GOODNESS-OF-FIT VIA COUNT STATISTICS IN DENSE RANDOM SIMPLICIAL COMPLEXES

Tadas Temčinas[✉1], Vidit Nanda[✉*2] and Gesine Reinert[✉1]

[1]Department of Statistics, University of Oxford, United Kingdom

[2]Mathematical Institute, University of Oxford, United Kingdom

(Communicated by Henry Adams)

ABSTRACT. A key object of study in stochastic topology is a random simplicial complex. In this work we study a multi-parameter random simplicial complex model, where the probability of including a k-simplex, given the lower dimensional structure, is fixed. This leads to a conditionally independent probabilistic structure. This model includes the Erdős–Rényi random graph, the random clique complex as well as the Linial-Meshulam complex as special cases. The model is studied from both probabilistic and statistical points of view. We prove multivariate central limit theorems with bounds and known limiting covariance structure for the subcomplex counts and the number of critical simplices under a lexicographical acyclic partial matching. We use the CLTs to develop a goodness-of-fit test for this random model and evaluate its empirical performance. In order for the test to be applicable in practice, we also prove that the MLE estimators are asymptotically unbiased, consistent, uncorrelated and normally distributed.

1. **Introduction.** While complex data are often represented as graphs or networks [43], there is a growing interest in modelling complex systems beyond pairwise interactions between nodes [9]. Simplicial complexes, often used in topological data analysis (TDA), provide a rich mathematical representation of higher-order networks [8]; other representations are detailed for example in [5]. As in network science, the statistical analysis of simplicial complex data relies on random complexes as null models. In this paper, statistical understanding is achieved through a central limit theorem (CLT). Stochastic topology, which studies such random complexes, is partly motivated by its use in TDA [11, 12]. Despite the growing popularity of TDA, random simplicial complexes are rarely used in practice, partly because of the lack of results in parameter estimation and asymptotic distribution of relevant statistics. This work aims to fill in the gap by providing a probabilistic and statistical understanding of the so-called multiparameter random simplicial complex model $\mathbf{X}(n, \mathbf{p})$, which is rather general and includes several important special cases such as the Linial-Meshulam complex, the random clique complex $\mathbf{X}(n, p)$ and the Bernoulli random graph $\mathbf{G}(n, p)$. Here $\mathbf{p} = (p_1, p_2, \ldots, p_{n-1})$ is a vector of probabilities; p_1 is the probability of an edge, p_2 is the probability of a triangle to form a 2-simplex,

and in general, p_k is the probability of a k-vertex clique to form a $(k-1)-$simplex. In particular $\mathbf{G}(n,p)$ is a special case of $\mathbf{X}(n,\mathbf{p})$ with $\mathbf{p} = (p,0,\ldots,0)$. A formal definition is given in Section 2.1.

The motivation of this work is three-fold:

1. establishing results that allow the use of $\mathbf{X}(n,\mathbf{p})$ as a null-model in practical TDA applications;
2. establishing an understanding of multivariate count statistics which have been instrumental in the study of algebraic invariants (such as Euler characteristic, homology groups, and the fundamental group) of random simplicial complexes; and finally,
3. quantifying the extent to which discrete Morse theory can simplify homology computation in a random setting.

We address these three points in the so-called *dense* regime, where the vector $\mathbf{p} \in (0,1)^{n-1}$ does not depend on n (although some relaxation of this assumption is also discussed). A key observation which we establish in this paper is that asymptotically the covariance matrix of all subcomplex counts in $\mathbf{X}(n,\mathbf{p})$ has rank one, which might be surprising. Even though the model has conditionally independent randomness in higher dimensions, this gets lost in the limit and so the behaviour of this random vector is analogous to a vector of subgraph counts in the $\mathbf{G}(n,p)$ random graph model. Intuitively, this means that one subcomplex count asymptotically determines all the other ones. We shall also establish that the behaviour of the second count statistic - critical simplex counts of a lexicographical acyclic partial matching, as arising in discrete Morse theory - is slightly different. Next we cover the three motivations for this paper in more detail.

1.1. **Multiparameter complex as a null model.** In order to further statistical understanding of the model, we study maximum likelihood estimation in the $\mathbf{X}(n,\mathbf{p})$ model as well as goodness-of-fit tests. We prove that the maximum likelihood estimator (MLE) for $\mathbf{p}$ is asymptotically unbiased and consistent. Also, we provide non-asymptotic bounds on the normal approximation error as well as the limiting covariance structure of the estimator. To our knowledge, this is the first study of the MLE for this random model. The classic abstract MLE results do not apply here because the observed simplex indicator variables are neither independent nor identically distributed; instead we employ results from [51] which are based on a probabilistic technique called *Stein's method*.

Unfortunately, the vector of subcomplex counts has limiting covariance matrix of rank one, which makes it an unsuitable candidate for standard goodness-of-fit tests. Instead, we propose critical simplex counts of a lexicographical acyclic partial matching as test statistic, because this multivariate statistic has a nontrivial correlation matrix also in the limit. We analyse its performance empirically in a simulation study and show that it is able to distinguish between $\mathbf{X}(n,\mathbf{p})$ and a selection of geometric simplicial complex models. This is the first work proposing a goodness-of-fit test for the $\mathbf{X}(n,\mathbf{p})$ model.

1.2. **Distributional approximation of count statistics.** This work contains several probabilistic results, which provide some understanding of the probabilistic structure and which also motivate the choice of statistics for goodness-of-fit tests. In the study of random graphs, distributional approximations of subgraph counts play an important role [47, 4]. Surprisingly, general subcomplex counts in $\mathbf{X}(n,\mathbf{p})$, have not been studied from a distributional approximation point of view before,

only from a large deviations point of view [48]. In the literature only CLTs for a special case of subcomplex counts - face counts - appear in $\mathbf{X}(n, \mathbf{p})$ and related models [45, 19]. In this work we prove a multivariate CLT for subcomplex counts with non-asymptotic bounds on convergence rates with respect to both smooth test functions as well as convex set indicator functions. Non-asymptotic bounds can be useful in applications since the observed data, however large, is never infinite. Moreover we prove a multivariate CLT with bounds on rates for critical simplex counts of a lexicographical acyclic partial matching and derive a formula for the asymptotic covariance matrix, which generically has full rank, in contrast to the subcomplex counts. In order to prove a multivariate CLT for the subcomplex counts, we establish a multivariate CLT for generalised U-statistics under some assumptions, which might be of independent interest.

Proving the existence of certain subcomplexes have been a crucial step in showing results about homology, persistent homology, and the fundamental group in the regimes when these algebraic invariants are non-trivial with high probability [1, 46, 42]. Distributional approximation of certain subcomplexes have also been used to deduce distributional approximation results for Betti number in the random clique complex [32]. We hope that understanding the multivariate counts of arbitrary subcomplexes in $\mathbf{X}(n, \mathbf{p})$ as well as critical simplex counts under a lexicographical matching will pave a way to understand the multivariate distribution of Betti numbers in the dense regime.

1.3. **Effectiveness of discrete Morse theory.** Discrete Morse theory [21] provides a powerful, flexible and widely-used mechanism for simplifying the machine computation of simplicial homology and related algebraic invariants [22, 41, 17, 6, 52]. The basic idea is to construct a *partial matching* which pairs adjacent simplices subject to a global acyclicity constraint; the homology of the original complex is then entirely determined by (a chain complex constructed from) the unpaired simplices, the so-called *critical* simplices. Several strategies have been proposed for constructing acyclic partial matchings which admit relatively few number of critical simplices, thus greatly easing the linear-algebraic burden of computing homology. Simultaneously, considerable efforts have been invested in understanding the efficacy of such reductions in terms of lowering the number of simplices to consider — the complexity of finding optimal matchings is discussed in [27] while certain empirical calculations have been described in [2]. In this paper we provide a multivariate normal approximation for vectors of critical simplex counts of *lexicographical* discrete Morse functions on $\mathbf{X}(n, p)$. This acyclic partial matching is a standard tool in TDA which is used in some of the popular TDA software packages, for example [6]. The multivariate normal approximation in our paper thus furnishes a convenient benchmark for testing the efficacy of homology-preserving reductions via discrete Morse theory on random simplicial complexes.

1.4. **Main results.** When presenting our main results here, we assume that the reader is familiar with random simplicial complexes and the multiparameter random simplicial complex $\mathbf{X}(n, \mathbf{p})$. The subsequent Section 2 provides the necessary background for the readers who are less familiar with these notions.

There are two central theoretical contributions of this work: one regarding the properties of the MLE in $\mathbf{X}(n, \mathbf{p})$ and one regarding the distributional approximation of the mentioned count statistics. We first give a simplified version of our MLE results; they are stated asymptotically for simplicity and clarity but we prove

the results non-asymptotically with explicit constants. To see the results in full generality and detail, we refer to Theorems 3.2 and 3.5 as well as Lemma 3.4.

Theorem 1.1. *Consider the random simplicial complex* $\mathbf{X}(n, \mathbf{p})$. *Assume there is a known and fixed integer* $d \geq 1$ *such that* $\mathbf{p} = (p_1, p_2, \ldots, p_{n-1})$ *is a vector of probabilities where* $p_i = 0$ *for* $i > d$ *and* $p_i \in (0, 1)$ *for* $i \leq d$. *Let* $\hat{p}_i$ *be the MLE of* p_i *for some fixed* $i \leq d$. *Let* $W_i = W_i(n) = \binom{n}{i+1}^{\frac{1}{2}} (\hat{p}_i - p_i)$ *and* $Z_i \sim \mathcal{N}\left(0, (1 - p_i) \prod_{j=1}^{i} p_j^{-\binom{i+1}{j+1}}\right)$. *Then*

1.

$$\lim_{n \to \infty} \mathbb{E}\{\hat{p}_i\} = p_i \ and \ \lim_{n \to \infty} \mathrm{Var}(\hat{p}_i) = 0;$$

2. for any fixed $1 \leq i < j \leq d$ *we have*

$$\lim_{n \to \infty} \mathrm{Cov}(W_i, W_j) = 0;$$

3. for any fixed $1 \leq i \leq d$,

$$\lim_{n \to \infty} \sup_{x \in \mathbb{R}} |\mathbb{P}(W_i \leq x) - \mathbb{P}(Z_i \leq x)| = 0.$$

Here we recognize

$$\sup_{x \in \mathbb{R}} |\mathbb{P}(W_i \leq x) - \mathbb{P}(Z_i \leq x)| = \sup_{x \in \mathbb{R}} |\mathbb{E}\{\mathbb{I}(W_i \in (-\infty, x])\} - \mathbb{E}\{\mathbb{I}(Z_i \in (-\infty, x])\}|$$

as the Kolmogorov-distance between the distribution of W_i and the distribution of Z_i, which is based on indicator test functions $\mathbb{I}(z \in (-\infty, x])$, taking the value 1 if $z \in (-\infty, x]$ and 0 otherwise. A natural multivariate generalisation of this distance is to take the supremum over indicator functions of convex sets. Similarly, taking instead the supremum over Lipschitz(1)-test functions yields Wasserstein distance.

Next we state a theorem that contains a simplified version of the multivariate normal approximation results, for both, simplex counts and and critical simplex counts. Even though the two random vectors are related from a topological point of view, a CLT for one of them does not automatically translate into a CLT for the other: the dependence structure in the two cases are quite different, requiring slightly different proofs. We quantify the approximation error in terms of both smooth test functions and convex set indicator functions. Even though an error vanishing asymptotically implies convergence in distribution in both cases, for finite n, the convex set indicator functions give us a stronger result that is also of interest in practice when, for example, estimating confidence regions. For the non-simplified version of the results we refer the reader to Theorems 5.2, 6.4, 6.6, Corollary 5.4, and Lemma 6.4. For the statement we let k' be the largest number such that for all $i \leq k'$ we have $p_i = 1$; $k' = 0$ if none of the probabilities equal 1.

Theorem 1.2. *Consider the random simplicial complex* $\mathbf{X}(n, \mathbf{p})$. *Let* $W^{(1)} \in \mathbb{R}^d$ *be an appropriately scaled and centered vector of subcomplex counts. Let* $W^{(2)} \in \mathbb{R}^d$ *be an appropriately scaled and centered count vector of simplices that are critical under a lexicographical acyclic partial matching. 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* $h : \mathbb{R}^d \to \mathbb{R}$ *be a three times partially differentiable function whose third partial derivatives are Lipschitz continuous and bounded. Let* $\mathcal{K}$ *be the class of convex sets in* $\mathbb{R}^d$.

1. *There is a constant $B_{5.4} > 0$, independent of n such that*

$$\left| \mathbb{E}h(W^{(1)}) - \mathbb{E}h(\Sigma_1^{\frac{1}{2}} Z) \right| \leq \sup_{i,j,k \in [d]} \left\| \frac{\partial^3 h}{\partial x_i \partial x_j \partial x_k} \right\|_\infty B_{5.4} n^{-\frac{k'+2}{2}}$$

and

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W^{(1)} \in A) - \mathbb{P}(\Sigma_1^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{5.4}^{\frac{1}{4}} n^{-\frac{k'+2}{8}}.$$

Moreover, for any $1 \leq i < j \leq d$ we have

$$\lim_{n \to \infty} (\Sigma_1)_{i,j} = 1.$$

2. *There is a constant $B_{6.6.1} > 0$ independent of n and a natural number $N_{6.6.1}$ such that for any $n \geq N_{6.6.1}$ we have*

$$\left| \mathbb{E}h(W^{(2)}) - \mathbb{E}h(\Sigma_2^{\frac{1}{2}} Z) \right| \leq B_{6.6.1} \sup_{i,j,k \in [d]} \left\| \frac{\partial^3 h}{\partial x_i \partial x_j \partial x_k} \right\|_\infty n^{-\frac{k'+2}{2}}.$$

Also, there is a constant $B_{6.6.2} > 0$ independent of n and a natural number $N_{6.6.2}$ such that for any $n \geq N_{6.6.2}$ we have

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W^{(2)} \in A) - \mathbb{P}(\Sigma_2^{\frac{1}{2}} Z \in A)| \leq B_{6.6.2} n^{-\frac{k'+2}{8}}.$$

Moreover, for any $1 \leq i < j \leq d$ there is a function $\sigma_\infty(i,j)$ such that

$$\lim_{n \to \infty} (\Sigma_2)_{i,j} = \sigma_\infty(i,j).$$

If $p_k \in (0,1)$ for all $k > k'$, then then $0 \leq \sigma_\infty(i,j) < 1$.

The constant $B_{5.4}$ is given explictly in Corollary 5.4, and the explicit expression of $\sigma_\infty(i,j)$ is found in Lemma 6.4. We also propose a testing procedure based on critical simplex counts and empirically evaluate its performance. The procedure is described in Algorithm 1 and the results are presented in Section 3.2.

1.5. **Related work.** Although subcomplex counts in $\mathbf{X}(n,\mathbf{p})$ are natural objects to consider, surprisingly, they have not yet been extensively studied. Nonetheless, a recent paper studies large deviation bounds for subcomplex counts in the $\mathbf{X}(n,\mathbf{p})$ model [48] and a CLT for face counts can be found in [45]. Another recent paper [19] proves a CLT for a special kind of subcomplex counts in a special case of the $\mathbf{X}(n,\mathbf{p})$ model, the number of isolated faces in the Linial-Meshulam random simplicial complex. As subcomplex counts can be seen as an example of generalised U-statistics, results from [26, Section 10.7] and [24, Chapter 11.3] could, with some work, in principle be adapted to prove convergence in distribution for the multivariate subcomplex counts; in [18] the author extends the results from [26] to study generalised U-statistics not only for graphs but also for hypergraphs and therefore for simplicial complexes in particular, but no non-asymptotic bounds on the approximation error are available. Zhang [53] used the exchangeable pair approach in Stein's method to derive a univariate CLT for the graph version of generalised U-statistics. These results only apply to graph statistics rather than the higher order type of statistics that we are interested in here.

Counts of simplices that are critical under a lexicographical acyclic partial matching have been studied for the random clique complex $\mathbf{X}(n,p)$ [51, Section 4]. For the $\mathbf{X}(n,\mathbf{p})$ model, only the expected value has been studied so far [7, Section 8].

There has been some limited work on statistics for random simplicial complexes. In [12, 54] the authors have noted a sufficient statistic for the $\mathbf{X}(n,\mathbf{p})$ model but the

MLE had not been investigated. Goodness-of-fit tests using count statistics have been proposed in the context of random graphs [33, 13, 39]. Topological goodness-of-fit tests have been suggested for point processes [10], sliced spatial data [14], and cylindrical networks [34], but not, to the best of our knowledge, for $\mathbf{X}(n, \mathbf{p})$.

1.6. **Organisation and notation.** In Section 2 we introduce topological definitions and the random models of interest. In Section 3 we study the properties of the MLE of the parameter $\mathbf{p}$ in $\mathbf{X}(n, \mathbf{p})$ and perform a simulation study to verify the validity of the proposed goodness-of-fit test. In Section 4 we introduce and prove probabilistic results that are needed to prove Theorem 1.2. It also contains a multivariate CLT for generalised U-statistics that might be of independent interest. In Section 5 we study multivariate subcomplex counts and prove a multivariate CLT with an explicit bound; we also describe the limiting covariance. In Section 6 we study critical simplex counts under a lexicographical acyclic partial matching. A multivariate CLT with an explicit bound is proved and the limiting covariance described. Finally, the Appendix A contains some technical computations that are needed for the proofs of Section 6.

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, $\mathcal{P}(X)$ for its powerset, and given a positive integer k we write $\binom{X}{k} = \{ t \in \mathcal{P}(X) \mid |t| = k \}$ for the collection of subsets of X 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 $|f|_k = \sup_{i_1, i_2, \ldots, i_k \in [d]} \|\partial_{i_1 i_2 \ldots i_k} f\|_\infty$ for any integer $k \geq 1$, as long as the quantities exist. Here $\| \cdot \|_\infty$ denotes the supremum norm. For a positive integer d we define a class of test functions $h : \mathbb{R}^d \to \mathbb{R}$, as follows. We say $h \in \mathcal{H}_d$ iff h is three times partially differentiable with third partial derivatives being Lipschitz and $|h|_3 < \infty$. The class of convex sets in $\mathbb{R}^d$ is denoted by $\mathcal{K}_d$. When the dimension of the space is clear from the context we may suppress the subscript d for these sets. 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 $[n]$, unless stated otherwise. We also use Bachmann-Landau asymptotic notation: we say $f(n) = O(g(n))$ iff $\limsup_{n \to \infty} \frac{|f(n)|}{g(n)} < \infty$, $f(n) = \Omega(g(n))$ iff $\liminf_{n \to \infty} \frac{f(n)}{g(n)} > 0$, and $f(n) = \omega(g(n))$ iff $\lim_{n \to \infty} \frac{f(n)}{g(n)} = \infty$. Given two real random variables X, Y we write $d_{\mathrm{KS}}(X, Y) := \sup_{x \in \mathbb{R}} |\mathbb{P}(X \leq x) - \mathbb{P}(Y \leq x)|$ for the Kolmogorov-Smirnov distance between their distributions. Moreover, we use the following bounds on binomial coefficients

$$\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \frac{n^k}{k!} < \left(\frac{ne}{k}\right)^k. \tag{1}$$

2. **Topological preliminaries.** For completeness, we include some topological preliminaries, see also [51].

2.1. **First definitions.** Firstly, we recall the notion of a simplicial complex [50, Ch 3.1]; these are special kinds of hypergraphs that provide higher-dimensional generalisations of a graph and constitute data structures of interest across algebraic, applied, and computational topology.

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

1. for each $v \in [n]$ the singleton $\{v\}$ lies in $\mathcal{L}$, and

2. if $t \in \mathcal{L}$ and $s \subseteq t$ then $s \in \mathcal{L}$.

The **dimension** of a simplicial complex $\mathcal{L}$ is $\dim(\mathcal{L}) := \max_{s \in \mathcal{L}} |s| - 1$. Elements of a simplicial complex are called **simplices**. If s is a simplex, then its dimension is $\dim(s) := |s| - 1$. A simplex of dimension k can be called a k-simplex. The k-skeleton of a simplicial complex $\mathcal{L}$ is the subcomplex $\mathcal{L}^{(k)} = \{ s \in \mathcal{L} \mid |s| \leq k + 1 \}$ of dimension k as long as $k < \dim(\mathcal{L})$. Note that the notion of a one-dimensional simplicial complex is equivalent to the notion of a graph, with the vertex set $[n]$ and edges as one-dimensional simplices.

Given $k_1, k_2 \in \mathbb{N}$ and two simplicial complexes $\mathcal{L}_1$ on $[k_1]$ and $\mathcal{L}_2$ on $[k_2]$, a **simplicial map** from $\mathcal{L}_1$ to $\mathcal{L}_2$ is a function $f : [k_1] \to [k_2]$ such that for any $s \in \mathcal{L}_1$ we have that $f(s) \in \mathcal{L}_2$. If f is bijective, it is called a **simplicial isomorphism**. For any finite simplicial complex $\mathcal{L}$ on $[n]$ we denote the set of simplicial complexes on $[n]$ that are isomorphic to $\mathcal{L}$ by $[\mathcal{L}]$. Note that $[\mathcal{L}]$ can be smaller than the number of automorphisms on $\mathcal{L}$. For example, if $\mathcal{L} = \{\{1\}, \{2\}, \{1, 2\}\}$ is just an edge then $[\mathcal{L}] = \{\mathcal{L}\}$. Even though the edge has two automorphisms (the identity and the one swapping the two vertices), both automorphisms give rise to the same simplicial complex. Indeed the swapping automorphism gives $\{\{2\}, \{1\}, \{2, 1\}\}$, which is the same set as $\mathcal{L}$ itself.

