Random Simplicial Complexes and Stein's Method

Tadas Temčinas

Keble College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2023

Abstract

The work presented in this thesis is largely motivated by the pressing need for statistical foundations in topological data analysis (TDA). By studying random simplicial complexes, which serve as null models in TDA, the thesis falls within the framework of stochastic topology. Despite various success stories involving applications of algebraic topology, there is a lack of statistical tools that allow to rigorously evaluate topological data under uncertainty. Here we use Stein's method for multivariate normal distributions to prove distributional approximation results for random variables that arise in the study of random simplicial complexes. We also apply the results to statistical inference and hypothesis testing using simplicial complex data. The main body of the work consists of two articles: one accepted journal publication, and one preprint, which has been submitted for publication.

The first main result of the thesis is an abstract multivariate central limit theorem with explicit bounds, established via Stein's method, for sums of locally dependent random variables, which provides the greatest generality for this type of dependence structure so far. The result is then applied to prove multivariate central limit theorems for variables that arise in the study of the clique complexes of Erdős–Rényi random graphs. The abstract central limit theorem applies to multivariate sums of random variables which contain different order of summands in each component, allowing it to be applied in the context of the random clique complexes.

The subsequent chapter extends and generalises the probabilistic results of the previous chapter and applies them to statistical inference and hypothesis testing. Here we work with a more general random simplicial complex model: the multi-parameter random complex $\mathbf{X}(n, \mathbf{p})$. The statistical applications include properties of the maximum likelihood estimator of $\mathbf{X}(n, p)$ given a single observed simplicial complex as well as using the central limit theorems to devise goodness-of-fit tests.

This thesis not only contributes to the Stein's method literature but also offers better understanding of multivariate statistics of random simplicial complexes from a probabilistic point of view as well as shows how such probabilistic understanding leads to useful statistical tools in practice.

Acknowledgements

First and foremost, I would like to thank the UK taxpayers, whose contributions have funded my research over the past four years via an EPSRC scholarship.

Secondly, I would like to thank everyone with whom I have had the pleasure to discuss mathematics in greater detail over the course of my degree. These are my supervisors Gesine Reinert and Vidit Nanda, who patiently guided me through this adventure and diligently held weekly meetings full of friendly and mathematically exciting discussions. Matt Kahle, whom I have been privileged to work with and learn from. Uzu Lim and Henry Adams, who have taught me a great deal about point clouds on circles. Wenkai Xu and Greg Terlov with whom I have had many caffeinated and passionate meetings chipping away at probability proofs. Christina Goldschmidt, Jon Keating, and Heather Harrington, who have not only examined my transfer and confirmation of status but also provided useful feedback and guidance. And many more people whom I have met at conferences, seminars, workshops, and reading groups at Oxford and beyond, who have kept me motivated and informed.

Finally, I would like to express my gratitude to the people who took care of me, encouraged me, and helped me stay sane while I was worried about the proofs. This is my fiancée Michelle as well as my family Sergejus, Violeta, Kotryna, and Dan. Also, my international group of friends in the UK, Lithuania, Switzerland, and Germany, whom I have too many to list here but you know who you are. Finally, the dancesport and Latin social dancing communities at Oxford, who kept my feet moving when I needed it so much.

There is no doubt that without these people - and many more whom I have not been able to mention - neither this work nor I would be in their current state.

Contents

1 Introduction

The starting point of topological data analysis (TDA) is a dataset that is (or can be morphed into) a topological space, most commonly, a simplicial complex. The standard way to study data arising from some random process employs probabilistic and statistical techniques, which provide a rigorous way to quantify and study observations under uncertainty. Often, probabilistic models are employed in order to make sense of the data: they serve as null models in the framework of hypothesis testing; they also can be used to make predictions about future observations. Motivated by bridging the gap between TDA on the one side, and probability as well as statistics on the other, this work studies probabilistic and statistical aspects of random simplicial complexes.

The study of algebraic invariants of random simplicial complexes falls under the notion of *stochastic topology*. It is an active area of research, the state of which is reviewed in the two surveys [3, 4]. Since classical methods of TDA such as persistent homology [19] and the Euler characteristic curve or transform [27] are built on algebraic invariants of simplicial complexes, results in stochastic topology are crucial in order to quantitatively interpret the output of the TDA algorithms. Proofs in stochastic topology often start with expressing an invariant of interest as a function of combinatorial variables or sandwiching an invariant between such functions. The combinatorial variables are usually different types of simplices. For example, the Euler characteristic can be expressed as an alternating sum of the number of simplices of different dimensions; the k-th Betti number over any field is upper bounded by the number of critical k-simplices with respect to any discrete Morse function. As we will see in later examples, the variables counting simplices of different types can often be analysed probabilistically and as a result some properties of the invariants of interest can be deduced.

Chapters 2 and 3 contribute to this field by proving multivariate central limit theorems (CLT) for counts of three types of simplices in random simplicial complex models. We study models related the Erdős–Rényi random graph $\mathbf{G}(n, p)$, where out of the n vertices any two are connected independently with probability p. The clique complex of the $\mathbf{G}(n, p)$ random graph is called the $\mathbf{X}(n, p)$ random simplicial complex, which sometimes is also called the random clique complex or the random flag complex, is the central object of study in Chapter 2. In Chapter 3 we consider a more general model $\mathbf{X}(n, \mathbf{p})$. Here $\mathbf{p} = (p_1, p_2, \dots, p_{n-1}) \in [0, 1]^n$ is a vector of probabilities; p_1 is the probability of connecting any two vertices by an edge, p_2 is the probability of any triangle to form a 2-simplex, and so on. This model gives rise to a conditionally independent probabilistic structure. Many other models such as the $\mathbf{G}(n,p)$, $\mathbf{X}(n,p)$ as well as the Linial-Meshulam random complexes fall under the framework of $\mathbf{X}(n,\mathbf{p})$ model as special cases.

Understanding the distributions of simplex counts of random simplicial complexes can also be directly utilised in statistical applications. Using the distribution of graph statistics such as small subgraph counts or the degree distribution to devise goodness of fit tests for different random graph models is a common practice in network analysis [14, 20]. Such tests typically have higher power and are computationally faster compared to Monte-Carlo tests that require extensive and computationally expensive sampling from the graph statistic distribution. However, for tests based on the distribution of the chosen statistics to be applied in the context of TDA and random simplicial complexes, the distributions under the null hypothesis have to be understood, and currently not enough such results are available. To help fill this gap, the paper in Chapter 3 uses distributional approximation results to devise a χ^2 goodness of fit test for the $\mathbf{X}(n,\mathbf{p})$ random simplicial complex model. To make the test useful in practice, the maximum likelihood estimator (MLE) of the $\mathbf{X}(n,\mathbf{p})$ model is also studied.

1.1 From algebraic invariants to subcomplex counts

Most commonly used tools in TDA are rooted in algebraic topology, which uses algebraic invariants to study topological spaces. Naturally, a goal of stochastic topology is to understand the distribution of such algebraic invariants under different random simplicial complex models. Studying these invariants directly is often unachievable from a probabilistic point of view. As a result, a common way to study them uses the connection between the algebraic invariants and combinatorial properties of simplicial complexes, an example of which are the so-called subcomplex counts. Here we review a couple of classic examples, which motivate the need for probabilistic understanding of subcomplex counts.

From the combinatorial point of view, one of the simplest algebraic invariants is the Euler characteristic. Given a finite simplicial complex $\mathcal{L}$ of dimension d, we write $e_k(\mathcal{L})$ for the number of k-simplices in $\mathcal{L}$. The Euler characteristic is defined as $\chi(\mathcal{L}) := \sum_{k=0}^{d}(-1)^k e_k(\mathcal{L})$. It is a linear combination of the number of simplices of different dimensions, which are a special case of more general subcomplex counts. Understanding the multivariate distribution of the number of simplices would directly translate into understanding the distribution of the Euler characteristic.

Combinatorially more complicated but potentially more useful algebraic invariants are the Betti numbers over a chosen field $\mathbb{F}$. We start with a simplicial complex $\mathcal{L}$ on a vertex set $[n] \coloneqq \{1, 2, \ldots, n\}$. Defining the Betti numbers is not as straightforward as defining the Euler characteristic. Readers interested in the more general homology theory are referred to [8, Chapter 2]. Intuitively, we can interpret the Betti numbers of dimension k as counting the number of k-dimensional holes within a simplicial complex. For example, the zeroth Betti number is just the number of connected components. Now we proceed with a more formal definition. Given a k-simplex $s \in \mathcal{L}$, we define s_{-i} as the subset of s with the $(i+1)$-st smallest vertex removed. The ordering of the vertices here is with respect to the natural ordering on $[n]$. So, for example, $s_{-0} = s \setminus \{\min(s)\}$ and $s_{-k} = s \setminus \{\max(s)\}$. The ordering of vertices also induces the lexicographical ordering on the simplices. Given an order number i and a dimension k, we write $s^{(i,k)}$ for the i-th k-simplex with respect to this ordering. The k-th boundary matrix ∂_k is

$$[\partial_k]_{i,j} \coloneqq \begin{cases} (-1)^m, & \text{if } s^{(i,k-1)} = s^{(j,k)}_{-m} \\ 0, & \text{otherwise} \end{cases}.$$

The algebraic expression for the boundary matrix can be interpreted over any field $\mathbb{F}$, and the k-th Betti number over a field $\mathbb{F}$ is defined as $\beta_k(\mathcal{L}) \coloneqq e_k(\mathcal{L}) - \operatorname{rank}_{\mathbb{F}}(\partial_k) - \operatorname{rank}_{\mathbb{F}}(\partial_{k+1})$. Standard theorems from homology theory show that the Betti numbers do not depend on the ordering on the vertices but that they can depend on the underlying field $\mathbb{F}$. In the case of random simplicial complexes, ∂_k is a highly structured sparse random matrix with dependent entries and as such, understanding the distribution of its rank is far from trivial. Even though random matrix theory studies the distributions of matrices with random entries, many of the random matrix theory results concentrate around understanding random matrices with a high level of independence of the entries that is not satisfied by ∂_k [24].

Luckily, there are some combinatorial relationships between β_k and different subcomplex counts that allow the study of the distribution of β_k. For example, using the trivial bounds on the ranks of the boundary matrices, we can see that

$$e_k(\mathcal{L}) - e_{k+1}(\mathcal{L}) - e_{k-1}(\mathcal{L}) \leq \beta_k(\mathcal{L}) \leq e_k(\mathcal{L}).$$

The bounds might seem not that useful but some combinatorial random simplicial complex models, such as the clique complex of the Erdős–Rényi random graph $\mathbf{X}(n,p)$ or the multi-parameter model $\mathbf{X}(n,\mathbf{p})$, have parameter regimes where $e_k(\mathcal{L})$ is much larger than the number of simplices in other dimensions. For example, this fact led to a univariate CLT for the Betti numbers over $\mathbb{Q}$ in the $\mathbf{X}(n, p)$ model [12, 13]. However, as noted in the erratum to the original paper [13], simply proving CLTs for the upper and lower bounds to the Betti numbers is not enough to deduce a CLT for the Betti numbers themselves.

Apart from the simplex counts, the number of subcomplexes that witness non-trivial Betti numbers also plays an important role in the study of random simplicial complexes. For example, in the $\mathbf{X}(n, p)$ model, understanding the number of octahedral spheres allows one to prove thresholds of emergence of Betti numbers and sometimes even their limit theorems. An octahedral k-sphere is a clique complex of a graph on $2k + 2$ vertices, where each of the vertices belongs to one of the two groups of size $k + 1$. So we can write the vertex set as $\{v_1, \dots v_{k+1}, w_1, \dots w_{k+1}\}$. The edge set consists of all possible edges except the pairs $\{v_i, w_i\}$ for all $i \in [k+1]$. Hence this graph has $\binom{2k+2}{2} - (k+1) = 2k^2 + 2k$ edges. The combinatorial fact that if a simplicial complex has an octahedral k-sphere with at least one isolated face as a subcomplex, then the k-th Betti number must be non-zero over any field $\mathbb{F}$ is used in [10] to find parameter regimes in which Betti number are non-zero in the $\mathbf{X}(n, p)$ model. More recently, a Poisson limit theorem for the number of octahedral k-spheres has been translated into a Poisson limit theorem for Betti numbers in the $\mathbf{X}(n, p)$ model [18]. In both applications, understanding the number of octahedral k-spheres, which is a special case of subcomplex counts, is crucial.

Such connections to algebraic invariants provide a probabilistic motivation for studying subcomplex counts in random simplicial complexes. In Chapter 2 we prove a multivariate CLT for simplex counts and simplex counts in links in the $\mathbf{X}(n, p)$ model via Stein's method. We bound the non-asymptotic error of the normal approximation as measured by both smooth and convex set indicator test functions. A favourable feature of the result is that the bounds are calculated explicitly in terms of the parameters of the model and the parameters of statistic in question (as opposed only the dependence on n). This allows one to explore the parameter regimes where some of the parameters have non-trivial dependence on n, which in our case is the number of vertices.

In Chapter 3 we prove a similar multivariate central limit theorem for subcomplex counts in the $\mathbf{X}(n, \mathbf{p})$ model, which could pave a way towards a more intricate understanding of multivariate distribution of Betti numbers as well as the Euler characteristic in this model and many of its important special cases.

