Y}_s = (Y_{s_1,s_2}, Y_{s_1,s_3}, \ldots Y_{s_{m-1},s_m})$. Recall that C_k denotes the set of subsets of $[n]$ which are of size k.

Definition 3.6 Given $1 \leq k \leq n$ and a measurable function $f : \mathscr{X}^k \times \mathscr{Y}^{\binom{k}{2}} \to \mathbb{R}$ define the associated generalised U-statistic by

$$S_{n,k}(f) = \sum_{s\in C_k} f(\mathscr{X}_s, \mathscr{Y}_s).$$

Let $\{k_i\}_{i\in[d]}$ be a collection of positive integers, each being at most n, and for each $i \in [d]$ let $f_i : \mathscr{X}^{k_i} \times \mathscr{Y}^{\binom{k_i}{2}} \to \mathbb{R}$ be a measurable function. We are interested in the joint distribution of the variables $S_{n,k_1}(f_1), S_{n,k_2}(f_2), \dots S_{n,k_d}(f_d)$, which are assumed to have finite mean and variance.

Fix $i \in [d]$. For $s \in \mathbb{I}_i := C_{k_i} \times \{i\}$ define $X_s = \sigma_i^{-1}\left(f_i(\mathscr{X}_s, \mathscr{Y}_s) - \mu_s\right)$, where $\mu_s = \mathbb{E}\{f_i(\mathscr{X}_s, \mathscr{Y}_s)\}$ and $\sigma_i^2 = \mathrm{Var}(S_{n,k_i}(f_i))$. Now let $W_i = \sum_{s\in\mathbb{I}_i} X_s$ be a random variable and write $W = (W_1, W_2, \dots W_d) \in \mathbb{R}^d$. By construction, W_i has mean 0 and variance 1.

Assumption 3.7 We assume that

(1) For any $i \in [d]$ there is some $\alpha_i > 0$ such that for all $s, t \in \mathbb{I}_i$, the variables $f_i(\mathscr{X}_s, \mathscr{Y}_s)$, $f_i(\mathscr{X}_t, \mathscr{Y}_t)$ are either independent or $\mathrm{Cov}(f_i(\mathscr{X}_s, \mathscr{Y}_s), f_i(\mathscr{X}_t, \mathscr{Y}_t)) > \alpha_i$.
(2) There is $\beta \geq 0$ such that for any $i, j, l \in [d]$ and any $s \in \mathbb{I}_i, t \in \mathbb{I}_j, u \in \mathbb{I}_l$ we have

$$\mathbb{E}\left|\{f_i(\mathscr{X}_s, \mathscr{Y}_s) - \mu_s\}\left\{f_j(\mathscr{X}_t, \mathscr{Y}_t) - \mu_t\right\}\{f_l(\mathscr{X}_u, \mathscr{Y}_u) - \mu_u\}\right| \leq \beta$$

as well as

$$\mathbb{E}\left|\{f_i(\mathscr{X}_s, \mathscr{Y}_s) - \mu_s\}\left\{f_j(\mathscr{X}_t, \mathscr{Y}_t) - \mu_t\right\}\right| \mathbb{E}\left|f_l(\mathscr{X}_u, \mathscr{Y}_u) - \mu_u\right| \leq \beta.$$

The first assumption is not necessary but very convenient and we use it to derive a lower bound for the variance σ_i^2. A normal approximation theorem can be proven in our framework when the assumption does not hold and a sufficiently large lower bound for the variance is acquired in a different way. Similarly, we use the second assumption to get a convenient bound on mixed moments. In order to maintain the generality and simplicity of the proofs, we work under Assumption 3.7.

We also consider the important special case that the functions in Definition 3.6 only depend on the second component, so that the sequence $\{\xi_i\}_{i\in[n]}$ can be ignored. Hence, we add an additional assumption.

Assumption 3.8 We assume that the functions f_i only depend on the variables $\left\{Y_{i,j}\right\}$ for $1 \leq i < j \leq n$. That is, we can write $f_i : \mathscr{Y}^{\binom{k_i}{2}} \to \mathbb{R}$.

Such functions appear naturally, for example, when counting subgraphs in an inhomogeneous Bernoulli random graph. An example of such generalised U-statistic is simplex counts in $\mathbf{X}(n, p)$ and is worked out in Sect. 6. We recall, for the purposes of the following result, that n is the number of variables in the sequence $\{\xi_i\}_{1\leq i\leq n}$ and $\binom{n}{2}$ is the number of variables in the sequence $\{Y_{i,j}\}_{1\leq i<j\leq n}$ from Definition 3.6.

Theorem 3.9 *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$ and let $h \in \mathscr{H}_d$. Assume W with covariance matrix Σ satisfies Assumption 3.7. Then*

$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \leq |h|_3\, B_{3.9} n^{-\gamma};$$

and

$$\sup_{A\in\mathscr{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{3.9}^{\frac{1}{4}} n^{-\frac{\gamma}{4}}.$$

Here,

$$B_{3.9} = \frac{2^{\delta}\beta}{3} \sum_{i,j,l=1}^{d} \frac{k_i^{\min(k_i,k_j)+1}}{k_i!\sqrt{\alpha_i\alpha_j\alpha_l}} \left(k_i^{\min(k_i,k_l)+1} + k_j^{\min(k_j,k_l)+1}\right) K_i K_j K_l$$

and

$$K_i = (2k_i^2 - k_i)^{-\frac{k_i}{2}+\frac{1}{2}}.$$

If only Assumption 3.7 is satisfied, then $\gamma = \frac{1}{2}$ *and* $\delta = 1$. *If additionally Assumption 3.8 is also satisfied, then* $\gamma = 1$ *and* $\delta = 4$.

4 Critical simplex counts for lexicographical Morse matchings

Now we attend to our motivating problem, critical simplex counts. Consider the random simplicial complex $\mathbf{X}(n, p)$. In this section we study the joint distribution of critical simplices in different dimensions with respect to the lexicographical matching on $\mathbf{X}(n, p)$. We start with the following lemma, which is an immediate consequence of Definition 2.3, allowing us to write down the variables of interest in terms of the edge indicators.

Lemma 4.1 *Let* $\mathscr{L}$ *be a simplicial complex endowed with the lexicographical acyclic partial matching, and consider a simplex* $t \in \mathscr{L}$ *with minimal vertex* $i \in [n]$. *Then,* t *is matched with*

(1) one of its co-faces if and only if there exists some $j < i$ *for which* $t \cup \{j\} \in \mathscr{L}$; *and,*
(2) one of its faces if and only for all $j < i$ *we have* $(t \setminus \{i\}) \cup \{j\} \notin \mathscr{L}$.

For any pair of integers $1 \leq i < j \leq n$ let $Y_{i,j} := \mathbb{1}(\{i, j\} \in \mathbf{X}(n, p))$ be the edge indicator. Fix $s \in C_k$. Define the variables $X_s^+ = \mathbb{1}$ (s matches with its coface given it is a simplex) and $X_s^- = \mathbb{1}$ (s matches with its face given it is a simplex). The events that the two variables indicate are disjoint. By Lemma 4.1 we can see that $X_s^+ = 1 - \prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b\in s} Y_{a,b}\right)$ and $X_s^- = \prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b\in s_-} Y_{a,b}\right)$, where $s_- := s\setminus\{\min(s)\}$. Hence,

$$\begin{aligned}
&\mathbb{1}(s \text{ is a critical simplex}) = \mathbb{1}(s \in \mathbf{X}(n, p))\,(1 - (X_s^+ + X_s^-)) \\
&= \prod_{i\neq j\in s} Y_{i,j} \left[\prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b\in s} Y_{a,b}\right) - \prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b\in s_-} Y_{a,b}\right)\right].
\end{aligned}$$

Thus, the random variable of interest, counting the number of $(k-1)$-simplices that are critical under the lexicographical matching, is

$$T_k = \sum_{s \in C_k} \prod_{i \neq j \in s} Y_{i,j} \left[\prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b \in s} Y_{a,b} \right) - \prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b \in s_-} Y_{a,b} \right) \right]. \quad (4.1)$$

Note that this random variable does not fit into the framework of generalised U-statistics because the summands in T_k depend not only on the variables that are indexed by the subset s. Therefore, Theorem 3.9 cannot be applied here.

4.1 Mean and variance

Lemma 4.2 *For any $1 \leq k \leq n-1$ we have:*

$$p^{\binom{k+1}{2}+k} \binom{n-2}{k} (1-p) \leq \mathbb{E}\{T_{k+1}\} \leq p^{\binom{k+1}{2}-k-1} \binom{n-1}{k} (1-p).$$

Proof

$$\begin{aligned}
\mathbb{E}\{T_{k+1}\} &= \sum_{l=1}^{n-k} \sum_{\substack{s \in C_{k+1} \\ \min(s)=l}} \mathbb{E}\left\{ \prod_{i \neq j \in s} Y_{i,j} \left[\prod_{a=1}^{l-1} \left(1 - \prod_{b \in s} Y_{a,b} \right) - \prod_{a=1}^{l-1} \left(1 - \prod_{b \in s_-} Y_{a,b} \right) \right] \right\} \\
&= p^{\binom{k+1}{2}} \sum_{l=1}^{n-k} \sum_{\substack{s \in C_{k+1} \\ \min(s)=l}} \left\{ (1-p^{k+1})^{l-1} - (1-p^k)^{l-1} \right\} \\
&= p^{\binom{k+1}{2}} \sum_{l=0}^{n-k-1} \binom{n-l-1}{k} \left\{ (1-p^{k+1})^{l} - (1-p^k)^{l} \right\} \\
&\leq p^{\binom{k+1}{2}} \binom{n-1}{k} \sum_{l=0}^{\infty} \left\{ (1-p^{k+1})^{l} - (1-p^k)^{l} \right\} \\
&= p^{\binom{k+1}{2}-k-1} \binom{n-1}{k} (1-p).
\end{aligned}$$

Moreover,

$$\begin{aligned}
\mathbb{E}\{T_{k+1}\} &= p^{\binom{k+1}{2}} \sum_{l=0}^{n-k-1} \binom{n-l-1}{k} \left\{ (1-p^{k+1})^{l} - (1-p^k)^{l} \right\} \\
&\geq p^{\binom{k+1}{2}} \binom{n-2}{k} \left\{ (1-p^{k+1})^{1} - (1-p^k)^{1} \right\} \\
&= p^{\binom{k+1}{2}+k} \binom{n-2}{k} (1-p).
\end{aligned}$$

□

In this example, bounding the variance is not immediate. The proof of the following Lemma 4.3 are long (and not particularly insightful) calculations, which are deferred to the Appendix. In Lemma 4.3 the constant could have been made explicit at the expense of a lengthy calculation while an explicit expression for the variance is given in Lemma A.1.

Lemma 4.3 *For a fixed integer $1 \leq k \leq n-1$ and $p \in (0,1)$ there is a constant $C_{p,k} > 0$ independent of n and a natural number $N_{p,k}$ such that for any $n \geq N_{p,k}$:*

$$\mathrm{Var}(T_{k+1}) \geq C_{p,k} n^{2k}.$$

Just knowing the expectation and the variance can already give us some information about the variable. For example, we obtain the following proposition. This proposition shows that considering only a subset of the simplices already gives a good approximation for the critical simplex counts.

Proposition 4.4 *Fix $k \in [n]$. Let $K \leq n-k$ and set the random variable:*

$$T_{k+1}^{K} := \sum_{\substack{s \in C_{k+1} \\ \min(s) \leq K}} \prod_{i \neq j \in s} Y_{i,j} \left[\prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b \in s} Y_{a,b} \right) - \prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b \in s_-} Y_{a,b} \right) \right].$$

If $K = K(n) = \omega(\ln^{1+\epsilon}(n))$ for any $\epsilon > 0$, then the variable $T_{k+1} - T_{k+1}^{K}$ vanishes with high probability, provided that p and k stay constant.