Assuming $k_1 \leq k_2$ it is also meaningful to count the number of copies of $\mathcal{L}_1$ in $\mathcal{L}_2$, which we define as the number of order-preserving injective simplicial maps from any element of $[\mathcal{L}_1]$ to $\mathcal{L}_2$. We call this the **subcomplex count** of $\mathcal{L}_1$ in $\mathcal{L}_2$. Subcomplex counts are investigated in Section 5 and concrete examples are given there.

Definition 2.1. The k-simplex is equal to the set $\mathcal{P}([k + 1]) \setminus \{\varnothing\}$, seen as a simplicial complex. That is, the vertices are $\{1, 2, \ldots, k + 1\}$ and any subset of the vertex set is a simplex. The **hollow** k-simplex is equal to the set $\mathcal{P}([k + 1]) \setminus \{[k + 1], \varnothing\}$ seen as a simplicial complex. That is, a hollow k-simplex is the $(k - 1)$-skeleton of the k-simplex and so its dimension is $k - 1$.

Definition 2.2. Let $[n]$ be the vertex set. Let $\mathbf{p} = (p_1, p_2, \ldots, p_{n-1}) \in [0, 1]^{n-1}$ be a finite sequence of parameters. We first build a random hypergraph H where any subset $s \subseteq [n]$ of size k is included independently of any other subsets with probability p_{k-1}. The **multi-parameter random simplicial complex** $\mathbf{X}(n, \mathbf{p})$ is the largest simplicial complex contained in H. That is, for any subset $s \subseteq [n]$, we have $s \in \mathbf{X}(n, \mathbf{p})$ iff $s \in H$ and for any $t \subseteq s$ we have $t \in H$.

The model $\mathbf{X}(n, \mathbf{p})$ is a general random simplicial complex model, defined by Kahle in [31] and extensively studied from a topological rather than combinatorial point of view by, amongst others, Costa and Farber in [15, 16]; it includes many families of random simplicial complexes as special cases. For example, if $\mathbf{p} = (p, 0, \ldots, 0)$, then we recover the $\mathbf{G}(n, p)$ random graph. If we set $\mathbf{p} = (p, 1, \ldots, 1)$, then we recover the random clique complex of $\mathbf{G}(n, p)$, that has been extensively studied, for example, in [28, 30, 32]. If we set $\mathbf{p} = (1, \ldots 1, p_k, 0, \ldots, 0)$, then we recover the k-dimensional Linial-Meshulam random complex that has also been extensively studied before [37, 38]. For a survey of stochastic topology, see [11, 12].

2.2. **Discrete Morse theory.** A **partial matching** on a simplicial complex $\mathcal{L}$ is a collection

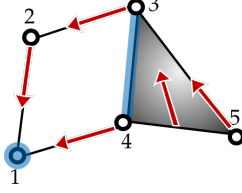$$V = \{ (s, t) \mid s \subseteq t \in \mathcal{L} \text{ and } |t| - |s| = 1 \}$$

FIGURE 1. Lexicographical matching given by the red arrows. Critical simplices are highlighted in blue.

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

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

We refer the interested reader to [21] for an introduction to discrete Morse theory and to [41] for an illustration of its use for simplifying computations of persistent homology. This work addresses how much computational improvement one should expect to gain on a random input when using a specific type of acyclic partial matching, defined below.

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

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

Now consider the pairings

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

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

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

Example 2.4. Consider the simplicial complex $\mathcal{L}$ depicted in Figure 1. 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 $\mathcal{L}$ is

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

Note that $\{3, 4\}$ cannot be matched because the set $I_{\mathcal{L}}(\{3, 4\})$ is empty. Also, in any lexicographical matching $\{1\}$ is always critical as there are no vertices with a

smaller label and hence the set $I_{\mathcal{L}}(\{1\})$ 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 $\mathcal{L}$ – a significant reduction in computational effort.

3. Maximum likelihood estimation and goodness-of-fit.

3.1. **Maximum likelihood estimation.** Recall the random simplicial complex model $\mathbf{X}(n, \mathbf{p})$ as given in Definition 2.2. Here we study a maximum likelihood estimator under this random model assuming that there is some known d such that $p_i \in (0, 1)$ if $i \leq d$ and $p_i = 0$ if $i > d$. Therefore, we have d parameters in this model to estimate, namely, $(p_1, p_2, \ldots, p_d) \in (0, 1)^d$. With slight abuse of notation, we write $\mathbf{p} = (p_1, p_2, \ldots, p_d)$. Let $\mathcal{L}$ be a fixed simplicial complex on the vertex set $[n]$, which we consider to be our observed data. Recall Definition 2.1, detailing the notions of a simplex and a hollow simplex. For each $i \in [n]$, let s_i be the number of i-simplices and h_i be the number of hollow i-simplices. Note that $s_i \leq h_i$ because every i-simplex contains a hollow i-simplex. Also note that $h_1 = \binom{n}{2}$ is just the number of pairs of vertices and hence it is a deterministic value.

As noted in [12, 54], for any $i \in [d]$ a sufficient statistic for p_i is the pair (s_i, h_i). We therefore can estimate p_i only for $i \leq I(\mathcal{L}) := \max\{i \in [d] \mid h_i \neq 0\}$. Note that if $h_i = 0$, then $h_j = 0$ and $s_j = 0$ for all $j \geq i$ and if $h_i \neq 0$, then $h_j \neq 0$ for all $j \leq i$. In the dense regime, if d is not too large, we can prove that $I(\mathcal{L}) = d$ with high probability (see Lemma 3.1). Assuming $\mathbf{p} \in (0, 1)^d$, the log-likelihood is well-defined and equal to

$$l(\mathbf{p}; \mathcal{L}) = \sum_{i=1}^{I(\mathcal{L})} \{s_i \ln(p_i) + (h_i - s_i) \ln(1 - p_i)\}.$$

For any $i \in [I(\mathcal{L})]$ we have $\frac{\partial l}{\partial p_i} = \frac{s_i}{p_i} - \frac{h_i - s_i}{1 - p_i}$. Solving the partials equal to zero and considering the boundary cases when $s_i = 0$ or $s_i = h_i$, we get $\hat{p}_i = \frac{s_i}{h_i}$ for $i \leq I(\mathcal{L})$. If $i > I(\mathcal{L})$, then $s_i = 0$ and so the log-likelihood does not depend on p_i. In this case $\hat{p}_i$ is not unique; we choose to set $\hat{p}_i = 0$. Summarising, we have

$$\hat{p}_i = \mathbb{1}(h_i \neq 0, s_i \neq h_i) \frac{s_i}{h_i}. \tag{2}$$

For any $i \neq j \in [I(\mathcal{L})]$, if $0 \neq h_i < s_i$ we have

$$\frac{\partial^2 l}{\partial p_i \partial p_j}(\hat{\mathbf{p}}) = 0; \qquad \frac{\partial^2 l}{\partial p_i^2}(\hat{\mathbf{p}}) = -h_i^2 \left(\frac{1}{s_i} + \frac{1}{h_i - s_i}\right) < 0.$$

For $i \neq j \in [I(\mathcal{L}) + 1, d]$ the corresponding second partial derivatives all vanish; the $I(\mathcal{L}) \times I(\mathcal{L})$ Hessian is negative-definite, and hence $\hat{\mathbf{p}}$ indeed is a local (and global) maximum of the likelihood. Classical theorems from the theory of MLE do not apply here because the simplex indicators are neither independent nor identically distributed. Hence we derive asymptotic results for the MLE in this paper.

For each $i \in [d]$ let H_i be the random variable counting the number of hollow i-simplices and let S_i be the random variable counting the number of solid i-simplices. Let h_i be an integer such that $1 \leq h_i \leq \binom{n}{i+1}$. Then by definition of $\mathbf{X}(n, \mathbf{p})$ for any $i \in [d]$ we have that $S_i | \{H_i = h_i\} \sim \text{Binomial}(h_i, p_i)$. Using this property of S_i being conditionally binomial we can prove that the estimator is asymptotically

unbiased and consistent. We start the investigation by proving concentration of H_i for $i \in [d]$ and hence that $I(\mathcal{L}) = d$ with high probability.

For the rest of the section we write $p^* := \min \{ p_j \mid j \in [d] \}$ and $P_i = \prod_{j=1}^{i-1} p_j^{\binom{i+1}{j+1}}$; then P_i is the probability of a hollow i-simplex to be present in $\mathbf{X}(n, \mathbf{p})$.

Lemma 3.1. *Fix $i \in [2, d]$ with $d \leq \log_2 \ln n + \log_2 \log_2 \ln n - \log_2(-\ln p^*)$ and $a \geq 0$, Then*

$$\mathbb{P}(H_i = 0) \leq \exp\left(-\frac{\binom{n}{i+1}(p^*)^{2^i - i - 1}}{4(i^2 - 1)n}\right);$$

$$\mathbb{P}(|H_i - \mu_i| \geq a) \leq 2\exp\left(-\frac{2a^2}{(i+1)^{i+2}\binom{n}{i-1}\binom{n}{i+1}}\right).$$

In particular, in this regime, $\mathbb{P}(H_i = 0) \leq \exp\left(-n^{\log_2 \log_2 \ln n(1+o(1))}\right)$ and $\lim_{n\to\infty} \mathbb{P}(I(\mathbf{X}(n, \mathbf{p})) = d) = 1$ as long as p^* is a positive constant. Here $o(1)$ relates to the limit $n \to \infty$.

Proof. Fix $i \in [2, d]$ and $\epsilon > 0$. For each $s \in \binom{[n]}{i+1}$ we define $X_s = \mathbb{1}(s \text{ spans a hollow } i\text{-simplex})$. Then

$$H_i = \sum_{s \in \binom{[n]}{i+1}} X_s \qquad \text{and hence} \qquad \mu_i := \mathbb{E}(H_i) = \binom{n}{i+1} P_i.$$

Write $\Delta_i = \sum_{s \neq t \in \binom{[n]}{i+1}} \mathbb{1}(|s \cap t| \geq 2) \mathbb{E}\{X_s X_t\}$. As long as $\Delta_i \geq \mu_i$, we can bound the denominator in Janson's inequality [23, Theorem 1] by $\mu_i + \Delta_i \leq 2\Delta_i$. Hence we get

$$\mathbb{P}(H_i = 0) \leq \exp\left(-\frac{\mu_i^2}{4\Delta_i}\right)$$

$$\leq \exp\left(-\frac{\binom{n}{i+1}^2 P_i^2}{4\sum_{m=2}^{i}\binom{n}{i+1}\binom{i+1}{m}\binom{n}{i+1-m}P_i^2 \prod_{j=1}^{m-1} p_j^{-\binom{m}{j+1}}}\right)$$

$$= \exp\left(-\frac{\binom{n}{i+1}}{4\sum_{m=2}^{i}\binom{i+1}{m}\binom{n}{i+1-m}\prod_{j=1}^{m-1} p_j^{-\binom{m}{j+1}}}\right)$$

$$\leq \exp\left(-\frac{\binom{n}{i+1}}{4\sum_{m=2}^{i}(p^*)^{-2^m+1+m}\binom{i+1}{m}\binom{n}{i+1-m}}\right).$$

To bound this expression we note that

$$\sum_{m=2}^{i}(p^*)^{-2^m+1+m}\binom{i+1}{m}\binom{n}{i+1-m}$$

$$\leq (i-1)\max_{2 \leq m \leq i}(p^*)^{-2^m+1+m}\binom{i+1}{m}\binom{n}{i+1-m}$$

$$= (i-1)(p^*)^{-2^i+1+i}n(i+1).$$

For fixed p^* this expression tends to 0 for $i = \log_2 \ln n + \log_2 \log_2 \ln n + c$. For such i we get the bound

$$\mathbb{P}(H_i = 0) \leq \exp\left(-\frac{\binom{n}{i+1}(p^*)^{2^i-i-1}}{4(i^2-1)n}\right).$$

As long as $c \leq -\log_2(-\ln p^*)$, the bound is at most $\exp\left(-n^{\log_2 \log_2 \ln n (1+o(1))}\right)$. The first inequality follows by recalling that $H_k = 0$ implies $H_j = 0$ for all $j \geq k$.

For concentration, we apply the large deviation result Corollary 2.2 in [25]. Fix a particular $s \in \binom{[n]}{i+1}$. Note that X_s and X_t are dependent iff $|s \cap t| \geq 2$. So the number of variables in the sum of H_i that X_s depends on is $\sum_{m=2}^{i+1} \binom{i+1}{m}\binom{n-i-1}{i+1-m} \leq (i+1)^{i+1}(i-1)\binom{n}{i-1} \leq (i+1)^{i+2}\binom{n}{i-1}$. Now applying the large deviation result gives the second inequality. $\square$

Theorem 3.2. *Let $i \in [2,d]$ with $d \leq \log_2 \ln n + \log_2 \log_2 \ln n - \log_2(-\ln p^*)$ and $\epsilon > 0$. Then*

1.
$$p_i\left(1 - \exp\left(-n^{\log_2 \log_2 \ln n (1+o(1))}\right)\right) \leq \mathbb{E}(\hat{p}_i) \leq p_i;$$

2.
$$\mathrm{Var}(\hat{p}_i) \leq 2p_i(1-p_i)\left((i+1)^{i+1}n^{-1+\epsilon}(p^*)^{-2^{i+1}+i+3} + \exp\left(-\frac{2n^{2\epsilon}(i-1)^{i-1}}{(i+1)e^{2i}}\right)\right)$$
$$+ p_i^2 \exp\left(-n^{\log_2 \log_2 \ln n (1+o(1))}\right).$$

In particular, for all $i \in [2,d]$, $\lim_{n\to\infty} \mathbb{E}(\hat{p}_i) = p_i$, and for $i < \log_2 \ln n + \log_2(1-\epsilon) - \log_2(-\ln(p^*))$ we have $\lim_{n\to\infty} \mathrm{Var}(\hat{p}_i) = 0$ as long as p^* is a positive constant.

Proof. To prove the first part of the theorem we use the law of total expectation and fact that S_i is conditionally binomial, to obtain

$$\mathbb{E}(\hat{p}_i) = \sum_{h_i=1}^{\binom{n}{i+1}} \frac{h_i p_i}{h_i}\mathbb{P}(H_i = h_i) = p_i\mathbb{P}(H_i \neq 0).$$

Applying the bounds from Lemma 3.1 the result follows. Turning to the variance of the estimator,

$$\mathrm{Var}(\hat{p}_i) = \mathbb{E}(\hat{p}_i^2) - \mathbb{E}(\hat{p}_i)^2 = \sum_{h_i=1}^{\binom{n}{i+1}} \left(\frac{p_i(1-p_i)}{h_i} + p_i^2\right)\mathbb{P}(H_i = h_i) - p_i^2\mathbb{P}(H_i \neq 0)^2$$

$$= \sum_{h_i=1}^{\binom{n}{i+1}} \frac{p_i(1-p_i)}{h_i}\mathbb{P}(H_i = h_i) + p_i^2\mathbb{P}(H_i \neq 0)(1 - \mathbb{P}(H_i \neq 0)).$$

For the second term we can immediately use the first inequality in Lemma 3.1. For the first term, we use the large deviation inequality from Lemma 3.1. Fix $\epsilon \in (0,1)$ and write $a = n^{i+\epsilon}$. Then we have

$$\sum_{h_i=1}^{\binom{n}{i+1}} \frac{p_i(1-p_i)}{h_i}\mathbb{P}(H_i = h_i)$$

$$= \sum_{h_i=1}^{\mu_i-a} \frac{p_i(1-p_i)}{h_i} \mathbb{P}(H_i = h_i) + \sum_{h_i=\mu_i-a+1}^{\mu_i+a-1} \frac{p_i(1-p_i)}{h_i} \mathbb{P}(H_i = h_i)$$

$$+ \sum_{h_i=\mu_i+a}^{\binom{n}{i+1}} \frac{p_i(1-p_i)}{h_i} \mathbb{P}(H_i = h_i)$$

$$\leq p_i(1-p_i) \left(\sum_{h_i=1}^{\mu_i-a} \mathbb{P}(H_i = h_i) + \sum_{h_i=\mu_i+a}^{\binom{n}{i+1}} \mathbb{P}(H_i = h_i) \right) + \sum_{h_i=\mu_i-a+1}^{\mu_i+a-1} \frac{p_i(1-p_i)}{\mu_i - a}$$

$$\leq 2p_i(1-p_i) \exp\left(-\frac{2n^{2\epsilon}(i-1)^{i-1}}{(i+1)e^{2i}} \right) + 2n^{i+\epsilon} \frac{p_i(1-p_i)}{\mu_i - \frac{1}{2}\mu_i}.$$

To finish the proof note that $\mu_i^{-1} n^{i+\epsilon} \leq (i+1)^{i+1} n^{-1+\epsilon} (p^*)^{-2^{i+1}+i+3}$. $\qquad\square$

Remark 3.3. Note that if $i = 1$ then $\hat{p}_i$ is just the MLE of the parameter p in a binomial distribution, which is very well studied and known to be unbiased and consistent. Also, [20, Theorem 3.1] shows that there is some constant c such that for $i \geq \log_2 \ln n + \log_2 \log_2 \ln n - c$ we have no i-simplices with high probability as long as the parameter $\mathbf{p} \in (0,1)^{n-1}$ does not depend on n. In light of this fact we can see that in Theorem 3.2 the upper bound on i is optimal up to an additive constant in the case of the expectation. In the case of the variance, the bound is off by a factor of $\log_2 \log_2 \ln n$, which is likely because the large deviation inequality is sub-optimal for large i.

Lemma 3.4. *For any* $1 \leq i < j \leq d \leq \log_2 \ln n + \log_2 \log_2 \ln n - \log_2(-\ln p^*)$ *we have*

$$- p_j \binom{n}{i+1}^{\frac{1}{2}} \binom{n}{j+1}^{\frac{1}{2}} \exp\left(-n^{\log_2 \log_2 \ln n (1+o(1))} \right)$$

$$\leq \mathrm{Cov}(W_i, W_j) \leq p_i p_j \binom{n}{i+1}^{\frac{1}{2}} \binom{n}{j+1}^{\frac{1}{2}} \exp\left(-n^{\log_2 \log_2 \ln n (1+o(1))} \right).$$

Moreover, we have $\lim_{n\to\infty} \mathrm{Cov}(W_i, W_j) = 0$ *as long as* p^* *is a positive constant.*

Proof. We will use a shorthand $F_n = \binom{n}{i+1}^{\frac{1}{2}} \binom{n}{j+1}^{\frac{1}{2}}$. Then

$$\mathrm{Cov}(W_i, W_j)$$

$$= F_n \mathrm{Cov}(\hat{p}_j, \hat{p}_j)$$

$$= F_n \left(\mathbb{E}\left\{ \frac{S_i}{H_i} \frac{S_j}{H_j} \mathbb{1}(H_i H_j \neq 0) \right\} - \mathbb{E}\left\{ \frac{S_i}{H_i} \mathbb{1}(H_i \neq 0) \right\} \mathbb{E}\left\{ \frac{S_j}{H_j} \mathbb{1}(H_j \neq 0) \right\} \right).$$

Assume $i < j$ and let us focus on $\mathbb{E}\left\{ \frac{S_i}{H_i} \frac{S_j}{H_j} \mathbb{1}(H_i H_j \neq 0) \right\}$. The main idea is to condition on the joint distribution of $(S_1, H_2, S_2, \ldots, H_{j-1}, S_{j-1}, H_j)$. Under this conditioning, S_j and H_j are deterministic so we can take it out of the conditional expectation and then use the fact that $S_j | H_j \sim \mathrm{Binomial}(H_j, p_j)$. The details are

as follows:

$$\mathbb{E}\left\{\frac{S_i}{H_i}\frac{S_j}{H_j}\mathbb{1}\left(H_iH_j \neq 0\right)\right\}$$

$$= \mathbb{E}\left\{\mathbb{E}\left(\frac{S_i}{H_i}\frac{S_j}{H_j}\mathbb{1}\left(H_iH_j \neq 0\right)\Big|(S_1, H_2, S_2, \ldots, H_{j-1}, S_{j-1}, H_j)\right)\right\}$$

$$= \mathbb{E}\left\{\frac{S_i}{H_i}\frac{1}{H_j}\mathbb{1}\left(H_iH_j \neq 0\right)\mathbb{E}\left(S_j\Big|(S_1, H_2, S_2, \ldots, H_{j-1}, S_{j-1}, H_j)\right)\right\}$$

$$= \mathbb{E}\left\{\frac{S_i}{H_i}\frac{1}{H_j}\mathbb{1}\left(H_iH_j \neq 0\right)\mathbb{E}\left(S_j\Big|H_j\right)\right\}$$

$$= \mathbb{E}\left\{\frac{S_i}{H_i}\frac{1}{H_j}\mathbb{1}\left(H_iH_j \neq 0\right)p_jH_j\right\}$$

$$= p_j\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i \neq 0\right)\mathbb{1}\left(H_j \neq 0\right)\right\}.$$

Now we can bound this quantity from above and below. Let us start with an upper bound:

$$p_j\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i \neq 0\right)\mathbb{1}\left(H_j \neq 0\right)\right\} \leq p_j\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i \neq 0\right)\right\} = p_ip_j\mathbb{P}(H_i \neq 0).$$

For the lower bound, we use the following argument:

$$p_j\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i \neq 0\right)\mathbb{1}\left(H_j \neq 0\right)\right\}$$

$$= p_j\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i \neq 0\right)\left(1 - \mathbb{1}\left(H_j = 0\right)\right)\right\}$$

$$\geq p_ip_j\mathbb{P}(H_i \neq 0) - p_j\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i \neq 0 \cap H_j = 0\right)\right\}$$

$$\geq p_ip_j\mathbb{P}(H_i \neq 0) - p_j\mathbb{P}(H_j = 0).$$

Here the last inequality follows from the fact that $S_i \leq H_i$.