The relation between popular algebraic invariants and subcomplex counts as well as the importance of subgraph counts in random graph models suggests that these are important statistics that would characterise a simplicial complex model well and as such could be useful in statistical applications. Therefore goodness-of-fit tests based on subcomplex counts ought to be useful in practice. As it is discussed in Chapter 3, the appropriately centered and scaled subcomplex counts are asymptotically perfectly correlated in the $\mathbf{X}(n, \mathbf{p})$ model, despite the independent randomness in each dimension. As a result, all of them asymptotically capture the same aspect of the model. Also, they correlate with simplex counts and hollow simplex counts, the ratio of which is the MLE for the parameter $\mathbf{p}$. Accounting for the asymptotically perfect correlation between statistics used for parameter estimation and statistics used for goodness-of-fit tests can be very hard. Therefore, we turn to discussing a different statistic, based on discrete Morse theory, which lends itself to goodness-of-fit testing.

The remainder of this introduction discusses discrete Morse theory in the context of random simplicial complexes as well as Stein's method - two topics that are central in this work.

1.2 Discrete Morse theory for random simplicial complexes

Discrete Morse theory, as pioneered by Forman [6], has many applications in both the pure study of simplicial complexes (and more general so-called CW complexes) as well as TDA. The goal of discrete Morse theory is to compress topological information of a simplicial complex so that the relevant algebraic invariants can be exactly calculated faster from the compressed version of the complex. Originally it was introduced to study CW complexes but the theory has later been extended to filtered simplicial complexes [16] and as a result used to speed up persistent homology computations in popular software packages [2].

A concept central to discrete Morse theory is that of an acyclic partial matching on a simplicial complex. Intuitively, an acyclic partial matching pairs up simplices whose dimensions differ by one in such a way that the topological information relevant to the algebraic invariants is encoded in the unmatched - critical - simplices and the paths between them. For a rigorous introduction to the topic we refer the reader to the classical paper by Forman [6]; topological preliminaries in Chapters 2 and 3 also contain the relevant definitions.

There are a few reasons why studying discrete Morse theory on random simplicial complexes is a useful endeavour. Because of the usefulness of acyclic partial matchings in reducing homology computations, finding matchings that have few critical simplices on large classes of simplicial complexes is desirable. However, for a given simplicial complex, finding an acyclic partial matching with the smallest number of critical simplices is known to be NP-hard and so computationally intractable for large complexes [9]. Therefore, instead of optimising a matching on each simplicial complex, one could compare a few potential matchings on random simplicial complex models. If a matching has very few critical simplices on a random model, it is reasonable to believe it would work relatively well on real datasets and so it could be promising to test this hypothesis out empirically. Such a matching then has potential to turn into a useful heuristic in TDA. Therefore, understanding how different acyclic partial matchings behave on random simplicial complexes can lead to insights which are useful in applications.

From a stochastic topology point of view, the behaviour of acyclic partial matchings on random simplicial complexes have lead to theorems about their algebraic invariants. For example, [10, Theorem 3.9] shows that the expected k-th Betti number is much smaller than the expected number of k-simplices in the $\mathbf{X}(n,p)$ model. Moreover, in [11, Theorem 5.1] the asymptotic behaviour of the expected k-th Betti number of a Vietoris-Rips complex formed from a binomial process on a smooth convex body in $\mathbb{R}^d$ is analysed via discrete Morse theory. In both cases, a lexicographical acyclic partial matching has been used. This matching is often not only a useful tool in the proofs of theorems in stochastic topology but it is also used as a popular heuristic in software packages to reduce computational complexity of homology calculations [2]. In Chapter 2 we contribute towards understanding a lexicographical acyclic partial matching in the $\mathbf{X}(n,p)$ model by proving a multivariate CLT for the number of critical simplices under the matching. In Chapter 3 the multivariate CLT for the number of critical simplices under the lexicographical matching is generalised to the multi-parameter $\mathbf{X}(n,\mathbf{p})$ model.

One aspect of discrete Morse theory for random simplicial complexes that has been under-explored is its use in statistical applications. The aforementioned results in [10, 11], show that the behaviour of a lexicographical matching on a random geometric complex like Vietoris-Rips and a combinatorial one like the $\mathbf{X}(n,p)$ is vastly different. In particular, the number of critical simplices under a lexicographical matching is much smaller in the geometric case. After noticing this difference one might ask if it can be used to differentiate the two models in the context of goodness-of-fit tests. In Chapter 3 we show that we can use the aforementioned CLTs for the number of critical simplices under a lexicographical matching to distinguish between a purely combinatorial model like $\mathbf{X}(n,\mathbf{p})$ and models that are either purely geometric or partially combinatorial and partially geometric.

1.3 Stein's method for the multivariate normal distribution

Stein's method is the main probabilistic tool we use in order to establish the CLTs in the context of random simplicial complexes. The method, pioneered by Charles Stein [22, 23], allows explicitly bounding the error - as measured by a class of test functions of choice - when we approximate an unknown distribution with a well-known target distribution. In our work, the target is always a multivariate normal distribution and the unknown distribution is that of a multivariate statistic of a random simplicial complex. Here we briefly introduce the basics of Stein's method for the multivariate normal distribution. A more extensive treatment of the subject can be found in [5, Chapter 12].

We denote the multivariate normal distribution with a mean vector μ and a positive semi-definite covariance matrix Σ by $\mathrm{MVN}(\mu, \Sigma)$. Also, we write ∇ for the gradient operator. The starting point of Stein's method is the following fact (see [15, Lemma 1]). Given a random vector $Z \in \mathbb{R}^d$ the following are equivalent:

1. $Z \sim \mathrm{MVN}(0, \Sigma)$ for some non-negative definite Σ.
2. $\mathbb{E}\left\{\nabla^T \Sigma \nabla f(Z) - Z^T \nabla f(Z)\right\} = 0$ for every twice continuously differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ for which the expectation exists.

The main idea behind Stein's method is to relax this equivalence and make the following heuristic rigorous: if $\mathbb{E}\left\{\nabla^T \Sigma \nabla f(Z) - Z^T \nabla f(Z)\right\} \approx 0$ for every suitable $f : \mathbb{R}^d \to \mathbb{R}$, then Z can be well approximated by a multivariate normal distribution with mean zero and covariance Σ. The method allows us to bound the non-asymptotic approximation error, which can be particularly useful in applications, when the size of a system is large but finite.

In order to be precise about the approximation error one needs to choose how the error is measured. Stein's method lends itself very well to so-called integral probability metrics [17, 21]. Given a random variable W having a distribution $\mathcal{L}(W)$, a target $Z \sim \mathrm{MVN}(0, \Sigma)$, and a class of test functions $\mathcal{H}$ the corresponding quantity between the law of W and $\mathrm{MVN}(0, \Sigma)$ is defined as

$$d_{\mathcal{H}}(\mathcal{L}(W), \mathrm{MVN}(0, \Sigma)) \coloneqq \sup_{h \in \mathcal{H}} \left|\mathbb{E}\left\{h(W)\right\} - \mathbb{E}\left\{h(Z)\right\}\right|.$$

For carefully chosen test functions the corresponding quantities can be proven to be distances. Many metrics used in literature can be fit into the framework of integral probability metrics. For example, by taking $\mathcal{H}$ to be the class of indicators all measurable sets, the total variation distance is recovered. By taking $\mathcal{H}$ to be the class of all Lipschitz functions, the Wasserstein-1 distance is obtained. By taking $\mathcal{H}$ to be the class of indicators of products of lower half-lines, we get the multivariate Kolmogorov distance. In this work we take $\mathcal{H}$ to either be the class of three times partially differentiable functions with the third partial derivatives being Lipschitz and bounded, or the class of the indicator functions of convex sets in $\mathbb{R}^d$. Both test classes characterise convergence in distribution, in the sense that if $\mathcal{L}(X_n)$ and $\mathcal{L}(Z)$ denote the laws of random variables X_n and Z respectively, then $d_{\mathcal{H}}(\mathcal{L}(X_n), \mathcal{L}(Z)) \to 0$ implies that the sequence of random variables (X_n) converges in distribution to the target Z as long as $\mathcal{L}(Z)$ is a continuous distribution with bounded density. One can see that by realising that the multivariate Kolmogorov distance characterises convergence in distribution [28, Example 1.3.5] and that the Kolmogorov distance can be upper bounded by a function of either of the distances we use in this thesis [7, Proposition 2.4]. However, the latter test function class proves to be more useful in practice as confidence regions of interest in statistics tend to be convex sets.

Once we have chosen a class of test functions $\mathcal{H}$, the usual way to proceed is to solve the so-called Stein equation. Given $h \in \mathcal{H}$ we want to find a function $f : \mathbb{R}^d \to \mathbb{R}$ such that for all $x \in \mathbb{R}^d$ the following equation holds

$$h(x) - \mathbb{E}\{h(Z)\} = \nabla^T \Sigma \nabla f(x) - x^T \nabla f(x).$$

This gives rise to the so-called Stein class $\mathcal{F}(\mathcal{H})$, which contains all the solutions f that correspond to some test function $h \in \mathcal{H}$. As a result, Stein's method allows us to prove the following equality:

$$d_{\mathcal{H}}(\mathcal{L}(W), \mathrm{MVN}(0, \Sigma)) = \sup_{f \in \mathcal{F}(\mathcal{H})} \left| \mathbb{E}\left\{\nabla^T \Sigma \nabla f(W) - W^T \nabla f(W)\right\}\right|.$$

Usefulness of this equality might not seem obvious but it should be noted that the expectation on the right hand side is only taken with respect to the law of W, making it a much easier expression to bound. Indeed, significant amount of research in Stein's method deals with bounding the right hand side of the equation under different assumptions on the structure of W and as a result, this is a well studied problem. In Chapter 2 we contribute to this arsenal of results by providing a way to bound the right hand side of this equality in the case when W is a vector of dissociated sums.

One of the strengths of Stein's method in the setting of a sum of random variables is that the method allows one to deal with dependent summands as long as the dependence is local, as formalised by the notion of a dissociated sum. Our Stein's method result in Chapter 2 is a multivariate generalisation of the univariate CLT for dissociated sums [1]. Given a sum of random variables $W = \sum_{i=1}^{n} X_i$ we call it a dissociated sum iff for every $i \in [n]$ there is a set K_i such that $W - \sum_{j \in K_i} X_j$ is independent of X_i and for every $i \neq j \in [n]$ the variable $W - \sum_{k \in K_i \cup K_j} X_k$ is independent of the pair (X_i, X_j). The following is a CLT for dissociated sums.

Theorem 1.1 ([1])**.** *Assume that $W = \sum_{i=1}^{n} X_i$ is a mean zero and unit variance dissociated sum and assume for each i the third absolute moment of X_i is finite. Let $Z \sim \mathcal{N}(0,1)$. Then for every bounded function $h : \mathbb{R} \to \mathbb{R}$ with bounded first derivatives we have:*

$$|\mathbb{E}\{h(W)\} - \mathbb{E}\{h(Z)\}| \leq 4 \sup_{x \in \mathbb{R}} |h'(x)| \left(\sum_{i \in [n]} \sum_{j,k \in K_i} \mathbb{E}\{|X_i X_j X_k|\} + \mathbb{E}\{|X_i X_j|\} \mathbb{E}\{|X_k|\} \right).$$

To better understand such results, we can see how this theorem applies in the classic CLT setting where we have $W = \sum_{i=1}^{n} \frac{X_i}{\sqrt{n}\sigma}$ and the collection $(X_i)_{i \in [n]}$ is i.i.d. with $\mathbb{E}\{X_i\} = 0$, $\mathrm{Var}(X_i) = \sigma^2$ as well as $\mathbb{E}\{|X_1^3|\} < \infty$ for all $i \in [n]$. In this case we see that W is a dissociated sum by taking $K_i = \{i\}$ for all $i \in [n]$. Applying the theorem and Cauchy-Schwarz inequality gives for any bounded function $h : \mathbb{R} \to \mathbb{R}$ with bounded first derivatives

$$|\mathbb{E}\{h(W)\} - \mathbb{E}\{h(Z)\}| \leq 4 \sup_{x \in \mathbb{R}} |h'(x)| \frac{1}{\sqrt{n}\sigma^{3/2}} \left(\mathbb{E}\{|X_1^3|\} + \sigma^2 \right).$$

As expected, the bound is of the order $n^{-\frac{1}{2}}$ assuming the second and third moment does not depend on n. This dependence on n is optimal in the classic CLT setting. Also, unsurprisingly, the bound depends on the second and third moments of the underlying distribution.

1.4 Organisation

The main body of this work is divided into two self-contained parts:

- Chapter 2 contains an accepted paper Temčinas T, Nanda V, Reinert G. "Multivariate Central Limit Theorems for Random Clique Complexes" [26].

- Chapter 3 contains a preprint by the same authors titled “Goodness-of-fit via Count Statistics in Dense Random Simplicial Complexes” [25].

Finally, in Chapter 4 we conclude the thesis and present ideas for future work.

References

[1] Andrew D Barbour, Michal Karoński, and Andrzej Ruciński. “A central limit theorem for decomposable random variables with applications to random graphs”. In: *Journal of Combinatorial Theory, Series B* 47.2 (1989), pp. 125–145.