Proof A similar calculation to that for Lemma 4.2 shows that:

$$\begin{aligned}
\mathbb{E}\left\{ T_{k+1} - T_{k+1}^{K} \right\} &= \sum_{i=K+1}^{n-k} \binom{n-i}{k} p^{\binom{k+1}{2}} \left\{ (1-p^{k+1})^{i-1} - (1-p^{k})^{i-1} \right\} \\
&\leq \binom{n}{k} p^{\binom{k+1}{2}} (1-p^{k+1})^{K} \sum_{i=0}^{\infty} (1-p^{k+1})^{i} \\
&\leq p^{\binom{k+1}{2}-k-1} \frac{n^k}{k!} (1-p^{k+1})^{K}.
\end{aligned}$$

Using Markov's inequality, we get:

$$\mathbb{P}(T_{k+1} - T_{k+1}^{K} \geq 1) \leq p^{\binom{k+1}{2}-k-1} \frac{n^k}{k!} (1-p^{k+1})^{K},$$

which asymptotically vanishes as long as $K = \omega(\ln^{1+\epsilon}(n))$. □

4.2 Approximation theorem

For $i \in [d]$, recall a random variable counting i-simplices in $\mathbf{X}(n, p)$ that are critical under the lexicographical matching, as given in (4.1). We write for the i-th index set $\mathbb{I}_i := C_{i+1} \times \{i\}$. For $s = (\phi, i) \in \mathbb{I}_i$ we write

$$\mu_s = p^{\binom{i+1}{2}} \left((1 - p^{i+1})^{\min(\phi)-1} - (1 - p^i)^{\min(\phi)-1} \right)$$

and $\sigma_i = \sqrt{\mathrm{Var}(T_{i+1})}$. Let

$$X_s = \sigma_i^{-1} \left\{ \prod_{i \neq j \in \phi} Y_{i,j} \left[\prod_{a=1}^{\min(\phi)-1} \left(1 - \prod_{b \in \phi} Y_{a,b} \right) - \prod_{a=1}^{\min(\phi)-1} \left(1 - \prod_{b \in \phi_-} Y_{a,b} \right) \right] - \mu_s \right\}.$$

Let $W_i = \sum_{s \in \mathbb{I}_i} X_s$ and $W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d$. For bounds that asymptotically go to zero for this example, we use Theorems 3.2 and 3.3 directly: the uniform bounds from Corollary 3.4 are not fine enough here. We note that here, due to the requirement of criticality, two summands X_s and X_u become dependent as soon as the corresponding subsets share a vertex. This is in contrast to simplex counts, which required an overlap of at least two vertices.

Theorem 4.5 *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$ and Σ be the covariance matrix of W.*

(1) Let $h \in \mathscr{H}_d$. Then there is a constant $B_{4.5.1} > 0$ independent of n and a natural number $N_{4.5.1}$ such that for any $n \geq N_{4.5.1}$ we have

$$\left| \mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z) \right| \leq B_{4.5.1} \, |h|_3 \, n^{-1}.$$

(2) There is a constant $B_{4.5.2} > 0$ independent of n and a natural number $N_{4.5.2}$ such that for any $n \geq N_{4.5.2}$ we have

$$\sup_{A \in \mathscr{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq B_{4.5.2} n^{-\frac{1}{4}}.$$

Proof It is clear that W satisfies the conditions of Theorems 3.2 and 3.3 for any $s = (\phi, i) \in \mathbb{I}_i$ setting

$$\mathbb{D}_j(s) = \left\{ (\psi, j) \in \mathbb{I}_j \mid |\phi \cap \psi| \geq 1 \right\}.$$

We apply Theorems 3.2 and 3.3. For the bounds on the quantity $B_{3.2}$ from Theorems 3.2 and 3.3 we use Lemma 3.5 and Lemma 4.3. We write C for an unspecified positive constant that does not depend on n. Also, we assume here that n is large enough for the bound in Lemma 4.3 to apply. Let $\mu(i, a) = p^{\binom{i+1}{2}} \left((1 - p^{i+1})^{a-1} - (1 - p^i)^{a-1} \right)$.

Then we have:

$$
\begin{aligned}
B_{3.2} &\le \frac{1}{3}\sum_{i,j,k=1}^{d}\sum_{a=1}^{n-i}\sum_{\substack{\phi\in C_{i+1}\\ \min(\phi)=a}}\sum_{b=1}^{n-j}\sum_{\substack{(\psi,j)\in\mathbb{D}_j((\phi,i))\\ \min(\psi)=b}}\\
&\Bigg\{\sum_{r\in\mathbb{D}_k((\phi,i))}\frac{3}{2}(\sigma_i\sigma_j\sigma_k)^{-1}\{\mu(i,a)\mu(j,b)(1-\mu(i,a))(1-\mu(j,b))\}^{\frac{1}{2}}\\
&\quad+\sum_{r\in\mathbb{D}_k((\psi,j))}(\sigma_i\sigma_j\sigma_k)^{-1}\{\mu(i,a)\mu(j,b)(1-\mu(i,a))(1-\mu(j,b))\}^{\frac{1}{2}}\Bigg\}\\
&\le \sum_{i,j,k=1}^{d}\sum_{a=1}^{n-i}\sum_{b=1}^{n-j}Cn^{i+j-1}n^{k}n^{-i-j-k}\\
&\quad\times\left\{(1-p^{i+1})^{a-1}(1-p^{j+1})^{b-1}+(1-p^{i})^{a-1}(1-p^{j})^{b-1}\right\}^{\frac{1}{2}}\\
&\le Cn^{-1}\sum_{i,j,k=1}^{d}\sum_{a=1}^{\infty}\sum_{b=1}^{\infty}\Bigg[\left\{(1-p^{i+1})^{a-1}(1-p^{j+1})^{b-1}\right\}^{\frac{1}{2}}\\
&\quad+\left\{(1-p^{i})^{a-1}(1-p^{j})^{b-1}\right\}^{\frac{1}{2}}\Bigg]\\
&\le Cn^{-1}d^{3}\left\{\frac{1}{(1-\sqrt{1-p^{d+1}})^{2}}+\frac{1}{(1-\sqrt{1-p^{d}})^{2}}\right\}\le Cn^{-1}.
\end{aligned}
$$

□

Remark 4.6 The relevance of understanding the number of critical simplices in the context of applied and computational topology is as follows. We assume that $p \in (0, 1)$ and $k \in \{1, 2, \ldots\}$ are constants.

(1) As seen in Lemma 4.2, the expected number of critical k-simplices under the lexicographical matching is one power of n smaller than the total number of k-simplices in $\mathbf{X}(n, p)$.
(2) From Proposition 4.4, it is with high probability that in $\mathbf{X}(n, p)$ all k-simplices $s \in \mathbf{X}(n, p)$ with $\min(s) = \omega(\ln^{1+\epsilon}(n))$ for any fixed $\epsilon > 0$ are not critical.

5 Simplex counts in links

Consider the random simplicial complex $\mathbf{X}(n, p)$. For $1 \le i < j \le n$ define the edge indicator $Y_{i,j} := \mathbb{1}(\{i, j\} \in \mathbf{X}(n, p))$. In this section we study the count of $(k-1)$-simplices that would be in the link of a fixed subset $t \subseteq [n]$ if the subset spanned a simplex in $\mathbf{X}(n, p)$. Given that t is a simplex, the variable counts the number of $(k-1)$-simplices in $\mathbf{lk}(t)$. Thus, the random variable of interest is

$$T_k^t = \sum_{s \in C_k} \left\{ \mathbb{1}\left(t \cap s = \varnothing\right) \prod_{i \neq j \in s} Y_{i,j} \prod_{i \in s, j \in t} Y_{i,j} \right\}. \tag{5.1}$$

Note that the product $\prod_{i \in s, j \in t} Y_{i,j}$ ensures that $t \cup s$ is a simplex if t spans a simplex.

Remark 5.1 The random variable T_k^t does not fit into the framework of generalised U-statistics, because the summands depend not only on the variables that are indexed by the subset s and we do not sum over all subsets s but rather only the ones that do not intersect t. Hence, Theorem 3.9 does not apply here.

Moreover, note that given the number of vertices of the link of a simplex t, the conditional distribution of the link of t is again $\mathbf{X}(n, p)[n'][p]$, where n' is a random variable equal to the number of vertices in the link. If we are interested in such a conditional distribution, the results proved later in Sect. 6 apply. However, in this section we study the number of simplices in the link of t given that t is a simplex rather than given the number of vertices of the link of t. Such a random variable behaves differently from the simplex counts in $\mathbf{X}(n, p)$. For example, the summands of T_k^t have a different dependence structure compared to the summands of T_k (see Eq. 6.1 below). As a result, the approximation bounds are of different order.

It is natural to ask whether the results obtained in this section follow from those of Sect. 6 below. This might well be the case, but the answer is not straightforward. One could derive an approximation for the number of simplices in $\mathbf{lk}(t)$ given the number of vertices in the link; the variable T_k^t could then be approximated by a mixture, induced by the distribution of the number of vertices in the link (which is binomial). However, applying this approach naïvely yields bounds that do not converge to zero. While it is certainly possible that a different approach would succeed, we prefer to prove the approximation directly.

5.1 Mean and variance

It is easy to see that for any positive integer k and $t \subseteq [n]$,

$$\mathbb{E}\{T_{k+1}^t\} = \binom{n - |t|}{k+1} p^{\binom{k+1}{2} + |t|(k+1)} =: \binom{n - |t|}{k+1} \mu_{k+1}^t$$

since there are $\binom{n-|t|}{k+1}$ choices for $s \in C_{k+1}$ such that $s \cap t = \varnothing$. Next we derive a lower bound on the variance.

Lemma 5.2 *For any fixed $1 \leq k \leq n-1$ and $t \subseteq [n]$ we have:*

$$\mathrm{Var}(T_{k+1}^t) \geq (k+1) \binom{n - |t|}{2k+1} \binom{2k+1}{k} (\mu_{k+1}^t)^2 \left\{ p^{-|t|} - 1 \right\}.$$

Proof First let us calculate $\mathrm{Cov}(T_{k+1}^t, T_{l+1}^t)$. For fixed subsets $s \in C_{k+1}$ and $u \in C_{l+1}$ if $|s \cap u| = 0$, then the corresponding variables $\prod_{i \neq j \in s} Y_{i,j} \prod_{i \in s, j \in t} Y_{i,j}$ and $\prod_{i \neq j \in u} Y_{i,j} \prod_{i \in u, j \in t} Y_{i,j}$ are independent and so have zero covariance.

For $1 \le m \le l+1$, the number of pairs of subsets $s \in C_{k+1}$ and $u \in C_{l+1}$ such that $s \cap t = \varnothing = u \cap t$ and $|s \cap u| = m$ is $\binom{n-|t|}{k+1}\binom{k+1}{m}\binom{n-|t|-k-1}{l+1-m}$. Since each summand is non-negative, we lower bound by the $m = 1$ summand and get (with $\binom{1}{2} := 0$)

$$\begin{aligned}
&\mathrm{Cov}(T^t_{k+1}, T^t_{l+1}) \\
&= \sum_{m=1}^{l+1} \binom{n-|t|}{k+1}\binom{k+1}{m}\binom{n-|t|-k-1}{l+1-m}\left\{\mu^t_{k+1}\mu^t_{l+1}p^{-\binom{m}{2}}p^{-|t|m} - \mu^t_{k+1}\mu^t_{l+1}\right\} \\
&\ge \binom{n-|t|}{k+1}(k+1)\binom{n-|t|-k-1}{l}\mu^t_{k+1}\mu^t_{l+1}\left\{p^{-|t|}-1\right\} \\
&= (k+1)\binom{n-|t|}{l+k+1}\binom{l+k+1}{l}\mu^t_{k+1}\mu^t_{l+1}\left\{p^{-|t|}-1\right\}.
\end{aligned}$$

Taking $l = k$ completes the proof. □

5.2 Approximation theorem

For a multivariate normal approximation of counts given in Equation (5.1), we write $\sigma_i = \sqrt{\mathrm{Var}(T^t_{i+1})}$ and $C^t_{i+1} = \{\, \phi \in C_{i+1} \mid \phi \cap t = \varnothing \,\}$, as well as $\mathbb{I}_i := C^t_{i+1} \times \{i\}$. For $s = (\phi, i) \in \mathbb{I}_i$ define

$$X_s = \sigma_i^{-1}\Big(\prod_{i \ne j \in \phi} Y_{i,j} \prod_{a \in \phi, b \in t} Y_{a,b} - \mu^t_{i+1}\Big).$$

It is clear that $\mathbb{E}\{X_s\} = 0$. Let $W^t_i = \sum_{s \in \mathbb{I}_i} X_s$ and $W^t = (W^t_1, W^t_2, \ldots, W^t_d) \in \mathbb{R}^d$. Then we have the following approximation theorem.

Theorem 5.3 *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$ and Σ be the covariance matrix of W^t.*

(1) Let $h \in \mathscr{H}_d$. Then

$$\left|\mathbb{E}h(W^t) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \le |h|_3\, B_{5.3}(n-|t|)^{-\frac{1}{2}}.$$

(2) Moreover,

$$\sup_{A \in \mathscr{K}} |\mathbb{P}(W^t \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}}Z \in A)| \le 2^{\frac{7}{2}}3^{-\frac{3}{4}}d^{\frac{3}{16}}B^{\frac{1}{4}}_{5.3}(n-|t|)^{-\frac{1}{8}}.$$

Here

$$B_{5.3} = \frac{7}{6}(2d+1)^{5d+\frac{17}{2}}(p^{-|t|}-1)^{-\frac{3}{2}}p^{-(d+1)(d+2|t|)}.$$

Proof It is clear that W^t satisfies the conditions of Corollary 3.4 with the dependency neighbourhood $\mathbb{D}_j(s) = \left\{\, (\psi, j) \in \mathbb{I}_j \mid |\phi \cap \psi| \ge 1 \,\right\}$ for any $s = (\phi, i) \in \mathbb{I}_i$. So we aim to bound the quantity $B_{3.4}$ from the corollary.