Therefore, we get a bound on the covariance:

$$\mathbb{P}(H_i \neq 0) - \frac{1}{p_i}\mathbb{P}(H_j = 0) - \mathbb{P}(H_i \neq 0)\mathbb{P}(H_j \neq 0)$$

$$\leq \frac{\text{Cov}(W_i, W_j)}{F_np_ip_j}$$

$$\leq \mathbb{P}(H_i \neq 0) - \mathbb{P}(H_i \neq 0)\mathbb{P}(H_j \neq 0).$$

The bound can be rewritten as:

$$\mathbb{P}(H_i \neq 0)\mathbb{P}(H_j = 0) - \frac{1}{p_i}\mathbb{P}(H_j = 0) \leq \frac{\text{Cov}(W_i, W_j)}{F_np_ip_j} \leq \mathbb{P}(H_i \neq 0)\mathbb{P}(H_j = 0).$$

In fact, the following simpler bound suffices to prove the assertion:

$$-\frac{1}{p_i}\mathbb{P}(H_j = 0) \leq \frac{\text{Cov}(W_i, W_j)}{F_np_ip_j} \leq \mathbb{P}(H_j = 0).$$

Since $1 \leq i < j$, we have $j \geq 2$ and so the following bound from Lemma 3.1 give us the result

$$\mathbb{P}(H_j = 0) \leq \exp\left(-n^{\log_2 \log_2 \ln n(1+o(1))}\right).$$

$\square$

Theorem 3.5. *Let* $i \in [2, d]$ *with* $d \leq \log_2 \ln n + \log_2 \log_2 \ln n - \log_2(-\ln p^*)$. *Also, let* $W_i = \sqrt{\binom{n}{i+1} P_i}(\hat{p}_i - p_i)$ *and* $Z_i \sim \mathcal{N}(0, p_i(1 - p_i))$. *Then*

$$d_{\mathrm{KS}}(W_i, Z_i) \leq \frac{\sqrt{10} + 3}{3\sqrt{\pi}} (p_i(1 - p_i)P_i)^{-\frac{1}{2}} \binom{n}{i+1}^{-\frac{1}{2}}$$
$$+ 2(n - i)^{-2} P_i^{-1} (i+1)^{\frac{i+6}{2}} \ln^{\frac{1}{2}} \binom{n}{i+1} + 2\binom{n}{i+1}^{-2}.$$

Remark 3.6. Theorem 3.5 gives in particular, for $i < \log_2 \ln n + 1 - \log_2(-\ln(p^*))$ we have $\lim_{n\to\infty} d_{\mathrm{KS}}(W_i, Z_i) = 0$ as long as p^* is a positive constant. This assumption can be relaxed; it suffices that P_i and $p_j, j \leq d$ depend on n in such a way that the expression in Theorem 3.5 tends to 0 with $n \to \infty$ while ensuring that $p_i(1 - p_i)$ is bounded away from 0.

Proof. Let $Z \sim \mathcal{N}(0, 1)$. Recall that H_i is the number of hollow i-simplices in $\mathbf{X}(n, \mathbf{p})$. Define the interval $I_i := (\mathbb{E}\{H_i\} - a, \mathbb{E}\{H_i\} + a)$, where $a > 0$ is to be chosen later. Let h be an indicator function of an interval $(-\infty, b)$ for some $b \in \mathbb{R}$. Set $c(x, i) = x^{\frac{1}{2}} \binom{n}{i+1}^{-\frac{1}{2}} P_i^{-\frac{1}{2}} (p_i(1 - p_i))^{-\frac{1}{2}}$. Recall that given two real random variables X, Y we write $d_{\mathrm{KS}}(X, Y) := \sup_{x \in \mathbb{R}} |\mathbb{P}(X \leq x) - \mathbb{P}(Y \leq x)|$ for the Kolmogorov-Smirnov distance.

To prove the theorem we condition on H_i. We use large deviation bounds Lemma 3.1 and inside the high probability region use a classic CLT comparing a binomial to a normal. For comparing a binomial variable to a normal we use [49, Theorem 1] and for comparing a normal to a normal we use [36, Section 6.2].

$\mathbb{E}\{|h(W_i) - h(Z_i)|\}$

$$= \sum_{x \in I_i} \mathbb{E}\{|h(W_i) - h(Z_i)| \, | H_i = x\} \mathbb{P}(H_i = x)$$

$$+ \mathbb{E}\{|h(W_i) - h(Z_i)| \, | H_i \notin I_i\} \mathbb{P}(H_i \notin I_i)$$

$$\leq \sum_{x \in I_i} d_{\mathrm{KS}}(W_i | H_i = x, Z_i) \mathbb{P}(H_i = x) + 2\exp\left(-\frac{2a^2}{(i+1)^{i+2}\binom{n}{i-1}\binom{n}{i+1}}\right)$$

$$\leq \sum_{x \in I_i} d_{\mathrm{KS}}(c(x, i)W_i | H_i = x, c(x, i)Z_i) \mathbb{P}(H_i = x) + 2\exp\left(-\frac{2a^2}{(i+1)^{i+2}\binom{n}{i-1}\binom{n}{i+1}}\right)$$

$$\leq \sum_{x \in I_i} (d_{\mathrm{KS}}(c(x, i)W_i | H_i = x, Z) + d_{\mathrm{KS}}(c(x, i)Z_i, Z)) \mathbb{P}(H_i = x)$$

$$+ 2\exp\left(-\frac{2a^2}{(i+1)^{i+2}\binom{n}{i-1}\binom{n}{i+1}}\right)$$

Note that $c(x, i)W_i|H_i = x$ is a binomial random variable with $n = x$ and $p = p_i$, which has been centered and scaled to have unit variance. Therefore, we can apply a classic CLT to bound the Kolmogorov-Smirnov distance between a standard normal and a centered and scaled binomial [49, Theorem 1]. This gives the bound

$$\sum_{x \in I_i} d_{\mathrm{KS}}\left(c(x, i)W_i|H_i = x, Z\right) \mathbb{P}(H_i = x)$$

$$\leq \sum_{x \in I_i} \frac{\sqrt{10} + 3}{6\sqrt{2\pi}} \cdot \frac{p_i^2 + (1 - p_i)^2}{\sqrt{xp_i(1 - p_i)}} \mathbb{P}(H_i = x)$$

$$\leq \frac{\sqrt{10} + 3}{6\sqrt{2\pi}} \cdot \frac{p_i^2 + (1 - p_i)^2}{\sqrt{(\mathbb{E}\{H_i\} - a)p_i(1 - p_i)}} \mathbb{P}(H_i \in I_i).$$

Now to compare two normal variables, namely, $c(x, i)Z_i$ and Z, we note that $\mathrm{Var}(c(x, i)Z_i) = \frac{x}{\mathbb{E}\{H_i\}}$ and $\mathrm{Var}(Z) = 1$, and we use [36, Section 6.2] (with $f_1 = f_2 = 1$) to obtain

$$\sum_{x \in I_i} d_{\mathrm{KS}}\left(c(x, i)Z_i, Z\right) \mathbb{P}(H_i = x)$$

$$\leq \sum_{x \in I_i} \left|\frac{x}{\mathbb{E}\{H_i\}} - 1\right| \sqrt{\frac{\mathbb{E}\{H_i\}}{x}} \mathbb{P}(H_i = x)$$

$$\leq \max\left\{\left|\frac{\mathbb{E}\{H_i\} + a}{\mathbb{E}\{H_i\}} - 1\right|, \left|\frac{\mathbb{E}\{H_i\} - a}{\mathbb{E}\{H_i\}} - 1\right|\right\} \sqrt{\frac{\mathbb{E}\{H_i\}}{\mathbb{E}\{H_i\} - a}} \mathbb{P}(H_i \in I_i)$$

$$\leq \frac{a}{\mathbb{E}\{H_i\} - a}.$$

Putting everything together, we get

$$\mathbb{E}\{|h(W_i) - h(Z_i)|\}$$

$$\leq \frac{\sqrt{10} + 3}{6\sqrt{\pi}} \frac{p_i^2 + (1 - p_i)^2}{\sqrt{\mathbb{E}\{H_i\}\, p_i(1 - p_i)}} + \frac{a}{\mathbb{E}\{H_i\} - a} + 2\exp\left(-\frac{2a^2}{(i + 1)^{i+2}\binom{n}{i-1}\binom{n}{i+1}}\right).$$

Now for a bound in Kolmogorov distance we take the supremum over all half-open interval indicator functions h and recall that $\mathbb{E}\{H_i\} = \binom{n}{i+1}P_i$. Picking $a = \left(\binom{n}{i+1}\binom{n}{i-1}(i + 1)^{i+2}\ln\binom{n}{i+1}\right)^{\frac{1}{2}}$ to approximately balance the last two summands gives the assertion. □

3.2. A goodness of fit test.
Theorem 6.6 provides a bound on the normal approximation error of critical simplex counts with respect to the lexicographical matching. In network analysis goodness of fit tests based on asymptotic distributions of random graph statistics are available, see for example [35, 44]. Inspired by this approach, we develop a goodness of fit test based on the asymptotic distribution of the critical simplex counts. It is known that in certain geometric random simplicial complexes the lexicographical matching behaves differently compared to the $\mathbf{X}(n, \mathbf{p})$ model, see [29, proof of Theorem 5.1]. With this in mind, we investigate if the statistic can distinguish between geometric models and $\mathbf{X}(n, \mathbf{p})$.

The starting point for the testing procedure is an observed simplicial complex $\mathcal{L}$. The null hypothesis is that the observed data is a sample from the model family $\mathbf{X}(n, \mathbf{p})$, which is tested against the general alternative in which the model family is not specified. For this test let $\hat{\mathbf{p}}_0$ be the observed value of the MLE under the null hypothesis, given in (2). For $T = (T_1, \ldots, T_n)$ the vector of critical simplex counts in $\mathbf{X}(n, \mathbf{p})$, we calculate the standardised statistics $W_i = W(T_i) = \mathrm{Var}(T_i)^{-\frac{1}{2}}(T_i - \mathbb{E}\{T_i\})$ for $i \in [n]$. Theorem 1.2 combined with Lemma 3.4 give that $W = (W_1, \ldots, W_n) \approx \mathrm{MVN}(0, \Sigma)$, where Σ is the limiting, diagonal covariance matrix of W. Hence, in the $\mathbf{X}(n, \mathbf{p})$ model, $W^T \Sigma^{-1} W$ is approximately chi-square distributed with n degrees of freedom. Based on this asymptotic result, as a proof of concept we propose and empirically study the following simple testing procedure that is described in Algorithm 1:

1. We observe a simplicial complex $\mathcal{L}$ and compute the corresponding value $\hat{\mathbf{p}}_0$ of the estimator $\hat{\mathbf{p}}$ from (2).
2. We compute the observed value $\mathbf{t} = (t_1, t_2, \ldots, t_n)$ of the critical simplex counts $T = (T_1, T_2, \ldots, T_n)$.
3. Based on $\hat{\mathbf{p}}_0$ we estimate $\mathbb{E}\{T_i\}$ and $\mathrm{Var}(T_i)$ by replacing $\mathbf{p}$ by $\hat{\mathbf{p}}_0$ in the explicit formulas. We denote these estimators $\widehat{\mathbb{E}\{T_i\}}$ and $\widehat{\mathrm{Var}(T_i)}$.
4. We calculate $\mathbf{w} = (w_1, w_2, \ldots, w_n)$ where $w_i = \widehat{\mathrm{Var}(T_i)}^{-\frac{1}{2}}(t_i - \widehat{\mathbb{E}\{T_i\}})$ for all $i \in [n]$.
5. We reject the null at approximate significance level $\alpha \in (0, 1)$ iff $w^T \Sigma^{-1} w > \chi_n^2(1 - \alpha)$, where $\chi_n^2(p)$ is the quantile function of the chi-squared distribution with n degrees of freedom.

Algorithm 1 Goodness-of-fit test

1: **procedure** PERFORMTEST($\mathcal{L}, \alpha, T, \Sigma$)   ▷ $\mathcal{L}$ is a finite observed simplicial complex, $\alpha \in (0, 1)$ is significance level, T is a test statistic, Σ is the limiting covariance of a centered and normalised version of T
2:        $\hat{\mathbf{p}}_0 := \text{COMPUTEMLE}(\mathcal{L})$ ▷ Computes the observed value $\hat{\mathbf{p}}_0$ of the MLE $\hat{\mathbf{p}}$
3:        $(t_1, t_2, \ldots, t_k) := \text{COMPUTESTATISTIC}(\mathcal{L}, T)$ ▷ Computes the observed value of a chosen T
4:      **for all** $i \in [k]$ **do**
5:            $\widehat{\mathbb{E}\{T_i\}} := \text{ESTIMATEMEAN}(\hat{\mathbf{p}}_0)$ ▷ Estimates the mean by replacing $\mathbf{p}$ with $\hat{\mathbf{p}}_0$
6:            $\widehat{\mathrm{Var}(T_i)} := \text{ESTIMATEVARIANCE}(\hat{\mathbf{p}}_0)$ ▷ Estimates the variance by replacing $\mathbf{p}$ with $\hat{\mathbf{p}}_0$
7:            $w_i := \widehat{\mathrm{Var}(T_i)}^{-\frac{1}{2}}(t_i - \widehat{\mathbb{E}\{T_i\}})$ ▷ Normalises the statistic
8:      **end for**
9:      $w := (w_1, w_2, \ldots, w_k)$
10:      **if** $w^T \Sigma^{-1} w > \chi_k^2(1 - \alpha)$ **then**   ▷ $\chi_k^2(p)$ is the quantile function of the χ_k^2 distribution
11:            **return** False   ▷ The null is rejected
12:      **else**
13:            **return** True   ▷ Insufficient evidence to reject the null
14:      **end if**
15: **end procedure**

Remark 3.7. This testing procedure can be improved in the following ways. First, it ignores the parameter estimation error, second, it ignores the estimation error of the mean and variances, and third, it ignores the normal approximation error. By using Theorem 3.5, one could account for the parameter estimation error. It is possible to account for estimation error when estimating the mean and the variance of the test statistic as well. Moreover, Theorem 6.6 gives explicit bounds on the quality of the normal approximation and hence it is also possible to account for this error. These improvements are outside the scope of this paper and are deferred to future work.

3.2.1. *Empirical simulation study.* Here we empirically study the goodness of fit test to distinguish between $\mathbf{X}(n, \mathbf{p})$ and a random geometric simplicial complex model, where geometry is present only in a particular dimension. We start by defining a soft geometric random simplicial complex model.

Definition 3.8. Let $\mathcal{X}_n$ be a set of n points that are i.i.d. in the unit d-cube $[0,1]^d \subseteq \mathbb{R}^d$. For $i \in [n-1]$ let $\phi_i : ([0,1]^d)^{i+1} \to [0,1]$ be a symmetric function. We inductively define a **soft random geometric simplicial complex** $\mathbb{X}(\mathcal{X}_n, \phi_1, \phi_2, \ldots \phi_{n-1})$ as the random simplicial complex with vertex set $\mathcal{X}_n$ such that any i-simplex $s = \{s_1, s_2, \ldots, s_{i+1}\} \subseteq \mathcal{X}_n$ is included with probability $\phi_i(s_1, s_2, \ldots, s_{i+1})$ if every proper subset of s is already in the complex. Note that if we take the functions ϕ_i to be constant functions that do not depend on the vertex locations for all $i \in [n-1]$, then we recover the $\mathbf{X}(n, \mathbf{p})$ model.

In the simulation study we compare $\mathbf{X}(n, \mathbf{p})$ with three geometric models by defining an interpolation process that starts with $\mathbf{X}(n, \mathbf{p})$ and finishes at each of the models. We are interested to see at which point in the interpolation there is enough geometry in the model for the statistical test to pick it up. We use the following three models:

1. Let $\mathcal{X}_{150}$ be a set of 150 points that are i.i.d. in the unit 7-cube $[0,1]^7 \subseteq \mathbb{R}^7$. Given four points $x_1, x_2, x_3, x_4 \in [0,1]^7$ define $A(x_1, x_2, x_3, x_4)$ to be the volume of the tetrahedron defined by the four points. The model is $\mathbb{X}(\mathcal{X}_{150}, \phi_1, \phi_2, \ldots \phi_{149})$ where $\phi_1(x_1, x_2) = 0.5$, $\phi_2(x_1, x_2, x_3) = 0.5$, $\phi_3(x_1, x_2, x_3, x_4) = \mathbb{1}\left(A(x_1, x_2, x_3, x_4) \leq 0.09\right)$ and $\phi_i(x_1, x_2, \ldots, x_{i+1}) = 0$ for all $i \neq 1, 2, 3$. This is just a 3-dimensional $\mathbf{X}(n, \mathbf{p})$ model with the exception that only tetrahedra of sufficiently small volume appear. We later refer to this model as the **tetrahedron model**.

2. Let $\mathcal{X}_{75}$ be a set of 75 points that are i.i.d. in the unit 3-cube $[0,1]^3 \subseteq \mathbb{R}^3$. Given three points $x_1, x_2, x_3 \in [0,1]^3$ define $A(x_1, x_2, x_3)$ to be the area of the triangle defined by the three points. The model is $\mathbb{X}(\mathcal{X}_{75}, \phi_1, \phi_2, \ldots \phi_{74})$ where $\phi_1(x_1, x_2) = 0.5$, $\phi_2(x_1, x_2, x_3) = \mathbb{1}\left(A(x_1, x_2, x_3) \leq 0.09\right)$ as well as $\phi_i(x_1, x_2, \ldots, x_{i+1}) = 0$ for all $i \neq 1, 2$. This is just a 2-dimensional $\mathbf{X}(n, \mathbf{p})$ model with the exception that only triangles of sufficiently small area appear. We later refer to this model as the **triangle model**.

3. Let $\mathcal{X}_{75}$ be a set of 75 points that are i.i.d. in the unit 3-cube $[0,1]^3 \subseteq \mathbb{R}^3$. The model here is $\mathbb{X}(\mathcal{X}_{75}, \phi_1, \phi_2, \ldots \phi_{74})$ where $\phi_1(x_1, x_2) = \mathbb{1}\left(\|x_1 - x_2\|_2 \leq 0.4924\right)$ and $\phi_i(x_1, x_2, \ldots, x_{i+1}) = 0.5$ for all $i \neq 1$. The distance threshold is chosen so that the edge density of the geometric graph is 0.5. This model is just a classical geometric random graph with higher simplices filled in combinatorially just like in $\mathbf{X}(n, \mathbf{p})$. We later refer to this model as the **edge model**.

In the case of the tetrahedron model, we define a sequence of models interpolating between the geometric model and $\mathbf{X}(n, \mathbf{p})$ as follows. Given $0 \le \epsilon_1 \le \epsilon_2$ we define the model $\mathbb{X}(\mathcal{X}_{150}, \phi_1, \phi_2, \phi_3^{(\epsilon_1, \epsilon_2)}, \ldots, \phi_{149})$ where we set $\phi_1(x_1, x_2) = 0.5$, $\phi_2(x_1, x_2, x_3) = 0.5$, $\phi_3(x_1, x_2, x_3, x_4) = \mathbb{1}\left(A(x_1, x_2, x_3, x_4) \le \epsilon_1\right) + 0.5 \times$ $\mathbb{1}\left(A(x_1, x_2, x_3, x_4) \in (\epsilon_1, \epsilon_2)\right)$ and $\phi_i(x_1, x_2, \ldots, x_{i+1}) = 0$ for all $i \ne 1, 2, 3$. That is, we connect any two vertices independently with probability 0.5, and any triangle becomes a 2-simplex, also independently, with probability 0.5. Finally, any hollow 3-simplex is filled in with probability 1 if its volume is at most ϵ_1 and, independently, with probability 0.5 if its volume is between ϵ_1 and ϵ_2. No hollow 3-simplex with volume at least ϵ_2 gets filled in. Note that if we have $\epsilon_1 = \epsilon_2 = 0.09$, then we recover the geometric model and if $\epsilon_1 = 0$, $\epsilon_2 = 1$, then we recover the $\mathbf{X}(150, (0.5, 0.5, 0.5, 0))$ model. We start the interpolation with $\epsilon_1 = 0$ and $\epsilon_2 = 0.5$, which is very close to the $\mathbf{X}(150, (0.5, 0.5, 0.5, 0))$ model. We increase ϵ_1 and decrease ϵ_2 at each step until we reach $\epsilon_1 = \epsilon_2 = 0.09$. We sample 100 simplicial complexes at each step of the interpolation and report the number of times the null hypothesis is not rejected at 0.95 significance level.

In the edge model, we define a very similar interpolation to the tetrahedron model interpolation. Given $0 \le \epsilon_1 \le \epsilon_2$ we define $\mathbb{X}(\mathcal{X}_{75}, \phi_1, \phi_2^{(\epsilon_1, \epsilon_2)}, \ldots \phi_{74})$ where $\phi_1(x_1, x_2) = 0.5$, $\phi_2(x_1, x_2, x_3) = \mathbb{1}\left(A(x_1, x_2, x_3) \le \epsilon_1\right) + 0.5 \times \mathbb{1}\left(A(x_1, x_2, x_3) \in (\epsilon_1, \epsilon_2)\right)$ and $\phi_i(x_1, x_2, \ldots, x_{i+1}) = 0$ for all $i \ne 1, 2$. We start the interpolation at $\epsilon_1 = 0$ and $\epsilon_2 = 0.66$, which is close to the $\mathbf{X}(75, (0.5, 0.5, 0))$ model. We increase ϵ_1 and decrease ϵ_2 at each step until we reach $\epsilon_1 = \epsilon_2 = 0.09$. We sample 100 simplicial complexes at each step of the interpolation and report the number of times the null hypothesis is not rejected at 0.95 approximate significance level.