[2] Ulrich Bauer. “Ripser: efficient computation of Vietoris–Rips persistence barcodes”. In: *Journal of Applied and Computational Topology* 5 (2021), pp. 391–423.

[3] Omer Bobrowski and Matthew Kahle. “Topology of random geometric complexes: a survey”. In: *Journal of Applied and Computational Topology* 1.3 (2018), pp. 331–364.

[4] Omer Bobrowski and Dmitri Krioukov. “Random simplicial complexes: models and phenomena”. In: *Higher-Order Systems*. Springer, 2022, pp. 59–96.

[5] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein’s Method*. Springer, 2011.

[6] Robin Forman. “A user’s guide to discrete Morse theory”. In: *Séminaire Lotharingien de Combinatoire* 48 (2002), B48c.

[7] Robert E Gaunt and Siqi Li. “Bounding Kolmogorov distances through Wasserstein and related integral probability metrics”. In: *Journal of Mathematical Analysis and Applications* 522.1 (2023), p. 126985.

[8] Allen Hatcher. *Algebraic Topology*. Cambridge, UK: Cambridge University Press, 2002.

[9] M Joswig and M Pfetsch. “Computing optimal discrete Morse functions”. In: *SIAM J. Discrete Math.* 20.1 (2004), pp. 11–25.

[10] Matthew Kahle. “Topology of random clique complexes”. In: *Discrete mathematics* 309.6 (2009), pp. 1658–1671.

[11] Matthew Kahle. “Random geometric complexes”. In: *Discrete & Computational Geometry* 45.3 (2011), pp. 553–573.

[12] Matthew Kahle and Elizabeth Meckes. “Limit theorems for Betti numbers of random simplicial complexes”. In: *Homology, Homotopy and Applications* 15.1 (2013), pp. 343–374.

[13] Matthew Kahle and Elizabeth Meckes. “Erratum: Limit theorems for Betti numbers of random simplicial complexes”. In: *arXiv preprint arXiv:1501.03759* (2015).

[14] Jing Lei. “A goodness-of-fit test for stochastic block models”. In: *The Annals of Statistics* 44.1 (2016), pp. 401–424.

[15] Elizabeth Meckes. “On Stein’s method for multivariate normal approximation”. In: *High dimensional probability V: the Luminy volume*. Institute of Mathematical Statistics, 2009, pp. 153–178.

[16] Konstantin Mischaikow and Vidit Nanda. “Morse theory for filtrations and efficient computation of persistent homology”. In: *Discrete and computational geometry* 50.2 (2013), pp. 330–353.

[17] Alfred Müller. “Integral probability metrics and their generating classes of functions”. In: *Advances in applied probability* 29.2 (1997), pp. 429–443.

[18] Andrew Newman. “One-sided sharp thresholds for homology of random flag complexes”. In: *arXiv preprint arXiv:2108.04299* (2021).

[19] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. “A roadmap for the computation of persistent homology”. In: *EPJ Data Science* 6 (2017), pp. 1–38.

[20] Sarah Ouadah, Stéphane Robin, and Pierre Latouche. “Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases”. In: *Scandinavian Journal of Statistics* 47.1 (2020), pp. 156–181.

[21] Svetlozar Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, 1991.

[22] Charles Stein. “A bound for the error in the normal approximation to the distribution of a sum of dependent random variables”. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. Vol. 6. University of California Press. 1972, pp. 583–603.

[23] Charles Stein. “Approximate computation of expectations”. In: IMS. 1986.

[24] Terence Tao. *Topics in random matrix theory*. Vol. 132. American Mathematical Soc., 2012.

[25] Tadas Temčinas, Vidit Nanda, and Gesine Reinert. "Goodness-of-fit via Count Statistics in Dense Random Simplicial Complexes". In: *arXiv preprint arXiv:2309.14017* (2023).

[26] Tadas Temčinas, Vidit Nanda, and Gesine Reinert. "Multivariate Central Limit Theorems for Random Clique Complexes". In: *Journal of Applied and Computational Topology* (2023+).

[27] Katharine Turner, Sayan Mukherjee, and Doug M Boyer. "Persistent homology transform for modeling shapes and surfaces". In: *Information and Inference: A Journal of the IMA* 3.4 (2014), pp. 310–344.

[28] A. W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes : with applications to statistics*. 1st ed. 1996. Springer, 1996.

2 Multivariate Central Limit Theorems for Random Clique Complexes

Multivariate Central Limit Theorems for Random Clique Complexes

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

[1]Department of Statistics, University of Oxford

[2]Mathematical Institute, University of Oxford

Abstract

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

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

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

Keywords: Stein's method, multivariate normal approximation, discrete Morse theory, random graphs, random simplicial complexes

MSC: 60F05 Central limit and other weak theorems, 60D05 Geometric probability, stochastic geometry, random sets, 05C80 Random graphs (graph-theoretic aspects).

1 Introduction

Methods from applied and computational algebraic topology have recently found substantial applications in the analysis of nonlinear and unstructured datasets [18, 8]. The modus operandi of topological data analysis is to first build a nested family of simplicial complexes around the elements of a dataset, and to then compute the associated persistent homology barcodes [12]. Of central interest, when testing hypotheses under this paradigm, is the question of what homology groups to expect when the input data are randomly generated. Significant efforts have therefore been devoted to answering this question for various models of noise, giving rise to the field of *stochastic topology* [23, 7, 22, 1, 11]. Our work here is a contribution to this area at the interface between probability theory and algebraic topology.

Distributional approximations provide a way of understanding random variables in cases where closed-form distributions cannot be easily obtained. This paper establishes the first multivariate normal approximations to three important counting problems in stochastic topology; as these approximations are based on Stein's method, explicit bounds on the approximation errors are provided. Our starting point is the ubiquitous graph model $\mathbf{G}(n,p)$; a graph G chosen from this model has as its vertex set $[n] = \{1, 2, \ldots, n\}$, and each of its possible $\binom{n}{2}$ edges is included independently with probability $p \in [0,1]$. A natural higher-order generalisation of $\mathbf{G}(n,p)$ is furnished by the random clique complex model $\mathbf{X}(n,p)$, whose constituent complexes $\mathcal{L}$ are constructed as follows. One first selects an underlying graph $G \sim \mathbf{G}(n,p)$, and then deterministically fills out all k-cliques in G with $(k-1)$-dimensional simplices for $k \geq 3$. Higher connectivity is measured by the Betti numbers $\beta_k(\mathcal{L})$, which are ranks of rational homology groups $\mathrm{H}_k(\mathcal{L};\mathbb{Q})$ — in particular, $\beta_0(\mathcal{L})$ equals the number of connected components of the underlying random graph G. In [24], Kahle proved the following far-reaching generalisation of the Erdős-Rényi connectivity result: for each $k \geq 1$ and $\epsilon > 0$,

1. if
$$p \geq \left[\left(\frac{k}{2} + 1 + \epsilon\right) \cdot \frac{\log(n)}{n}\right]^{1/(k+1)},$$
then $\beta_k(\mathcal{L}) = 0$ with high probability; and moreover,

2. if
$$\left[\frac{k+1+\epsilon}{n}\right]^{1/k} \leq p \leq \left[\left(\frac{k}{2} + 1 - \epsilon\right) \cdot \frac{\log(n)}{n}\right]^{1/(k+1)},$$
then $\beta_k(\mathcal{L}) \neq 0$ with high probability.

Unlike Kahle's result, we study $\mathbf{X}(n,p)$ in the regime where p is a constant. In that case, which is in contrast to the case discussed above where $\lim_{n\to\infty} p = 0$ and k is constant, we may have $\beta_k(\mathbf{X}(n,p)) \neq 0$ for multiple values of k that depend on n. Hence, studying the multivariate distribution of the Betti numbers is of interest in this regime. Since many results about Betti numbers in the univariate case are based on counting simplices, we hope that understanding the multivariate simplex counts will facilitate multivariate understanding of the Betti numbers. With this result in mind, we motivate and describe three random vectors pertaining to $\mathcal{L} \sim \mathbf{X}(n,p)$; the normal approximation of these three random vectors will be our focus in this paper. All three are denoted $T = (T_1, \ldots, T_d)$ for an integer $d > 0$. For the purpose of this introduction we add a superscript $(1), (2)$ or (3) to indicate the particular vector.

Random Vector 1: Critical Simplex Counts

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

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

Here Ch_k is a vector space whose dimension equals the number of k-simplices in $\mathcal{L}$, while $d_k : \mathrm{Ch}_k \to \mathrm{Ch}_{k-1}$ is an incidence matrix encoding which $(k-1)$-simplices lie in the boundary of a given k-simplex. These matrices satisfy the property that every successive composite $d_{k+1} \circ d_k$ equals zero, and $\beta_k(\mathcal{L})$ is the dimension of the quotient vector space $\ker d_k / \mathrm{img}\, d_{k+1}$. In order to calculate $\beta_k(\mathcal{L})$, one is required to put the matrices $\{d_k : \mathrm{Ch}_k \to \mathrm{Ch}_{k-1}\}$ in reduced echelon form, which is a straightforward task in principle. Unfortunately, Gaussian elimination on an $m \times m$ matrix incurs an $O(m^3)$ cost, which becomes prohibitive when facing simplicial complexes built around large data sets [36]. The standard remedy is to construct a much smaller chain complex which has the same homology groups, and by far the most fruitful mechanism for achieving such homology-preserving reductions is **discrete Morse theory** [15, 34, 19, 29].

The key structure here is that of an *acyclic partial matching*, which pairs together certain adjacent simplices of $\mathcal{L}$; and the homology groups of $\mathcal{L}$ may be recovered from a chain complex whose vector spaces are spanned by unpaired, or *critical*, simplices. One naturally seeks an optimal acyclic partial matching on $\mathcal{L}$ which admits the fewest possible critical simplices. Unfortunately, the optimal matching problem is computationally intractable to solve [21] even approximately [5] for large $\mathcal{L}$. Our first random vector $T^{(1)}$ is obtained by letting $T_k^{(1)}$ equal the number of critical k-simplices for a specific type of acyclic partial matching on $\mathcal{L}$, called the *lexicographical* matching. Knowledge of this random vector serves to simultaneously quantify the benefit of using discrete Morse theoretic reductions on random simplicial complexes and to provide a robust null model by which to measure their efficacy on general (i.e., not necessarily random) simplicial complexes. This is the first time the asymptotic distribution of this random vector is studied.

Random Vector 2: Link Simplex Counts

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

There are compelling reasons to better understand the combinatorics and topology of such links from a probabilistic viewpoint. For instance, the fact that the link of a k-simplex in a triangulated n-manifold is always a triangulated sphere of dimension $(n - k - 1)$ has been exploited to produce canonical stratifications of simplicial complexes into homology manifolds [2, 35]. Knowledge of simplex counts (and hence, Betti numbers) of links would therefore form an essential first step in any systematic study involving canonical stratifications of random clique complexes.

Random Vector 3: Total Simplex Counts

The strategy employed in Kahle's proof of the second assertion above involves first checking that the expected number of k-simplices in $\mathcal{L} \sim \mathbf{X}(n, p)$ is much larger than the expected number of simplices of dimensions $k \pm 1$ whenever p lies in the range indicated by (2). Therefore, one may combine the Morse inequalities with the linearity of expectation in order to guarantee that the expected $\beta_k(\mathcal{L})$ is nonzero — see [24, Section 4] for details. To facilitate more refined analysis and estimates of this sort, the third random vector $T^{(3)}$ we study in this paper is obtained by letting $T_k^{(3)}$ equal the total number of k-dimensional simplices in $\mathcal{L}$.

Since $T_k^{(3)}$ is precisely the number of $(k+1)$-cliques in $G \sim \mathbf{G}(n,p)$, this random vector falls within the purview of *generalised U-statistics*. We extend results from [20] to show not only distributional convergence asymptotically but a stronger result, detailing explicit non-asymptotic bounds on the approximation. Several interesting problems can be seen as special cases — these include classical U-statistics [30, 28], monochromatic subgraph counts of inhomogeneous random graphs with independent random vertex colours, and the number of overlapping patterns in a sequence of independent Bernoulli trials. To the best of our knowledge, this is the first multivariate normal approximation result with explicit bounds where the sizes of the subgraphs are permitted to increase with n.

Main Results

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

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

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

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

1. *There is a constant $B_{1.1.1} > 0$ independent of n and a natural number $N_{1.1.1}$ such that for any*

$n \geq N_{1.1.1}$ we have

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

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

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

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

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

and

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

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

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

and

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

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

1. The expected number of critical k-simplices is one order of n smaller that the expected total number of k-simplices; see Lemma 4.2.

2. The variance of the number of critical k-simplices is at least of the order n^{2k}, as shown in Lemma 4.3. An upper bound of the same order can be proved similarly. The variance of the total number of k-simplices is also of the same order.

3. Knowing the expected value and the variance one can prove concentration results using different concentration inequalities, for example, Chebyshev's inequality. This would show that not only

the expected value of the number of critical simplices is smaller compared to all simplices but also that large deviations from the mean are unlikely, hence implying that the substantial improvement of one order of n is not only expected but also likely.

4. For counting critical simplices to high accuracy in probability, it is not necessary to check every simplex. Certain simplices have a very small chance of being critical, and can be safely ignored. The probability of this omission causing an error is vanishingly small asymptotically; see Proposition 4.4.