Given $\phi \in C^t_{i+1}$ and $m \leq \min(i+1, j+1)$ there are $\binom{i+1}{m}\binom{n-|t|-i-1}{j+1-m}$ subsets $\psi \in C^t_{j+1}$ such that $|\phi \cap \psi| = m$. Therefore, for any $i, j \in [d]$ and $s \in \mathbb{I}_i$ we have

$$\begin{aligned}|\mathbb{D}_j(s)| &= \sum_{m=1}^{\min(i,j)+1} \binom{i+1}{m}\binom{n-|t|-i-1}{j+1-m} \\ &\leq (i+1)^{\min(i,j)+2}(n-|t|)^j \\ &\leq (d+1)^{d+2}(n-|t|)^j \end{aligned} \tag{5.2}$$

giving a bound for α_{ij}. For a bound on β_{ijk}, applying Lemma 3.5, for any $i, j, k \in [d]$ and $s \in \mathbb{I}_i, u \in \mathbb{I}_j, v \in \mathbb{I}_k$ we get

$$\mathbb{E}\,|X_s X_u X_v| \leq (\sigma_i\sigma_j\sigma_k)^{-1}\left\{\mu^t_{i+1}\mu^t_{j+1}(1-\mu^t_{i+1})(1-\mu^t_{j+1})\right\}^{\frac{1}{2}}; \tag{5.3}$$

$$\mathbb{E}\,|X_s X_u|\,\mathbb{E}\,|X_v| \leq (\sigma_i\sigma_j\sigma_k)^{-1}\left\{\mu^t_{i+1}\mu^t_{j+1}(1-\mu^t_{i+1})(1-\mu^t_{j+1})\right\}^{\frac{1}{2}}. \tag{5.4}$$

Now we apply Corollary 5.2 and get

$$\begin{aligned}\sigma_i^2 &\geq (i+1)\binom{n-|t|}{2i+1}\binom{2i+1}{i}(\mu^t_{k+1})^2\left\{p^{-|t|}-1\right\} \\ &\geq \frac{(n-|t|)^{2i+1}}{(2d+1)^{d+1}d^d}(\mu^t_{k+1})^2\left\{p^{-|t|}-1\right\}.\end{aligned}$$

Taking both sides of the inequality to the power of $-\frac{1}{2}$ we get for any $i \in [d]$

$$\sigma_i^{-1} \leq (n-|t|)^{-i-\frac{1}{2}}(2d+1)^{\frac{d+1}{2}}d^{\frac{d}{2}}(\mu^t_{k+1})^{-1}\left\{p^{-|t|}-1\right\}^{-\frac{1}{2}}. \tag{5.5}$$

Using Eqs. (5.2)–(5.5) to bound $B_{3.4}$ from Corollary 3.4 we get:

$$\begin{aligned}B_{3.4} &\leq \frac{7}{6}\sum_{i,j,k=1}^{d}\binom{n-|t|}{i+1}(d+1)^{2d+4}(n-|t|)^{j+k}(\sigma_i\sigma_j\sigma_k)^{-1} \\ &\quad\left\{\mu^t_{i+1}\mu^t_{j+1}(1-\mu^t_{i+1})(1-\mu^t_{j+1})\right\}^{\frac{1}{2}} \\ &\leq \frac{7}{6}\sum_{i,j,k=1}^{d}(n-|t|)^{i+j+k+1}(d+1)^{2d+4}(n-|t|)^{-i-j-k-\frac{3}{2}}(2d+1)^{\frac{3d+3}{2}}d^{\frac{3d}{2}} \\ &\quad(p^{-|t|}-1)^{-\frac{3}{2}}(\mu^t_{k+1}\mu^t_{i+1}\mu^t_{j+1})^{-1}\left\{\mu^t_{i+1}\mu^t_{j+1}(1-\mu^t_{i+1})(1-\mu^t_{j+1})\right\}^{\frac{1}{2}} \\ &\leq (n-|t|)^{-\frac{1}{2}}\frac{7}{6}(2d+1)^{5d+\frac{11}{2}}(p^{-|t|}-1)^{-\frac{3}{2}}\sum_{i,j,k=1}^{d}\left((\mu^t_{i+1}\mu^t_{j+1})^{-1}(\mu^t_{k+1})^{-2}\right)^{\frac{1}{2}}\end{aligned}$$

$$\leq \left\{ \frac{7}{6}(2d+1)^{5d+\frac{17}{2}}(p^{-|t|}-1)^{-\frac{3}{2}}p^{-(d+1)(d+2|t|)} \right\}(n-|t|)^{-\frac{1}{2}}.$$

□

Remark 5.4 Recall that $\mathbb{E}\left\{T^t_{k+1}\right\} = \binom{n-|t|}{k+1}p^{\binom{k+1}{2}+|t|(k+1)}$. By Stirling's approximation, if $p \in (0,1)$ is a constant, then $\max(k,|t|) = \Omega(\ln^{1+\epsilon}(n))$ for any positive ϵ forces the expectation to go to 0 asymptotically. Hence, by Markov's inequality, with high probability there are no k-simplices in the link of t as long as $\max(k,|t|)$ is of order $\ln^{1+\epsilon}(n)$ or larger for any $\epsilon > 0$ for a constant p.

Recall that in Theorem 5.3 we count all simplices up to dimension d in the link of t. Note that if $\max(d^2, d|t|) = O(\ln^{1-\epsilon}(n))$ for any $\epsilon > 0$, then the bounds in Theorem 5.3 tend to 0 as n tends to infinity as long as $p \in (0,1)$ stays constant. In particular, if d is a constant, Theorem 5.3 gives an approximation for all sizes of t for which the approximation is needed.

6 Simplex counts in $\mathbf{X}(n,p)$

In this section we apply Theorem 3.9 to approximate simplex counts. Consider $G \sim \mathbf{G}(n,p)$. For $1 \leq x < y \leq n$ let $Y_{x,y} := \mathbb{1}(x \sim y)$ be the edge indicator. In this section we are interested in the $(i+1)$-clique count in $\mathbf{G}(n,p)$ or, equivalently, the i-simplex count in $\mathbf{X}(n,p)$, given by

$$T_{i+1} = \sum_{s \in C_{i+1}} \prod_{x \neq y \in s} Y_{x,y}. \tag{6.1}$$

Let $\mathscr{Y}^{i+1} = \{0,1\}^{i+1}$ and let $f_i : \mathscr{Y}^{i+1} \to \mathbb{R}$ be the function

$$f_i(\mathscr{Y}_s) = \prod_{Y_{x,y} \in \mathscr{Y}_s} Y_{x,y}.$$

Then the associated generalised U-statistic $S_{n,i+1}(f_i)$ equals the $(i+1)$-clique count T_{i+1}, as given by Eq. (6.1). To apply Theorem 3.9 we need to center and rescale our variables. It is easy to see that $\mathbb{E}\{f_i(\mathscr{Y}_\phi)\} = p^{\binom{i+1}{2}}$ if $\phi \in C_{i+1}$. We let $I_i := C_{i+1} \times \{i\}$ and for $s = (\phi, i) \in \mathbb{I}_i$ we define $X_s := \sigma^{-1}\left(f_i(\mathscr{Y}_\phi) - p^{\binom{i+1}{2}}\right)$ and $W_i = \sum_{s \in \mathbb{I}_i} X_s$. Now the vector of interest is $W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d$. This brings us to the next approximation theorem.

Corollary 6.1 *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$ and Σ be the covariance matrix of W.*

(1) Let $h \in \mathscr{H}_d$. Then

$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \leq |h|_3 \, B_{6.1} n^{-1}.$$

(2) Moreover,

$$\sup_{A\in\mathscr{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{6.1}^{\frac{1}{4}} n^{-\frac{1}{4}}.$$

Here

$$B_{6.1} = \frac{16}{3} d^{2d+5} p^{-3\binom{d+1}{2}+1} (1 - p^{\binom{d+1}{2}})(p^{-1} - 1)^{-\frac{3}{2}}.$$

Proof Firstly, observe that for any $\phi, \psi \in C_{i+1}$ for which $|\phi \cap \psi| \leq 1$ the covariance vanishes, while if $|\phi \cap \psi| \geq 2$ the covariance is non-zero, and we have

$$\mathrm{Cov}(f_i(\mathscr{Y}_\phi), f_i(\mathscr{Y}_\psi)) = p^{2\binom{i+1}{2}-\binom{|\phi\cap\psi|}{2}} - p^{2\binom{i+1}{2}} \geq p^{2\binom{i+1}{2}}(p^{-1} - 1).$$

For $s = (\phi, i) \in \mathbb{I}_i$ write $\hat{X}_s = f_i(\mathscr{Y}_\phi) - p^{\binom{i+1}{2}}$. Then by Lemma 3.5 we get:

$$\mathbb{E}\left|\hat{X}_s \hat{X}_t\right| \mathbb{E}\left|\hat{X}_u\right| \leq \left\{p^{\binom{i+1}{2}+\binom{j+1}{2}}(1 - p^{\binom{i+1}{2}})(1 - p^{\binom{j+1}{2}})\right\}^{\frac{1}{2}};$$
$$\mathbb{E}\left|\hat{X}_s \hat{X}_t \hat{X}_u\right| \leq \left\{p^{\binom{i+1}{2}+\binom{j+1}{2}}(1 - p^{\binom{i+1}{2}})(1 - p^{\binom{j+1}{2}})\right\}^{\frac{1}{2}}.$$

Since $\left\{p^{\binom{i+1}{2}+\binom{j+1}{2}}(1 - p^{\binom{i+1}{2}})(1 - p^{\binom{j+1}{2}})\right\}^{\frac{1}{2}} \leq p(1 - p^{\binom{d+1}{2}})$, we see that Assumption 3.7 holds. Assumption 3.8 also holds and therefore we can apply Theorem 3.9 with $k_i = i+1$, $K_i = (2(i+1)^2 - 2(i+1))^{-\frac{1}{2}(i+1)+1}$, $\alpha_i = p^{2\binom{i+1}{2}}(p^{-1} - 1)$, and $\beta = p(1 - p^{\binom{d+1}{2}})$. Using the bounds $K_i \leq 1$ as well as $2 \leq k_i^{\min(k_i,k_j)+1} \leq d^{d+1}$, and $\sqrt{\alpha_i} \geq p^{\binom{d+1}{2}}\sqrt{p^{-1} - 1}$ finishes the proof. □

Remark 6.2 It is easy to show that with high probability there are no large cliques in $\mathbf{G}(n, p)$ for $p < 1$ constant. To see this, the expectation of the number of k-cliques is $\binom{n}{k} p^{\binom{k}{2}}$. By Stirling's approximation, $k = \Omega(\ln^{1+\epsilon}(n))$ for any positive ϵ forces the expectation to go to 0 asymptotically. Hence, by Markov's inequality, for any $\epsilon > 0$, with high probability there are no cliques of order $\ln^{1+\epsilon}(n)$ or larger.

Recall that in Corollary 6.1 the size of the maximal clique we count is $d + 1$. Note that if $d = O(\ln^{\frac{1}{2}-\epsilon}(n))$ for any $\epsilon > 0$, then the bounds in Corollary 6.1 tend to 0 as n tends to infinity as long as $p \in (0, 1)$ stays constant. This value might seem quite small but in the light of there not being any cliques of order $\ln^{1+\epsilon}(n)$ with high probability, this is meaningfully large.

Remark 6.3 Note that in Corollary 6.1 we use a multivariate normal distribution with covariance Σ, which is the covariance of W when n is finite and it differs from the limiting covariance, as mentioned in Reinert and Röllin (2010). To approximate W with the limiting distribution, in the spirit of Reinert and Rollin (2010, Proposition 3) one could proceed in two steps: use the existing theorems to approximate W with ΣZ

and then approximate ΣZ with $\Sigma_L Z$ where Σ_L is the limiting covariance, which is non-invertible, as observed in Janson and Nowicki (1991).