Finally, in the edge model we perform an analogous interpolation again. Given $0 \le \epsilon_1 \le \epsilon_2$ we define $\mathbb{X}(\mathcal{X}_{75}, \phi_1, \phi_2^{(\epsilon_1, \epsilon_2)}, \ldots \phi_{74})$ where $\phi_1(x_1, x_2) = \mathbb{1}\left(\|x_1 - x_2\|_2 \le \epsilon_1\right)$ $+ 0.5 \times \mathbb{1}\left(\|x_1 - x_2\|_2 \in (\epsilon_1, \epsilon_2)\right)$ as well as $\phi_i(x_1, x_2, \ldots, x_{i+1}) = 0.5$ for all $i \ne 1$. We start the interpolation at $\epsilon_1 = 0$ and $\epsilon_2 = \sqrt{3}$, which is exactly the $\mathbf{X}(75, (0.5, 0.5, 0.5 \ldots, 0.5))$ model. We increase ϵ_1 and decrease ϵ_2 at each step until we reach $\epsilon_1 = \epsilon_2 = 0.4924$.

For each of these three perturbation models we sample 100 simplicial complexes at each step of the interpolation and report the number of times the null hypothesis of an $\mathbf{X}(\mathbf{n}, \mathbf{p})$ model is not rejected (the *number of passes*) at 0.95 approximate significance level. Ideally the number of passes should be 0 when $\Delta\epsilon = 0$, corresponding to the purely geometric model, and increase sharply with increasing $\Delta\epsilon$, to 100 when $\Delta\epsilon = 1$, corresponding to the $\mathbf{X}(\mathbf{n}, \mathbf{p})$ model.

We compare goodness of fit tests based on three statistics: triangle counts, centered triangle counts [33, 13], and the critical simplex counts. $\Delta\epsilon$. Because the triangle counts and centered triangle counts only rely on 1-dimensional information, we expect these statistics to detect the difference only when comparing $\mathbf{X}(n, \mathbf{p})$ to the last model. The results for the tetrahedron, triangle, and edge models are summarised in Figure 2, 3, and 4, respectively; while the results depend on the particular values for ϵ_1 and ϵ_2, for ease of presentation we plot the test results against $\Delta\epsilon := \epsilon_2 - \epsilon_1$. Table 1, 2, and 3 in Appendix D contain the full set of results for different ϵ_1 and ϵ_2. For these tables we start with a regular grid and refine it when the number of passes of the goodness of fit test based on critical simplex counts drops below 99.
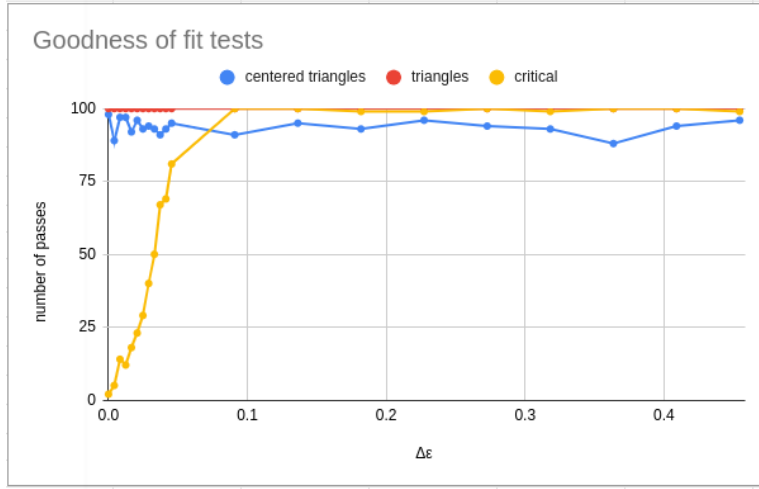
FIGURE 2. Goodness of fit testing results of the tetrahedron model. The x-axis corresponds to $\Delta\epsilon := \epsilon_2 - \epsilon_1$.
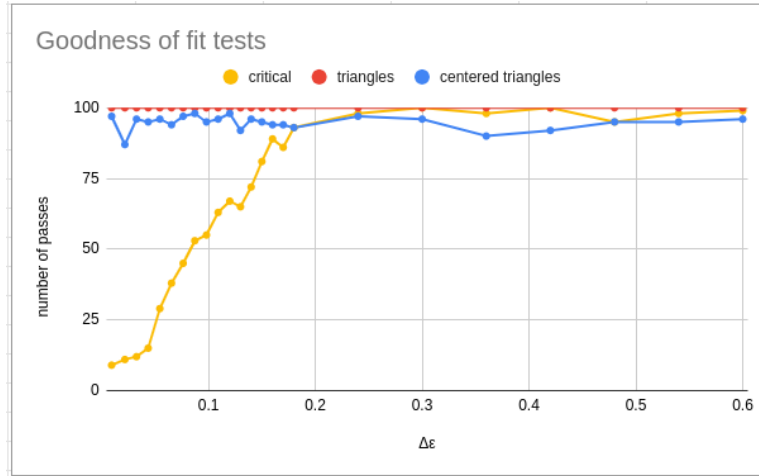

FIGURE 3. Goodness of fit testing results of the triangle model. The x-axis corresponds to $\Delta\epsilon := \epsilon_2 - \epsilon_1$.

The simulation results. We find that critical simplex counts under the lexicographical matching can differentiate between the $\mathbf{X}(n, \mathbf{p})$ model, which is fully combinatorial, and models which depend on geometric information only in a particular dimension. Even if the model is far from a fully geometric one (like a Vietoris-Rips or Čech complex of a point process in $\mathbb{R}^d$ would be), we can still distinguish it from the $\mathbf{X}(n, \mathbf{p})$ based on the critical simplex counts. In dimension 1 it is relatively easy to distinguish between a low-dimensional geometric random graph and $\mathbf{G}(n, p)$ as it has been noted by others [13, 39]. In this case, as seen in Figure 4, the centered triangle counts can tolerate more combinatorial noise and it starts detecting geometry earlier than the number of critical simplices. However, we have noticed based on
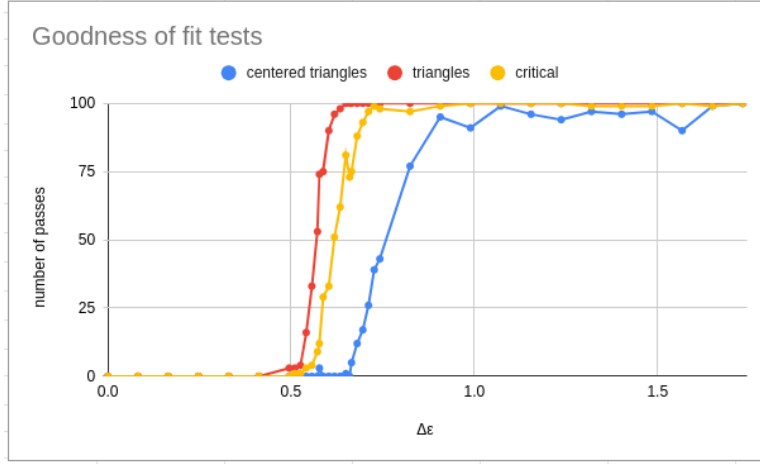
FIGURE 4. Goodness of fit testing results of the edge model. The x-axis corresponds to $\Delta\epsilon := \epsilon_2 - \epsilon_1$.

the test simulations that the centered triangle counts are less stable and affected by parameter estimation errors. Quite often we would see less than 90% passes in the cases where the null hypothesis is true at approximate 0.95 significance level. As an aside, triangles in random geometric graphs behave very differently to triangles in Bernoulli random graphs, and hence it is no surprise that for small $\Delta\epsilon$ it is easily detected that the simplicial complex is not a $\mathbf{X}(n, \mathbf{p})$ complex. In contrast, in the $\mathbf{X}(n, \mathbf{p})$ models, the triangle statistic does not depend on higher-order simplices and hence cannot detect the presence or absence of geometry.

Generalising centered subgraph counts to centered subcomplex counts and investigating their use in goodness-of-fit tests would be a fruitful research project. An advantage of critical simplex counts, beyond their performance as a test statistic, is that when performing persistent homology computations in practice one quite often tries to reduce the computational cost by applying a lexicographical acyclic partial matching and so a statistical test based on the number of critical simplices can be performed at very little added cost.

4. **Probabilistic tools.** In this section we present the probabilistic tools that we use to prove the multivariate CLTs for the subcomplex counts and the number of critical simplices under the lexicographical matching.

4.1. **Multivariate CLT for generalised U-statistics.** To prove the results in this section we rely on the theorems from [51], which are briefly introduced in Appendix B for completeness.

Assume we have k not necessarily independent collections of independent random elements, where the i-th collection is indexed by $\binom{[n]}{i}$, the set of all subsets of $[n]$ of size i. That is, the i-th collection is given by

$$\left\{ \xi^{(i)}_\alpha \right\}_{\alpha \in \binom{[n]}{i}}.$$

Here i is any positive integer not bigger than n. For example, one can think of $\xi_\alpha^{(i)}$ as an indicator of a hyper-edge α of size i in some random hypergraph model in which the hyperedges occur independently.

Now we consider the sequence of collections

$$\left\{\xi_\alpha^{(1)}\right\}_{\alpha\in\binom{[n]}{1}}, \left\{\xi_\alpha^{(2)}\right\}_{\alpha\in\binom{[n]}{2}}, \ldots, \left\{\xi_\alpha^{(k)}\right\}_{\alpha\in\binom{[n]}{k}}.$$

Assume that the random elements $\left\{\xi_\alpha^{(i)}\right\}_{\alpha\in\binom{[n]}{i}}$ take values in a measurable set $\mathbb{X}^{(i)}$ for all $i \in [k]$ and that there is some measurable set $\mathbb{X}$ such that for all $i \in [k]$ we have $\mathbb{X}^{(i)} \subseteq \mathbb{X}$.

Remark 4.1. Given a subset $s \subseteq [n]$ of size $m \geq k$ let $\binom{s}{i}$ be the set of i-subsets of s. This set of subsets may be lexicographically ordered as follows. Consider $j := \{j_1, j_2, \ldots, j_i\} \subset s$ with $j_1 < j_2 < \cdots < j_i$ and similarly $j' = \{j'_1, \ldots, j'_i\}$. Then, j strictly precedes j' in the lexicographical ordering if and only if there exists some $\nu \in \{0, 1, \ldots, i-1\}$ such that $j_\mu = j'_\mu$ for all $\mu \leq \nu$ and $j_{\nu+1} < j'_{\nu+1}$. It is readily checked that this notion of precedence yields a total ordering on the elements of $\binom{s}{i}$.

Define a sequence (with respect to the lexicographical ordering on $\binom{s}{i}$) of the random elements

$$\mathcal{X}_s^{(i)} = (\xi_\alpha^{(i)})_{\alpha\in\binom{s}{i}}. \tag{3}$$

Definition 4.2. Given two integers $1 \leq k, m \leq n$ and a measurable function

$$f : \prod_{i=1}^k (\mathbb{X}^{(i)})^{\binom{m}{i}} \to \mathbb{R}$$

define the associated **generalised U-statistic** of order k to be

$$S_{n,m}^{(k)}(f) = \sum_{s\in\binom{[n]}{m}} f(\mathcal{X}_s^{(1)}, \mathcal{X}_s^{(2)}, \ldots, \mathcal{X}_s^{(k)}).$$

To clarify this definition, the aim is to be able to input all the variables corresponding to subsets of size i of a base subset s, which we sum over. There are $\binom{m}{i}$ such subsets and since there is a 1-1 correspondence between the subsets and the random elements, this is how many of those random elements we want to be able to input into our function. Finally we take a product over all $i \in [k]$ because we want the function to be capable of depending of subsets of different size (up to size k).

Example 4.3. Generalised U-statistics of order $k > 2$ covers many random variables of interest in the context of random hypergraphs and random simplicial complexes. Such statistics have been studied in [26, Section 10.7], [18], [24, Chapter 11.3]. For example, consider a hypergraph on the vertex set $[n]$ where a hyper-edge $\alpha \subseteq n$ is present independently of all other hyper-edges with probability p. Here we assume that $|\alpha| \geq 2$ so that the number of vertices is fixed. For any $\alpha \subseteq [n]$, such that $|\alpha| > 1$, let $\xi_\alpha^{(|\alpha|)}$ be the hyper-edge indicator. Then the sequences

$$\left\{\xi_\alpha^{(2)}\right\}_{\alpha\in\binom{[n]}{2}}, \ldots, \left\{\xi_\alpha^{(k)}\right\}_{\alpha\in\binom{[n]}{k}}$$

are all i.i.d. Bernoulli(p) but they are of different size. Because all vertices are present in this hypergraph model, we set $\xi_\alpha^{(1)} = 1$ for all $\alpha \in \binom{[n]}{1}$, which makes the first sequence of variables deterministic. Define a function $f : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R} \to \mathbb{R}$ by

$$f(x_1, x_2, x_3, y_{1,2}, y_{1,3}, y_{2,3}, z_{1,2,3}) = z_{1,2,3}(1 - y_{1,2})(1 - y_{1,3})(1 - y_{2,3}).$$

Then the associated U-statistic of order 3, $S_{n,3}^{(3)}$, is the number of hyper-edges of size 3, which do not contain any hyper-edge of smaller size in this random hypergraph.

Fix positive integers $\{m_i\}_{i \in [d]}$ and let $\{f_i\}_{i \in [d]}$ be a collection of measurable functions

$$f_i : \prod_{t=1}^{k} (\mathbb{X}^{(t)})^{\binom{m_i}{t}} \to \mathbb{R}$$

and consider the collection of associated generalised U-statistics of order k

$$\left\{ S_{n,m_i}^{(k)}(f_i) \right\}_{i \in [d]}.$$

For any $i \in [d]$ let $\mathbb{I}_i := \binom{[n]}{m_i} \times \{i\}$. For $s = (\phi, i) \in \mathbb{I}_i$ define $Y_s = f_i(\mathcal{X}_\phi^{(1)}, \mathcal{X}_\phi^{(2)}, \ldots, \mathcal{X}_\phi^{(k)})$ and

$$X_s = \sigma_i^{-1} \left\{ Y_s - \mu_s \right\},$$

where $\mu_s = \mathbb{E}\{Y_s\}$ and $\sigma_i = \sqrt{\text{Var}(S_{n,m_i}^{(k)}(f_i))}$, assuming the relevant expectations are finite and variances are all non-zero and finite.

Let $W_i = \sum_{s \in \mathbb{I}_i} X_s$ and define a random vector $W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d$. We are interested in the distribution of W. However, to apply Corollary B.3 we need some assumptions.

Let k' be the largest non-negative integer such that the random elements

$$\left\{ \xi_\alpha^{(1)} \right\}_{\alpha \in \binom{[n]}{1}}, \left\{ \xi_\alpha^{(2)} \right\}_{\alpha \in \binom{[n]}{2}}, \ldots, \left\{ \xi_\alpha^{(k')} \right\}_{\alpha \in \binom{[n]}{k'}}$$

are deterministic. Such k' always exists and if there are no deterministic variables, we will say that $k' = 0$. It is convenient to have such k' as a parameter when studying Linial-Meshulam random simplicial complexes; there we take a $(k'-1)$-skeleton that is complete and put in higher simplices with different probabilities. For other classical models like $\mathbf{X}(n, p)$ we usually have $k' = 1$ since there is no randomness at the level of vertices.

Assumption 4.4. *We use the following two assumptions:*

1. *For any $i \in [d]$ there is $\alpha_i > 0$ such that for all pairs $s, u \in \mathbb{I}_i$ for which Y_s, Y_u are dependent, we have*

$$\text{Cov}(Y_s, Y_u) \geq \alpha_i.$$

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*

$$\mathbb{E}\left|\{Y_s - \mu_s\}\{Y_t - \mu_t\}\{Y_u - \mu_u\}\right| \leq \beta$$

as well as

$$\mathbb{E}\left|\{Y_s - \mu_s\}\{Y_t - \mu_t\}\right| \mathbb{E}\left|Y_u - \mu_u\right| \leq \beta.$$

The assumptions used here are far from necessary for the normal approximation to hold. However, they hold in the settings studied in this work and adopting them makes the bounds quite convenient.

Theorem 4.5. *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d \times d})$ and Σ be the covariance matrix of W, where W satisfies Assumption 4.4.*

1. *Let $h \in \mathcal{H}_d$. Then*

$$\left| \mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z) \right| \leq |h|_3 B_{4.5} n^{-\frac{k'+1}{2}}.$$

2.

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{4.5}^{\frac{1}{4}} n^{-\frac{k'+1}{8}},$$

where

$$B_{4.5} = \frac{4}{3} \frac{\beta(k'+1)^{\frac{3}{2}(k'+1)}}{((k'+1)!)^2} (\sqrt{2}m)^{5m+2-3(k'+1)} \sum_{(i,j,k)}^{d} \frac{1}{\sqrt{\alpha_i \alpha_j \alpha_k}}$$

and $m = \max_{i \in [d]} m_i$.

Proof. Fix $i, j \in [d]$ and $s = (\phi, i) \in \mathbb{I}_i$. Then using the definition of k', we see that one choice of dependency neighbourhoods is $\mathbb{D}_j(s) = \{ (\psi, j) \in \mathbb{I}_j \mid |\phi \cap \psi| \geq k' + 1 \}$. This follows because in our setup the sequences of random elements $\xi_\alpha^{(l)}$ are independent and so the summands associated to two subsets can only be dependent if they share a random element. By definition of k', the smallest such random element that is not deterministic will be associated to a subset of size $k' + 1$. Now we see that Corollary B.3 applies.

Looking at the size of $\mathbb{D}_j(s)$, with (1) we have

$$|\mathbb{D}_j(s)| = \sum_{m=k'+1}^{\min(m_i, m_j)} \binom{m_i}{m} \binom{n - m_i}{m_j - m} \leq n^{m_j - 1 - k'} \frac{m_i^{m_i+1}}{(k'+1)!}. \tag{4}$$

Let us now calculate a lower bound for the variance using the second part of Assumption 4.4, noting that $|\mathbb{I}_i| = \binom{n}{m_i}$:

$$\begin{aligned}
\sigma_i^2 &= \mathrm{Var}(S_{n,m_i}^{(k)}(f_i)) \\
&= \sum_{s \in \mathbb{I}_i} \sum_{u \in D_i(s)} \mathrm{Cov}(Y_s, Y_u) \\
&= \sum_{s \in \mathbb{I}_i} \sum_{m=k'+1}^{m_i} \sum_{\substack{u \in \binom{[n]}{m_i} \\ |u \cap s| = m}} \mathrm{Cov}(Y_s, Y_u) \\
&\geq \sum_{m=k'+1}^{m_i} \binom{n}{m_i} \binom{m_i}{m} \binom{n - m_i}{m_i - m} \alpha_i \\
&= \sum_{m=k'+1}^{m_i} \binom{n}{2m_i - m} \binom{2m_i - m}{m_i} \binom{m_i}{m} \alpha_i \\
&\geq \binom{n}{2m_i - 1 - k'} \binom{2m_i - 1 - k'}{m_i} \binom{m_i}{k'+1} \alpha_i \\
&\geq \frac{n^{2m_i - 1 - k'}}{(2m_i - 1 - k')^{2m_i - 1 - k'}} \frac{(2m_i - 1 - k')^{m_i}}{m_i^{m_i}} \frac{m_i^{k'+1}}{(k'+1)^{k'+1}} \alpha_i \\
&= \frac{n^{2m_i - 1 - k'} m_i^{k'+1-m_i}}{(2m_i - 1 - k')^{m_i - 1 - k'} (k'+1)^{k'+1}} \alpha_i.
\end{aligned}$$

Taking the inequality to the power of $-\frac{1}{2}$ we get

$$\sigma_i^{-1} \le n^{-m_i + \frac{k'+1}{2}} (2m_i^2 - m_i(1+k'))^{\frac{m_i}{2} - \frac{k'+1}{2}} (k'+1)^{\frac{k'+1}{2}} \alpha_i^{-\frac{1}{2}}.$$

Using the language of Corollary B.3 and using the last part of Assumption 4.4 we can say that $\beta_{ijl} \le \beta(\sigma_i \sigma_j \sigma_l)$. Moreover we can bound α_{ij} in Corollary B.3 by (4). The bound on the quantity $B_{B.3}$ from the corollary is

$B_{B.3}$

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

$$\le \frac{1}{3} \sum_{(i,j,k)} \frac{n^{m_i}}{m_i!} n^{m_j - 1 - k'} \frac{m_i^{m_i+1}}{(k'+1)!} \left(\frac{3}{2} n^{m_k - 1 - k'} \frac{m_i^{m_i+1}}{(k'+1)!} + 2n^{m_k - 1 - k'} \frac{m_j^{m_j+1}}{(k'+1)!} \right) \beta_{ijk}$$

$$\le \frac{2\beta(k'+1)^{\frac{3}{2}(k'+1)}}{3((k'+1)!)^2} \sum_{(i,j,k)} n^{m_i + m_j + m_k - 2(k'+1)}$$

$$\frac{m_i^{m_i+1}}{m_i! \sqrt{\alpha_i \alpha_j \alpha_k}} \left(m_i^{m_i+1} + m_j^{m_j+1} \right) n^{-m_i - m_j - m_k + \frac{3}{2}(k'+1)} M_{i,j,k}$$

$$= \frac{2\beta(k'+1)^{\frac{3}{2}(k'+1)}}{3((k'+1)!)^2} \sum_{(i,j,k)} n^{-\frac{k'+1}{2}} \frac{m_i^{m_i+1}(m_i^{m_i+1} + m_j^{m_j+1})}{m_i! \sqrt{\alpha_i \alpha_j \alpha_k}} M_{i,j,k}$$

$$= \left\{ \frac{2}{3} \frac{\beta(k'+1)^{\frac{3}{2}(k'+1)}}{((k'+1)!)^2} \sum_{(i,j,k)} \frac{m_i^{m_i+1}(m_i^{m_i+1} + m_j^{m_j+1})}{m_i! \sqrt{\alpha_i \alpha_j \alpha_k}} M_{i,j,k} \right\} n^{-\frac{k'+1}{2}},$$

where $M_{i,j,k} = (2m_i^2 - m_i(1+k'))^{\frac{m_i}{2} - \frac{k'+1}{2}} (2m_j^2 - m_j(1+k'))^{\frac{m_j}{2} - \frac{k'+1}{2}} (2m_k^2 - m_k(1+k'))^{\frac{m_k}{2} - \frac{k'+1}{2}}$.