The main ingredient in establishing such results is often an abstract approximation theorem that can be applied to the random variables of interest. While there is no shortage of multivariate normal approximation theorems [14, 39, 33, 10], the existing ones are not sufficiently fine-grained for proving multivariate normal approximations to the random vectors studied here. We therefore return to the pioneering work of Barbour, Karoński, and Ruciński [4], who proved a univariate central limit theorem (CLT) for a decomposable sum of random variables using Stein's method, treating the case of dissociated sums as a special case. Our approximation result (Theorem 3.2) forms an extension of their ideas to the multivariate setting.

Related Work

There are different versions of distributional approximation results for subgraph counts in $\mathbf{G}(n,p)$ that can be interpreted as simplex counts in $\mathbf{X}(n,p)$. For example, a multivariate central limit theorem for *centered* subgraph counts in the more general setting of a random graph associated to a graphon can be found in [27]. That proof is based on Stein's method via a Stein coupling. Translating this result for uncentered subgraph counts would yield an approximation by a function of a multivariate normal. In [40], an exchangeable pair coupling led to [40, Proposition 2] which can be specialised to joint counts of edges and triangles; our approximation significantly generalises this result beyond the case where $k \in \{1,2\}$. Several univariate normal approximation theorems for subgraph counts are available; recent developments in this area include [38], which uses Malliavin calculus together with Stein's method, and [13], which uses the Stein-Tikhomirov method. Stein's method is used in [25] to show a CLT for the Betti numbers of $\mathbf{X}(n,p)$ in a sparse regime; in [37] limit theorems for Betti numbers and Euler characteristic are proven in a dynamical random simplicial complex model, also using Stein's method.

Theorem 3.2 is not the first generalisation of the results in [4] to a multivariate setting, see for example [14, 39]. The key advantage of our approach is that it allows for bounds which are non-uniform in each component of the vector W. This is useful when, for example, the number of summands in each component are of different order or when the sizes of dependency neighbourhoods in each component are of different order. The applications considered here are precisely of this type, where the non-uniformity of the bounds is crucial. Moreover, we do not require the covariance matrix Σ to be invertible, and can therefore accommodate degenerate multivariate normal distributions.

Organisation

In Section 2 we recall concepts from the theory of simplicial complexes, which we later use. In Section 3 we state the theorems that serve as main tools in proving the CLTs. In order to maintain focus on the main results, we defer the proofs of this Section until the end of the paper. In Section 4 we prove an approximation theorem for critical simplex counts of lexicographical matchings. Two technical computations required in this Section have been consigned to the Appendix. In Section 5 we prove an approximation theorem for count variables of simplices that are in the link of a fixed simplex. In Section 6 we study simplex counts in the random clique complex and prove a CLT for this random variable. This CLT is a corollary of a multivariate normal approximation of generalised U-statistics, which might be of independent interest. Finally, in Section 7 we prove our main tools: the abstract approximation theorem (Theorem 3.2) as well as the approximation theorem for U-statistics (Theorem 3.9). We first use smooth test functions and then extend the results to convex set indicators using a smoothing technique from [16].

Notation

Throughout this paper we use the following notation. Given positive integers n, m we write $[m, n]$ for the set $\{m, m+1, \dots, n\}$ 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 $C_k = \{ t \in \mathcal{P}([n]) \mid |t| = k \}$ for the collection of subsets of $[n]$ which are of size k. For a function $f : \mathbb{R}^d \to \mathbb{R}$ we write $\partial_{ij} f = \frac{\partial^2 f}{\partial x_i \partial x_j}$ and $\partial_{ijk} f = \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}$. Also, we write $|f|_k = \sup_{i_1, i_2, \dots, i_k \in [d]} \|\partial_{i_1 i_2 \dots i_k} f\|_\infty$ for any integer $k \geq 1$, as long as the quantities exist. Here $|| \cdot ||_\infty$ denotes the supremum norm while $|| \cdot ||_2$ denotes the Euclidean norm. The notation ∇ denotes the gradient operator in $\mathbb{R}^d$. The notation $\mathrm{Id}_{d \times d}$ denotes the $d \times d$ identity matrix. The vertex set of all graphs and simplicial complexes is assumed to be $[n]$. We also use Bachmann-Landau asymptotic notation: we say $f(n) = O(g(n))$ iff $\limsup_{n\to\infty} \frac{|f(n)|}{g(n)} < \infty$ and $f(n) = \Omega(g(n))$ iff $\liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0$. The notation that $f(n) = \omega(g(n))$ indicates that $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$.

Acknowledgements

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

2 Simplicial Complex Preliminaries

2.1 First definitions

Firstly, we recall the notion of a simplicial complex [42, Ch 3.1]; these provide higher-dimensional generalisations of a graph and constitute data structures of interest across algebraic topology in general as well as applied and computational topology in particular.

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

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

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

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

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

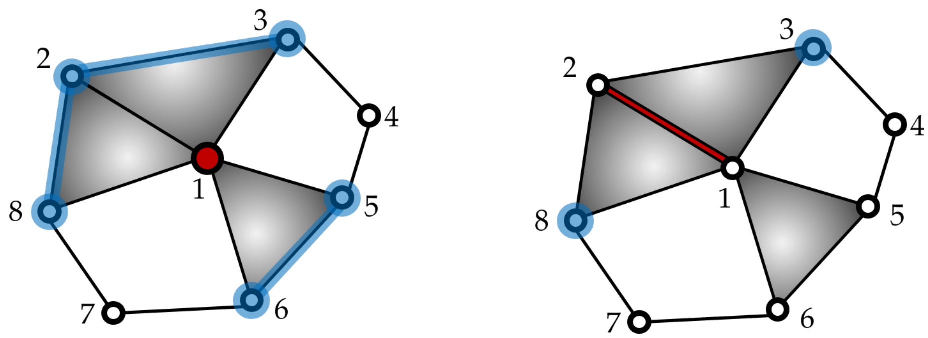

Figure 1: Left: the link (highlighted in blue) of the vertex 1 (highlighted in red). Right: the link (highlighted in blue) of the edge $\{1, 2\}$ (highlighted in red). The two-dimensional simplices are shaded in grey.

Recall that $\mathbf{G}(n, p)$ is a random graph on n vertices where each pair of vertices is connected with probability p, independently of any other pair. The $\mathbf{X}(n, p)$ random simplicial complex is the clique complex of the $\mathbf{G}(n, p)$ random graph, which is a random model studied in stochastic topology [22, 24]. Note that $t \in \mathcal{X}$ if and only if the vertices of t span a clique in G. Thus, elements in $\mathbf{X}(n, p)$ are cliques in $\mathbf{G}(n, p)$.

2.2 Links

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

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

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

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

2.3 Discrete Morse theory

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

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

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

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

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

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

We refer the interested reader to [15] for an introduction to discrete Morse theory and to [34] for seeing how it is used to reduce computations in the persistent homology algorithm. In this work we aim to understand how much improvement one would likely get on a random input when using a specific type of acyclic partial matching, defined below.

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

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

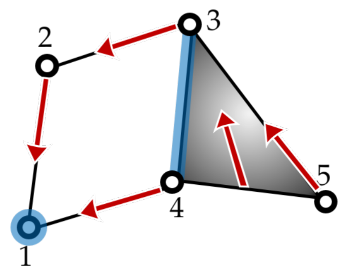

Figure 2: Lexicographical matching given by the red arrows. Critical simplices are highlighted in blue.

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

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

Note that $\{3, 4\}$ cannot be matched because the set $I_{\mathcal{L}}(\{3, 4\})$ 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 improvement.

3 Probabilistic tools

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

3.1 CLT for Dissociated Sums

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

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

Consider the resulting random vector $W = (W_1, \ldots, W_d) \in \mathbb{R}^d$. The following notion is a natural multivariate generalisation of the dissociated sum from [32]; see also [4].

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

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

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

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

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

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

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

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

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

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

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

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

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

Corollary 3.4. *We have the following two bounds:*

1. *Under the assumptions of Theorem 3.2,*

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

2. *Assuming the hypotheses of Theorem 3.3,*

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

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

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

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

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

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

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

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

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

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

$$\mathbb{E}\left|X_1 X_2\right| \mathbb{E}\left|X_3\right| \leq c_3 \mathbb{E}\left|X_1 X_2\right|;$$

$$\mathbb{E}\left|X_1 X_2 X_3\right| \leq c_3 \mathbb{E}\left|X_1 X_2\right|.$$

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

$$\mathbb{E}\left|X_1 X_2\right| \leq \left\{\mathbb{E}\left\{X_1^2\right\} \mathbb{E}\left\{X_2^2\right\}\right\}^{\frac{1}{2}} = c_1 c_2 \left\{\mu_1 \mu_2 (1-\mu_1)(1-\mu_2)\right\}^{\frac{1}{2}},$$

which finishes the proof. □

3.2 CLT for U-statistics

Here we consider generalised U-statistics, which were first introduced in [20]. The result we derive could be of independent interest but, most importantly, the approximation theorem for simplex counts follows as a consequence. Let $\{\xi_i\}_{1 \leq i \leq n}$ be a sequence of of independent random variables taking values in a measurable subset $\mathcal{X} \subseteq \mathcal{U}$ and let $\{Y_{i,j}\}_{1 \leq i < j \leq n}$ be an array of of independent random variables taking values in a measurable subset $\mathcal{Y} \subseteq \mathcal{U}$ which is independent of $\{\xi_i\}_{1 \leq i \leq n}$. We use the convention that $Y_{i,j} = Y_{j,i}$ for any $i < j$. For example, one can think of X_i as a random label of a vertex i in a random graph where $Y_{i,j}$ is the indicator for the edge connecting i and j. Given a subset $s \subseteq [n]$ of size m, write $s = \{s_1, s_2, \dots, s_m\}$ such that $s_1 < s_2 < \dots < s_m$ and set $\mathcal{X}_s = (\xi_{s_1}, \xi_{s_2}, \dots, \xi_{s_m})$ and $\mathcal{Y}_s = (Y_{s_1,s_2}, Y_{s_1,s_3}, \dots Y_{s_{m-1},s_m})$. Recall that C_k denotes the set of subsets of $[n]$ which are of size k.

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

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

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

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

Assumption 3.7. We assume that

1. For any $i \in [d]$ there is some $\alpha_i > 0$ such that for all $s, t \in \mathbb{I}_i$, the variables $f_i(\mathcal{X}_s, \mathcal{Y}_s), f_i(\mathcal{X}_t, \mathcal{Y}_t)$ are either independent or $\mathrm{Cov}(f_i(\mathcal{X}_s, \mathcal{Y}_s), f_i(\mathcal{X}_t, \mathcal{Y}_t)) > \alpha_i$.

2. There is $\beta \geq 0$ such that for any $i, j, l \in [d]$ and any $s \in \mathbb{I}_i, t \in \mathbb{I}_j, u \in \mathbb{I}_l$ we have

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

   as well as

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

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

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

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

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

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

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

and

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

Here,

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

and

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

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

4 Critical Simplex Counts for Lexicographical Morse Matchings

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

Lemma 4.1. *Let $\mathcal{L}$ 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 for all $j < i$ we have $(t \setminus \{i\}) \cup \{j\} \notin \mathcal{L}$.*

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

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

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

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

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

4.1 Mean and variance

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

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

Proof.

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

Moreover,

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

□

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

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

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

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

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

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

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

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

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

Using Markov's inequality, we get:

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

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

4.2 Approximation theorem

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}
B_{3.2} \leq \frac{1}{3} \sum_{i,j,k=1}^{d} \sum_{a=1}^{n-i} \sum_{\substack{\phi \in C_{i+1} \\ \min(\phi)=a}} \sum_{b=1}^{n-j} \sum_{\substack{(\psi,j) \in \mathbb{D}_j((\phi,i)) \\ \min(\psi)=b}} \\
\Bigg\{ \sum_{r \in \mathbb{D}_k((\phi,i))} \frac{3}{2} (\sigma_i \sigma_j \sigma_k)^{-1} \left\{ \mu(i,a)\mu(j,b)(1-\mu(i,a))(1-\mu(j,b)) \right\}^{\frac{1}{2}} \\
+ \sum_{r \in \mathbb{D}_k((\psi,j))} (\sigma_i \sigma_j \sigma_k)^{-1} \left\{ \mu(i,a)\mu(j,b)(1-\mu(i,a))(1-\mu(j,b)) \right\}^{\frac{1}{2}} \Bigg\}
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i,j,k=1}^{d} \sum_{a=1}^{n-i} \sum_{b=1}^{n-j} C n^{i+j-1} n^{k} n^{-i-j-k} \left\{ (1-p^{i+1})^{a-1} (1-p^{j+1})^{b-1} + (1-p^{i})^{a-1} (1-p^{j})^{b-1} \right\}^{\frac{1}{2}} \\
&\leq C n^{-1} \sum_{i,j,k=1}^{d} \sum_{a=1}^{\infty} \sum_{b=1}^{\infty} \left[\left\{ (1-p^{i+1})^{a-1} (1-p^{j+1})^{b-1} \right\}^{\frac{1}{2}} + \left\{ (1-p^{i})^{a-1} (1-p^{j})^{b-1} \right\}^{\frac{1}{2}} \right] \\
&\leq C n^{-1} d^{3} \left\{ \frac{1}{(1-\sqrt{1-p^{d+1}})^{2}} + \frac{1}{(1-\sqrt{1-p^{d}})^{2}} \right\} \leq C n^{-1}.
\end{aligned}
$$

□

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

1. As seen in Lemma 4.2, the expected number of critical k-simplices under the lexicographical matching is one power of n smaller than the total number of k-simplices in $\mathbf{X}(n,p)$.