Remark 6.4 Corollary 6.1 generalises the result (Reinert and Röllin 2010, Proposition 2) beyond the case when $d = 2$ and we get a bound of the same order of n. Kaur and Roollin (2021, Theorem 3.1) considers centered subgraph counts in a random graph associated to a graphon. If we take the graphon to be constant, the associated random graph is just $\mathbf{G}(n, p)$. Compared to Kaur and Roollin (2021, Theorem 3.1) we place weaker smoothness conditions on our test functions. However, we make use of the special structure of cliques whereas Kaur and Roollin (2021, Theorem 3.1) applies to any centered subgraph counts. Translating Kaur and Roollin (2021, Theorem 3.1) into a result for uncentered subgraph counts, as we provide here in the special case of clique counts, is not trivial for general d.

However, it should be possible to extend our results, using the same abstract approximation theorem, beyond the random clique complex to Linial-Meshulam random complexes (Linial and Meshulam 2006) or even more general multiparameter random complexes (Costa and Farber 2016). We shall consider this conjecture in future work.

7 A proof of the multivariate CLT for dissociated Sums and *U*-statistics

7.1 A proof of the multivariate CLT

Throughout this subsection, $W \in \mathbb{R}^d$ is a vector of dissociated sums in the sense of Definition 3.1, with covariance matrix whose entries are $\Sigma_{ij} = \mathrm{Cov}(W_i, W_j)$ for $(i, j) \in [d]^2$. For each $s \in \mathbb{I}$ we denote by $\mathbb{D}(s) \subset \mathbb{I}$ the disjoint union $\bigcup_{j=1}^d \mathbb{D}_j(s)$. For each triple $(s, t, j) \in \mathbb{I}^2 \times [d]$ we write the set-difference $\mathbb{D}_j(t) \backslash \mathbb{D}_j(s)$ as $\mathbb{D}_j(t; s)$, with $\mathbb{D}(t; s) \subset \mathbb{I}$ denoting the disjoint union of such differences over $j \in [d]$.

7.1.1 Smooth test functions

To prove Theorem 3.2 we use Stein's method for multivariate normal distributions; for details see for example Chapter 12 in Chen (2011). Our proof of Theorem 3.2 is based on the *Stein characterization* of the multivariate normal distribution: $Z \in \mathbb{R}^d$ is a multivariate normal $\mathrm{MVN}(0, \Sigma)$ if and only if the identity

$$\mathbb{E}\left\{\nabla^T \Sigma \nabla f(Z) - Z^T \nabla f(Z)\right\} = 0 \tag{7.1}$$

holds for all twice continuously differentiable $f : \mathbb{R}^d \to \mathbb{R}$ for which the expectation exists. In particular, we will use the following result based on Meckes (2009, Lemma 1 and Lemma 2). As Lemma 1 and Lemma 2 in Meckes (2009) are stated there only for infinitely differentiable test functions, we give the proof here for completeness.

Lemma 7.1 (Lemma 1 and Lemma 2 in Meckes (2009)) *Fix $n \geq 2$. Let $h : \mathbb{R}^d \to \mathbb{R}$ be n times continuously differentiable with n-th partial derivatives being Lipschitz and*

$Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$. Then, if $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric positive semidefinite, there exists a solution $f : \mathbb{R}^d \to \mathbb{R}$ to the equation

$$\nabla^T \Sigma \nabla f(w) - w^T \nabla f(w) = h(w) - \mathbb{E}h\left(\Sigma^{1/2} Z\right), \quad w \in \mathbb{R}^d, \tag{7.2}$$

such that f is n times continuously differentiable and we have for every $k = 1, \ldots, n$:

$$|f|_k \leq \frac{1}{k} |h|_k .$$

Proof Let h be as in the assertion. It is shown in Lemma 2.1 in Chatterjee and Meckes (2008), which is based on a reformulation of Eq. (2.20) in Barbour (1990), that a solution of (7.2) for h is given by $f(x) = f_h(x) = \int_0^1 \frac{1}{2t}\mathbb{E}\{h(Z_{x,t})\}\, dt$, with $Z_{x,t} = \sqrt{t}x + \sqrt{1-t}\Sigma^{1/2}Z$. As h has n-th partial derivatives being Lipschitz and hence for differentiating f we can bring the derivative inside the integral, it is straightforward to see that the solution f is n times continuously differentiable.

The bound on $|f|_k$ is a consequence of

$$\frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}}(x) = \int_0^1 (2t)^{-1} t^{k/2} \mathbb{E}\left\{ \frac{\partial^k h}{\partial x_{i_1} \cdots \partial x_{i_k}} \left(Z_{x,t}\right) \right\} dt$$

for any $i_1, i_2, \ldots, i_k$; see, for example, Equation (10) in Meckes (2009). Taking the sup-norm on both sides and bounding the right hand side of the equation gives

$$\left\| \frac{\partial^k f}{\partial x_{i_1} \cdots \partial x_{i_k}} \right\|_\infty \leq \left\| \frac{\partial^k h}{\partial x_{i_1} \cdots \partial x_{i_k}} \right\|_\infty \int_0^1 (2t)^{-1} t^{k/2} dt \leq \frac{1}{k} |h|_k .$$

□

Proof of Theorem 3.2 To prove Theorem 3.2, we replace w by W in Eq. (7.2) and take the expected value on both sides. As a result, we aim to bound the expression

$$\left| \mathbb{E}\left\{ \nabla^T \Sigma \nabla f(W) - W^T \nabla f(W) \right\} \right| = \left| \mathbb{E}\left\{ \sum_{i,j=1}^d \partial_{ij} f(W) \Sigma_{ij} - \sum_{i=1}^d W_i \partial_i f(W) \right\} \right| \tag{7.3}$$

where f is a solution to the Stein equation (7.2) for the test function h. Since the variables $\{X_s \mid s \in \mathbb{I}\}$ are centered and as X_t is independent of X_s if $t \notin \mathbb{D}(s)$, for each $(i, j) \in [d]^2$ we have

$$\Sigma_{ij} = \mathrm{Cov}(W_i, W_j) = \sum_{s\in\mathbb{I}_i} \sum_{t\in\mathbb{D}_j(s)} \mathbb{E}\{X_s X_t\} . \tag{7.4}$$

We now use the decomposition of Σ_{ij} from (7.4) in the expression (7.3). For each pair $(s, j) \in \mathbb{I} \times [d]$ and $t \in \mathbb{D}(s)$ we set $\mathbb{D}_j(t; s) = \mathbb{D}_j(t) \backslash \mathbb{D}_j(s)$ and

$$U_j^s := \sum_{u \in \mathbb{D}_j(s)} X_u; \quad W_j^s := W_j - U_j^s, \quad \text{and} \quad V_j^{s,t} := \sum_{v \in \mathbb{D}_j(t;s)} X_v; \quad W_j^{s,t} := W_j^s - V_j^{s,t}. \tag{7.5}$$

By Definition 3.1, W_j^s is independent of X_s, while $W_j^{s,t}$ is independent of the pair (X_s, X_t).

Next we decompose the r.h.s. of (7.3);

$$\left| \mathbb{E} \left\{ \sum_{i=1}^d W_i \partial_i f(W) - \sum_{i,j=1}^d \partial_{ij} f(W) \Sigma_{ij} \right\} \right| = |R_1 + R_2 + R_3| ;$$

with

$$R_1 = \sum_{i=1}^d \mathbb{E} \{ W_i \partial_i f(W) \} - \sum_{s \in \mathbb{I}} \sum_{j=1}^d \mathbb{E} \left\{ X_s U_j^s \partial_{|s|j} f(W^s) \right\}, \tag{7.6}$$

$$R_2 = \sum_{s \in \mathbb{I}} \sum_{j=1}^d \mathbb{E} \left\{ X_s U_j^s \partial_{|s|j} f(W^s) \right\} - \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \mathbb{E} \{ X_s X_t \} \mathbb{E} \partial_{|s||t|} f(W^{s,t}), \text{ and} \tag{7.7}$$

$$R_3 = \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \mathbb{E} \{ X_s X_t \} \left(\mathbb{E} \partial_{|s||t|} f(W^{s,t}) - \mathbb{E} \partial_{|s||t|} f(W) \right). \tag{7.8}$$

Here we recall that if $s = (k, i)$ then $|s| = i \in [d]$. □

As with the vector of dissociated sums $W \in \mathbb{R}^d$ itself, we can assemble these differences into random vectors. Thus, $W^s \in \mathbb{R}^d$ is $(W_1^s, \ldots, W_d^s)$, and similarly $W^{s,t} = (W_1^{s,t}, \ldots W_d^{s,t})$. In the next three claims, we provide bounds on R_i for $i \in [3]$.

Claim 7.2 *The absolute value of the expression R_1 from (7.6) is bounded above by*

$$|R_1| \leq \left(\frac{1}{2} \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(s)} \mathbb{E} |X_s X_t X_u| \right) |f|_3 .$$

Proof Note that

$$\begin{aligned} R_1 &= \sum_{i=1}^d \sum_{s \in \mathbb{I}_i} \mathbb{E} \{ X_s \partial_i f(W) \} - \sum_{s \in \mathbb{I}} \sum_{j=1}^d \mathbb{E} \left\{ X_s U_j^s \partial_{|s|j} f(W^s) \right\} \\ &= \sum_{i=1}^d \sum_{s \in \mathbb{I}_i} \left(\mathbb{E} \{ X_s \partial_i f(W) \} - \sum_{j=1}^d \mathbb{E} \left\{ X_s U_j^s \partial_{ij} f(W^s) \right\} \right). \end{aligned}$$

For each $s \in \mathbb{I}_i$, it follows from (7.5) that $W = U^s + W^s$. Using the Lagrange form of the remainder term in Taylor's theorem, we obtain

$$\partial_i f(W) = \partial_i f(W^s) + \sum_{j=1}^{d} \partial_{ij} f(W^s) U_j^s + \frac{1}{2} \sum_{j,k=1}^{d} \partial_{ijk} f(W^s + \theta_s U^s) U_j^s U_k^s$$

for some random $\theta_s \in (0, 1)$. Using this Taylor expansion in the expression for R_1, we get the following four-term summand $S_{i,s}$ for each $i \in [d]$ and $s \in \mathbb{I}_i$:

$$\begin{aligned} S_{i,s} = {} & \mathbb{E}\left\{X_s \partial_i f(W^s)\right\} + \sum_{j=1}^{d} \mathbb{E}\left\{X_s \partial_{ij} f(W^s) U_j^s\right\} \\ & + \frac{1}{2} \sum_{j,k=1}^{d} \mathbb{E}\left\{X_s \partial_{ijk} f(W^s + \theta_s U^s) U_j^s U_k^s\right\} - \sum_{j=1}^{d} \mathbb{E}\left\{X_s \partial_{ij} f(W^s) U_j^s\right\}. \end{aligned}$$

The second and fourth terms cancel each other. Recalling that X_s is centered by definition and independent of W^s by Definition 3.1, the first term also vanishes and

$$R_1 = \sum_{i=1}^{d} \sum_{s \in \mathbb{I}_i} S_{i,s} = \frac{1}{2} \sum_{i,j,k=1}^{d} \sum_{s \in \mathbb{I}_i} \mathbb{E}\left\{X_s \partial_{ijk} f(W^s + \theta_s U^s) U_j^s U_k^s\right\}.$$

Recalling that $\left\|\partial_{ijk} f\right\|_\infty \le |f|_3$ and that $U_j^s = \sum_{t \in \mathbb{D}_j(s)} X_t$, we have:

$$\begin{aligned} |R_1| &\le \frac{1}{2} \sum_{i,j,k=1}^{d} \sum_{s \in \mathbb{I}_i} \mathbb{E}\left|X_s \partial_{ijk} f(W^s + \theta_s U^s) U_j^s U_k^s\right| \\ &\le \frac{|f|_3}{2} \sum_{i,j,k=1}^{d} \sum_{s \in \mathbb{I}_i} \mathbb{E}\left|X_s \sum_{t \in \mathbb{D}_j(s)} X_t \sum_{u \in \mathbb{D}_k(s)} X_u\right| \\ &\le \frac{|f|_3}{2} \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(s)} \mathbb{E}\left|X_s X_t X_u\right|, \end{aligned}$$

as desired. □

Claim 7.3 *The absolute value of the expression R_2 from (7.7) is bounded above by*

$$|R_2| \le \left(\sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(t;s)} \mathbb{E}\left|X_s X_t X_u\right| \right) |f|_3 .$$

Proof Recalling that $U_j^s = \sum_{t\in\mathbb{D}_j(s)} X_t$ and $\mathbb{D}(s) = \bigcup_{j=1}^d \mathbb{D}_j(s)$,

$$R_2 = \sum_{s\in\mathbb{I}} \sum_{t\in\mathbb{D}(s)} \left\{\mathbb{E}\left\{X_s X_t \partial_{|s||t|} f(W^s)\right\} - \mathbb{E}\left\{X_s X_t\right\} \mathbb{E}\left\{\partial_{|s||t|} f(W^{s,t})\right\}\right\}.$$