Now set $m = \max_{i \in [d]} m_i$. Then we have $(2m_i^2 - m_i(1+k'))^{\frac{m_i}{2} - \frac{k'+1}{2}} \le (2m^2)^{\frac{m}{2} - \frac{k'+1}{2}}$ and so

$$B_{B.3} \le \left\{ \frac{4}{3} \frac{\beta(k'+1)^{\frac{3}{2}(k'+1)}}{((k'+1)!)^2} (\sqrt{2}m)^{5m+2-3(k'+1)} \sum_{(i,j,k)}^{d} \frac{1}{\sqrt{\alpha_i \alpha_j \alpha_k}} \right\} n^{-\frac{k'+1}{2}}.$$

$\square$

5. Subcomplex counts.

Let $\mathcal{L}$ be a fixed connected simplicial complex on a vertex set $[k+1]$. Let $\mathbf{p} = (p_1, p_2, \ldots, p_{n-1}) \in \mathbb{R}^{n-1}$ be a vector where for any $i \in [n-1]$ we have $p_i \in [0,1]$. Let $k' \in \mathbb{N}$ be such that for any $i \in [k']$ we have $p_i = 1$. We are interested in subcomplex counts of $\mathcal{L}$ in $\mathbf{X}(n, \mathbf{p})$ as defined in Section 2. This is a natural generalisation of subgraph counts.

Let us discuss this variable from a probabilistic perspective. Consider the random hypergraph model from Definition 2.2. For any $k \in [n-1]$ and $\alpha \in \binom{[n]}{k+1}$ let $\xi_\alpha^{(k+1)} \sim$ Bernoulli(p_k) be the hyperedge indicators in the random hypergraph model. Hence the random variables in this collection are independent by construction. Then the

indicator of the k-simplex α in the random simplicial complex model can be written as a product of the hyperedge indicators:

$$\mathbb{1}\left(\alpha \in \mathbf{X}(n, \mathbf{p})\right) = \prod_{i=k'+1}^{k} \prod_{\substack{\beta \subseteq \alpha \\ |\beta|=i+1}} \xi_\beta^{(i+1)}. \tag{5}$$

Note that by independence we have

$$\mathbb{E}\left\{\mathbb{1}\left(\alpha \in \mathbf{X}(n, \mathbf{p})\right)\right\} = \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}}.$$

Let $[\mathcal{L}]$ be the isomorphism class of $\mathcal{L}$ in the set of all simplicial complexes on the vertex set $[k+1]$. That is, $[\mathcal{L}]$ is the set of all simplicial complexes on the vertex set $[k+1]$ that are simplicially isomorphic to $\mathcal{L}$. For any $s \in \binom{[n]}{k+1}$, we write $s = \{s_1, s_2, \ldots, s_{k+1}\}$ such that $s_1 < s_2 < \ldots < s_{k+1}$. Then we define $\mathcal{L}[s]$ to be the complex on the vertex set s such that any $\alpha = \{s_{i_1}, s_{i_2}, \ldots s_{i_l}\} \subseteq s$ is a simplex in $\mathcal{L}[s]$ iff $\{i_1, i_2, \ldots, i_l\} \in \mathcal{L}$. We are interested in the following variable:

$$T_\mathcal{L} := \sum_{s \in \binom{[n]}{k+1}} \sum_{\mathcal{L}' \in [\mathcal{L}]} \prod_{i=k'+1}^{k} \prod_{\substack{\alpha \in \mathcal{L}'[s] \\ |\alpha|=i+1}} \xi_\alpha^{(i+1)}. \tag{6}$$

Note that this random variable, which counts the number of copies of $\mathcal{L}$ in $\mathbf{X}(n, \mathbf{p})$, can be studied in the framework of generalised U-statistics. In particular, if the complex $\mathcal{L}$ is of dimension d, then $T_\mathcal{L}$ is a U-statistic of order $d+1$.

To fit the random variable $T_\mathcal{L}$ in the framework of generalised U-statistics, we say that the collections of sequences

$$\left\{\xi_\alpha^{(1)}\right\}_{\alpha \in \binom{[n]}{1}}, \left\{\xi_\alpha^{(2)}\right\}_{\alpha \in \binom{[n]}{2}}, \ldots, \left\{\xi_\alpha^{(d+1)}\right\}_{\alpha \in \binom{[n]}{d+1}}$$

are the underlying variables from which we can build simplex indicators as it is done in (5). That is, we have $\xi_\alpha^{(1)} = 1$ for any $\alpha \in \binom{[n]}{1}$. For any $1 < k \leq d+1$ and $\alpha \in \binom{[n]}{k+1}$ we have $\xi_\alpha^{(k+1)} \sim \text{Bernoulli}(p_k)$. All the sequences are independent and within each sequence the variables are i.i.d. Recall that $\mathcal{X}_s^{(i+1)}$ from (3) is a sequence of the random elements, corresponding to the $(i+1)$-subsets of s, ordered lexicographically as described in Remark 4.1. We write $\mathcal{X}_s^{(i+1)}[j]$ for the j-th element of the sequence. Let $o(\alpha)$ be the position of $\alpha \in \binom{s}{i+1}$ in this lexicographical ordering. Now let us define the associated function $f : \prod_{k=2}^{d+1}(\{0,1\})^{\binom{d+1}{k}} \to \mathbb{R}$ by

$$f(\mathcal{X}_s^{(1)}, \mathcal{X}_s^{(2)}, \ldots, \mathcal{X}_s^{(d+1)}) = \sum_{\mathcal{L}' \in [\mathcal{L}]} \prod_{i=k'+1}^{d} \prod_{\substack{\alpha \in \mathcal{L}'[s] \\ |\alpha|=i+1}} \mathcal{X}_s^{(i+1)}[o(\alpha)]. \tag{7}$$

Now it is easy to see that for $T_\mathcal{L}$ given in (6),

$$T_\mathcal{L} = \sum_{s \in \binom{[n]}{k+1}} f(\mathcal{X}_s^{(1)}, \mathcal{X}_s^{(2)}, \ldots, \mathcal{X}_s^{(d+1)})$$

and hence the random variable in question is a generalised U-statistic of order $d+1$.

5.1. **Moments and the limiting covariance.** Let $e_i(\mathcal{L})$ be the number of i-simplices in $\mathcal{L}$. Then by independence we have

$$\mathbb{E}\{T_\mathcal{L}\} = \binom{n}{k+1}|[\mathcal{L}]|\prod_{i=1}^{k}p_i^{e_i(\mathcal{L})}.$$

Consider, for example, the simplicial complex on four vertices given by

$$\mathcal{L} = \{1, 2, 3, 4, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3\}, \{2, 3, 4\}\}.$$

This consists of two 2-simplices $\{1, 2, 3\}$ and $\{2, 3, 4\}$ which share the edge $\{2, 3\}$. For this simplicial complex, $\mathbb{E}\{T_\mathcal{L}\} = 3\binom{n}{4}p_1^5p_2^2$. Here 1 and 4 can be exchanged, 2 and 3 can be exchanged, and there are no further simplicial isomorphisms and hence $|[\mathcal{L}]| = 3$. The following lemma is used in our limiting covariance calculations.

Lemma 5.1. *Let l, m, a be positive integers independent of $n \in \mathbb{N}$ such that $l, m > \frac{a}{2}$. Also, let $N(n)$ be a sequence of integers such that $N(n) = \Theta(n)$. We write N instead of $N(n)$. Then*

$$\lim_{n\to\infty}\binom{N}{l+m-a}\binom{l+m-a}{l}\binom{l}{a}$$

$$\left\{\binom{N}{2l-a}\binom{2l-a}{l}\binom{l}{a}\binom{N}{2m-a}\binom{2m-a}{m}\binom{m}{a}\right\}^{-\frac{1}{2}} = 1.$$

Proof.

$$\lim_{n\to\infty}\binom{N}{l+m-a}\binom{l+m-a}{l}\binom{l}{a}$$

$$\left\{\binom{N}{2l-a}\binom{2l-a}{l}\binom{l}{a}\binom{N}{2m-a}\binom{2m-a}{m}\binom{m}{a}\right\}^{-\frac{1}{2}}$$

$$= \lim_{n\to\infty}\frac{N!}{(N-(l+m-a))!}\frac{1}{(l!)^{\frac{1}{2}}(m-a)!}\left(\frac{1}{a!(l-a)!}\right)^{\frac{1}{2}}$$

$$\left\{\frac{N!}{(N-(2l-a))!}\frac{1}{l!(l-a)!}\frac{N!}{(N-(2m-a))!}\frac{1}{((m-a)!)^2}\frac{1}{a!}\right\}^{-\frac{1}{2}}$$

$$= \lim_{n\to\infty}\frac{((N-(2m-a))!(N-(2l-a))!)^{\frac{1}{2}}}{(N-(l+m-a))!} = 1.$$

$\square$

Theorem 5.2 shows that, asymptotically, similarly to the subgraph counts in $G(n, p)$ for a constant $p \in (0, 1)$, subcomplex counts are perfectly correlated.

Theorem 5.2. *Let $\mathcal{L}, \mathcal{M}$ be two simplicial complexes of dimension at least $k' + 1$ on the vertex sets $[l]$ and $[m]$ respectively. Let $\sigma_\mathcal{L}^2 = \text{Var}(T_\mathcal{L})$ and $\sigma_\mathcal{M}^2 = \text{Var}(T_\mathcal{M})$. Then assuming that for all i the parameters p_i stay constant, for any fixed $m, l \in \mathbb{N}$ we have*

$$\lim_{n\to\infty}\sigma_\mathcal{L}^{-1}\sigma_\mathcal{M}^{-1}\text{Cov}(T_\mathcal{L}, T_\mathcal{M}) = 1.$$

Proof. Assume w.l.o.g. that $m \leq l$. For $s \in \binom{[n]}{l}$ set the variable

$$X_s^{\mathcal{L}} = \sum_{\mathcal{L}' \in [\mathcal{L}]} \prod_{i=k'+1}^{k} \prod_{\substack{\alpha \in \mathcal{L}'[s] \\ |\alpha|=i+1}} \xi_\alpha^{(i+1)},$$

so that $T_{\mathcal{L}} = \sum_{s \in \binom{[n]}{l}} X_s^{\mathcal{L}}$, and analogously $X_u^{\mathcal{M}}$ for $u \in \binom{[n]}{m}$. Note that $X_s^{\mathcal{L}}$ does not depend on n. If one would like to relate this expression back to the U-statistics framework, one needs to note that if f is a function that gives us $T_{\mathcal{L}}$ as a U-statistic, then for any $s \in \binom{[n]}{l}$ we have $X_s^{\mathcal{L}} = f(\mathcal{X}_s^{(1)}, \mathcal{X}_s^{(2)}, \ldots, \mathcal{X}_s^{(l)})$.

Note that for any $s \in \binom{[n]}{l}$, $u \in \binom{[n]}{m}$ if $|s \cap u| < k'+2$, then the variables $X_s^{\mathcal{L}}$ and $X_u^{\mathcal{M}}$ are independent, since the intersection of the underlying subsets is not large enough to contain a random element of the form $\xi_\alpha^{(i+1)}$ for $i+1 \geq k'+2$, which would create dependence. For s of size l, u of size m, with intersection of size k, $c_k(l, m) := \mathrm{Cov}(X_s^{\mathcal{L}}, X_u^{\mathcal{M}})$ does not depend on n. Hence we see that

$$\mathrm{Cov}(T_{\mathcal{L}}, T_{\mathcal{M}}) = \sum_{s \in \binom{[n]}{l}} \sum_{k=k'+2}^{m} \sum_{\substack{u \in \binom{[n]}{m} \\ |s \cap u|=k}} \mathrm{Cov}(X_s^{\mathcal{L}}, X_u^{\mathcal{M}}) = \sum_{s \in \binom{[n]}{l}} \sum_{k=k'+2}^{m} \sum_{\substack{u \in \binom{[n]}{m} \\ |s \cap u|=k}} c_k(l, m).$$

The leading term of $\mathrm{Cov}(T_{\mathcal{L}}, T_{\mathcal{M}})^2$ and of $\sigma_{\mathcal{L}}^2 \sigma_{\mathcal{M}}^2$, as polynomials in n, is $n^{2(l+m-(k'+2))}$. This follows because the number of pairs $s \in \binom{[n]}{l}$, $u \in \binom{[n]}{m}$ where $|s \cap u| = k$ is $\binom{n}{l}\binom{l}{k}\binom{n-l}{m-k} = \binom{n}{l+m-k}\binom{l+m-k}{l}\binom{l}{k}$, which is of the order n^{l+m-k}. As the highest power of the nominator and the denominator in $\sigma_{\mathcal{L}}^{-1} \sigma_{\mathcal{M}}^{-1} \mathrm{Cov}(T_{\mathcal{L}}, T_{\mathcal{M}})$ are the same, the limit of the ratio is determined by the ratio of the highest order term from the numerator and the highest order term from the denominator. We write $L_{k'+1}$ for the number of elements of $[\mathcal{L}]$ that contain a particularly chosen $(k'+1)$−simplex and $M_{k'+1}$ for the analogous quantity corresponding to $\mathcal{M}$.

We use Lemma 5.1 with $N = n$ and $a = k'+2$ to get:

$$\lim_{n \to \infty} \sigma_{\mathcal{L}}^{-1} \sigma_{\mathcal{M}}^{-1} \mathrm{Cov}(T_{\mathcal{L}}, T_{\mathcal{M}})$$

$$= \lim_{n \to \infty} \binom{n}{l+m-2-k'} \binom{l+m-2-k'}{l} \binom{l}{k'+2} L_{k'+1} M_{k'+1} (p_{k'+1}^{-1} - 1) \prod_{i=k'+1}^{l-1} p_i^{e_i(\mathcal{L})+e_i(\mathcal{M})}$$

$$\left\{ \binom{n}{2l-k'-2} \binom{2l-k'-2}{l} \binom{l}{k'+2} L_{k'+1}^2 (p_{k'+1}^{-1} - 1) \prod_{i=k'+1}^{l-1} p_i^{2e_i(\mathcal{L})} \right\}^{-\frac{1}{2}}$$

$$\left\{ \binom{n}{2m-k'-2} \binom{2m-k'-2}{m} \binom{m}{k'+2} M_{k'+1}^2 (p_{k'+1}^{-1} - 1) \prod_{i=k'+1}^{m-1} p_i^{2e_i(\mathcal{M})} \right\}^{-\frac{1}{2}}$$

$$= \lim_{n \to \infty} L_{k'+1} M_{k'+1} (p_{k'+1}^{-1} - 1) \prod_{i=k'+1}^{l-1} p_i^{e_i(\mathcal{L})+e_i(\mathcal{M})}$$

$$\left\{ L_{k'+1}^2 (p_{k'+1}^{-1} - 1) \prod_{i=k'+1}^{l-1} p_i^{2e_i(\mathcal{L})} \prod_{i=k'+1}^{m-1} M_{k'+1}^2 (p_{k'+1}^{-1} - 1) \prod_{i=k'+1}^{m-1} p_i^{2e_i(\mathcal{M})} \right\}^{-\frac{1}{2}}$$

$$= 1.$$

$\square$

5.2. **Approximation theorem.** Let $\{\mathcal{L}_i\}_{i\in[d]}$ be a collection of finite connected simplicial complexes. Let $[m_i]$ be the vertex set of $\mathcal{L}_i$ and let $k_i \geq k'+1$ be its dimension. Let f_i be the function associated to $\mathcal{L}_i$ as set out in (7). For any $i \in [d]$, let $\mathbb{I}_i = \binom{[n]}{m_i} \times \{i\}$. For $s = (\phi, i) \in \mathbb{I}_i$ define $Y_s = f_i(\mathcal{X}_\phi^{(1)}, \mathcal{X}_\phi^{(2)}, \ldots, \mathcal{X}_\phi^{(k_i+1)})$ and

$$X_s = \sigma_i^{-1}\{Y_s - \mu_s\},$$

where $\mu_s = \mathbb{E}\{Y_s\}$ and $\sigma_i = \sqrt{\mathrm{Var}(T_{\mathcal{L}_i})}$.

Let $W_i = \sum_{s\in\mathbb{I}_i} X_s$ and define a random vector $W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d$. We are interested in the distribution of W. To apply the approximation Theorem 4.5, we need to show that Assumption 4.4 holds and find the quantities α_i and β from the assumption. The following proposition, the proof of which is trivial and therefore does not appear here, lets us do that.

Proposition 5.3. *Let $\mathcal{L}$ be a connected simplicial complex of dimension $d \geq m$ on the vertex set $[k]$. Consider $s, u \in \binom{[n]}{k}$ such that $|s \cap u| \geq m+1$. Then there exists $\mathcal{L}', \mathcal{L}'' \in [\mathcal{L}]$ such that $\mathcal{L}'[s]$ and $\mathcal{L}''[u]$ share at least one m-dimensional simplex.*

Now we can give a normal approximation for subcomplex counts, based on Theorem 4.5. Note that using the definition of k' in this section means that the variables

$$\left\{\xi_\alpha^{(1)}\right\}_{\alpha\in\binom{[n]}{1}}, \left\{\xi_\alpha^{(2)}\right\}_{\alpha\in\binom{[n]}{2}}, \ldots, \left\{\xi_\alpha^{(k'+1)}\right\}_{\alpha\in\binom{[n]}{k'+1}}$$

are deterministic and equal to 1, which means that when we apply Theorem 4.5, we should use $k'+1$ instead of k'.

Corollary 5.4. *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$ and Σ be the covariance matrix of W.*

1. *Let $h \in \mathcal{H}$. Then*

$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \leq |h|_3\, B_{5.4}\, n^{-\frac{k'+2}{2}}.$$

2. *Let $m = \max_{i\in[d]} m_i$. Then*

$$\sup_{A\in\mathcal{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}}Z \in A)| \leq 2^{\frac{7}{2}}3^{-\frac{3}{4}}d^{\frac{3}{16}}B_{5.4}^{\frac{1}{4}}n^{-\frac{k'+2}{8}},$$

where

$$B_{5.4} = \frac{4}{3}\frac{(k'+2)^{\frac{3}{2}(k'+2)}}{((k'+2)!)^2}(\sqrt{2}m)^{5m+2-3(k'+2)}d^3\gamma$$

and

$$\gamma = \left(\prod_{k=k'+1}^{\min_{i\in[d]}(k_i)} p_k^{\min_{i\in[d]}(e_k(\mathcal{L}_i))}\right)^{-\frac{1}{2}}\left(1 - \prod_{k=k'+1}^{\max_{i\in[d]}(k_i)} p_k^{\max_{i\in[d]}(e_k(\mathcal{L}_i))}\right)\left(p_{k'+1}^{-1} - 1\right)^{-\frac{3}{2}}.$$

Before embarking on the proof we would like to point out that Corollary 5.4 allows for the case that $p_k = n^{-\alpha_k}$; in this case we bound

$$\gamma \leq \left(\prod_{k=k'+1}^{\min_{i\in[d]}(k_i)} n^{\frac{1}{2}\alpha_k \min_{i\in[d]}(e_k(\mathcal{L}_i))}\right)n^{\frac{2}{3}\alpha_{k'+1}}.$$

The bounds in Corollary 5.4 then tend to 0 with $n \to \infty$ as long as $k' + 2 \geq \sum_{k=k'+1}^{\min_{i\in[d]}(k_i)} \alpha_k \min_{i\in[d]}(e_k(\mathcal{L}_i)) + \frac{4}{3}\alpha_{k'+1}$. However when p_k decreases to 0 fast, then instead of a normal approximation, a multivariate Poisson approximation would be more appropriate. Deriving bounds on such an approximation would be a considerable endeavour; a starting point could be Theorem 10.A in [3].

Proof. Take $s = (\phi, i), u = (\psi, i) \in \mathbb{I}_i$ such that $|\phi \cap \psi| \geq k' + 2$. We have

$$\text{Cov}(Y_s, Y_u) = \sum_{\mathcal{L}' \in [\mathcal{L}_i]} \sum_{\mathcal{L}'' \in [\mathcal{L}_i]} \mathbb{E} \prod_{i=k'+1}^{k} \prod_{j=k'+1}^{k} \prod_{\substack{\alpha \in \mathcal{L}'[s] \\ |\alpha|=i+1}} \prod_{\substack{\beta \in \mathcal{L}''[u] \\ |\beta|=u+1}} \xi_\alpha^{(i+1)} \xi_\beta^{(j+1)}$$

$$- |[\mathcal{L}_i]|^2 \prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)}.$$

Using Proposition 5.3, there is at least one pair $\mathcal{L}'[s], \mathcal{L}''[u]$ that share at least one $k'+1$-simplex. This pair has a contribution of at least $p_{k'+1}^{2e_{k'+1}(\mathcal{L}_i)-1} \prod_{j=k'+2}^{k_i} p_j^{2e_j(\mathcal{L}_i)}$ to the sum. The contribution of any other pair is bounded below by $\prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)}$. Hence,

$$\text{Cov}(Y_s, Y_u) \geq p_{k'+1}^{2e_{k'+1}(\mathcal{L}_i)-1} \prod_{j=k'+2}^{k_i} p_j^{2e_j(\mathcal{L}_i)}$$

$$+ (|[\mathcal{L}_i]|^2 - 1) \prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)} - |[\mathcal{L}_i]|^2 \prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)}$$

$$= \prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)} \left\{ p_{k'+1}^{-1} - 1 \right\} > 0.$$

Hence, we see that $\alpha_i = \prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)} \left\{ p_{k'+1}^{-1} - 1 \right\}$, using the notation of Assumption 4.4.