2. From Proposition 4.4, it is with high probability that in $\mathbf{X}(n,p)$ all k-simplices $s \in \mathbf{X}(n,p)$ with $\min(s) = \omega(\ln^{1+\epsilon}(n))$ for any fixed $\epsilon > 0$ are not critical.

5 Simplex Counts in Links

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

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

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

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

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

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

5.1 Mean and variance

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

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

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

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

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

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

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

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

Taking $l = k$ completes the proof. □

5.2 Approximation theorem

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

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

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

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

1. *Let $h \in \mathcal{H}_d$. Then*
$$
\left|\mathbb{E}h(W^t) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \leq |h|_3\, B_{5.3}(n-|t|)^{-\frac{1}{2}}.
$$

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

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

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

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

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

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

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

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

Now we apply Corollary 5.2 and get

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

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

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

Using Equations (5.2) - (5.5) to bound $B_{3.4}$ from Corollary 3.4 we get:

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

$$
\begin{aligned}
\leq & \frac{7}{6}\sum_{i,j,k=1}^{d}(n-|t|)^{i+j+k+1}(d+1)^{2d+4}(n-|t|)^{-i-j-k-\frac{3}{2}}(2d+1)^{\frac{3d+3}{2}}d^{\frac{3d}{2}} \\
& (p^{-|t|}-1)^{-\frac{3}{2}}(\mu^t_{k+1}\mu^t_{i+1}\mu^t_{j+1})^{-1}\left\{\mu^t_{i+1}\mu^t_{j+1}(1-\mu^t_{i+1})(1-\mu^t_{j+1})\right\}^{\frac{1}{2}} \\
\leq & (n-|t|)^{-\frac{1}{2}}\frac{7}{6}(2d+1)^{5d+\frac{11}{2}}(p^{-|t|}-1)^{-\frac{3}{2}}\sum_{i,j,k=1}^{d}\left((\mu^t_{i+1}\mu^t_{j+1})^{-1}(\mu^t_{k+1})^{-2}\right)^{\frac{1}{2}} \\
\leq & \left\{\frac{7}{6}(2d+1)^{5d+\frac{17}{2}}(p^{-|t|}-1)^{-\frac{3}{2}}p^{-(d+1)(d+2|t|)}\right\}(n-|t|)^{-\frac{1}{2}}.
\end{aligned}
$$

□

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

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

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

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

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

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

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

Then the associated generalised U-statistic $S_{n,i+1}(f_i)$ equals the $(i+1)$-clique count T_{i+1}, as given by

Equation (6.1). To apply Theorem 3.9 we need to center and rescale our variables. It is easy to see that $\mathbb{E}\{f_i(\mathcal{Y}_\phi)\} = p^{\binom{i+1}{2}}$ if $\phi \in C_{i+1}$. We let $I_i \coloneqq C_{i+1} \times \{i\}$ and for $s = (\phi, i) \in \mathbb{I}_i$ we define $X_s \coloneqq \sigma^{-1}\left(f_i(\mathcal{Y}_\phi) - p^{\binom{i+1}{2}}\right)$ and $W_i = \sum_{s\in\mathbb{I}_i} X_s$. Now the vector of interest is $W = (W_1, W_2, \ldots, W_d) \in \mathbb{R}^d$. This brings us to the next approximation theorem.

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

1. *Let $h \in \mathcal{H}_d$. Then*
$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z)\right| \le |h|_3\, B_{6.1} n^{-1}.$$

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

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

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

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

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

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

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

Remark 6.2. It is easy to show that with high probability there are no large cliques in $\mathbf{G}(n, p)$ for $p < 1$ constant. To see this, the expectation of the number of k-cliques is $\binom{n}{k} p^{\binom{k}{2}}$. By Stirling's approximation,

$k = \Omega(\ln^{1+\epsilon}(n))$ for any positive ϵ forces the expectation to go to 0 asymptotically. Hence, by Markov's inequality, for any $\epsilon > 0$, with high probability there are no cliques of order $\ln^{1+\epsilon}(n)$ or larger.

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

Remark 6.3. Note that in Corollary 6.1 we use a multivariate normal distribution with covariance Σ, which is the covariance of W when n is finite and it differs from the limiting covariance, as mentioned in [40]. To approximate W with the limiting distribution, in the spirit of [40, Proposition 3] one could proceed in two steps: use the existing theorems to approximate W with ΣZ and then approximate ΣZ with $\Sigma_L Z$ where Σ_L is the limiting covariance, which is non-invertible, as observed in [20].

Remark 6.4. Corollary 6.1 generalises the result [40, Proposition 2] beyond the case when $d = 2$ and we get a bound of the same order of n. [27, Theorem 3.1] considers centered subgraph counts in a random graph associated to a graphon. If we take the graphon to be constant, the associated random graph is just $\mathbf{G}(n,p)$. Compared to [27, Theorem 3.1] we place weaker smoothness conditions on our test functions. However, we make use of the special structure of cliques whereas [27, Theorem 3.1] applies to any centered subgraph counts. Translating [27, Theorem 3.1] into a result for uncentered subgraph counts, as we provide here in the special case of clique counts, is not trivial for general d.

However, it should be possible to extend our results, using the same abstract approximation theorem, beyond the random clique complex to Linial-Meshulam random complexes [31] or even more general multiparameter random complexes [11]. We shall consider this conjecture in future work.

7 A Proof of the Multivariate CLT for Dissociated Sums and U-statistics

7.1 A Proof of the Multivariate CLT

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

$\mathbb{D}_j(t) \setminus \mathbb{D}_j(s)$ as $\mathbb{D}_j(t;s)$, with $\mathbb{D}(t;s) \subset \mathbb{I}$ denoting the disjoint union of such differences over $j \in [d]$.

7.1.1 Smooth Test Functions

To prove Theorem 3.2 we use Stein's method for multivariate normal distributions; for details see for example Chapter 12 in [10]. Our proof of Theorem 3.2 is based on the **Stein characterization** of the multivariate normal distribution: $Z \in \mathbb{R}^d$ is a multivariate normal $\mathrm{MVN}(0, \Sigma)$ if and only if the identity

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

holds for all twice continuously differentiable $f : \mathbb{R}^d \to \mathbb{R}$ for which the expectation exists. In particular, we will use the following result based on [33, Lemma 1 and Lemma 2]. As Lemma 1 and Lemma 2 in [33] are stated there only for infinitely differentiable test functions, we give the proof here for completeness.

Lemma 7.1 (Lemma 1 and Lemma 2 in [33])**.** *Fix $n \geq 2$. Let $h : \mathbb{R}^d \to \mathbb{R}$ be n times continuously differentiable with n-th partial derivatives being Lipschitz and $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$. Then, if $\Sigma \in \mathbb{R}^{d\times d}$ is symmetric positive semidefinite, there exists a solution $f : \mathbb{R}^d \to \mathbb{R}$ to the equation*

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

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

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

Proof. Let h be as in the assertion. It is shown in Lemma 2.1 in [9], which is based on a reformulation of Eq. (2.20) in [3], that a solution of (7.2) for h is given by $f(x) = f_h(x) = \int_0^1 \frac{1}{2t} \mathbb{E}\{h(Z_{x,t})\}\, dt$, with $Z_{x,t} = \sqrt{t}x + \sqrt{1-t}\Sigma^{1/2} Z$. As h has n-th partial derivatives being Lipschitz and hence for differentiating f we can bring the derivative inside the integral, it is straightforward to see that the solution f is n times continuously differentiable.

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

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

for any $i_1, i_2, \ldots, i_k$; see, for example, Equation (10) in [33]. Taking the sup-norm on both sides and bounding the right hand side of the equation gives

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

$\square$

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

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

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

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

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

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

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

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

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

with

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

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

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

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

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

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

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

Proof. Note that

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

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

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

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

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

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

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

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

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

as desired. □

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

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

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

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

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

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

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

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

$$- \mathbb{E}\left\{X_s X_t\right\} \mathbb{E}\left\{\partial_{|s||t|} f(W^{s,t})\right\}.$$

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

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

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

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

as required.

□

Claim 7.4.

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

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

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

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

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

$$-\sum_{s\in\mathbb{I}}\sum_{t\in\mathbb{D}(s)}\sum_{u\in\mathbb{D}(t;s)}\mathbb{E}\left\{X_sX_t\right\}\mathbb{E}\left\{X_u\partial_{|s||t|k}f(W-\rho_{s,t}(U^s+V^{s,t}))\right\}.$$

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

$$|R_3|\leq |f|_3\sum_{s\in\mathbb{I}}\sum_{t\in\mathbb{D}(s)}\sum_{u\in\mathbb{D}(s)}\mathbb{E}\left|X_sX_t\right|\mathbb{E}\left|X_u\right| + |f|_3\sum_{s\in\mathbb{I}}\sum_{t\in\mathbb{D}(s)}\sum_{u\in\mathbb{D}(t;s)}\mathbb{E}\left|X_sX_t\right|\mathbb{E}\left|X_u\right|,$$

as required. □

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

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

□

7.1.2 Non-smooth Test Functions

Convex set indicator test functions provide a stronger distance between probability distributions. Also, from a non-asymptotic perspective, the distance might be more useful in statistical applications: for example, when estimating confidence regions, which are often convex sets. Here we follow [27, Section 5.3] very closely to derive a bound on the convex set distance between a vector of dissociated sums $W \in \mathbb{R}^d$ with covariance matrix Σ and a target multivariate normal distribution $\Sigma^{\frac{1}{2}}Z$, where $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$. The smoothing technique used here is introduced in [16]. However, a better (polylogarithmic) dependence on d could potentially be achieved using a recent result [17, Proposition 2.6], at the expense of larger constants. The recursive approach from [41, 26] usually yields better dependence on n; however, this requires the target normal distribution to have an invertible covariance matrix. Since this property does not always hold in our applications of interest, we do not use the recursive approach here.

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

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

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

Let $\mathcal{H}_{\epsilon,A} \coloneqq \left\{ h_{\epsilon,A} : \mathbb{R}^d \to [0,1]; A \in \mathcal{K} \right\}$ be a class of functions such that $h_{\epsilon,A}(x) = 1$ for $x \in A$ and 0 for $x \notin A^{\epsilon}$. Then, by [6, Lemma 2.1] as well as inequalities (1.2) and (1.4) from [6], for any $\epsilon > 0$ we have

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

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

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

Set $\delta = \frac{\epsilon}{16\sqrt{d}}$ and $h_{\epsilon,A} = S_\delta^4 I_{A^{\epsilon/4}}$, where $I_{A^{\epsilon/4}}$ is the indicator function of the subset $A^{\epsilon/4} \subseteq \mathbb{R}^d$. By [16, Lemma 3.9] we have that $h_{\epsilon,A}$ is bounded and is in $\mathcal{H}_d$. Moreover, the following bounds hold:

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

Note that $h_{\epsilon,A} = S_\delta^4 I_{A^{\epsilon/4}} \in \mathcal{H}_{\epsilon,A}$ and hence [6, Lemma 2.1] applies. Using this with Theorem 3.2 we get

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

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

7.2 A Proof of the CLT for U-statistics

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

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

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

Note that

$$\mathbb{E}\,|X_s X_t X_u| = (\sigma_i\sigma_j\sigma_k)^{-1}\mathbb{E}\,|\{f_i(\mathcal{X}_s,\mathcal{Y}_s)-\mu_s\}\{f_j(\mathcal{X}_t,\mathcal{Y}_t)-\mu_t\}\{f_l(\mathcal{X}_u,\mathcal{Y}_u)-\mu_u\}|$$

as well as

$$\mathbb{E}\,|X_s X_t|\,\mathbb{E}\,|X_u| = (\sigma_i\sigma_j\sigma_k)^{-1}\mathbb{E}\,|\{f_i(\mathcal{X}_s,\mathcal{Y}_s)-\mu_s\}\{f_j(\mathcal{X}_t,\mathcal{Y}_t)-\mu_t\}|\,\mathbb{E}\,|f_l(\mathcal{X}_u,\mathcal{Y}_u)-\mu_u|\,.$$

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

$$\mathbb{E}\,|X_s X_t X_u| \le (\sigma_i\sigma_j\sigma_k)^{-1}\beta \ \text{ and } \ \mathbb{E}\,|X_s X_t|\,\mathbb{E}\,|X_u| \le (\sigma_i\sigma_j\sigma_k)^{-1}\beta. \tag{7.10}$$

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

$$\begin{aligned}
\operatorname{Var}(S_{n,k_i}(f_i)) &= \sum_{s\in C_{k_i}}\sum_{t\in\mathbb{D}_i(s)} \operatorname{Cov}(f_i(\mathcal{X}_s,\mathcal{Y}_s), f_i(\mathcal{X}_t,\mathcal{Y}_t)) \\
&= \sum_{m=1}^{k_i}\sum_{s\in C_{k_i}}\sum_{\substack{t\in C_{k_i}\\ |s\cap t|=m}} \operatorname{Cov}(f_i(\mathcal{X}_s,\mathcal{Y}_s), f_i(\mathcal{X}_t,\mathcal{Y}_t)) \\
&\ge \binom{n}{k_i}\sum_{m=1}^{k_i}\binom{k_i}{m}\binom{n-k_i}{k_i-m}\alpha_i = \alpha_i\sum_{m=1}^{k_i}\binom{n}{2k_i-m}\binom{2k_i-m}{k_i}\binom{k_i}{m}
\end{aligned}$$

$$\begin{aligned}
&\geq \alpha_i k_i \binom{n}{2k_i - 1}\binom{2k_i-1}{k_i} \geq \alpha_i k_i \frac{n^{2k_i-1}}{(2k_i-1)^{2k_i-1}} \frac{(2k_i-1)^{k_i}}{k_i^{k_i}} \\
&= \alpha_i \frac{n^{2k_i-1}}{(2k_i^2 - k_i)^{k_i-1}}.
\end{aligned}$$

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

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

Using Equations (7.9) - (7.11) to bound the quantity $B_{3.4}$ from Corollary 3.4 we get

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

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

Now Equation (7.9) becomes

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

Equation (7.11) becomes

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

Using the adjusted bounds in Corollary 3.4 gives the result. $\square$

References

[1] Robert J Adler, Omer Bobrowski, and Shmuel Weinberger. "Crackle: the homology of noise". In: *Discrete and Computational Geometry* (2014), pp. 680–704.