Fix $s \in \mathbb{I}$ and $t \in \mathbb{D}_j(s)$. Recall that by (7.5), $W^s = W^{s,t} + V^{s,t}$. Using the Lagrange form of the remainder term in Taylor's theorem, we obtain:

$$\partial_{|s||t|} f(W^s) = \partial_{|s||t|} f(W^{s,t}) + \sum_{k=1}^d \partial_{|s||t|k} f(W^{s,t} + \theta_{s,t} V^{s,t}) V_k^{s,t}$$

for some random $\theta_{s,t} \in (0, 1)$. Using this Taylor expansion in the expression for R_2, we get the following three-term summand $S_{s,t}$ for each pair $(s, t) \in \mathbb{I} \times \mathbb{D}_j(s)$:

$$\begin{aligned} S_{s,t} = {} & \mathbb{E}\left\{X_s X_t \partial_{|s||t|} f(W^{s,t})\right\} + \sum_{k=1}^d \mathbb{E}\left\{X_s X_t \partial_{|s||t|k} f(W^{s,t} + \theta_{s,t} V^{s,t}) V_k^{s,t}\right\} \\ & - \mathbb{E}\left\{X_s X_t\right\} \mathbb{E}\left\{\partial_{|s||t|} f(W^{s,t})\right\}. \end{aligned}$$

Recalling that $W^{s,t}$ is independent of the pair (X_s, X_t) the first and the last terms cancel each other and only the sum over k is left:

$$R_2 = \sum_{s\in\mathbb{I}} \sum_{t\in\mathbb{D}(s)} S_{s,t} = \sum_{s\in\mathbb{I}} \sum_{t\in\mathbb{D}(s)} \sum_{k=1}^d \mathbb{E}\left\{X_s X_t \partial_{|s||t|k} f(W^{s,t} + \theta_{s,t} V^{s,t}) V_k^{s,t}\right\}.$$

Recalling that $\left\|\partial_{ijk} f\right\|_\infty \le |f|_3$ and that $V_k^{s,t} = \sum_{v\in\mathbb{D}_k(t;s)} X_v$ we have:

$$\begin{aligned} |R_2| &\le \sum_{s\in\mathbb{I}} \sum_{t\in\mathbb{D}(s)} \sum_{k=1}^d \sum_{v\in\mathbb{D}_k(t;s)} \mathbb{E}\left|X_s X_t X_v \partial_{|s||t|k} f(W^{s,t} + \theta_{s,t} V^{s,t})\right| \\ &\le |f|_3 \sum_{s\in\mathbb{I}} \sum_{t\in\mathbb{D}(s)} \sum_{u\in\mathbb{D}(t;s)} \mathbb{E}\left|X_s X_t X_u\right|, \end{aligned}$$

as required. □

Claim 7.4

$$|R_3| \le \left(\sum_{s\in\mathbb{I}} \sum_{t\in\mathbb{D}(s)} \left\{\sum_{u\in\mathbb{D}(s)} \mathbb{E}\left|X_s X_t\right| \mathbb{E}\left|X_u\right| + \sum_{u\in\mathbb{D}(t;s)} \mathbb{E}\left|X_s X_t\right| \mathbb{E}\left|X_u\right|\right\}\right) |f|_3.$$

Proof Fix $(s,t) \in \mathbb{I} \times \mathbb{D}_j(s)$. Recall that by (7.5), $W^{s,t} = W - U^s - V^{s,t}$. Using the Lagrange form of the remainder term in Taylor's theorem, we obtain

$$\partial_{|s||t|} f(W^{s,t}) = \partial_{|s||t|} f(W) - \sum_{k=1}^{d} \partial_{|s||t|k} f(W - \rho_{s,t}(U^s + V^{s,t}))(U_k^s + V_k^{s,t})$$

for some random $\rho_{s,t} \in (0,1)$. Recalling that $U_k^s = \sum_{t \in \mathbb{D}_k(s)} X_t$ and $V_k^{s,t} = \sum_{u \in \mathbb{D}_j(t;s)} X_u$,

$$\begin{aligned} R_3 &= -\sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{k=1}^{d} \mathbb{E}\{X_s X_t\} \mathbb{E}\left\{\partial_{|s||t|k} f(W - \rho_{s,t}(U^s + V^{s,t}))(U_k^s + V_k^{s,t})\right\} \\ &= -\sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(s)} \mathbb{E}\{X_s X_t\} \mathbb{E}\left\{X_u \partial_{|s||t|k} f(W - \rho_{s,t}(U^s + V^{s,t}))\right\} \\ &\quad - \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(t;s)} \mathbb{E}\{X_s X_t\} \mathbb{E}\left\{X_u \partial_{|s||t|k} f(W - \rho_{s,t}(U^s + V^{s,t}))\right\}. \end{aligned}$$

Recalling that $\left\|\partial_{ijk} f\right\|_\infty \le |f|_3$ we bound:

$$|R_3| \le |f|_3 \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(s)} \mathbb{E}|X_s X_t| \mathbb{E}|X_u| + |f|_3 \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(t;s)} \mathbb{E}|X_s X_t| \mathbb{E}|X_u|,$$

as required. □

Take any $h \in \mathscr{H}_d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be the associated solution from Lemma 7.1. Combining Claims 7.1–7.4 and using Lemma 7.1 we have:

$$\begin{aligned} &\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z)\right| \\ &\le \left|\mathbb{E}\left\{\nabla^T \Sigma \nabla f(W) - W^T \nabla f(W)\right\}\right| \le |R_1| + |R_2| + |R_3| \\ &\le |f|_3 \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(s)} \left(\frac{1}{2}\mathbb{E}|X_s X_t X_u| + \mathbb{E}|X_s X_t| \mathbb{E}|X_u|\right) \\ &\quad + |f|_3 \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{u \in \mathbb{D}(t;s)} (\mathbb{E}|X_s X_t X_u| + \mathbb{E}|X_s X_t| \mathbb{E}|X_u|) \\ &\le \frac{1}{3} |h|_3 B_{3.2}. \end{aligned}$$

7.1.2 Non-smooth test functions

Convex set indicator test functions provide a stronger distance between probability distributions. Also, from a non-asymptotic perspective, the distance might be more

useful in statistical applications: for example, when estimating confidence regions, which are often convex sets. Here we follow (Kaur and Röllin 2021, Section 5.3) very closely to derive a bound on the convex set distance between a vector of dissociated sums $W \in \mathbb{R}^d$ with covariance matrix Σ and a target multivariate normal distribution $\Sigma^{\frac{1}{2}} Z$, where $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$. The smoothing technique used here is introduced in Gan et al. (2017). However, a better (polylogarithmic) dependence on d could potentially be achieved using a recent result (Gaunt and Li 2023, Proposition 2.6), at the expense of larger constants. The recursive approach from Schulte and Yukich (2019), Kasprzak and Peccati (2022) usually yields better dependence on n; however, this requires the target normal distribution to have an invertible covariance matrix. Since this property does not always hold in our applications of interest, we do not use the recursive approach here.

Proof of Theorem 3.3 Fix $A \in \mathscr{K}$, $\epsilon > 0$ and define

$$A^{\epsilon} = \left\{ y \in \mathbb{R}^d : d(y, A) < \epsilon \right\}, \quad \text{and} \quad A^{-\epsilon} = \left\{ y \in \mathbb{R}^d : B(y; \epsilon) \subseteq A \right\}$$

where $d(y, A) = \inf_{x\in A} \|x - y\|_2$ and $B(y; \epsilon) = \left\{ z \in \mathbb{R}^d \mid \|y - z\|_2 \leq \epsilon \right\}$.

Let $\mathscr{H}_{\epsilon,A} := \left\{ h_{\epsilon,A} : \mathbb{R}^d \to [0, 1]; A \in \mathscr{K} \right\}$ be a class of functions such that $h_{\epsilon,A}(x) = 1$ for $x \in A$ and 0 for $x \notin A^{\epsilon}$. Then, by Bentkus (2003, Lemma 2.1) as well as inequalities (1.2) and (1.4) from Bentkus (2003), for any $\epsilon > 0$ we have

$$\sup_{A\in\mathscr{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq 4d^{\frac{1}{4}}\epsilon + \sup_{A\in\mathscr{K}} \left| \mathbb{E}h_{\epsilon,A}(W) - \mathbb{E}h_{\epsilon,A}(\Sigma^{\frac{1}{2}} Z) \right|.$$

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a bounded Lebesgue measurable function, and for $\delta > 0$ let

$$(S_\delta f)(x) = \frac{1}{(2\delta)^d} \int_{x_1-\delta}^{x_1+\delta} \cdots \int_{x_d-\delta}^{x_d+\delta} f(z) dz_d \ldots dz_1.$$

Set $\delta = \frac{\epsilon}{16\sqrt{d}}$ and $h_{\epsilon,A} = S_\delta^4 I_{A^{\epsilon/4}}$, where $I_{A^{\epsilon/4}}$ is the indicator function of the subset $A^{\epsilon/4} \subseteq \mathbb{R}^d$. By Gan et al. (2017, Lemma 3.9) we have that $h_{\epsilon,A}$ is bounded and is in $\mathscr{H}_d$.

Moreover, the following bounds hold:

$$\left\|h_{\epsilon,A}\right\|_\infty \leq 1, \quad \left|h_{\epsilon,A}\right|_2 \leq \frac{1}{\epsilon^2}, \quad \left|h_{\epsilon,A}\right|_3 \leq \frac{1}{\epsilon^3}.$$

Note that $h_{\epsilon,A} = S_\delta^4 I_{A^{\epsilon/4}} \in \mathscr{H}_{\epsilon,A}$ and hence (Bentkus 2003, Lemma 2.1) applies. Using this with Theorem 3.2 we get

$$\begin{aligned}\sup_{A\in\mathscr{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| &\leq 4d^{\frac{1}{4}}\epsilon + \sup_{A\in\mathscr{K}} \left| \mathbb{E}h_{\epsilon,A}(W) - \mathbb{E}h_{\epsilon,A}(\Sigma^{\frac{1}{2}} Z) \right| \\ &\leq 4d^{\frac{1}{4}}\epsilon + \frac{1}{3\epsilon^3} B_{3.2}.\end{aligned}$$

Since this bound works for every $\epsilon > 0$, we minimise it by using $\epsilon = \left(\frac{3B_{3.2}}{4d^{\frac{1}{4}}}\right)^{\frac{1}{4}}$. □

7.2 A proof of the CLT for *U*-statistics

Proof of Theorem 3.9 Note that if $s = (\phi, i) \in \mathbb{I}_i$ and $u = (\psi, j) \in \mathbb{I}_j$ are chosen such that $\phi \cap \psi = \varnothing$, then the corresponding variables X_s and X_u are independent since $f_i(\mathscr{X}_s, \mathscr{Y}_s)$ and $f_j(\mathscr{X}_u, \mathscr{Y}_u)$ do not share any random variables from the sets $\{\xi_i\}_{1\le i\le n}$ and $\{Y_{i,j}\}_{1\le i<j\le n}$. Hence, if for any $s = (\phi, i) \in \mathbb{I}_i$ we set $\mathbb{D}_j(s) = \left\{(\psi, j) \in \mathbb{I}_j \mid |\phi \cap \psi| \ge 1\right\}$, then W satisfies the assumptions of Corollary 3.4. It remains to bound the quantity $B_{3.4}$.