Recall that $e_k(\mathcal{L}_i)$ is the number of k-simplices in $\mathcal{L}$. 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 using Lemma B.4 with $\mu_i = \prod_{k=k'+1}^{k_i} p_k^{e_k(\mathcal{L}_i)}$,

$$\mathbb{E} \left| \{Y_s - \mu_s\} \{Y_t - \mu_t\} \{Y_u - \mu_u\} \right|$$

$$\leq \left\{ \prod_{k=k'+1}^{k_i} p_k^{e_k(\mathcal{L}_i)} \prod_{k=k'+1}^{k_j} p_k^{e_k(\mathcal{L}_j)} \left(1 - \prod_{k=k'+1}^{k_i} p_k^{e_k(\mathcal{L}_i)} \right) \left(1 - \prod_{k=k'+1}^{k_j} p_k^{e_k(\mathcal{L}_j)} \right) \right\}^{\frac{1}{2}}$$

$$\leq \prod_{k=k'+1}^{\min_{i \in [d]}(k_i)} p_k^{\min_{i \in [d]}(e_k(\mathcal{L}_i))} \left(1 - \prod_{k=k'+1}^{\max_{i \in [d]}(k_i)} p_k^{\max_{i \in [d]}(e_k(\mathcal{L}_i))} \right).$$

So $\beta = \prod_{k=k'+1}^{\min_{i \in [d]}(k_i)} p_k^{\min_{i \in [d]}(e_k(\mathcal{L}_i))} \left(1 - \prod_{k=k'+1}^{\max_{i \in [d]}(k_i)} p_k^{\max_{i \in [d]}(e_k(\mathcal{L}_i))} \right)$.

Now we can apply Theorem 4.5 and bound the quantity $B_{4.5}$:

$$B_{4.5} \leq \frac{4}{3} \frac{(k'+2)^{\frac{3}{2}(k'+2)}}{((k'+2)!)^2} (\sqrt{2}m)^{5m+2-3(k'+2)} d^3 \beta \left\{ (p_{k'+1}^{-1} - 1) \right.$$

$$\left. \prod_{k=k'+1}^{\min_{i \in [d]}(k_i)} p_k^{\min_{i \in [d]}(e_k(\mathcal{L}_i))} \right\}^{-\frac{3}{2}}$$

$$\leq \frac{4}{3} \frac{(k'+2)^{\frac{3}{2}(k'+2)}}{((k'+2)!)^2} (\sqrt{2}m)^{5m+2-3(k'+2)} d^3 \gamma.$$

$\square$

Remark 5.5. Note that from Corollary 5.2 it follows that the limiting covariance matrix of W, assuming the parameters p_i and the number of vertices m_i are constant for all i, is equal to the $d \times d$ matrix with every entry equal to 1. Hence the limiting distribution in that setting is a degenerate multivariate normal with covariance matrix having rank 1.

Also note that if we pick d, $p_i \in (0,1)$ for $i > k'$, k_i for all i as well as $e_k(\mathcal{L}_i)$ for all k and all i to to be constant, then the bound still goes to zero asymptotically as long as $m = \max_{i \in [d]} m_i$ is of the order $o(\ln^{1-\epsilon}(n))$ for a fixed $\epsilon \in (0,1)$. By inspecting the inequalities it is possible to vary the parameter setting further since the theorem allows the quantities in the inequalities to depend on each other.

6. **Critical simplex counts in lexicographical matchings.** The following lemma is an immediate consequence of Definition 2.3; it is Lemma 4.1 from [51].

Lemma 6.1. *Let $\mathcal{L}$ be a simplicial complex endowed with the lexicographical acyclic partial matching, and consider a simplex $t \in \mathcal{L}$ 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 $t \cup \{j\} \in \mathcal{L}$; and,*
2. *one of its faces if and only if for all $j < i$ we have $(t \setminus \{i\}) \cup \{j\} \notin \mathcal{L}$.*

Let $\mathbf{p} = (p_1, p_2, \ldots, p_{n-1}) \in [0,1]^{n-1}$ and $k' \in \mathbb{N}$ be such that for any $i \in [k']$ we have $p_i = 1$. Consider the multi-parameter random simplicial complex $\mathbf{X}(n, \mathbf{p})$ as defined in [12, Section 2.2]. For any $k \in [n-1]$ and $\alpha \in \binom{[n]}{k+1}$ let $\xi_\alpha^{(k+1)} \sim$ Bernoulli(p_k).

Fix $s \in \binom{[n]}{k}$. Define the random variables

$$X_s^+ = 1 - \prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j=k'+1}^{k} \prod_{\substack{\alpha \subseteq s \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) \text{ and } X_s^- = \prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j=k'+1}^{k-1} \prod_{\substack{\alpha \subseteq s_- \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right),$$

where $s_- := s \setminus \{\min(s)\}$. By Lemma 6.1, we can interpret X_s^+ as the event that s matches with its coface, given it is a simplex, and similarly X_s^- is the event that s matches with its face given it is a simplex. The events that the two variables indicate are disjoint. Hence,

$$\mathbb{1}\left(s \text{ is a critical simplex}\right) = \mathbb{1}\left(s \in \mathbf{X}(n, \mathbf{p})\right)\left(1 - (X_s^+ + X_s^-)\right)$$

$$= \prod_{j=k'+1}^{k-1} \prod_{\substack{\alpha \subseteq s \\ |\alpha|=j+1}} \xi_\alpha^{(j+1)} \left[\prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j=k'+1}^{k} \prod_{\substack{\beta \subseteq s \\ |\beta|=j}} \xi_{\beta \cup \{i\}}^{(j+1)} \right) \right.$$

$$\left. - \prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j=k'+1}^{k-1} \prod_{\substack{\beta \subseteq s_- \\ |\beta|=j}} \xi_{\beta \cup \{i\}}^{(j+1)} \right) \right].$$

Thus, the random variable of interest is the number of k-simplices that are critical under the lexicographical matching,

$$T_k = \sum_{s \in \binom{[n]}{k}} \prod_{j=k'+1}^{k-1} \prod_{\substack{\alpha \subseteq s \\ |\alpha|=j+1}} \xi_\alpha^{(j+1)} \left[\prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j=k'+1}^{k} \prod_{\substack{\beta \subseteq s \\ |\beta|=j}} \xi_{\beta \cup \{i\}}^{(j+1)} \right) \right.$$

$$\left. - \prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j=k'+1}^{k-1} \prod_{\substack{\beta \subseteq s_- \\ |\beta|=j}} \xi_{\beta \cup \{i\}}^{(j+1)} \right) \right]. \tag{8}$$

Note that this random variable does not fit into the framework of generalised U-statistics as in Definition 4.2 because the summands depend not only on the variables that are indexed by the subset s. In the complex counts we have $\min(s)$ in the expression, which means that the function depends on the index s itself.

6.1. Moments and the limiting covariance.

Lemma 6.2. *For $k' + 1 \le k \le n - 1$ we have*

$$\binom{n-2}{k} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}+\binom{k}{i}} \left(1 - \prod_{i=k'+1}^{k+1} p_i^{\binom{k}{i-1}} \right)$$

$$\le \mathbb{E}\{T_{k+1}\}$$

$$\le \binom{n-1}{k} \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i+1}-\binom{k+1}{i}} \left(1 - \prod_{i=k'+1}^{k} p_i^{\binom{k}{i-1}} \right).$$

Proof. We have

$$\mathbb{E}\{T_{k+1}\}$$

$$= \sum_{l=1}^{n-k} \sum_{\substack{s \in \binom{[n]}{k+1} \\ \min(s)=l}} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}} \left[\left(1 - \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}} \right)^{l-1} - \left(1 - \prod_{i=k'+1}^{k} p_i^{\binom{k}{i}} \right)^{l-1} \right]$$

$$= \sum_{l=1}^{n-k} \binom{n-l}{k} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}} \left[\left(1 - \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}} \right)^{l-1} - \left(1 - \prod_{i=k'+1}^{k} p_i^{\binom{k}{i}} \right)^{l-1} \right]$$

$$\le \binom{n-1}{k} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}} \sum_{l=0}^{\infty} \left[\left(1 - \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}} \right)^{l} - \left(1 - \prod_{i=k'+1}^{k} p_i^{\binom{k}{i}} \right)^{l} \right]$$

$$= \binom{n-1}{k} \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i+1}-\binom{k+1}{i}} \left(1 - \prod_{i=k'+1}^{k} p_i^{\binom{k}{i-1}} \right).$$

For the lower bound, since all the summands are non-negative, we bound the expectation by the $l = 2$ term:

$$\mathbb{E}\{T_{k+1}\} \geq \binom{n-2}{k} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}} \left[\left(1 - \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}}\right) - \left(1 - \prod_{i=k'+1}^{k} p_i^{\binom{k}{i}}\right) \right]$$

$$= \binom{n-2}{k} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}+\binom{k}{i}} \left(1 - \prod_{i=k'+1}^{k+1} p_i^{\binom{k}{i-1}}\right).$$

$\square$

The following two lemmas follow similar arguments as in Lemma 6.2 but are more involved. The proofs are long calculations which are not particularly insightful and so they are deferred to the Appendix A.

Lemma 6.3. *For a fixed integer $k' + 1 \leq k \leq n - 1$ and a fixed sequence of probabilities $p_i \in [0,1]$ there is a constant $C > 0$ independent of n and a natural number $N_{p,k}$ such that for any $n \geq N_{p,k}$:*

$$\mathrm{Var}(T_{k+1}) \geq Cn^{2k-k'}.$$

Lemma 6.4. *Let $\sigma_{i+k'}^2 = \mathrm{Var}(T_{i+k'+1})$ for any $i \geq 1$. Let Σ be a $d \times d$ symmetric matrix such that $\Sigma_{i,j} = \sigma_{i+k'}^{-1}\sigma_{j+k'}^{-1}\mathrm{Cov}(T_{i+k'+1}, T_{j+k'+1})$. Then assuming that for all l the parameters p_l do not depend on n, for any fixed $i < j$ we have*

$$\lim_{n\to\infty} \Sigma_{i,j}$$

$$= \frac{\rho(k+1)\rho(r+1)\sqrt{(2-\rho(k+1))(2-\rho(r+1))(2-\rho(k+1)\rho(k'+1)^{-1})(2-\rho(r+1)\rho(k'+1)^{-1})}}{(\rho(k+1)+\rho(r+1)-\rho(k+1)\rho(r+1)\rho(k'+1)^{-1})(\rho(k+1)+\rho(r+1)-\rho(k+1)\rho(r+1))},$$

where $k' + i = k$, $k' + j = r$, and

$$\rho(a) := \prod_{i=k'+1}^{a} p_i^{\binom{a}{i}}.$$

Remark 6.5. The square root in Lemma 6.4 is understood to be the positive square root. As $0 \leq \rho(a) \leq 1$ the limiting expression in Lemma 6.4 is always non-negative. Moreover we can bound the limiting covariance from above by $\frac{4\rho(k+1)\rho(r+1)}{(\rho(k+1)+\rho(r+1))^2}$. Note that $\frac{4\rho(k+1)\rho(r+1)}{(\rho(k+1)+\rho(r+1))^2} < 1$ as long as $\rho(k+1) \neq \rho(r+1)$. In particular, if $p_k \in (0,1)$ for all $k > k'$, then $\rho(k+1) \neq \rho(r+1)$ and so the limiting covariance is strictly smaller than 1.

6.2. **Approximation theorem.** For $i \in [d]$, recall a random variable counting i-simplices in $\mathbf{X}(n,\mathbf{p})$ that are critical under the lexicographical matching, as given in (8). We write for the i-th index set $\mathbb{I}_i := \binom{[n]}{i+1} \times \{i\}$. For $s = (\phi, i) \in \mathbb{I}_i$ we have

$$\mu_s = \prod_{j=k'+1}^{k} p_j^{\binom{k+1}{j+1}} \left[\left(1 - \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}}\right)^{\min(\phi)-1} - \left(1 - \prod_{i=k'+1}^{k} p_i^{\binom{k}{i}}\right)^{\min(\phi)-1} \right]$$

and $\sigma_i = \sqrt{\mathrm{Var}(T_{i+1})}$. Note that $\mathbb{E}\{T_{i+1}\} = \sum_{s \in \mathbb{I}_i} \mu_s$. Let

$$X_s = \sigma_i^{-1} \left\{ \prod_{j=k'+1}^{k-1} \prod_{\substack{\alpha \subseteq \phi \\ |\alpha|=j+1}} \xi_\alpha^{(j+1)} \left[\prod_{i=1}^{\min(\phi)-1} \left(1 - \prod_{j=k'+1}^{k} \prod_{\substack{\alpha \subseteq \phi \\ |\alpha|=j}} \xi_{\alpha\cup\{i\}}^{(j+1)}\right) \right. \right.$$

$$- \prod_{i=1}^{\min(\phi)-1} \left(1 - \prod_{j=k'+1}^{k-1} \prod_{\substack{\alpha \subseteq \phi_- \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) \Bigg] - \mu_s \bigg\}.$$

Let $W_i = \sum_{s \in \mathbb{I}_i} X_s$ and $W = (W_{k'+1}, W_{k'+2}, \dots, W_{k'+d}) \in \mathbb{R}^d$. For bounds that asymptotically go to zero for this example, we use Theorem B.2 directly: the uniform bounds from Corollary B.3 are not fine enough.

Theorem 6.6. *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d \times d})$ and Σ be the covariance matrix of W.*

1. *Let $h \in \mathcal{H}$. Then there is a constant $B_{6.6.1} > 0$ independent of n and a natural number $N_{6.6.1}$ such that for any $n \geq N_{6.6.1}$ we have*

$$\left| \mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z) \right| \leq B_{6.6.1} \, |h|_3 \, n^{-1-\frac{k'}{2}}.$$

2. *There is a constant $B_{6.6.2} > 0$ independent of n and a natural number $N_{6.6.2}$ such that for any $n \geq N_{6.6.2}$ we have*

$$\sup_{A \in \mathcal{K}} |\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}} Z \in A)| \leq B_{6.6.2} n^{-\frac{1}{4} - -\frac{k'}{8}}.$$

Proof. It is clear that W satisfies the conditions of Theorem B.2 for any $s = (\phi, i) \in \mathbb{I}_i$ setting

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

We apply Theorem B.2. The existence of $B_{6.6.1}$ and $B_{6.6.2}$ is proved by applying Theorem B.2 and bounding the quantity $B_{B.2}$ from the theorem. We use the notation $\mu(a) = \mu_s$ for $s = (\phi, i)$ with $a = \min(\phi)$. Then

$$B_{B.2} \leq \frac{1}{3} \sum_{i,j,\ell=1}^{d} \sum_{a=1}^{n-i} \sum_{\substack{\phi \in \binom{[n]}{i+1} \\ \min(\phi)=a}} \sum_{b=1}^{n-j} \sum_{\substack{(\psi,j) \in \mathbb{D}_j((\phi,i)) \\ \min(\psi)=b}}$$

$$\left\{ \sum_{r \in \mathbb{D}_\ell((\phi,i))} \frac{3}{2} (\sigma_i \sigma_j \sigma_\ell)^{-1} \{\mu(a)\mu(b)(1-\mu(a))(1-\mu(b))\}^{\frac{1}{2}} \right.$$

$$\left. + \sum_{r \in \mathbb{D}_\ell((\psi,j))} (\sigma_i \sigma_j \sigma_\ell)^{-1} \Big\{ \mu(a)\mu(b)(1-\mu(a))(1-\mu(b)) \Big\}^{\frac{1}{2}} \right\}.$$

We write C for an unspecified positive constant that does not depend on n. We set $p_{\max} = \max_{k'+1 \leq j \leq d} p_j$, so that by construction we have $0 < p_{\max} < 1$, and $\mu(b) \leq p_{\max}^{\sum_{i=k'+1}^{b} \binom{b}{i}} \leq p_{\max}^b$. Moreover, the number of ways of choosing a set ψ of size $j+1$, with smallest element b and sharing at least $k'+1$ given elements with ϕ, is $\binom{n-b}{j+1-(k'+1)-1}$. Then with Claim A.2 from Appendix A,

$$\sum_{b=1}^{n-j} \sum_{\substack{(\psi,j) \in \mathbb{D}_j((\phi,i)) \\ \min(\psi)=b}} \sqrt{\mu(b)} \leq \sum_{b=1}^{n-j} \binom{n-b}{j+1-(k'+1)-1} (\sqrt{p_{\max}})^b \leq Cn^{j-k'-1}.$$

Similarly,

$$\sum_{a=1}^{n-i} \sum_{\substack{(\phi,i) \in \mathbb{I}_i \\ \min(\phi)=a}} \sqrt{\mu(a)} \leq \sum_{a=1}^{n-i} \binom{n-a}{i} (\sqrt{p_{\max}})^b \leq Cn^i.$$

Next we use that $|\mathbb{D}_\ell(s)| \leq Cn^{\ell-k'}$. Also, we assume here that n is large enough for the bound in Lemma 6.3 to apply; this yields the bound $\sigma_i \geq n^{i-\frac{k'}{2}}$. Collecting these bounds gives

$$B_{B.2} \leq \sum_{i,j,\ell=1}^{d} Cn^{i+j+\ell-2k'-1} n^{\frac{3}{2}k'-i-j-\ell} \leq Cn^{-1-\frac{k'}{2}} d^3.$$

$\square$

REFERENCES

[1] A. Ababneh and M. Kahle, Maximal persistence in random clique complexes, *Journal of Applied and Computational Topology*, **8** (2014), 1449-1463.

[2] K. A. Adiprasito, B. Benedetti and F. H. Lutz, Extremal examples of collapsible complexes and random discrete Morse theory, *Discrete & Computational Geometry*, **57** (2017), 824-853.

[3] A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford Stud. Probab., 2, Oxford Sci. Publ., The Clarendon Press, Oxford University Press, New York, 1992.

[4] A. D. Barbour, M. Karoński and A. Ruciński, A central limit theorem for decomposable random variables with applications to random graphs, *Journal of Combinatorial Theory, Series B*, **47** (1989), 125-145.

[5] M. Barthelemy, Class of models for random hypergraphs, *Physical Review E*, **106** (2022), Paper No. 064310, 10 pp.

[6] U. Bauer, Ripser: efficient computation of Vietoris–Rips persistence barcodes, *Journal of Applied and Computational Topology*, **5** (2021), 391-423.

[7] U. Bauer and A. Rathod, Parameterized inapproximability of Morse matching, *Comput. Geom.*, **126** (2025), Paper No. 102148, 35 pp.

[8] G. Bianconi, *Higher-Order Networks*, Cambridge University Press, 2021.

[9] C. Bick, E. Gross, H. A. Harrington and M. T. Schaub, What are higher-order networks?, *SIAM Review*, **65** (2023), 686-731.

[10] C. A. N. Biscio, N. Chenavier, C. Hirsch and A. M. Svane, Testing goodness of fit for point processes via topological data analysis, *Electronic Journal of Statistics*, **14** (2020), 1024-1074.

[11] O. Bobrowski and M. Kahle, Topology of random geometric complexes: A survey, *Journal of Applied and Computational Topology*, **1** (2018), 331-364.

[12] O. Bobrowski and D. Krioukov, Random simplicial complexes: Models and phenomena, in *Higher-Order Systems*, Springer, Cham, 2022, 59-96.

[13] S. Bubeck, J. Ding, R. Eldan and M. Z. Rácz, Testing for high-dimensional geometry in random graphs, *Random Structures & Algorithms*, **49** (2016), 503-532.

[14] A. Cipriani, C. Hirsch and M. Vittorietti, Topology-based goodness-of-fit tests for sliced spatial data, *Computational Statistics & Data Analysis*, **179** (2023), 107655.

[15] A. Costa and M. Farber, Large random simplicial complexes, I, *Journal of Topology and Analysis*, **8** (2016), 399-429.

[16] A. Costa and M. Farber, Large random simplicial complexes, iii the critical dimension, *Journal of Knot Theory and Its Ramifications*, **26** (2017), 1740010, 26pp.

[17] J. Curry, R. Ghrist and V. Nanda, Discrete Morse theory for computing cellular sheaf cohomology, *Foundations of Computational Mathematics*, **16** (2016), 875-897.

[18] P. de Jong, A central limit theorem with applications to random hypergraphs, *Random Structures & Algorithms*, **8** (1996), 105-120.

[19] P. Eichelsbacher, B. Rednoß, C. Thäle and G. Zheng, A simplified second-order gaussian poincaré inequality in discrete setting with applications, in *Annales de l'Institut Henri Poincare (B) Probabilites et Statistiques*, Institut Henri Poincaré, **59** (2023), 271-302.

[20] M. Farber and L. Mead, Random simplicial complexes in the medial regime, *Topology and its Applications*, **272** (2020), 107065, 22pp.

[21] R. Forman, A user's guide to discrete morse theory, *Séminaire Lotharingien de Combinatoire*, **48** (2002), B48c, 35pp.

[22] S. Harker, K. Mischaikow, M. Mrozek and V. Nanda, Discrete Morse theoretic algorithms for computing homology of complexes and maps, *Foundations of Computational Mathematics*, **14** (2014), 151-184.

[23] S. Janson, Poisson approximation for large deviations, *Random Structures & Algorithms*, **1** (1990), 221-229.

[24] S. Janson, *Gaussian Hilbert Spaces*, Cambridge Tracts in Math., 129, Cambridge University Press, Cambridge, 1997.

[25] S. Janson, Large deviations for sums of partly dependent random variables, *Random Structures & Algorithms*, **24** (2004), 234-248.

[26] S. Janson and K. Nowicki, The asymptotic distributions of generalized u-statistics with applications to random graphs, *Probability Theory and Related Fields*, **90** (1991), 341-375.

[27] M. Joswig and M. E. Pfetsch, Computing optimal discrete Morse functions, *SIAM J. Discrete Math.*, **20** (2006), 11-25.

[28] M. Kahle, Topology of random clique complexes, *Discrete Mathematics*, **309** (2009), 1658-1671.

[29] M. Kahle, Random geometric complexes, *Discrete & Computational Geometry*, **45** (2011), 553-573.