[2] Ryo Asai and Jay Shah. "Algorithmic canonical stratifications of simplicial complexes". In: *Journal of Pure and Applied Algebra* 226.9 (2022), p. 107051.

[3] Andrew D Barbour. "Stein's method for diffusion approximations". In: *Probability Theory and Related Fields* 84.3 (1990), pp. 297–322.

[4] Andrew D Barbour, Michal Karoński, and Andrzej Ruciński. "A central limit theorem for decomposable random variables with applications to random graphs". In: *Journal of Combinatorial Theory, Series B* 47.2 (1989), pp. 125–145.

[5] Ulrich Bauer and Abhishek Rathod. "Hardness of approximation for Morse matching". In: *SODA '19: Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* (2019), pp. 2663–2674.

[6] Vidmantas Bentkus. "On the dependence of the Berry–Esseen bound on dimension". In: *Journal of Statistical Planning and Inference* 113.2 (2003), pp. 385–402.

[7] Omer Bobrowski and Matthew Kahle. "Topology of random geometric complexes: a survey". In: *Journal of Applied and Computational Topology* 1.3 (2018), pp. 331–364.

[8] Gunnar Carlsson, Afra Zomorodian, Anne Collins, and Leonidas J Guibas. "Persistence barcodes for shapes". In: *International Journal of Shape Modeling* 11.02 (2005), pp. 149–187.

[9] Sourav Chatterjee and Elizabeth Meckes. "Multivariate normal approximation using exchangeable pairs". In: *Alea* 4 (2008), pp. 257–283.

[10] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal Approximation by Stein's Method*. Springer, 2011.

[11] A Costa and M Farber. “Large random simplicial complexes, I”. In: *Journal of Topology and Analysis* 8.03 (2016), pp. 399–429.

[12] Herbert Edelsbrunner and John Harer. *Computational Topology: An Introduction*. Providence, USA: American Mathematical Society, 2010.

[13] Peter Eichelsbacher and Benedikt Rednoß. “Kolmogorov bounds for decomposable random variables and subgraph counting by the Stein–Tikhomirov method”. In: *Bernoulli* 29.3 (2023), pp. 1821–1848.

[14] Xiao Fang. “A multivariate CLT for bounded decomposable random vectors with the best known rate”. In: *Journal of Theoretical Probability* 29.4 (2016), pp. 1510–1523.

[15] Robin Forman. “A user’s guide to discrete Morse theory”. In: *Séminaire Lotharingien de Combinatoire* 48 (2002), B48c.

[16] Han L Gan, Adrian Röllin, and Nathan Ross. “Dirichlet approximation of equilibrium distributions in Cannings models with mutation”. In: *Advances in Applied Probability* 49.3 (2017), pp. 927–959.

[17] Robert E Gaunt and Siqi Li. “Bounding Kolmogorov distances through Wasserstein and related integral probability metrics”. In: *Journal of Mathematical Analysis and Applications* (2023), p. 126985.

[18] Robert Ghrist. “Barcodes: the persistent topology of data”. In: *Bull. Amer. Math. Soc. (N.S.)* 45.1 (2008), pp. 61–75.

[19] Gregory Henselman-Petrusek and Robert Ghrist. “Matroid Filtrations and Computational Persistent Homology”. In: *arXiv:1606.00199* (2016).

[20] Svante Janson and Krzysztof Nowicki. “The asymptotic distributions of generalized U-statistics with applications to random graphs”. In: *Probability Theory and Related Fields* 90.3 (1991), pp. 341–375.

[21] Michael Joswig and Marc E Pfetsch. “Computing optimal Morse matchings”. In: *SIAM Journal on Discrete Mathematics* 20.1 (2006), pp. 11–25.

[22] Matthew Kahle. “Topology of random clique complexes”. In: *Discrete Mathematics* 309.6 (2009), pp. 1658–1671.

[23] Matthew Kahle. “Random geometric complexes”. In: *Discrete & Computational Geometry* 45.3 (2011), pp. 553–573.

[24] Matthew Kahle. "Sharp vanishing thresholds for cohomology of random flag complexes". In: *Annals of Mathematics* (2014), pp. 1085–1107.

[25] Matthew Kahle and Elizabeth Meckes. "Limit theorems for Betti numbers of random simplicial complexes". In: *Homology, Homotopy and Applications* 15.1 (2013), pp. 343–374.

[26] Mikołaj J Kasprzak and Giovanni Peccati. "Vector-valued statistics of binomial processes: Berry-Esseen bounds in the convex distance". In: *arXiv preprint arXiv:2203.13137* (2022).

[27] Gursharn Kaur and Adrian Röllin. "Higher-order fluctuations in dense random graph models". In: *Electronic Journal of Probability* 26 (2021), pp. 1–36.

[28] Vladimir S Korolyuk and Yu V Borovskich. *Theory of U-statistics*. Vol. 273. Springer Science & Business Media, 2013.

[29] Leon Lampret. "Chain complex reduction via fast digraph traversal". In: *arXiv:1903.00783* (2019).

[30] A J Lee. *U-statistics: Theory and Practice*. Taylor & Francis, 1990.

[31] Nathan Linial and Roy Meshulam. "Homological connectivity of random 2-complexes". In: *Combinatorica* 26.4 (2006), pp. 475–487.

[32] WG McGinley and Robin Sibson. "Dissociated random variables". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 77. 1. Cambridge University Press. 1975, pp. 185–188.

[33] Elizabeth Meckes. "On Stein's method for multivariate normal approximation". In: *High dimensional probability V: the Luminy volume*. Institute of Mathematical Statistics, 2009, pp. 153–178.

[34] Konstantin Mischaikow and Vidit Nanda. "Morse theory for filtrations and efficient computation of persistent homology". In: *Discrete & Computational Geometry* 50.2 (2013), pp. 330–353.

[35] Vidit Nanda. "Local Cohomology and Stratification". In: *Foundations of Computational Mathematics* 20 (2020), pp. 195–222.

[36] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. "A roadmap for the computation of persistent homology". In: *EPJ Data Science* 6 (2017), pp. 1–38.

[37] Takashi Owada, Gennady Samorodnitsky, and Gugan Thoppe. "Limit theorems for topological invariants of the dynamic multi-parameter simplicial complex". In: *Stochastic Processes and their Applications* 138 (2021), pp. 56–95.

[38] Nicolas Privault and Grzegorz Serafin. “Normal approximation for sums of weighted U-statistics– application to Kolmogorov bounds in random subgraph counting”. In: *Bernoulli* 26.1 (2020), pp. 587–615.

[39] Martin Raič. “A multivariate CLT for decomposable random vectors with finite second moments”. In: *Journal of Theoretical Probability* 17.3 (2004), pp. 573–603.

[40] Gesine Reinert and Adrian Röllin. “Random subgraph counts and U-statistics: multivariate normal approximation via exchangeable pairs and embedding”. In: *Journal of Applied Probability* 47.2 (2010), pp. 378–393.

[41] Matthias Schulte and Joseph E Yukich. “Multivariate second order Poincaré inequalities for Poisson functionals”. In: *Electronic Journal of Probability* 24 (2019), pp. 1–42.

[42] Edwin Spanier. *Algebraic Topology*. McGraw-Hill, 1966.

A Proof of Lemma 4.3

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

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

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

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

where

$$\begin{aligned}
V_1 =& \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}\binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m-1}\binom{j-i+1}{q-1}\\
&\Big\{\theta(i,j,q,m,1)\big[(1-2p^{k+1}+p^{2k+2-m})^{i-1}-(1-p^{k+1}-p^k+p^{2k+1-m})^{i-1}\big]\\
&+\theta(i,j,q,m,0)\big[(1-2p^k+p^{2k+1-m})^{i-1}-(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\big]-\eta(i)\eta(j)\Big\};\\
V_2 =& \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}\binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k}{m}\binom{j-i+1}{q-1}\\
&\Big\{\theta(i,j,q,m,1)\big[(1-2p^{k+1}+p^{2k+2-m})^{i-1}-(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\big]\\
&+\theta(i,j,q,m,0)\big[(1-2p^k+p^{2k-m})^{i-1}-(1-p^{k+1}-p^k+p^{2k+2-m})^{i-1}\big]-\eta(i)\eta(j)\Big\};\\
V_3 =& \sum_{i=1}^{n-k}\sum_{m=1}^{k}\binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1}\\
&\Big\{p^{-\binom{m}{2}}\Big[(1-2p^{k+1}+p^{2k+2-m})^{i-1}+\\
&(1-2p^k+p^{2k+1-m})^{i-1}-2(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1}\Big]-\eta(i)^2\Big\};\\
V_4 =& \sum_{i=1}^{n-k}\binom{n-i}{k}\Big\{\eta(i)-p^{\binom{k+1}{2}}\eta(i)^2\Big\}.
\end{aligned}$$

Here we have used the following notation:

$$\eta(a) \coloneqq (1-p^{k+1})^{a-1} - (1-p^k)^{a-1};$$
$$\theta(i,j,q,m,\delta) \coloneqq p^{-\binom{m}{2}}(1-p^{k+\delta})^{j-i-q}(1-p^{k+\delta-m})^q.$$

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

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

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

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

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

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

$$\begin{aligned}\sum_{s\in C_{k+1}} \operatorname{Var}(Z_sY_s) &= \sum_{i=1}^{n} \sum_{\substack{s\in C_{k+1}\\ \min(s)=i}} \left(\mathbb{E}\left\{(Z_sY_s)^2\right\} - \mathbb{E}\left\{Z_sY_s\right\}^2\right)\\ &= \sum_{i=1}^{n-k}\binom{n-i}{k}\left(\mathbb{P}(Z_sY_s=1) - \mathbb{P}(Z_sY_s=1)^2\right)\\ &= \sum_{i=1}^{n-k}\binom{n-i}{k}\left\{\mu(i)-\mu(i)^2\right\} = p^{\binom{k+1}{2}}V_4.\end{aligned}$$

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

Fix $s,t \in C_{k+1}$ and assume $|s\cap t| = m$ where $1 \le m \le k$. Note that because $m \neq k+1$, we have

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

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

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

$$
\begin{aligned}
&\mathbb{P}[(1 - \prod_{j \in s} Y_{i,j})(1 - \prod_{j \in t} Y_{i,j}) = 1] \\
&\qquad = \mathbb{P}[\neg(\prod_{j \in s} Y_{i,j} = 1 \cup \prod_{j \in t} Y_{i,j} = 1)] \\
&\qquad = 1 - \left\{ \mathbb{P}(\prod_{j \in s} Y_{i,j} = 1) + \mathbb{P}(\prod_{j \in t} Y_{i,j} = 1) - \mathbb{P}(\prod_{j \in s} Y_{i,j} = 1 \cap \prod_{j \in t} Y_{i,j} = 1) \right\} \\
&\qquad = 1 - (2p^{k+1} - p^{2k+2-m}).
\end{aligned}
$$

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

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

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

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

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

Using the described strategy we get:

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

Now we are ready to calculate the covariance:

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

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

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

as well as

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

and

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

Next we argue that

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

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

An analogous argument shows that

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

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

$$\begin{aligned}
&\sum_{\substack{s \neq t \in C_{k+1} \\ \min(s) \neq \min(t)}} \mathrm{Cov}(Z_sY_s, Z_tY_t) \\
&= \sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t)\in\Gamma^+_{k+1}(i,j,m,q)} \mathrm{Cov}(Z_sY_s, Z_tY_t) \\
&\quad + \sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} \sum_{(s,t)\in\Gamma^-_{k+1}(i,j,m,q)} \mathrm{Cov}(Z_sY_s, Z_tY_t) \\
&\quad + \sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)} \sum_{(s,t)\in\Gamma^+_{k+1}(j,i,m,q)} \mathrm{Cov}(Z_sY_s, Z_tY_t) \\
&\quad + \sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)} \sum_{(s,t)\in\Gamma^-_{k+1}(j,i,m,q)} \mathrm{Cov}(Z_sY_s, Z_tY_t) \\
&= \sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} |\Gamma^+_{k+1}(i,j,m,q)| \Big\{ p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(i,j,k+1,k+1,m,m,q)
\end{aligned}$$

$$
\begin{aligned}
&+\pi(i,j,k,k,m-1,m-1,q)-\pi(i,j,k,k+1,m-1,m-1,q)\\
&-\pi(i,j,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&+\sum_{i=1}^{n-k}\sum_{j=i+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)}|\Gamma^-_{k+1}(i,j,m,q)|\Big\{p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(i,j,k+1,k+1,m,m,q)\\
&+\pi(i,j,k,k,m,m,q)-\pi(i,j,k,k+1,m,m,q)\\
&-\pi(i,j,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&+\sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)}|\Gamma^+_{k+1}(j,i,m,q)|\Big\{p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(j,i,k+1,k+1,m,m,q)\\
&+\pi(j,i,k,k,m-1,m-1,q)-\pi(j,i,k,k+1,m-1,m-1,q)\\
&-\pi(j,i,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&+\sum_{j=1}^{n-k}\sum_{i=j+1}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,i-j)}|\Gamma^-_{k+1}(j,i,m,q)|\Big\{p^{2\binom{k+1}{2}-\binom{m}{2}}(\pi(j,i,k+1,k+1,m,m,q)\\
&+\pi(j,i,k,k,m,m,q)-\pi(j,i,k,k+1,m,m,q)\\
&-\pi(j,i,k+1,k,m,m,q))-\mu(i)\mu(j)\Big\}\\
&=2p^{2\binom{k+1}{2}}V_1+2p^{2\binom{k+1}{2}}V_2.
\end{aligned}
$$

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

$$
\begin{aligned}
\sum_{\substack{s\neq t\in C_{k+1}\\ \min(s)=\min(t)}}\operatorname{Cov}(Z_sY_s,Z_tY_t)&=\sum_{i=1}^{n-k}\sum_{m=1}^{k}\sum_{(s,t)\in\Gamma_{k+1}(i,i,m,0)}\operatorname{Cov}(Z_sY_s,Z_tY_t)\\
=&\sum_{i=1}^{n-k}\sum_{m=1}^{k}\binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1}\\
&\Big\{p^{-\binom{m}{2}}\Big[(1-2p^{k+1}+p^{2k+2-m})^{i-1}+(1-2p^k+p^{2k+1-m})^{i-1}\\
&-2(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1}\Big]-((1-p^{k+1})^{i-1}-(1-p^k)^{i-1})^2\Big\}\\
=&p^{2\binom{k+1}{2}}V_3.
\end{aligned}
$$

□

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

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

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

$$
\begin{aligned}
R_1 &:= \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} \binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k+1}{m}\binom{j-i+1}{q-1} \\
&(1-p^{k+1})^{i+j-2}; \\
R_2 &:= \sum_{i<j}^{n-k}\sum_{m=1}^{k}\sum_{q=1}^{\min(k+1,j-i)} \binom{n-j}{2k+1-m-q}\binom{2k+1-m-q}{k}\binom{k+1}{m}\binom{j-i+1}{q-1} \\
&p^{-\binom{m}{2}}(1-p^{k+1})^{j-i}(1-p^{k+1}-p^{k}+p^{2k+1-m})^{i-1}.
\end{aligned}
$$

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

$$
\begin{aligned}
V_3 \ge &\sum_{i=1}^{n-k}\binom{n-i}{2k}\binom{2k}{k}\Big\{(1-2p^{k+1}+p^{2k+1})^{i-1} - (1-p^{k+1})^{2i-2}\Big\} \\
&-\sum_{i=1}^{n-k}\sum_{m=2}^{k}\binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1}
\end{aligned}
$$

$$\left\{2p^{-\binom{m}{2}}(1-p^k-p^{k+1}+p^{2k+2-m})^{i-1}+2(1-p^{k+1})^{2i-2}\right\}$$
$$=R_4-R_3;$$

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

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

Let us now upper bound R_1:

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

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

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

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

$$R_3 \leq \sum_{i=1}^{n-k}\sum_{m=2}^{k}\binom{n-i}{2k+1-m}\binom{2k+1-m}{k}\binom{k}{m-1}2(p^{-\binom{m}{2}}+1)(1-p^{k+1})^{i-1}$$

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

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

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

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

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

$\square$

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Multivariate Central Limit Theorems for Random Clique Complexes
Publication Status	□Published X Accepted for Publication □Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	T. Temčinas, V. Nanda, G. Reinert "Multivariate Central Limit Theorems for Random Clique Complexes", accepted for publication in the Journal of Applied and Computational Topology.

Student Confirmation

Student Name:	Tadas Temčinas		
Contribution to the Paper	Under the guidance of my supervisors, who are the co-authors, I have proved the main results and wrote most of the manuscript (excluding the introduction).		
Signature		Date	14/09/2023

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof Gesine Reinert		
Supervisor comments The description is accurate.		
Gesine Reinert	Date	14/09/2023

This completed form should be included in the thesis, at the end of the relevant chapter.

3 Goodness-of-fit via Count Statistics in Dense Random Simplicial Complexes

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

[2]Mathematical Institute, University of Oxford

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.

Keywords: Multivariate normal approximation, Goodness-of-fit tests, Random simplicial complexes

1 Introduction

While complex data are often represented as graphs or networks [42], there is a growing interest in modelling complex systems beyond pairwise interactions between nodes [8]. Simplicial complexes, often used in topological data analysis (TDA), provide a rich mathematical representation of higher-order networks [7]; other representations are detailed for example in [4]. 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 [10, 11]. 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.

Multiparameter complex as a null model

In order to further statistical understanding of the model, we study maximum likelihood estimation in $\mathbf{X}(n,\mathbf{p})$ 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 [49] 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 candidates 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.

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 [45, 3]. Surprisingly, 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 [46]. 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, 44, 41]. Distributional approximation of certain subcomplexes have also been used to deduce distributional approximation results for Betti number in the random clique complex [31]. 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.

Effectiveness of discrete Morse theory

Discrete Morse theory [19] provides a powerful, flexible and widely-used mechanism for simplifying the machine computation of simplicial homology and related algebraic invariants [20, 40, 16, 5, 50]. 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 [26] 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 [5]. 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.

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 assume we observe a simplicial complex on the vertex set $\{1, \dots n\}$, which can be seen as a set of indicator variables - one for each subset of vertices - which would equal one iff that subset is a simplex. Then the MLE is seen as a function of those indicator variables. We first state a simplified version of our MLE results here. We state the results 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 $\mathbf{p} = (p_1, p_2, \dots, p_{n-1})$ 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. $$\lim_{n\to\infty} \mathbb{E}\{\hat{p}_i\} = p_i \text{ and } \lim_{n\to\infty} \mathrm{Var}(\hat{p}_i) = 0;$$

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* $1 \le i \le d$,

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

Here we recognize $\sup_{x\in\mathbb{R}} |\mathbb{P}(W_i \le x) - \mathbb{P}(Z_i \le x)| = \sup_{x\in\mathbb{R}} |\mathbb{E}\{\mathbb{I}(W_i \in (-\infty, x])\} - -\mathbb{E}\{\mathbb{I}(Z_i \in (-\infty, x])\}|$ 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 $i \le k'$ 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 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 an explicit constant* $B_{5.4} > 0$, *independent of* n *such that*

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

and

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

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

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

2. *There is an implicit 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 an implicit 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 an explicit 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 $0 \leq \sigma_\infty(i,j) < 1$.

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.

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 [46]. Another recent paper [17] 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 [24, Section 10.7] and [22, Chapter 11.3] could, with some work, in principle be adapted to prove convergence in distribution for the multivariate subcomplex counts; in [25] the author extends the results from [24] 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. Recently Zhang [51] 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 $\mathbf{X}(n, p)$ [49, Section 4]. For the $\mathbf{X}(n, \mathbf{p})$ model, only the expected value has been studied so far [6, Section 8].

There has been some limited work on statistics for random simplicial complexes. In [11, 52] 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 [32, 12, 38]. Topological goodness-of-fit tests have been suggested for point processes [9], sliced spatial data [13], and cylindrical networks [33], but not, to the best of our knowledge, for $\mathbf{X}(n, \mathbf{p})$.

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 $d_{\mathrm{KS}}(X, Y) \coloneqq \sup_{x \in \mathbb{R}} |\mathbb{P}(X \leq x) - \mathbb{P}(Y \leq x)|$ 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 [49].

2.1 First definitions

Firstly, we recall the notion of a simplicial complex [48, 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 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 [30] and extensively studied from a topological rather than combinatorial point of view by, amongst others, Costa and Farber in [14, 15]; 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 [27, 29, 31]. 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 [36, 37]. For a survey of stochastic topology, see [10, 11].

2.2 Discrete Morse theory

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

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

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 [19] for an introduction to discrete Morse theory and to [40] for an illustration

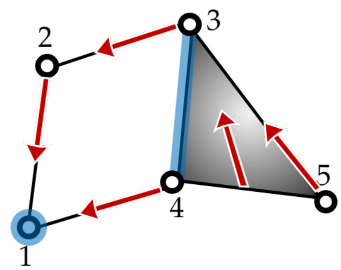

Figure 1: Lexicographical matching given by the red arrows. Critical simplices are highlighted in blue.

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 $[n] = \{1, \dots, n\}$. For each simplex $s \in \mathcal{L}$ define

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

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 it has been noted in [11, 52], 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 $i \leq I(\mathcal{L})$. 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 $j \leq i$. Write $I(\mathcal{L}) = \max\{\, i \in [d] \mid h_i \neq 0 \,\}$; 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

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

For any $i \neq j \in [I(\mathcal{L})]$ 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 $1 \le h_i \le \binom{n}{i+1}$. 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 $P_i = \prod_{j=1}^{i-1} p_j^{\binom{i+1}{j+1}}$.

Lemma 3.1. *Fix $i \in [2, d]$ with $d \le \log_2 \ln n + \log_2 \log_2 \ln n - \log_2(-\ln p^*)$ and $a \ge 0$, then:*

$$\mathbb{P}(H_i = 0) \le \exp\left(-n^{\log_2 \log_2 \ln n (1+o(1))}\right) \qquad \mathbb{P}(|H_i - \mu_i| \ge a) \le 2\exp\left(-\frac{2a^2}{(i+1)^{i+2}\binom{n}{i-1}\binom{n}{i+1}}\right).$$

In particular, $\lim_{n\to\infty} \mathbb{P}(I(\mathbf{X}(n, \mathbf{p})) = d) = 1$ as long as p^* is a positive constant.

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 $\Delta_i = \sum_{s \neq t \in \binom{[n]}{i+1}} \mathbb{1}\,(|s \cap t| \ge 2)\, \mathbb{E}\{X_s X_t\}$. As long as $\Delta_i \ge \mu_i$, we can bound the denominator in Janson's inequality [21, Theorem 1] by $\mu_i + \Delta_i \le 2\Delta_i$. Hence we get

$$\begin{aligned}\mathbb{P}(H_i = 0) \le \exp\left(-\frac{\mu_i^2}{4\Delta_i}\right) &\le \exp\left(-\frac{\binom{n}{i+1}^2 P_i^2}{4\sum_{m=2}^{i}\binom{n}{i+1}\binom{i+1}{m}\binom{n}{i+1-m} P_i^2 \prod_{j=1}^{i-1} p_j^{-\binom{m}{j+1}}}\right)\\ &= \exp\left(-\frac{\binom{n}{i+1}}{4\sum_{m=2}^{i}\binom{i+1}{m}\binom{n}{i+1-m}\prod_{j=1}^{i-1} p_j^{-\binom{m}{j+1}}}\right).\end{aligned}$$

We are interested in the case when $i = \log_2 \ln n + \log_2 \log_2 \ln n + c$ for some constant $c \in \mathbb{R}$. In this case, by examining the ratio of two consecutive summands in the denominator sum, we can see that the largest summand will correspond to $m = i$. Therefore, we get the bound

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

As long as $c \le -\log_2(-\ln p^*)$, 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 $j \ge k$.