First, to find α_{ij} as in Corollary 3.4, given $\phi \in C_{k_i}$ and if $k_i, k_j \ge m$ then there are $\binom{k_i}{m}\binom{n-k_i}{k_j-m}$ subsets $\psi \in C_{k_j}$ such that $|\phi \cap \psi| = m$. Therefore, we have for any $i, j \in [d]$ and $s \in \mathbb{I}_i$

$$|\mathbb{D}_j(s)| = \sum_{m=1}^{\min(k_i,k_j)} \binom{k_i}{m}\binom{n-k_i}{k_j-m} = \alpha_{ij} \le k_i^{\min(k_i,k_j)+1}(n-k_i)^{k_j-1}. \quad (7.9)$$

Note that

$$\begin{aligned}&\mathbb{E}\,|X_sX_tX_u| \\ &= (\sigma_i\sigma_j\sigma_k)^{-1}\mathbb{E}\left|\{f_i(\mathscr{X}_s, \mathscr{Y}_s) - \mu_s\}\left\{f_j(\mathscr{X}_t, \mathscr{Y}_t) - \mu_t\right\}\{f_l(\mathscr{X}_u, \mathscr{Y}_u) - \mu_u\}\right|\end{aligned}$$

as well as

$$\begin{aligned}&\mathbb{E}\,|X_sX_t|\,\mathbb{E}\,|X_u| \\ &= (\sigma_i\sigma_j\sigma_k)^{-1}\mathbb{E}\left|\{f_i(\mathscr{X}_s, \mathscr{Y}_s) - \mu_s\}\left\{f_j(\mathscr{X}_t, \mathscr{Y}_t) - \mu_t\right\}\right|\mathbb{E}\,|f_l(\mathscr{X}_u, \mathscr{Y}_u) - \mu_u|\,.\end{aligned}$$

Using Assumption 3.7, for any $i, j, l \in [d]$ and $s \in \mathbb{I}_i, t \in \mathbb{I}_j, u \in \mathbb{I}_l$

$$\mathbb{E}\,|X_sX_tX_u| \le (\sigma_i\sigma_j\sigma_k)^{-1}\beta \text{ and } \mathbb{E}\,|X_sX_t|\,\mathbb{E}\,|X_u| \le (\sigma_i\sigma_j\sigma_k)^{-1}\beta. \quad (7.10)$$

To take care of the variance terms, we lower bound the variance using Assumption 3.7;

$$\begin{aligned}\mathrm{Var}(S_{n,k_i}(f_i)) &= \sum_{s\in C_{k_i}}\sum_{t\in\mathbb{D}_i(s)}\mathrm{Cov}(f_i(\mathscr{X}_s, \mathscr{Y}_s), f_i(\mathscr{X}_t, \mathscr{Y}_t)) \\ &= \sum_{m=1}^{k_i}\sum_{s\in C_{k_i}}\sum_{\substack{t\in C_{k_i}\\ |s\cap t|=m}}\mathrm{Cov}(f_i(\mathscr{X}_s, \mathscr{Y}_s), f_i(\mathscr{X}_t, \mathscr{Y}_t))\end{aligned}$$

$$\geq \binom{n}{k_i}\sum_{m=1}^{k_i}\binom{k_i}{m}\binom{n-k_i}{k_i-m}\alpha_i = \alpha_i\sum_{m=1}^{k_i}\binom{n}{2k_i-m}\binom{2k_i-m}{k_i}\binom{k_i}{m}$$
$$\geq \alpha_i k_i\binom{n}{2k_i-1}\binom{2k_i-1}{k_i} \geq \alpha_i k_i\frac{n^{2k_i-1}}{(2k_i-1)^{2k_i-1}}\frac{(2k_i-1)^{k_i}}{k_i^{k_i}}$$
$$= \alpha_i\frac{n^{2k_i-1}}{(2k_i^2-k_i)^{k_i-1}}.$$

Here the second-to-last inequality follows by taking only the term for $m = 1$. Now we take both sides of the inequality to the power of $-\frac{1}{2}$ to get that for any $i \in [d]$

$$\sigma_i^{-1} \leq n^{-k_i+\frac{1}{2}}\alpha_i^{-\frac{1}{2}}(2k_i^2-k_i)^{-\frac{k_i}{2}+\frac{1}{2}}. \tag{7.11}$$

Using Eqs. (7.9)–(7.11) to bound the quantity $B_{3.4}$ from Corollary 3.4 we get

$$B_{3.4} \leq \frac{2}{3}\sum_{i,j,l=1}^{d}(\sigma_i\sigma_j\sigma_k)^{-1}\beta\binom{n}{k_i}k_i^{\min(k_i,k_j)+1}(n-k_i)^{k_j-1}$$
$$\left\{k_i^{\min(k_i,k_l)+1}(n-k_i)^{k_l-1}+k_j^{\min(k_j,k_l)+1}(n-k_j)^{k_l-1}\right\}$$
$$\leq \frac{2}{3}\sum_{i,j,l=1}^{d}n^{k_i+k_j+k_l-2}\frac{k_i^{\min(k_i,k_j)+1}}{k_i!}\left\{k_i^{\min(k_i,k_l)+1}+k_j^{\min(k_j,k_l)+1}\right\}\beta$$
$$\left(n^{-k_i-k_j-k_l+\frac{3}{2}}(\alpha_i\alpha_j\alpha_l)^{-\frac{1}{2}}(2k_i^2-k_i)^{-\frac{k_i}{2}+\frac{1}{2}}(2k_j^2-k_j)^{-\frac{k_j}{2}+\frac{1}{2}}(2k_l^2-k_l)^{-\frac{k_l}{2}+\frac{1}{2}}\right)$$
$$\leq \left\{\frac{2\beta}{3}\sum_{i,j,l=1}^{d}\frac{k_i^{\min(k_i,k_j)+1}}{k_i!\sqrt{\alpha_i\alpha_j\alpha_l}}\left(k_i^{\min(k_i,k_l)+1}+k_j^{\min(k_j,k_l)+1}\right)K_iK_jK_l\right\}n^{-\frac{1}{2}}.$$

Now further assume that Assumption 3.8 hold. The key difference in this case is that the dependency neighbourhoods become smaller: now the subsets need to overlap in at least 2 elements for the corresponding summands to share at least one variable $Y_{i,j}$ and hence become dependent. This makes both the variance and the size of dependency neighbourhoods smaller. In the context of Theorem 3.2, the trade-off works out in our favour to give smaller bounds, as follows. For any $s = (\phi, i) \in \mathbb{I}_i$ we set $\mathbb{D}_j(s) = \left\{(\psi, j) \in \mathbb{I}_j \mid |\phi \cap \psi| \geq 2\right\}$, so that W, under the additional Assumption 3.8, satisfies the assumptions of Corollary 3.4.

Now Eq. (7.9) becomes

$$|\mathbb{D}_j(s)| \leq k_i^{\min(k_i,k_j)+1}(n-k_i)^{k_j-2}.$$

Equation (7.11) becomes

$$\sigma_i^{-1} \leq 2n^{-k_i+1}\alpha_i^{-\frac{1}{2}}(2k_i^2 - 2k_i)^{-\frac{k_i}{2}+1}.$$

Using the adjusted bounds in Corollary 3.4 gives the result. □

Acknowledgements TT acknowledges funding from EPSRC studentship 2275810. VN is supported by the EPSRC grant EP/R018472/1 and the US AFOSR grant FA9550-22-1-0462. GR is funded in part by the EPSRC grants EP/T018445/1 and EP/R018472/1. We would like to thank Xiao Fang, Matthew Kahle, Heather Harrington, Adrian Röllin, and Christina Goldschmidt for helpful discussions Moreover, we are grateful to two anonymous referees whose comments helped improve this paper.

Declarations

Conflicts of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Appendix A. Proof of Lemma 4.3

We recall that T_k, counting the number of $(k-1)$-simplices that are critical under the lexicographical matching, is given by

$$T_k = \sum_{s\in C_k} \prod_{i\neq j\in s} Y_{i,j} \left[\prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b\in s} Y_{a,b}\right) - \prod_{a=1}^{\min(s)-1} \left(1 - \prod_{b\in s_-} Y_{a,b}\right)\right].$$

Lemma A.1 *For any integer $1 \leq k \leq n-1$ we have:*

$$\mathrm{Var}\{T_{k+1}\} = 2p^{2\binom{k+1}{2}}V_1 + 2p^{2\binom{k+1}{2}}V_2 + p^{\binom{k+1}{2}}V_3 + p^{\binom{k+1}{2}}V_4,$$

where

$$\begin{aligned} V_1 = \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} &\binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m-1}\binom{j-i+1}{q-1} \\ &\Big\{\theta(i,j,q,m,1)\big[(1-2p^{k+1}+p^{2k+2-m})^{i-1} - (1-p^{k+1}-p^k+p^{2k+1-m})^{i-1}\big] \\ &+\theta(i,j,q,m,0)\big[(1-2p^k+p^{2k+1-m})^{i-1} - (1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\big] \end{aligned}$$

$$
\begin{aligned}
&\quad - \eta(i)\eta(j)\Big\};\\
V_2 &= \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} \binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m}\binom{j-i+1}{q-1}\\
&\quad \Big\{\theta(i,j,q,m,1)\big[(1-2p^{k+1}+p^{2k+2-m})^{i-1}-(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\big]\\
&\quad + \theta(i,j,q,m,0)\big[(1-2p^{k}+p^{2k-m})^{i-1}-(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\big]\\
&\quad - \eta(i)\eta(j)\Big\};\\
V_3 &= \sum_{i=1}^{n-k}\sum_{m=1}^{k}\binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1}\\
&\quad \Big\{p^{-\binom{m}{2}}\big[(1-2p^{k+1}+p^{2k+2-m})^{i-1}+\\
&\quad (1-2p^{k}+p^{2k+1-m})^{i-1}-2(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1}\big]-\eta(i)^2\Big\};\\
V_4 &= \sum_{i=1}^{n-k}\binom{n-i}{k}\Big\{\eta(i)-p^{\binom{k+1}{2}}\eta(i)^2\Big\}.
\end{aligned}
$$

Here we have used the following notation:

$$
\begin{aligned}
\eta(a) &:= (1-p^{k+1})^{a-1}-(1-p^k)^{a-1};\\
\theta(i,j,q,m,\delta) &:= p^{-\binom{m}{2}}(1-p^{k+\delta})^{j-i-q}(1-p^{k+\delta-m})^{q}.
\end{aligned}
$$

Also, $\sum_{i<j}^{n-k}$ *stands for* $\sum_{i=1}^{n-k-1}\sum_{j=i+1}^{n-k}$.

Proof of Lemma A.1 For $s \in C_{k+1}$ recall that $s_- = s \setminus \{\min(s)\}$. We write:

$$
\begin{aligned}
Y_s^+ &= \prod_{i=1}^{\min(s)-1}\left(1-\prod_{j\in s}Y_{i,j}\right), \qquad & Y_s^- &= \prod_{i=1}^{\min(s)-1}\left(1-\prod_{j\in s_-}Y_{i,j}\right),\\
Z_s &= \prod_{i\neq j\in s}Y_{i,j}, & Y_s &= Y_s^+ - Y_s^-.
\end{aligned}
$$

Then Z_s and Y_s are independent and $T_{k+1} = \sum_{s\in C_{k+1}} Z_s Y_s$. Consider the variance:

$$
\begin{aligned}
\mathrm{Var}(T_{k+1}) &= \sum_{s\in C_{k+1}}\mathrm{Var}(Z_sY_s) + \sum_{\substack{s\neq t\in C_{k+1}\\ \min(s)\neq\min(t)}}\mathrm{Cov}(Z_sY_s, Z_tY_t)\\
&\quad + \sum_{\substack{s\neq t\in C_{k+1}\\ \min(s)=\min(t)}}\mathrm{Cov}(Z_sY_s, Z_tY_t). \qquad (A.1)
\end{aligned}
$$

For the first term in (A.1), writing $\mathbb{P}(Z_s Y_s = 1) = \mu(i) := p^{\binom{k+1}{2}}((1-p^{k+1})^{i-1} - (1-p^k)^{i-1})$ we see:

$$\sum_{s\in C_{k+1}} \mathrm{Var}(Z_s Y_s) = \sum_{i=1}^{n} \sum_{\substack{s\in C_{k+1} \\ \min(s)=i}} \left(\mathbb{E}\left\{(Z_s Y_s)^2\right\} - \mathbb{E}\left\{Z_s Y_s\right\}^2\right)$$

$$= \sum_{i=1}^{n-k} \binom{n-i}{k} \left(\mathbb{P}(Z_s Y_s = 1) - \mathbb{P}(Z_s Y_s = 1)^2\right) = \sum_{i=1}^{n-k} \binom{n-i}{k} \left\{\mu(i) - \mu(i)^2\right\} = p^{\binom{k+1}{2}} V_4.$$

Now consider the covariance terms in (A.1), the expansion of the variance. Note that for any $s, t \in C_{k+1}$ if $s \cap t = \varnothing$, then the variables $Z_s Y_s$ and $Z_t Y_t$ can be written as functions of two disjoint sets of independent edge indicators and hence have zero covariance.