[30] M. Kahle, Sharp vanishing thresholds for cohomology of random flag complexes, *Annals of Mathematics*, **179** (2014), 1085-1107.

[31] M. Kahle, Topology of random simplicial complexes: A survey, *AMS Contemp. Math*, **620** (2014), 201-222.

[32] M. Kahle and E. Meckes, Limit theorems for betti numbers of random simplicial complexes, *Homology, Homotopy and Applications*, **15** (2013), 343-374.

[33] G. Kaur and A. Röllin, Higher-order fluctuations in dense random graph models, *Electronic Journal of Probability*, **26** (2021), 1-36.

[34] J. Krebs and C. Hirsch, Functional central limit theorems for persistent betti numbers on cylindrical networks, *Scandinavian Journal of Statistics*, **49** (2022), 427-454.

[35] J. Lei, A goodness-of-fit test for stochastic block models, *The Annals of Statistics*, **44** (2016), 401-424.

[36] C. Ley, G. Reinert and Y. Swan, Stein's method for comparison of univariate distributions, *Probability Surveys*, **14** (2017), 1-52.

[37] N. Linial and R. Meshulam, Homological connectivity of random 2-complexes, *Combinatorica*, **26** (2006), 475-487.

[38] N. Linial and Y. Peled, On the phase transition in random simplicial complexes, *Annals of Mathematics*, **184** (2016), 745-773.

[39] S. Liu and M. Z. Rácz, Phase transition in noisy high-dimensional random geometric graphs, *Electron. J. Stat.*, **17** (2023), 3512-3574.

[40] W. McGinley and R. Sibson, Dissociated random variables, in *Mathematical Proceedings of the Cambridge Philosophical Society*, Cambridge University Press, **77** (1975), 185-188.

[41] K. Mischaikow and V. Nanda, Morse theory for filtrations and efficient computation of persistent homology, *Discrete and Computational Geometry*, **50** (2013), 330-353.

[42] A. Newman, One-sided sharp thresholds for homology of random flag complexes, *J. Lond. Math. Soc. (2)*, **109** (2024), Paper No. e12872, 26 pp.

[43] M. Newman, *Networks*, Oxford University Press, Oxford, 2018.

[44] S. Ouadah, S. Robin and P. Latouche, Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases, *Scandinavian Journal of Statistics*, **47** (2020), 156-181.

[45] T. Owada, G. Samorodnitsky and G. Thoppe, Limit theorems for topological invariants of the dynamic multi-parameter simplicial complex, *Stochastic Processes and their Applications*, **138** (2021), 56-95.

[46] A. Roy and D. Yogeshwaran, Random clique complex process inside the critical window, arXiv preprint, `arXiv:2303.17535`.

[47] A. Ruciński, When are small subgraphs of a random graph normally distributed?, *Probability Theory and Related Fields*, **78** (1988), 1-10.

[48] G. Samorodnitsky and T. Owada, Large deviations for subcomplex counts and betti numbers in multiparameter simplicial complexes, *Random Structures Algorithms*, **63** (2023), 533-556.

[49] J. Schulz, *The Optimal Berry-Esseen Constant in the Binomial Case*, PhD thesis, University of Trier, 2016.

[50] E. Spanier, *Algebraic Topology*, McGraw-Hill Book Co., New York-Toronto-London, 1966.

[51] T. Temčinas, V. Nanda and G. Reinert, Multivariate central limit theorems for random clique complexes, *Journal of Applied and Computational Topology*, **8** (2024), 1837-1880.

[52] N. Yerolemou and V. Nanda, Morse theory for complexes of groups, *J. Pure Appl. Algebra*, **228** (2024), Paper No. 107606, 31 pp.

[53] Z.-S. Zhang, Berry–esseen bounds for generalized u-statistics, *Electronic Journal of Probability*, **27** (2022), 1-36.

[54] K. Zuev, O. Eisenberg and D. Krioukov, Exponential random simplicial complexes, *Journal of Physics A: Mathematical and Theoretical*, **48** (2015), 465002, 25pp.

Appendix A. Proof of Lemmas 6.3 and 6.4.

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

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

where

V_1

$$= 2 \sum_{l < m}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)} \binom{n-m}{2k+1-j-q} \binom{2k+1-j-q}{k} \binom{k}{j-1} \binom{m-l+1}{q-1}$$

$$\left\{ \rho^+(k+1)^2 \rho^+(j)^{-1} \tau(l,m,q,k,j-1) \right.$$

$$\left[\left(1 - 2\rho(k) + \rho(k)^2 \rho(j-1)^{-1}\right)^{l-1} - \left(1 - \rho(k+1) - \rho(k) + \rho(k+1)\rho(k)\rho(j-1)^{-1}\right)^{l-1} \right]$$

$$+ \rho^+(k+1)^2 \rho^+(j)^{-1} \tau(l,m,q,k+1,j)$$

$$\left[\left(1 - 2\rho(k+1) + \rho(k+1)^2 \rho(j)^{-1}\right)^{l-1} - \left(1 - \rho(k+1) - \rho(k) + \rho(k)\rho(k+1)\rho(j)^{-1}\right)^{l-1} \right]$$

$$\left. - \mu(l)\mu(m) \right\};$$

V_2

$$= 2 \sum_{l < m}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)} \binom{n-m}{2k+1-j-q} \binom{2k+1-j-q}{k} \binom{k}{j} \binom{m-l+1}{q-1}$$

$$\left\{ \rho^+(k+1)^2 \rho^+(j)^{-1} \tau(l,m,q,k,j) \right.$$

$$\left[\left(1 - 2\rho(k) + \rho(k)^2 \rho(j)^{-1}\right)^{l-1} - \left(1 - \rho(k+1) - \rho(k) + \rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1} \right]$$

$$+ \rho^+(k+1)^2 \rho^+(j)^{-1} \tau(l,m,q,k+1,j)$$

$$\left[\left(1 - 2\rho(k+1) + \rho(k+1)^2 \rho(j)^{-1}\right)^{l-1} - \left(1 - \rho(k+1) - \rho(k) + \rho(k)\rho(k+1)\rho(j)^{-1}\right)^{l-1} \right]$$

$$\left. - \mu(l)\mu(m) \right\};$$

$$V_3 = \sum_{l=1}^{n-k} \sum_{j=k'+1}^{k} \binom{n-l}{2k+1-j}\binom{2k+1-j}{k}\binom{k}{j-1}\Big\{\rho^+(k+1)^2\rho^+(j)^{-1}$$

$$\Big[\big(1-2\rho(k+1)+\rho(k+1)^2\rho(j)^{-1}\big)^{l-1} + \big(1-2\rho(k)+\rho(k)^2\rho(j-1)^{-1}\big)^{l-1}$$

$$- 2\big(1-\rho(k)-\rho(k+1)+\rho(k)\rho(k+1)\rho(j-1)^{-1}\big)^{l-1}\Big] - \mu(l)^2\Big\};$$

$$V_4 = \sum_{l=1}^{n-k} \binom{n-l}{k}\big\{\mu(l)-\mu(l)^2\big\}.$$

Here we have used the following abbreviations:

$$\rho(a) := \prod_{i=k'+1}^{a} p_i^{\binom{a}{i}}; \qquad \tau(l,m,q,a,b) := (1-\rho(a))^{m-l-q}\big(1-\rho(a)\rho(b)^{-1}\big)^q;$$

$$\rho^+(a) := \prod_{i=k'+1}^{a} p_i^{\binom{a}{i+1}}; \quad \mu(a) := \rho^+(k+1)\Big((1-\rho(k+1))^{a-1}-(1-\rho(k))^{a-1}\Big).$$

Also, the notation $\sum_{l<m}^{n-k}$ stands for $\sum_{l=1}^{n-k-1}\sum_{m=l+1}^{n-k}$.

Proof of Lemma A.1. Here, we adapt and generalise the proof of [51, Lemma 34]. For $s \in \binom{[n]}{k+1}$ recall that $s_- = s \setminus \{\min(s)\}$. We write:

$$Y_s^+ = \prod_{i=1}^{\min(s)-1}\left(1-\prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq s\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right),$$

$$Y_s^- = \prod_{i=1}^{\min(s)-1}\left(1-\prod_{j=k'+1}^{k}\prod_{\substack{\alpha\subseteq s_-\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right),$$

$$Z_s = \prod_{j=k'+1}^{k}\prod_{\substack{\alpha\subseteq s\\|\alpha|=j+1}}\xi_\alpha^{(j+1)}, \qquad\qquad Y_s = Y_s^+ - Y_s^-.$$

Then Z_s and Y_s are independent and $T_{k+1} = \sum_{s\in\binom{[n]}{k+1}} Z_s Y_s$. Consider the variance

$$\mathrm{Var}(T_{k+1}) = \sum_{s\in\binom{[n]}{k+1}} \mathrm{Var}(Z_s Y_s) + \sum_{\substack{s\neq t\in\binom{[n]}{k+1}\\\min(s)\neq\min(t)}}\mathrm{Cov}(Z_s Y_s, Z_t Y_t)$$

$$+ \sum_{\substack{s\neq t\in\binom{[n]}{k+1}\\\min(s)=\min(t)}}\mathrm{Cov}(Z_s Y_s, Z_t Y_t). \tag{9}$$

For the first term in (9), writing $\min(s) = l$, we get

$$\mathbb{P}(Z_s Y_s = 1) = \mu(l) = \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}}\left[\left(1-\prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}}\right)-\left(1-\prod_{i=k'+1}^{k} p_i^{\binom{k}{i}}\right)^{l-1}\right].$$

Now we see

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

$$
= \sum_{l=1}^{n-k} \binom{n-l}{k} \left(\mathbb{P}(Z_s Y_s = 1) - \mathbb{P}(Z_s Y_s = 1)^2 \right)
$$

$$
= \sum_{l=1}^{n-k} \binom{n-l}{k} \left\{ \mu(l) - \mu(l)^2 \right\} = V_4.
$$

Now consider the covariance terms in (9), the expansion of the variance. Note that for any $s, t \in \binom{[n]}{k+1}$ if $|s \cap t| < k' + 1$, then the variables $Z_s Y_s$ and $Z_t Y_t$ can be written as functions of two disjoint sets of independent random variables for the form $\xi_\alpha^{(l)}$ for $\alpha \in \binom{[n]}{l}$ and hence have zero covariance.

Fix $s, t \in \binom{[n]}{k+1}$ and assume $k' + 1 \leq |s \cap t| \leq k$. Note that because $|s \cap t| \neq k+1$, we have $s \neq t$. There are $2\binom{k+1}{i} - \binom{|s \cap t|}{i}$ distinct simplices of size i in s and t combined and hence $\mathbb{P}(Z_s Z_t = 1) = \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1} - \binom{|s \cap t|}{i}}$. 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 $\min(s) \leq \min(t)$. Then we have for $Y_s^+ Y_t^+$:

$$
Y_s^+ Y_t^+ = \prod_{i=1}^{\min(t)-1} \left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq t \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) \prod_{i=1}^{\min(s)-1} \left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq s \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right)
$$

$$
= \prod_{i=1}^{\min(s)-1} \left\{ \left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq s \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) \left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq t \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) \right\}
$$

$$
\prod_{\substack{i=\min(s) \\ i \in s}}^{\min(t)-1} \left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq t \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) \prod_{\substack{i=\min(s) \\ i \notin s}}^{\min(t)-1} \left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq t \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right).
$$

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

$$
\mathbb{P}\left[\left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq s \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) \left(1 - \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq t \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} \right) = 1 \right]
$$

$$
= \mathbb{P}\left[\neg \left(\prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq s \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} = 1 \cup \prod_{j=k'+1}^{k+1} \prod_{\substack{\alpha \subseteq t \\ |\alpha|=j}} \xi_{\alpha \cup \{i\}}^{(j+1)} = 1 \right) \right]
$$

$$
= 1 - 2 \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}} + \prod_{i=k'+1}^{k+1} p_i^{2\binom{k+1}{i} - \binom{|s \cap t|}{i}}.
$$

Here $\cup$ indicates that either or both of these events may happen.

Moreover, the variable

$\prod_{i=1}^{\min(s)-1}\left(1 - \prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq s\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right)\left(1 - \prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq t\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right)$ and the vari-

able $\prod_{\substack{i=\min(s)\\i\notin s}}^{\min(t)-1}\left(1 - \prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq t\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right)$ are independent of Z_sZ_t. Recall the

notation $[a,b] = \{a, a+1, \ldots, b\}$ for two positive integers $a \leq b$. Setting $q :=$
$|s \cap [\min(s), \min(t) - 1]|$,

$$\mathbb{P}\left(Y_s^+Y_t^+ = 1|Z_sZ_t = 1\right)$$

$$=\mathbb{P}\left(\prod_{i=1}^{\min(s)-1}\left(1 - \prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq s\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right)\left(1 - \prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq t\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right) = 1\right)$$

$$\mathbb{P}\left(\prod_{\substack{i=\min(s)\\i\notin s}}^{\min(t)-1}\left(1 - \prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq t\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right) = 1\right)$$

$$\mathbb{P}\left(\prod_{\substack{i=\min(s)\\i\in s}}^{\min(t)-1}\left(1 - \prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq t\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right) = 1\middle|Z_sZ_t = 1\right)$$

$$=\left(1 - 2\prod_{i=k'+1}^{k+1}p_i^{\binom{k+1}{i}} + \prod_{i=k'+1}^{k+1}p_i^{2\binom{k+1}{i}-\binom{|s\cap t|}{i}}\right)^{\min(s)-1}$$

$$\left(1 - \prod_{i=k'+1}^{k+1}p_i^{\binom{k+1}{i}}\right)^{\min(t)-\min(s)-q}\left(1 - \prod_{i=k'+1}^{k+1}p_i^{\binom{k+1}{i}-\binom{|s\cap t|}{i}}\right)^q.$$

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 $l = \min(s)$, $m = \min(t)$. Also,
we set

$$\pi(l, m, a, b, d_1, d_2, q) := \left(1 - \prod_{i=k'+1}^{a}p_i^{\binom{a}{i}} - \prod_{i=k'+1}^{b}p_i^{\binom{b}{i}} + \prod_{i=k'+1}^{\max(a,b)}p_i^{\binom{a}{i}+\binom{b}{i}-\binom{d_1}{i}}\right)^{l-1}$$

$$\left(1 - \prod_{i=k'+1}^{a}p_i^{\binom{a}{i}}\right)^{m-l-q}\left(1 - \prod_{i=k'+1}^{a}p_i^{\binom{a}{i}-\binom{d_2}{i}}\right)^q.$$

Using the described strategy we get

$$\mathbb{P}\left(Y_s^+Y_t^+ = 1|Z_sZ_t = 1\right) = \pi(l, m, k+1, k+1, |s\cap t|, |s\cap t|, q)$$

$$\mathbb{P}\left(Y_s^-Y_t^- = 1|Z_sZ_t = 1\right) = \pi(l, m, k, k, |s_-\cap t_-|, |s\cap t_-|, q)$$

$$\mathbb{P}\left(Y_s^+Y_t^- = 1|Z_sZ_t = 1\right) = \pi(l, m, k, k+1, |s\cap t_-|, |s\cap t_-|, q)$$

$$\mathbb{P}\left(Y_s^-Y_t^+ = 1|Z_sZ_t = 1\right) = \pi(l, m, k+1, k, |s_-\cap t|, |s\cap t|, q).$$

Now we are ready to calculate the covariance:

$$\text{Cov}(Z_s Y_s, Z_t Y_t)$$

$$= \mathbb{E}\left\{Z_s Z_t Y_s^+ Y_t^+\right\} + \mathbb{E}\left\{Z_s Z_t Y_s^- Y_t^-\right\} - \mathbb{E}\left\{Z_s Z_t Y_s^+ Y_t^-\right\}$$

$$- \mathbb{E}\left\{Z_s Z_t Y_s^- Y_t^+\right\} - \mathbb{E}\left\{Z_s Y_s\right\} \mathbb{E}\left\{Z_t Y_t\right\}$$

$$= \mathbb{P}(Z_s Z_t = 1)\Big\{\mathbb{P}\left(Y_s^+ Y_t^+ = 1 | Z_s Z_t = 1\right) + \mathbb{P}\left(Y_s^- Y_t^- = 1 | Z_s Z_t = 1\right)$$

$$- \mathbb{P}\left(Y_s^+ Y_t^- = 1 | Z_s Z_t = 1\right) - \mathbb{P}\left(Y_s^- Y_t^+ = 1 | Z_s Z_t = 1\right)\Big\}$$

$$- \mathbb{P}(Z_s Y_s = 1)\mathbb{P}(Z_t Y_t = 1)$$

$$= \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1} - \binom{|s \cap t|}{i+1}} (\pi(l, m, k+1, k+1, |s \cap t|, |s \cap t|, q)$$

$$+ \pi(l, m, k, k, |s_- \cap t_-|, |s \cap t_-|, q)$$

$$- \pi(l, m, k, k+1, |s \cap t_-|, |s \cap t_-|, q)$$

$$- \pi(l, m, k+1, k, |s_- \cap t|, |s \cap t|, q)) - \mu(l)\mu(m).$$

Next we consider the two covariance sums in (9) separately. First assume that $\min(s) \neq \min(t)$. Given $l, m \in [n-k]$, $j \in [k]$, and $q \in [\min(k+1, |m-l|)]$ define the set

$$\Gamma_{k+1}(l, m, j, q)$$

$$= \left\{(s, t) \mid s, t \in \binom{[n]}{k+1}, \min(s) = l, \min(t) = m, |s \cap t| = j, \max(q_{s,t}, q_{t,s}) = q\right\}$$

as well as

$$\Gamma_{k+1}^+(l, m, j, q) = \left\{(s, t) \in \Gamma_{k+1}(l, m, j, q) \mid \min(t) \in s\right\}$$

and

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

Here $q_{s,t} = |s \cap [\min(s), \min(t) - 1]|$, $q_{t,s} = |t \cap [\min(t), \min(s) - 1]|$. From the proof of Lemma 4.3 in [51] we know that

$$|\Gamma_{k+1}^+(l, m, j, q)| = \binom{n-m}{2k+1-j-q}\binom{2k+1-j-q}{k}\binom{k}{j-1}\binom{m-l+1}{q-1};$$

and

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

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

$$\sum_{\substack{s \neq t \in \binom{[n]}{k+1} \\ \min(s) \neq \min(t)}} \text{Cov}(Z_s Y_s, Z_t Y_t)$$

$$= \sum_{l=1}^{n-k} \sum_{m=i+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)} \sum_{(s,t) \in \Gamma_{k+1}^{+}(l,m,j,q)} \text{Cov}(Z_s Y_s, Z_t Y_t)$$

$$+ \sum_{l=1}^{n-k} \sum_{m=i+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)} \sum_{(s,t) \in \Gamma_{k+1}^{-}(l,m,j,q)} \text{Cov}(Z_s Y_s, Z_t Y_t)$$

$$+ \sum_{m=1}^{n-k} \sum_{l=j+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,l-m)} \sum_{(s,t) \in \Gamma_{k+1}^{+}(m,l,j,q)} \text{Cov}(Z_s Y_s, Z_t Y_t)$$

$$+ \sum_{m=1}^{n-k} \sum_{l=j+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,l-m)} \sum_{(s,t) \in \Gamma_{k+1}^{-}(m,l,j,q)} \text{Cov}(Z_s Y_s, Z_t Y_t)$$

$$= \sum_{l=1}^{n-k} \sum_{m=i+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)} |\Gamma_{k+1}^{+}(l,m,j,q)|$$

$$\Big\{ \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1} - \binom{j}{i+1}} (\pi(l,m,k+1,k+1,j,j,q)$$

$$+ \pi(l,m,k,k,j-1,j-1,q) - \pi(l,m,k,k+1,j-1,j-1,q)$$

$$- \pi(l,m,k+1,k,j,j,q)) - \mu(l)\mu(m) \Big\}$$

$$+ \sum_{l=1}^{n-k} \sum_{m=i+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)} |\Gamma_{k+1}^{-}(l,m,j,q)|$$

$$\Big\{ \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1} - \binom{j}{i+1}} (\pi(l,m,k+1,k+1,j,j,q)$$

$$+ \pi(l,m,k,k,j,j,q) - \pi(l,m,k,k+1,j,j,q)$$

$$- \pi(l,m,k+1,k,j,j,q)) - \mu(l)\mu(m) \Big\}$$

$$+ \sum_{m=1}^{n-k} \sum_{l=j+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,l-m)} |\Gamma_{k+1}^{+}(m,l,j,q)|$$

$$\Big\{ \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1} - \binom{j}{i+1}} (\pi(m,l,k+1,k+1,j,j,q)$$

$$+ \pi(m,l,k,k,j-1,j-1,q) - \pi(m,l,k,k+1,j-1,j-1,q)$$

$$- \pi(m,l,k+1,k,j,j,q)) - \mu(l)\mu(m) \Big\}$$

$$+ \sum_{m=1}^{n-k} \sum_{l=j+1}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,l-m)} |\Gamma_{k+1}^{-}(m,l,j,q)|$$

$$\left\{ \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1}-\binom{j}{i+1}} (\pi(m,l,k+1,k+1,j,j,q) \right.$$

$$+ \pi(m,l,k,k,j,j,q) - \pi(m,l,k+1,j,j,q)$$

$$\left. - \pi(m,l,k+1,k,j,j,q)) - \mu(l)\mu(m) \right\} = V_1 + V_2.$$

Similarly, we calculate the remaining term in the expansion of the variance (9). We notice that if $l = m$, then $q = 0$ and we have $\Gamma_{k+1}(l,l,j,0) = \Gamma_{k+1}^{+}(l,l,j,0)$. Hence, $|\Gamma_{k+1}(l,l,j,0)| = \binom{n-l}{2k+1-j}\binom{2k+1-j}{k}\binom{k}{j-1}$, and

$$\sum_{\substack{s\neq t\in\binom{[n]}{k+1} \\ \min(s)=\min(t)}} \mathrm{Cov}(Z_sY_s, Z_tY_t) = \sum_{l=1}^{n-k} \sum_{j=k'+1}^{k} \sum_{(s,t)\in\Gamma_{k+1}(l,l,j,0)} \mathrm{Cov}(Z_sY_s, Z_tY_t)$$

$$= \sum_{l=1}^{n-k} \sum_{j=k'+1}^{k} \binom{n-l}{2k+1-j}\binom{2k+1-j}{k}\binom{k}{j-1}\left\{ \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1}-\binom{j}{i+1}} \right.$$

$$[\pi(l,l,k+1,k+1,j,0,0) + \pi(l,l,k,k,j-1,0,0) - 2\pi(l,l,k+1,k,j-1,0,0)]$$

$$\left. - \mu(l)^2 \right\}$$

$$= V_3.$$

$\square$

Proof for Lemma 6.3. Fix $k' + 1 \leq k \leq n - 1$ and $p \in (0,1)$, and consider the variance. From Lemma A.1 we have $\mathrm{Var}\{T_{k+1}\} = V_1 + V_2 + V_3 + V_4$. First we lower bound V_1 and V_2 by just the negative part of the sum:

$$V_1 \geq -2\rho^+(k+1)^2 \sum_{l<m}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)}$$

$$\binom{n-m}{2k+1-j-q}\binom{2k+1-j-q}{k}\binom{k}{j-1}\binom{m-l+1}{q-1}$$

$$\left\{ \rho^+(j)^{-1}\tau(l,m,q,k,j-1)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j-1)^{-1}\right)^{l-1} \right.$$

$$+ \rho^+(j)^{-1}\tau(l,m,q,k+1,j)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1}$$

$$\left. + (1-\rho(k+1))^{l+m-2} + (1-\rho(k))^{l+m-2} \right\};$$

$$V_2 \geq -2\rho^+(k+1)^2 \sum_{l<m}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)}$$

$$\binom{n-m}{2k+1-j-q}\binom{2k+1-j-q}{k}\binom{k}{j}\binom{m-l+1}{q-1}$$

$$\left\{\rho^+(j)^{-1}\tau(l,m,q,k,j)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1}\right.$$

$$+\rho^+(j)^{-1}\tau(l,m,q,k+1,j)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1}$$

$$\left.+(1-\rho(k+1))^{l+m-2}+(1-\rho(k))^{l+m-2}\right\}.$$

Now using that $\binom{k}{j}+\binom{k}{j-1}=\binom{k+1}{j}$ and $\rho(a)\le\rho(b)$ for any two integers $b<a$, it is easy to see that $V_1+V_2\ge-8\rho^+(k+1)^2R_1-8\rho^+(k+1)^2R_2$, where

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

$$\binom{k+1}{j}\binom{m-l+1}{q-1}(1-\rho(k+1))^{l+m-2};$$

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

$$\binom{2k+1-j-q}{k}\binom{k+1}{j}\binom{m-l+1}{q-1}$$

$$\rho^+(j)^{-1}(1-\rho(k+1))^{m-l}\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1}.$$

For V_3 we lower bound by terms with $m=k'+1$ and the negative parts of the other terms. Note that $\rho^+(k'+1)=1=\rho(k')$ and $\rho(k'+1)<1$. Now we can proceed:

$$V_3\ge\rho^+(k+1)^2\sum_{l=1}^{n-k}\binom{n-l}{2k-k'}\binom{2k-k'}{k}\binom{k}{k'}$$

$$\left\{(1-2\rho(k+1)+\rho(k+1)^2\rho(k'+1)^{-1})^{l-1}-(1-\rho(k+1))^{2l-2}\right\}$$

$$-2\rho^+(k+1)^2\sum_{l=1}^{n-k}\sum_{j=k'+2}^{k}\binom{n-l}{2k+1-j}\binom{2k+1-j}{k}\binom{k}{j-1}$$

$$\left\{\rho^+(j)^{-1}\left(1-\rho(k)-\rho(k+1)+\rho(k)\rho(k+1)\rho(j-1)^{-1}\right)^{l-1}\right.$$

$$\left.+(1-\rho(k+1))^{2l-2}\right\}$$

$$=\rho^+(k+1)^2R_4-2\rho^+(k+1)^2R_3;$$

here we call the first sum of the lower bound R_4 and the second sum R_3. For V_4 we use the trivial lower bound $V_4\ge0$. Hence, we have

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

Let us now upper bound R_1:

$$R_1 \leq \sum_{l<m}^{n-k} \sum_{j=k'+1}^{k} \sum_{q=1}^{\min(k+1,m-l)} \frac{(n-m)^{2k+1-j-q}}{(2k+1-j-q)!} \frac{(2k+1-j-q)^k}{k!}$$

$$\frac{(k+1)^j}{j!} \frac{(m-l+1)^{q-1}}{(q-1)!}(1-\rho(k+1))^{l+m-2}$$

$$\leq \sum_{l<m}^{n-k} \sum_{q=1}^{\min(k+1,m-l)} (k-k') \frac{n^{2k-k'-q}}{1} \frac{(2k-k'-1)^k}{k!}$$

$$\frac{(k+1)^k}{(k'+1)!} \frac{n^{q-1}}{1}(1-\rho(k+1))^{l+m-2}$$

$$\leq n^{2k-k'-1}(k-k')\frac{(2k-k'-1)^k}{k!}\frac{(k+1)^{k+1}}{(k'+1)!}\sum_{l<m}^{\infty}(1-\rho(k+1))^{l+m-2}$$

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

Noting that $(1-\rho(k+1))^{m-l}\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1} \leq (1-\rho(k+1))^{m-1}$ (since $\rho(k+1)\rho(j)^{-1} \leq 1$ for all $j \in [k]$), we can bound R_2 in an identical way:

$$R_2 \leq n^{2k-k'-1}(k-k')\frac{(2k-k'-1)^k}{k!}\frac{(k+1)^{k+1}}{(k'+1)!}\rho^+(k)^{-1}\sum_{l<m}^{\infty}(1-\rho(k+1))^{m-1}$$

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

Noting that $\left(1-\rho(k)-\rho(k+1)+\rho(k)\rho(k+1)\rho(j-1)^{-1}\right)^{l-1} \leq (1-\rho(k+1))^{l-1}$ and $(1-\rho(k+1))^{2l-2} \leq (1-\rho(k+1))^{l-1}$ we proceed to bound R_3:

$$R_3 \leq \sum_{l=1}^{n-k} \sum_{j=k'+2}^{k} \binom{n-l}{2k+1-j}\binom{2k+1-j}{k}\binom{k}{j-1}(\rho^+(j)^{-1}+1)(1-\rho(k+1))^{l-1}$$

$$\leq \sum_{l=1}^{n-k} \sum_{j=k'+2}^{k} \frac{n^{2k+1-j}}{(2k+1-j)!}\frac{(2k+1-j)^k}{k!}\frac{k^{j-1}}{(j-1)!}(\rho^+(k)^{-1}+1)(1-\rho(k+1))^{l-1}$$

$$\leq (k-k'-1)\frac{n^{2k-k'-1}}{(k+1)!}\frac{(2k-k'-1)^k}{k!}\frac{k^k}{(k'+1)!}(\rho^+(k)^{-1}+1)\sum_{l=1}^{\infty}(1-\rho(k+1))^{l-1}$$

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

To lower bound R_4 we just take the $l=2$ term as all summands are non-negative:

$$R_4 \geq \binom{n-2}{2k-k'}\binom{2k-k'}{k}\binom{k}{k'}\left\{\rho(k+1)^2\rho(k'+1)^{-1}-\rho(k+1)^2\right\}$$

$$\geq \frac{(n-2)^{2k-k'}}{(2k-k')^{2k-k'}}\binom{2k-k'}{k}\binom{k}{k'}\rho(k+1)^2(\rho(k'+1)^{-1}-1).$$

Since R_1, R_2, R_3 are all at most of the order $n^{2k-k'-1}$ and R_4 is at least of the order $n^{2k-k'}$, we have that for any fixed $k \geq k' + 1$ and a sequence of probabilities $p_i \in (0, 1)$ there exists a constant $C > 0$ independent of n and a natural number N such that for any $n \geq N$,

$$\text{Var}(T_{k+1}) \geq \rho^+(k+1)^2(R_4 - 8R_1 - 8R_2 - 2R_3) \geq Cn^{2k-k'}.$$

$\square$

Proof of Lemma 6.4. For $i < j$, write $k' + i = k$ and $k' + j = r$. Note that σ_k^2 and σ_r^2 are sums and so is $\text{Cov}(T_{k+1}, T_{r+1})$. Hence, to get the limit of the ratio we can just look at the ratio of the highest order terms from each sum. From the proof of Lemma 6.3 it is clear that the highest order term for σ_k^2 is the sum of the terms in V_3, where the index of the sum j has the value $k' + 1$. That is,

$$\sigma_k^2 = \sum_{l=1}^{n-k} \binom{n-l}{2k-k'} \binom{2k-k'}{k} \binom{k}{k'} \rho^+(k+1)^2$$

$$\left[(1 - 2\rho(k+1) + \rho(k+1)^2\rho(k'+1)^{-1})^{l-1} - (1 - 2\rho(k+1) + \rho(k+1)^2)^{l-1}\right]$$

$$+ o(n^{2k-k'}).$$

Analogously to the calculations in Lemmas A.1 and 6.3, one can perform calculations to find the highest order term of $\text{Cov}(T_{k+1}, T_{r+1})$. Similarly, like in the variance case, it will correspond to subsets $s \in \binom{[n]}{k+1}, t \in \binom{[n]}{r+1}$ such that $\min(s) = \min(t)$ and $|s \cap t| = k' + 1$. That is, borrowing the notation from the proof of Lemma A.1, we have

$$\text{Cov}(T_{k+1}, T_{r+1})$$

$$= \sum_{\substack{s \in \binom{[n]}{k+1}, t \in \binom{[n]}{r+1} \\ \min(s) = \min(t) \\ |s \cap t| = k' + 1}} \text{Cov}(Z_s Y_s, Z_t Y_t) + o(n^{k+r-k'})$$

$$= o(n^{k+r-k'}) + \sum_{l=1}^{n-r} \binom{n-l}{k+r-k'} \binom{k+r-k'}{k} \binom{k}{k'} \rho^+(k+1)\rho^+(r+1)$$

$$\left[(1 - \rho(k+1) - \rho(r+1) + \rho(k+1)\rho(r+1)\rho(k'+1)^{-1})^{l-1} - (1 - \rho(k+1)\right.$$

$$\left. - \rho(r+1) + \rho(k+1)\rho(r+1))^{l-1}\right].$$

To deal with those sums, we will use this auxiliary claim.

Claim A.2. *For any fixed positive integer k and $x \in (0, 1)$ we have*

$$\sum_{l=1}^{n} \binom{n-l}{k} x^{l-1} = \frac{n^k}{k!}(1-x)^{-1} + o(n^k).$$

A proof of this claim may be obtained by first noting that

$$\binom{n-l}{k} \leq \binom{n}{k} \leq \frac{n^k}{k!},$$

and then using the fact that $\sum_0^\infty x^k = (1-x)^{-1}$ holds because $|x| < 1$.

Now we can proceed with the limit, cancelling the binomial coefficients in a similar way to Lemma 5.1 and then using Claim A.2.

$$\lim_{n \to \infty} \Sigma_{i,j}$$

$$= \lim_{n \to \infty} \sum_{l=1}^{n-r} \binom{n-l}{k+r-k'} (k+r-k')! \big[(1 - \rho(k+1) - \rho(r+1)$$

$$+ \rho(k+1)\rho(r+1)\rho(k'+1)^{-1})^{l-1} - (1 - \rho(k+1) - \rho(r+1) + \rho(k+1)\rho(r+1))^{l-1} \big]$$

$$\left\{ \sum_{l=1}^{n-k} \binom{n-l}{2k-k'} (2k-k')! \left[(1 - 2\rho(k+1) + \rho(k+1)^2\rho(k'+1)^{-1})^{l-1} - (1 - 2\rho(k+1) + \rho(k+1)^2)^{l-1} \right] \right\}^{-\frac{1}{2}}$$

$$\left\{ \sum_{l=1}^{n-r} \binom{n-l}{2r-k'} (2r-k')! \left[(1 - 2\rho(r+1) + \rho(r+1)^2\rho(k'+1)^{-1})^{l-1} - (1 - 2\rho(r+1) + \rho(r+1)^2)^{l-1} \right] \right\}^{-\frac{1}{2}}$$

$$= \lim_{n \to \infty} n^{k+r-k'} \big[(\rho(k+1) + \rho(r+1) - \rho(k+1)\rho(r+1)\rho(k'+1)^{-1})^{-1}$$

$$- (\rho(k+1) + \rho(r+1) - \rho(k+1)\rho(r+1))^{-1} \big]$$

$$\left\{ n^{2k-k'} \left[(2\rho(k+1) - \rho(k+1)^2\rho(k'+1)^{-1})^{-1} - (2\rho(k+1) - \rho(k+1)^2)^{-1} \right] \right\}^{-\frac{1}{2}}$$

$$\left\{ n^{2r-k'} \left[(2\rho(r+1) - \rho(r+1)^2\rho(k'+1)^{-1})^{-1} - (2\rho(r+1) - \rho(r+1)^2)^{-1} \right] \right\}^{-\frac{1}{2}}$$

$$= \frac{\rho(k+1)\rho(r+1)\sqrt{(2 - \rho(k+1))(2 - \rho(r+1))(2 - \rho(k+1)\rho(k'+1)^{-1})(2 - \rho(r+1)\rho(k'+1)^{-1})}}{(\rho(k+1) + \rho(r+1) - \rho(k+1)\rho(r+1)\rho(k'+1)^{-1})(\rho(k+1) + \rho(r+1) - \rho(k+1)\rho(r+1))}.$$

□

Appendix B. A multivariate 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 $\mathbb{I} =: \bigcup_{i \in [d]} \mathbb{I}_i$. Let $X_s, s = (k, i) \in \mathbb{I}$ be a collection of real centered random variables which are defined on the same probability space. For each $i \in [d]$ form the sum

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

Our interest here is in 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 that has been studied in [51, 4, 40].

Definition B.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);
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 $s \in \mathbb{I}$, as $\mathbb{I}_j \cap \mathbb{I}_\ell = \emptyset$ for $j \neq \ell$ 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). $\mathbb{D}_j(s)$ is the dependency

neighbourhood of X_s in the j-th component of the vector W. We write $\mathbb{D}(s) = \bigcup_{j \in [d]} \mathbb{D}_j(s)$ for the disjoint union of these of dependency neighbourhoods.

Theorem B.2 (Theorems 2 and 9 in [51]). *Let $h : \mathbb{R}^d \to \mathbb{R}$ be any three times continuously differentiable function whose third partial derivatives are Lipschitz continuous and bounded. 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:*

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

2.
$$\sup_{A \in \mathcal{K}} \left|\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}}Z \in A)\right| \le 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{B.2}^{\frac{1}{4}},$$

where $B_{B.2} = B_{B.2.1} + B_{B.2.2}$ is the sum given by

$$B_{B.2.1} := \frac{1}{3} \sum_{s \in \mathbb{I}} \sum_{t,u \in \mathbb{D}(s)} \left(\frac{1}{2}\mathbb{E}\left|X_s X_t X_u\right| + \mathbb{E}\left|X_s X_t\right| \mathbb{E}\left|X_u\right|\right)$$

$$B_{B.2.2} := \frac{1}{3} \sum_{s \in \mathbb{I}} \sum_{t \in \mathbb{D}(s)} \sum_{v \in \mathbb{D}(t) \setminus \mathbb{D}(s)} \left(\mathbb{E}\left|X_s X_t X_v\right| + \mathbb{E}\left|X_s X_t\right| \mathbb{E}\left|X_v\right|\right).$$

Corollary B.3 (Corollaries 8 and 11 in [51]). *Let the assumptions of Theorem B.2 hold. If we further require that the random variables X_s and X_t are independent whenever t lies in $\mathbb{I} \setminus \mathbb{D}(s)$, then for any vector of dissociated sums $W \in \mathbb{R}^d$ with a positive semi-definite covariance matrix Σ,*

1.
$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \le B_{B.3}\left|h\right|_3;$$

2.
$$\sup_{A \in \mathcal{K}} \left|\mathbb{P}(W \in A) - \mathbb{P}(\Sigma^{\frac{1}{2}}Z \in A)\right| \le 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{B.3}^{\frac{1}{4}},$$

where $B_{B.3}$ is a sum over $(i, j, k) \in [d]^3$ of the form:

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

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

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

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

The following lemma is useful when we deal with Bernoulli variables.

Lemma B.4 (Lemma 9 in [51]). *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 deterministic quantities. Consider variables $X_i := c_i(\xi_i - \mu_i)$ for $i = 1, 2, 3$. Then we have*

$$\mathbb{E}\left|X_1 X_2 X_3\right| \le c_1 c_2 c_3 \left\{\mu_1 \mu_2 (1 - \mu_1)(1 - \mu_2)\right\}^{\frac{1}{2}};$$

$$\mathbb{E}\left|X_1 X_2\right| \mathbb{E}\left|X_3\right| \le c_1 c_2 c_3 \left\{\mu_1 \mu_2 (1 - \mu_1)(1 - \mu_2)\right\}^{\frac{1}{2}}.$$

Appendix C. **Testing calibration procedure on single maturities.**

ϵ_1	ϵ_2	centered triangles	triangles	critical
0.00818181818181818	0.462727272727272	96	100	99
0.0163636363636363	0.425454545454545	94	100	100
0.0245454545454545	0.388181818181818	88	100	100
0.0327272727272727	0.35090909090909	93	100	99
0.0409090909090909	0.313636363636363	94	100	100
0.049090909090909	0.276363636363636	96	100	99
0.0572727272727272	0.239090909090909	93	100	99
0.0654545454545454	0.201818181818181	95	100	100
0.0736363636363636	0.164545454545454	91	100	100
0.0818181818181818	0.127272727272727	95	100	81
0.0825619834710743	0.123884297520661	93	100	69
0.0833057851239669	0.120495867768595	91	100	67
0.0840495867768594	0.117107438016528	93	100	50
0.084793388429752	0.113719008264462	94	100	40
0.0855371900826446	0.110330578512396	93	100	29
0.0862809917355371	0.10694214876033	96	100	23
0.0870247933884297	0.103553719008264	92	100	18
0.0877685950413223	0.100165289256198	97	100	12
0.0885123966942148	0.0967768595041322	97	100	14
0.0892561983471074	0.0933884297520661	89	100	5
0.09	0.09	98	100	2

TABLE 1.   Goodness of fit testing results of the tetrahedron model.

ϵ_1	ϵ_2	centered triangles	triangles	critical
0.00818181818181818	0.608181818181818	96	100	99
0.0163636363636363	0.556363636363636	95	100	98
0.0245454545454545	0.504545454545454	95	100	95
0.0327272727272727	0.452727272727272	92	100	100
0.0409090909090909	0.40090909090909	90	100	98
0.049090909090909	0.349090909090909	96	100	100
0.0572727272727272	0.297272727272727	97	100	98
0.0654545454545454	0.245454545454545	93	100	93
0.06681818182	0.2368181819	91	100	86
0.06818181818	0.2281818182	94	100	89
0.069545454545	0.21954545455	95	100	81
0.07090909091	0.2109090909	96	100	72
0.072272727275	0.20227272725	92	100	65
0.0736363636363636	0.193636363636363	98	100	67
0.0751239669421487	0.184214876033057	96	100	63
0.0766115702479338	0.174793388429752	95	100	55
0.078099173553719	0.165371900826446	98	100	53
0.0795867768595041	0.15595041322314	97	100	45
0.0810743801652892	0.146528925619834	94	100	38
0.0825619834710743	0.137107438016528	96	100	28
0.0840495867768594	0.127685950413223	95	100	15
0.0855371900826446	0.118264462809917	96	100	12
0.0870247933884297	0.108842975206611	87	100	11
0.09	0.0994214876033057	97	100	9

TABLE 2.   Goodness of fit testing results of the triangle model.

ϵ_1	ϵ_2	centered triangles	triangles	critical
0	1.732050808	100	100	100
0.02344631004	1.673018508	99	100	99
0.04689262009	1.613986208	90	100	100
0.07033893013	1.554953908	97	100	99
0.09378524018	1.495921608	96	100	99
0.1172315502	1.436889308	97	100	99
0.1406778603	1.377857009	94	100	100
0.1641241703	1.318824709	96	100	100
0.1875704804	1.259792409	99	100	100
0.2110167904	1.200760109	91	100	100
0.2344631004	1.141727809	95	100	99
0.2579094105	1.082695509	77	100	97
0.2813557205	1.023663209	43	100	98
0.2857519	1.01259465	39	100	99
0.29014809	1.0015261	26	100	97
0.29454427	0.99045754	17	100	93
0.29894045	0.97938898	12	100	88
0.30333664	0.96832043	5	100	75
0.3048020306	0.9646309096	0	100	73
0.30773282	0.95725187	1	100	81
0.312129	0.94618332	0	98	62
0.31652519	0.93511476	0	96	51
0.32092137	0.9240462	0	90	33
0.32531755	0.91297765	0	75	29
0.3282483406	0.9055986098	3	74	12
0.32971373	0.90190909	0	53	9
0.33410992	0.89084053	0	33	4
0.3385061	0.87977198	0	16	3
0.34290228	0.86870342	0	4	1
0.34729847	0.85763487	0	3	1
0.3516946507	0.84656631	0	3	0
0.3751409607	0.7875340101	0	0	0
0.3985872707	0.7285017103	0	0	0
0.4220335808	0.6694694104	0	0	0
0.4454798908	0.6104371106	0	0	0
0.4689262009	0.5514048108	0	0	0
0.4923725109	0.4923725109	0	0	0

TABLE 3. Goodness of fit testing results of the edge model.