Fix $s, t \in C_{k+1}$ and assume $|s \cap t| = m$ where $1 \le m \le k$. Note that because $m \ne k+1$, we have $s \ne t$. There are $2\binom{k+1}{2} - \binom{m}{2}$ distinct edges in s and t combined and hence $\mathbb{P}(Z_s Z_t = 1) = p^{2\binom{k+1}{2} - \binom{m}{2}}$. Also, $Y_s Y_t = Y_s^+ Y_t^+ + Y_s^- Y_t^- - Y_s^+ Y_t^- - Y_s^- Y_t^+$. For the rest of the proof when calculating probabilities we assume w.l.o.g. that $\min(s) \le \min(t)$. Then we have for $Y_s^+ Y_t^+$:

$$\begin{aligned} Y_s^+ Y_t^+ &= \prod_{i=1}^{\min(t)-1} \left(1 - \prod_{j\in t} Y_{i,j}\right) \prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j\in s} Y_{i,j}\right) \\ &= \prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j\in s} Y_{i,j}\right) \left(1 - \prod_{j\in t} Y_{i,j}\right) \prod_{\substack{i=\min(s) \\ i\in s}}^{\min(t)-1} \left(1 - \prod_{j\in t} Y_{i,j}\right) \\ &\quad \prod_{\substack{i=\min(s) \\ i\notin s}}^{\min(t)-1} \left(1 - \prod_{j\in t} Y_{i,j}\right). \end{aligned}$$

Fix $i \in [\min(s) - 1]$. Then with $\neg$ denoting the complement

$$\begin{aligned} &\mathbb{P}\left[\left(1 - \prod_{j\in s} Y_{i,j}\right)\left(1 - \prod_{j\in t} Y_{i,j}\right) = 1\right] \\ &\qquad = \mathbb{P}\left[\neg\left(\prod_{j\in s} Y_{i,j} = 1 \cup \prod_{j\in t} Y_{i,j} = 1\right)\right] \\ &\qquad = 1 - \left\{\mathbb{P}\left(\prod_{j\in s} Y_{i,j} = 1\right) + \mathbb{P}\left(\prod_{j\in t} Y_{i,j} = 1\right)\right. \\ &\qquad\quad \left. - \mathbb{P}\left(\prod_{j\in s} Y_{i,j} = 1 \cap \prod_{j\in t} Y_{i,j} = 1\right)\right\} \end{aligned}$$

$$= 1 - (2p^{k+1} - p^{2k+2-m}).$$

Moreover, $\prod_{i=1}^{\min(s)-1}(1-\prod_{j\in s} Y_{i,j})(1-\prod_{j\in t} Y_{i,j})$ and $\prod_{\substack{i=\min(s)\\ i\notin s}}^{\min(t)-1}(1-\prod_{j\in t} Y_{i,j})$ are independent of Z_sZ_t.

Recall the notation $[a,b] = \{a, a+1, \dots, b\}$ for two positive integers $a \le b$. Setting $q_{s,t} := |s \cap [\min(s), \min(t)-1]|$,

$$\begin{aligned}
&\mathbb{P}\left(Y_s^+Y_t^+ = 1 | Z_sZ_t = 1\right)\\
&= \mathbb{P}\left(\prod_{i=1}^{\min(s)-1}(1-\prod_{j\in s} Y_{i,j})(1-\prod_{j\in t} Y_{i,j}) = 1\right)\mathbb{P}\left(\prod_{\substack{i=\min(s)\\ i\notin s}}^{\min(t)-1}(1-\prod_{j\in t} Y_{i,j}) = 1\right)\\
&\mathbb{P}\left(\prod_{\substack{i=\min(s)\\ i\in s}}^{\min(t)-1}(1-\prod_{j\in t} Y_{i,j}) = 1 \,\middle|\, \prod_{i\neq j\in s} Y_{i,j}\prod_{i\neq j\in t} Y_{i,j} = 1\right)\\
&= (1-2p^{k+1}+p^{2k+2-m})^{\min(s)-1}(1-p^{k+1})^{\min(t)-\min(s)-q_{s,t}}(1-p^{k+1-m})^{q_{s,t}}.
\end{aligned}$$

This strategy of splitting the product $Y_s^+Y_t^+$ into three products of independent variables, only one of which is dependent on Z_sZ_t works exactly in the same way for the variables $Y_s^-Y_t^+$, $Y_s^+Y_t^-$, $Y_s^-Y_t^-$. We write $i = \min(s)$, $j = \min(t)$, and q instead of $q_{s,t}$. Also, we set

$$\pi(i,j,a,b,d_1,d_2,q) := (1-p^a-p^b+p^{a+b-d_1})^{i-1}(1-p^a)^{j-i-q}(1-p^{a-d_2})^q.$$

Using the described strategy we get:

$$\begin{aligned}
\mathbb{P}\left(Y_s^-Y_t^- = 1 | Z_sZ_t = 1\right) &= \pi(i,j,k,k,|s_-\cap t_-|,|s_-\cap t_-|,q)\\
\mathbb{P}\left(Y_s^+Y_t^- = 1 | Z_sZ_t = 1\right) &= \pi(i,j,k,k+1,|s\cap t_-|,|s\cap t_-|,q)\\
\mathbb{P}\left(Y_s^-Y_t^+ = 1 | Z_sZ_t = 1\right) &= \pi(i,j,k+1,k,|s_-\cap t|,m,q).
\end{aligned}$$

Now we are ready to calculate the covariance:

$$\begin{aligned}
\mathrm{Cov}(Z_sY_s, Z_tY_t) &= \mathbb{E}\left\{Z_sZ_tY_s^+Y_t^+\right\} + \mathbb{E}\left\{Z_sZ_tY_s^-Y_t^-\right\} - \mathbb{E}\left\{Z_sZ_tY_s^+Y_t^-\right\}\\
&\quad - \mathbb{E}\left\{Z_sZ_tY_s^-Y_t^+\right\} - \mathbb{E}\{Z_sY_s\}\mathbb{E}\{Z_tY_t\}\\
&= \mathbb{P}(Z_sZ_t=1)\Big\{\mathbb{P}\left(Y_s^+Y_t^+ = 1|Z_sZ_t=1\right) + \mathbb{P}\left(Y_s^-Y_t^- = 1|Z_sZ_t=1\right)\\
&\quad - \mathbb{P}\left(Y_s^+Y_t^- = 1|Z_sZ_t=1\right) - \mathbb{P}\left(Y_s^-Y_t^+ = 1|Z_sZ_t=1\right)\Big\} - \mathbb{P}(Z_sY_s=1)\mathbb{P}(Z_tY_t=1)\\
&= p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(i,j,k+1,k+1,m,m,q) + \pi(i,j,k,k,|s_-\cap t_-|,|s_-\cap t_-|,q)\\
&\quad - \pi(i,j,k,k+1,|s\cap t_-|,|s\cap t_-|,q) - \pi(i,j,k+1,k,|s_-\cap t|,m,q)) - \mu(i)\mu(j).
\end{aligned}$$

Next we consider the two covariance sums in (A.1) separately. First let us assume that $\min(s) \neq \min(t)$. Given $i, j \in [n-k]$, $m \in [k]$, and $q \in [\min(k+1, |j-i|)]$ define the set

$$\begin{aligned}\Gamma_{k+1}(i,j,m,q) = \{\,(s,t) \mid s\,\}, t \in C_{k+1}, \min(s) = i, \min(t) = j, |s \cap t| \\ = m, \max(q_{s,t}, q_{t,s}) = q\end{aligned}$$

as well as

$$\Gamma^{+}_{k+1}(i,j,m,q) = \{\,(s,t) \in \Gamma_{k+1}(i,j,m,q) \mid \min(t) \in s\,\}$$

and

$$\Gamma^{-}_{k+1}(i,j,m,q) = \{\,(s,t) \in \Gamma_{k+1}(i,j,m,q) \mid \min(t) \notin s\,\}.$$

Next we argue that

$$|\Gamma^{+}_{k+1}(i,j,m,q)| = \binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m-1}\binom{j-i+1}{q-1}.$$

To see this, assume $i < j$. Note that to pick a pair $(s,t) \in \Gamma^{+}_{k+1}(i,j,m,q)$ with $\min(s) = i$ and $\min(t) = j$ we need to pick the $2k-m$ vertices in $s \cup t$. Firstly, we pick the vertices that are not included in $s \cap [\min(s), \min(t)-1] = s \cap [i, j-1]$. Since $\min(s) \in s \cap [\min(s), \min(t)-1]$, this amounts to choosing $2k-m-(q-1)$ vertices out of $n-j$. Then we decide which of the vertices that we have just picked will lie in t. This means we further need to choose k out of $2k+1-m-q$ vertices. Then we choose $m-1$ out of k vertices of t to lie in $s \cap t$ (under the assumption that we already have $\min(t) \in s$). Finally, we choose the set $s \cap [\min(s), \min(t)-1]$, which amounts to picking $q-1$ vertices out of $j-i+1$ possible choices. If any of the binomial coefficients are negative, we set them to 0. The case $j < i$ is analogous.

An analogous argument shows that

$$|\Gamma^{-}_{k+1}(i,j,m,q)| = \binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m}\binom{j-i+1}{q-1}.$$

Now using the covariance expression we have just derived, we get

$$\begin{aligned}&\sum_{\substack{s \neq t \in C_{k+1} \\ \min(s) \neq \min(t)}} \operatorname{Cov}(Z_s Y_s, Z_t Y_t) \\ &= \sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}\sum_{(s,t)\in\Gamma^{+}_{k+1}(i,j,m,q)} \operatorname{Cov}(Z_s Y_s, Z_t Y_t) \\ &\quad + \sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}\sum_{(s,t)\in\Gamma^{-}_{k+1}(i,j,m,q)} \operatorname{Cov}(Z_s Y_s, Z_t Y_t)\end{aligned}$$

$$
\begin{aligned}
&+\sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)}\sum_{(s,t)\in\Gamma_{k+1}^{+}(j,i,m,q)}\operatorname{Cov}(Z_sY_s,Z_tY_t)\\
&+\sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)}\sum_{(s,t)\in\Gamma_{k+1}^{-}(j,i,m,q)}\operatorname{Cov}(Z_sY_s,Z_tY_t)\\
&=\sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}|\Gamma_{k+1}^{+}(i,j,m,q)|\\
&\times\Big\{p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(i,j,k+1,k+1,m,m,q)\\
&+\pi(i,j,k,k,m-1,m-1,q)-\pi(i,j,k,k+1,m-1,m-1,q)\\
&-\pi(i,j,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&+\sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}|\Gamma_{k+1}^{-}(i,j,m,q)|\\
&\times\Big\{p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(i,j,k+1,k+1,m,m,q)\\
&+\pi(i,j,k,k,m,m,q)-\pi(i,j,k,k+1,m,m,q)\\
&-\pi(i,j,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&+\sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)}|\Gamma_{k+1}^{+}(j,i,m,q)|\\
&\times\Big\{p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(j,i,k+1,k+1,m,m,q)\\
&+\pi(j,i,k,k,m-1,m-1,q)-\pi(j,i,k,k+1,m-1,m-1,q)\\
&-\pi(j,i,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&+\sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)}|\Gamma_{k+1}^{-}(j,i,m,q)|\\
&\times\Big\{p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(j,i,k+1,k+1,m,m,q)\\
&+\pi(j,i,k,k,m,m,q)-\pi(j,i,k,k+1,m,m,q)\\
&-\pi(j,i,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&=2p^{2\binom{k+1}{2}}V_1+2p^{2\binom{k+1}{2}}V_2.
\end{aligned}
$$

Similarly, we calculate the remaining term in the expansion of the variance (A.1). We notice that if $i=j$, then $q=0$ and we have $\Gamma_{k+1}(i,i,m,0)=\Gamma_{k+1}^{+}(i,i,m,0)$.

Hence, $|\Gamma_{k+1}(i,i,m,0)| = \binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1}$, and

$$\sum_{\substack{s\neq t\in C_{k+1}\\ \min(s)=\min(t)}} \operatorname{Cov}(Z_sY_s, Z_tY_t) = \sum_{i=1}^{n-k}\sum_{m=1}^{k}\sum_{(s,t)\in\Gamma_{k+1}(i,i,m,0)} \operatorname{Cov}(Z_sY_s, Z_tY_t)$$

$$= \sum_{i=1}^{n-k}\sum_{m=1}^{k}\binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1}$$
$$\Big\{p^{-\binom{m}{2}}\Big[(1-2p^{k+1}+p^{2k+2-m})^{i-1} + (1-2p^k+p^{2k+1-m})^{i-1}$$
$$- 2(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1}\Big] - ((1-p^{k+1})^{i-1} - (1-p^k)^{i-1})^2\Big\}$$
$$= p^{2\binom{k+1}{2}}V_3.$$

□

Proof of Lemma 4.3 Fix $1 \le k \le n-1$ and $p \in (0,1)$, and consider the variance. From Lemma A.1 we have $\operatorname{Var}\{T_{k+1}\} = 2p^{2\binom{k+1}{2}}V_1 + 2p^{2\binom{k+1}{2}}V_2 + p^{2\binom{k+1}{2}}V_3 + p^{\binom{k+1}{2}}V_4$. First we lower bound V_1 and V_2 by just the negative part of the sum:

$$V_1 \ge -\sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}\binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m-1}\binom{j-i+1}{q-1}$$
$$\Big\{(1-p^{k+1})^{i+j-2} + (1-p^k)^{i+j-2}$$
$$+ p^{-\binom{m}{2}}(1-p^{k+1})^{j-i-q}(1-p^{k+1-m})^q(1-p^{k+1}-p^k+p^{2k+1-m})^{i-1}$$
$$+ p^{-\binom{m}{2}}(1-p^k)^{j-i-q}(1-p^{k-m})^q(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\Big\};$$
$$V_2 \ge -\sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}\binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m}\binom{j-i+1}{q-1}$$
$$\Big\{(1-p^{k+1})^{i+j-2} + (1-p^k)^{i+j-2}$$
$$+ p^{-\binom{m}{2}}(1-p^{k+1})^{j-i-q}(1-p^{k+1-m})^q(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}$$
$$+ p^{-\binom{m}{2}}(1-p^k)^{j-i-q}(1-p^{k-m})^q(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\Big\}.$$

Now using that $\binom{k}{m} + \binom{k}{m-1} = \binom{k+1}{m}$ and $(1-p^{k+1}) \ge (1-p^{k-m})$ for $m \ge 0$ it is easy to see that $V_1 + V_2 \ge -4R_1 - 4R_2$, where

$$R_1 := \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}\binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k+1}{m}\binom{j-i+1}{q-1}$$
$$(1-p^{k+1})^{i+j-2};$$

$$R_2 := \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} \binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k+1}{m}\binom{j-i+1}{q-1}$$
$$p^{-\binom{m}{2}}(1-p^{k+1})^{j-i}(1-p^{k+1}-p^k+p^{2k+1-m})^{i-1}.$$

For V_3 we lower bound by terms with $m = 1$ and the negative parts of the other terms:

$$\begin{aligned} V_3 &\geq \sum_{i=1}^{n-k}\binom{n-i}{2k}\binom{2k}{k}\left\{(1-2p^{k+1}+p^{2k+1})^{i-1}-(1-p^{k+1})^{2i-2}\right\} \\ &\quad -\sum_{i=1}^{n-k}\sum_{m=2}^{k}\binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1} \\ &\quad \left\{2p^{-\binom{m}{2}}(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1}+2(1-p^{k+1})^{2i-2}\right\} \\ &= R_4 - R_3; \end{aligned}$$

here we call the positive part of the lower bound R_4 and the negative part R_3. For V_4 we use the trivial lower bound $V_4 \geq 0$. Hence, we have:

$$\mathrm{Var}(T_{k+1}) \geq p^{2\binom{k+1}{2}}(R_4 - 8R_1 - 8R_2 - R_3).$$

Let us now upper bound R_1:

$$\begin{aligned} R_1 &\leq \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} \frac{(n-j)^{2k+1-m-q}}{(2k+1-m-q)!}\frac{(2k+1-m-q)^k}{k!}\frac{(k+1)^m}{m!} \\ &\quad \frac{(j-i+1)^{q-1}}{(q-1)!}(1-p^{k+1})^{i+j-2} \\ &\leq \sum_{i<j}^{n-k}\sum_{q=1}^{\min(k+1,j-i)} k\frac{n^{2k-q}}{1}\frac{(2k-1)^k}{k!}\frac{(k+1)^k}{1}\frac{n^{q-1}}{1}(1-p^{k+1})^{i+j-2} \\ &\leq (k+1)^{k+1}\frac{n^{2k-1}}{(k-1)!}(2k-1)^k\sum_{i<j}^{\infty}(1-p^{k+1})^{i+j-2} \\ &= \frac{n^{2k-1}(2k-1)^k(k+1)^{k+1}}{(k-1)!}\frac{1-p^{k+1}}{(2-p^{k+1})p^{2k+2}}. \end{aligned}$$

Noting that $(1-p^{k+1})^{j-i}(1-p^{k+1}-p^k+p^{2k+1-m})^{i-1} \leq (1-p^{k+1})^{j-1}$, we can bound R_2 in an identical way:

$$R_2 \leq \frac{n^{2k-1}(2k-1)^k(k+1)^{k+1}}{(k-1)!}p^{-\binom{k}{2}}\sum_{i<j}^{\infty}(1-p^{k+1})^{j-1}$$

$$= \frac{n^{2k-1}(2k-1)^k(k+1)^{k+1}}{(k-1)!} p^{-\binom{k}{2}} \frac{1-p^{k+1}}{p^{2k+2}}.$$

Noting that $(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1} \leq (1-p^{k+1})^{i-1}$ and $(1-p^{k+1})^{2i-2} \leq (1-p^{k+1})^{i-1}$ we proceed to bound R_3:

$$\begin{aligned}
R_3 &\leq \sum_{i=1}^{n-k} \sum_{m=2}^{k} \binom{n-i}{2k+1-m} \binom{2k+1-m}{k} \binom{k}{m-1} 2(p^{-\binom{m}{2}}+1)(1-p^{k+1})^{i-1} \\
&\leq \sum_{i=1}^{n-k} \sum_{m=2}^{k} \frac{n^{2k+1-m}(2k+1-m)^k k^{m-1}}{k!} 2(p^{-\binom{k}{2}}+1)(1-p^{k+1})^{i-1} \\
&\leq \sum_{i=1}^{n-k} k \frac{n^{2k+1-2}(2k+1-2)^k k^{k-1}}{k!} 2(p^{-\binom{k}{2}}+1)(1-p^{k+1})^{i-1} \\
&\leq \frac{n^{2k-1}(2k-1)^k k^k}{k!} 2(p^{-\binom{k}{2}}+1) \sum_{i=1}^{\infty} (1-p^{k+1})^{i-1} \\
&= \frac{n^{2k-1}(2k-1)^k k^k}{k!} 2(p^{-\binom{k}{2}}+1) p^{-k-1}.
\end{aligned}$$

To lower bound R_4 we just take the $i=2$ term:

$$\begin{aligned}
R_4 &\geq \binom{n-2}{2k} \binom{2k}{k} \left\{ (1-2p^{k+1}+p^{2k+1}) - (1-2p^{k+1}+p^{2k+2}) \right\} \\
&\geq \frac{(n-2)^{2k}}{(2k)^{2k}} \binom{2k}{k} p^{2k+1}(1-p).
\end{aligned}$$

Since R_1, R_2, R_3 are all at most of the order n^{2k-1} and R_2 is at least of the order n^{2k}, we have that for any fixed $k \geq 1$ and $p \in (0,1)$ there exists a constant $C_{p,k} > 0$ independent of n and a natural number $N_{p,k}$ such that for any $n \geq N_{p,k}$:

$$\mathrm{Var}(T_{k+1}) \geq p^{2\binom{k+1}{2}}(R_4 - 8R_1 - 8R_2 - R_3) \geq C_{p,k} n^{2k}.$$

□

References

Adler, R.J., Bobrowski, O., Weinberger, S.: Crackle: the homology of noise. In: Discrete and Computational Geometry, pp. 680–704 (2014)

Asai, R., Shah, J.: Algorithmic canonical stratifications of simplicial complexes. J. Pure Appl. Algebra **226**(9), 107051 (2022)

Barbour, A.D.: Stein's method for diffusion approximations. Probab. Theory Relat. Fields **84**(3), 297–322 (1990)

Barbour, A.D., Karoński, M., Ruciński, A.: A central limit theorem for decomposable random variables with applications to random graphs. J. Combin. Theory Ser. B **47**(2), 125–145 (1989)

Bauer, U., Rathod, A.: Hardness of approximation for Morse matching. In: SODA '19: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 2663–2674 (2019)
Bentkus, V.: On the dependence of the Berry–Esseen bound on dimension. J. Stat. Plann. Inference **113**(2), 385–402 (2003)
Bobrowski, O., Kahle, M.: Topology of random geometric complexes: a survey. J. Appl. Comput. Topol. **1**(3), 331–364 (2018)
Carlsson, G., et al.: Persistence barcodes for shapes. Int. J. Shape Model. **11**(02), 149–187 (2005)
Chatterjee, S., Meckes, E.: Multivariate normal approximation using exchangeable pairs. Alea **4**, 257–283 (2008)
Chen, L.H.Y., Goldstein, L., Shao, Q.M.: Normal Approximation by Stein's Method. Springer, Berlin (2011)
Costa, A., Farber, M.: Large random simplicial complexes I. J. Topol. Anal. **8**(03), 399–429 (2016)
Edelsbrunner, H., Harer, J.: Computational Topology: An Introduction. American Mathematical Society, Providence (2010)
Eichelsbacher, P., Rednoß, B.: Kolmogorov bounds for decomposable random variables and subgraph counting by the Stein–Tikhomirov method. Bernoulli **29**(3), 1821–1848 (2023)
Fang, X.: A multivariate CLT for bounded decomposable random vectors with the best known rate. J. Theor. Probab. **29**(4), 1510–1523 (2016)
Forman, R.: A user's guide to discrete Morse theor. Séminaire Lotharingien de Combinatoire **48**, B48c (2002)
Gan, H.L., Röllin, A., Ross, N.: Dirichlet approximation of equilibrium distributions in Cannings models with mutation. Adv. Appl. Probab. **49**(3), 927–959 (2017)
Gaunt, R.E., Li, S.: Bounding Kolmogorov distances through Wasserstein and related integral probability metrics. J. Math. Anal. Appl. **522**, 126985 (2023)
Ghrist, R.: Barcodes: the persistent topology of data. Bull. Amer. Math. Soc. (N.S.) **45**(1), 61–75 (2008)
Henselman-Petrusek, G., Ghrist, R.: Matroid filtrations and computational persistent homology (2016). arXiv:1606.00199
Janson, S., Nowicki, K.: The asymptotic distributions of generalized U-statistics with applications to random graphs. Probab. Theory Relat. Fields **90**(3), 341–375 (1991)
Joswig, M., Pfetsch, M.E.: Computing optimal Morse matchings. SIAM J. Discrete Math. **20**(1), 11–25 (2006)
Kahle, M.: Topology of random clique complexes. Discrete Math. **309**(6), 1658–1671 (2009)
Kahle, M.: Random geometric complexes. Discrete Comput. Geom. **45**(3), 553–573 (2011)
Kahle, M.: Sharp vanishing thresholds for cohomology of random flag complexes. Ann. Math. **25**, 1085–1107 (2014)
Kahle, M., Meckes, E.: Limit theorems for Betti numbers of random simplicial complexes. Homol. Homotopy Appl. **15**(1), 343–374 (2013)
Kasprzak, M.J., Peccati, G.: Vector-valued statistics of binomial processes: Berry–Esseen bounds in the convex distance (2022). arXiv preprint arXiv:2203.13137
Kaur, G., Röollin, A.: Higher-order fluctuations in dense random graph models. Electron. J. Probab. **26**, 1–36 (2021)
Korolyuk, V.S., Borovskich, Y.V.: Theory of U-Statistics, vol. 273. Springer, Berlin (2013)
Lampret, L.: Chain complex reduction via fast digraph traversal (2019). arXiv:1903.00783
Lee, A.J.: U-Statistics: Theory and Practice. Taylor & Francis (1990)
Linial, N., Meshulam, R.: Homological connectivity of random 2-complexes. Combinatorica **26**(4), 475–487 (2006)
McGinley, W.G., Sibson, R.: Dissociated random variables. In: Mathematical Proceedings of the Cambridge Philosophical Society, vol. 77, pp. 185–188. Cambridge University Press (1975)
Meckes, E.: On Stein's method for multivariate normal approximation. In: High Dimensional Probability V: The Luminy Volume, pp. 153–178. Institute of Mathematical Statistics (2009)
Mischaikow, K., Nanda, V.: Morse theory for filtrations and efficient computation of persistent homology. Discrete Comput. Geom. **50**(2), 330–353 (2013)
Nanda, V.: Local Cohomology and Stratification. Found. Comput. Math. **20**, 195–222 (2020)
Otter, N., et al.: A roadmap for the computation of persistent homology. EPJ Data Sci. **6**, 1–38 (2017)
Owada, T., Samorodnitsky, G., Thoppe, G.: Limit theorems for topological invariants of the dynamic multi-parameter simplicial complex. Stochast. Process. Appl. **138**, 56–95 (2021)
Privault, N., Serafin, G.: Normal approximation for sums of weighted U-statistics—application to Kolmogorov bounds in random subgraph counting. Bernoulli **26**(1), 587–615 (2020)

Raiíc, M.: A multivariate CLT for decomposable random vectors with finite second moments. J. Theor. Probab. **17**(3), 573–603 (2004)
Reinert, G., Röllin, A.: Random subgraph counts and U-statistics: multivariate normal approximation via exchangeable pairs and embedding. J. Appl. Probab. **47**(2), 378–393 (2010)
Schulte, M., Yukich, J.E.: Multivariate second order Poincaré inequalities for Poisson functionals. Electron. J. Probab. **24**, 1–42 (2019)
Spanier, E.: Algebraic Topology. McGraw-Hill (1966)