For concentration, we apply the large deviation result Corollary 2.2 in [23]. Fix a particular $s \in \binom{[n]}{i+1}$. Note that X_s and X_t are dependent iff $|s \cap t| \ge 2$. So the number of variables in the sum of H_i that X_s depends on is $\sum_{m=2}^{i+1}\binom{i+1}{m}\binom{n-i-1}{i+1-m} \le (i+1)^{i+1}(i-1)\binom{n}{i-1} \le (i+1)^{i+2}\binom{n}{i-1}$. Now applying the large deviation result gives the second inequality. $\square$

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

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

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

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,

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

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

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

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

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

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, [18, Theorem 3.1] shows that there is some constant c such that for $i\ge \log_2\ln n+\log_2\log_2\ln n-c$ 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* $1\le i<j\le d\le \log_2\ln n+\log_2\log_2\ln n-\log_2(-\ln p^*)$ *we have*

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

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

$$\mathrm{Cov}(W_i,W_j)=F_n\mathrm{Cov}(\hat{p}_j,\hat{p}_j)=F_n\left(\mathbb{E}\left\{\frac{S_i}{H_i}\frac{S_j}{H_j}\mathbb{1}\left(H_iH_j\neq 0\right)\right\}-\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i\neq 0\right)\right\}\mathbb{E}\left\{\frac{S_j}{H_j}\mathbb{1}\left(H_j\neq 0\right)\right\}\right).$$

Assume $i<j$ and let us focus on $\mathbb{E}\left\{\frac{S_i}{H_i}\frac{S_j}{H_j}\mathbb{1}\left(H_iH_j\neq 0\right)\right\}$. 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:

$$\begin{aligned}
\mathbb{E}\left\{\frac{S_i}{H_i}\frac{S_j}{H_j}\mathbb{1}\left(H_iH_j\neq 0\right)\right\}&=\mathbb{E}\left\{\mathbb{E}\left(\frac{S_i}{H_i}\frac{S_j}{H_j}\mathbb{1}\left(H_iH_j\neq 0\right)\middle|(S_1,H_2,S_2,\ldots,H_{j-1},S_{j-1},H_j)\right)\right\}\\
&=\mathbb{E}\left\{\frac{S_i}{H_i}\frac{1}{H_j}\mathbb{1}\left(H_iH_j\neq 0\right)\mathbb{E}\left(S_j\middle|(S_1,H_2,S_2,\ldots,H_{j-1},S_{j-1},H_j)\right)\right\}\\
&=\mathbb{E}\left\{\frac{S_i}{H_i}\frac{1}{H_j}\mathbb{1}\left(H_iH_j\neq 0\right)\mathbb{E}\left(S_j\middle|H_j\right)\right\}\\
&=\mathbb{E}\left\{\frac{S_i}{H_i}\frac{1}{H_j}\mathbb{1}\left(H_iH_j\neq 0\right)p_jH_j\right\}\\
&=p_j\mathbb{E}\left\{\frac{S_i}{H_i}\mathbb{1}\left(H_i\neq 0\right)\mathbb{1}\left(H_j\neq 0\right)\right\}.
\end{aligned}$$

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

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

For the lower bound, we use the following argument:

$$\begin{aligned}
&p_j \mathbb{E}\left\{\frac{S_i}{H_i} \mathbb{1}\left(H_i \neq 0\right) \mathbb{1}\left(H_j \neq 0\right)\right\} \\
&= p_j \mathbb{E}\left\{\frac{S_i}{H_i} \mathbb{1}\left(H_i \neq 0\right)\left(1 - \mathbb{1}\left(H_j = 0\right)\right)\right\} \\
&\geq p_i p_j \mathbb{P}(H_i \neq 0) - p_j \mathbb{E}\left\{\frac{S_i}{H_i} \mathbb{1}\left(H_i \neq 0 \cap H_j = 0\right)\right\} \\
&\geq p_i p_j \mathbb{P}(H_i \neq 0) - p_j \mathbb{P}(H_j = 0).
\end{aligned}$$

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

Therefore, we get a bound on the covariance:

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

The bound can be rewritten as:

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

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

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

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

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

$\square$

Recall that $p^* := \min\{\, p_j \mid j \in [d] \,\}$ and $P_i = \prod_{j=1}^{i-1} p_j^{\binom{i+1}{j+1}}$.

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 [47, Theorem 1] and for comparing a normal to a normal we use [35, Section 6.2].

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

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 [47, Theorem 1]. This gives the bound

$$\begin{aligned}
\sum_{x\in I_i} d_{\mathrm{KS}}(c(x,i)W_i|H_i = x, Z)\,\mathbb{P}(H_i = x) &\leq \sum_{x\in I_i} \frac{\sqrt{10}+3}{6\sqrt{2\pi}} \cdot \frac{p_i^2 + (1-p_i)^2}{\sqrt{xp_i(1-p_i)}}\mathbb{P}(H_i = x)\\
&\leq \frac{\sqrt{10}+3}{6\sqrt{2\pi}} \cdot \frac{p_i^2 + (1-p_i)^2}{\sqrt{(\mathbb{E}\{H_i\} - a)p_i(1-p_i)}}\mathbb{P}(H_i \in I_i).
\end{aligned}$$

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 [35, Section 6.2] (with $f_1 = f_2 = 1$) to obtain

$$\begin{aligned}\sum_{x\in I_i} d_{\mathrm{KS}}\left(c(x,i)Z_i, Z\right)\mathbb{P}(H_i = x) &\leq \sum_{x\in I_i}\left|\frac{x}{\mathbb{E}\{H_i\}} - 1\right|\sqrt{\frac{\mathbb{E}\{H_i\}}{x}}\mathbb{P}(H_i = x)\\ &\leq \max\left\{\left|\frac{\mathbb{E}\{H_i\}+a}{\mathbb{E}\{H_i\}} - 1\right|, \left|\frac{\mathbb{E}\{H_i\}-a}{\mathbb{E}\{H_i\}} - 1\right|\right\}\sqrt{\frac{\mathbb{E}\{H_i\}}{\mathbb{E}\{H_i\}-a}}\mathbb{P}(H_i \in I_i)\\ &\leq \frac{a}{\mathbb{E}\{H_i\}-a}.\end{aligned}$$

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 [34, 43]. 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 [28, 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 $\mathbf{X}(n,\hat{\mathbf{p}}_\mathbf{0})$, where $\hat{\mathbf{p}}_0$ is the observed value of the MLE of $\mathbf{p}$ as given in Equation 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 Equation 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

procedure PERFORMTEST($\mathcal{L}$, α, T, Σ) $\triangleright$ $\mathcal{L}$ 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
    $\hat{\mathbf{p}}_0 \coloneqq$ COMPUTEMLE($\mathcal{L}$) $\triangleright$ Computes the observed value $\hat{\mathbf{p}}_0$ of the MLE $\hat{\mathbf{p}}$
    $(t_1, t_2, \ldots, t_k) \coloneqq$ COMPUTESTATISTIC($\mathcal{L}, T$) $\triangleright$ Computes the observed value of a chosen T
    for all $i \in [k]$ **do**
        $\widehat{\mathbb{E}\{T_i\}} \coloneqq$ ESTIMATEMEAN($\hat{\mathbf{p}}_0$) $\triangleright$ Estimates the mean by replacing $\mathbf{p}$ with $\hat{\mathbf{p}}_0$
        $\widehat{\mathrm{Var}(T_i)} \coloneqq$ ESTIMATEVARIANCE($\hat{\mathbf{p}}_0$) $\triangleright$ Estimates the variance by replacing $\mathbf{p}$ with $\hat{\mathbf{p}}_0$
        $w_i \coloneqq \widehat{\mathrm{Var}(T_i)}^{-\frac{1}{2}}(t_i - \widehat{\mathbb{E}\{T_i\}})$ $\triangleright$ Normalises the statistic
    end for
    $w \coloneqq (w_1, w_2, \ldots, w_k)$
    if $w^T \Sigma^{-1} w > \chi_k^2(1-\alpha)$ **then** $\triangleright$ $\chi_k^2(p)$ is the quantile function of the χ_k^2 distribution
        return False $\triangleright$ The null is rejected
    else
        return True $\triangleright$ Insufficient evidence to reject the null
    end if
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** $\mathbb{X}(\mathcal{X}_n, \phi_1, \phi_2, \dots \phi_{n-1})$ as the random simplicial complex with vertex set $\mathcal{X}_n$ such that any i-simplex $s = \{s_1, s_2, \dots, s_{i+1}\} \subseteq \mathcal{X}_n$ is included with probability $\phi_i(s_1, s_2, \dots, s_{i+1})$ 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 $\mathbb{X}(\mathcal{X}_{150}, \phi_1, \phi_2, \dots \phi_{149})$ 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 $\phi_i(x_1, x_2, \dots, x_{i+1}) = 0$ 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 $\mathbb{X}(\mathcal{X}_{75}, \phi_1, \phi_2, \dots \phi_{74})$ where $\phi_1(x_1, x_2) = 0.5$, $\phi_2(x_1, x_2, x_3) = \mathbb{1}\left(A(x_1, x_2, x_3) \leq 0.09\right)$ and lastly we let $\phi_i(x_1, x_2, \dots, x_{i+1}) = 0$ 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. The area threshold 0.09 was chosen empirically as one example of a threshold where geometry detection is possible. 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 is $\mathbb{X}(\mathcal{X}_{75}, \phi_1, \phi_2, \dots \phi_{74})$ where $\phi_1(x_1, x_2) = \mathbb{1}\left(\|x_1 - x_2\|_2 \leq 0.4924\right)$ and $\phi_i(x_1, x_2, \dots, x_{i+1}) = 0.5$ 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 $0 \leq \epsilon_1 \leq \epsilon_2$ we define the model $\mathbb{X}(\mathcal{X}_{150}, \phi_1, \phi_2, \phi_3^{(\epsilon_1, \epsilon_2)}, \dots, \phi_{149})$ where we set the functions $\phi_1(x_1, x_2) = 0.5$, $\phi_2(x_1, x_2, x_3) = 0.5$, $\phi_3(x_1, x_2, x_3, x_4) = \mathbb{1}\left(A(x_1, x_2, x_3, x_4) \leq \epsilon_1\right) + 0.5 \times$

$\mathbb{1}\left(A(x_1,x_2,x_3,x_4) \in (\epsilon_1,\epsilon_2)\right)$ and $\phi_i(x_1,x_2,\ldots,x_{i+1}) = 0$ for all $i \neq 1,2,3$. 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 $0 \leq \epsilon_1 \leq \epsilon_2$ we define $\mathbb{X}(\mathcal{X}_{75},\phi_1,\phi_2^{(\epsilon_1,\epsilon_2)},\ldots\phi_{74})$ where $\phi_1(x_1,x_2) = 0.5$, $\phi_2(x_1,x_2,x_3) = \mathbb{1}\left(A(x_1,x_2,x_3) \leq \epsilon_1\right) + 0.5 \times \mathbb{1}\left(A(x_1,x_2,x_3) \in (\epsilon_1,\epsilon_2)\right)$ and $\phi_i(x_1,x_2,\ldots,x_{i+1}) = 0$ for all $i \neq 1,2$. 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 $0 \leq \epsilon_1 \leq \epsilon_2$ we define the model $\mathbb{X}(\mathcal{X}_{75},\phi_1,\phi_2^{(\epsilon_1,\epsilon_2)},\ldots\phi_{74})$ where $\phi_1(x_1,x_2) = \mathbb{1}\left(\|x_1 - x_2\|_2 \leq \epsilon_1\right) + 0.5 \times \mathbb{1}\left(\|x_1 - x_2\|_2 \in (\epsilon_1,\epsilon_2)\right)$ as well as $\phi_i(x_1,x_2,\ldots,x_{i+1}) = 0.5$ for all $i \neq 1$. 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 [32, 12], 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 can be seen in Figure 2, 3, and 4 respectively. The data used to generate the respective figures can be found in Table 1, 2, and 3 in Appendix C.

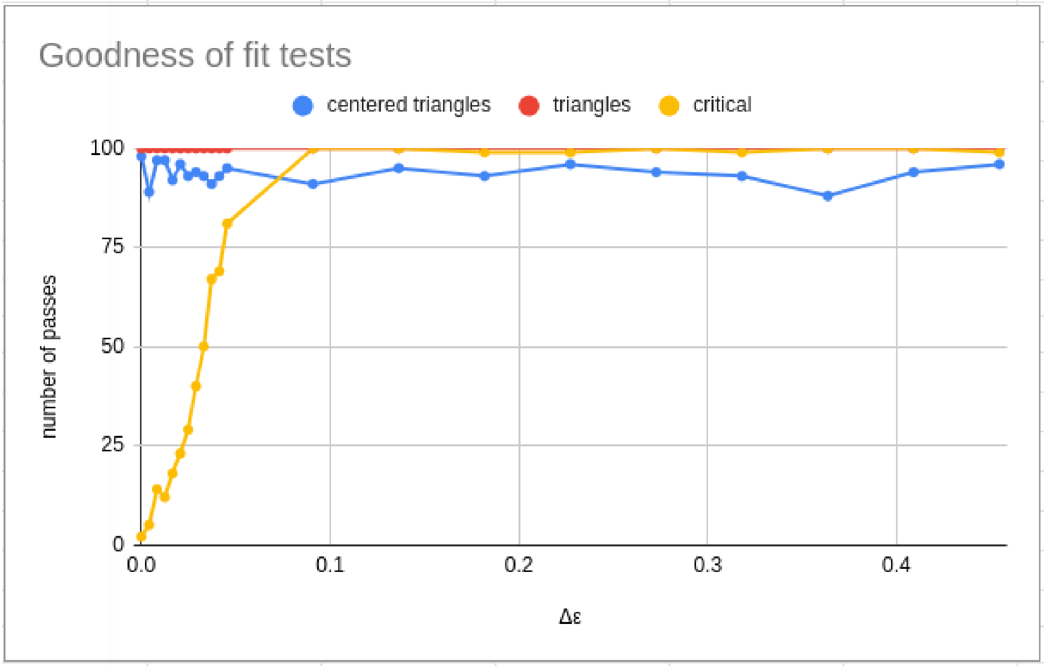

Figure 2: Goodness of fit testing results of the tetrahedron model. The x-axis corresponds to $\Delta\epsilon := \epsilon_2 - \epsilon_1$.

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

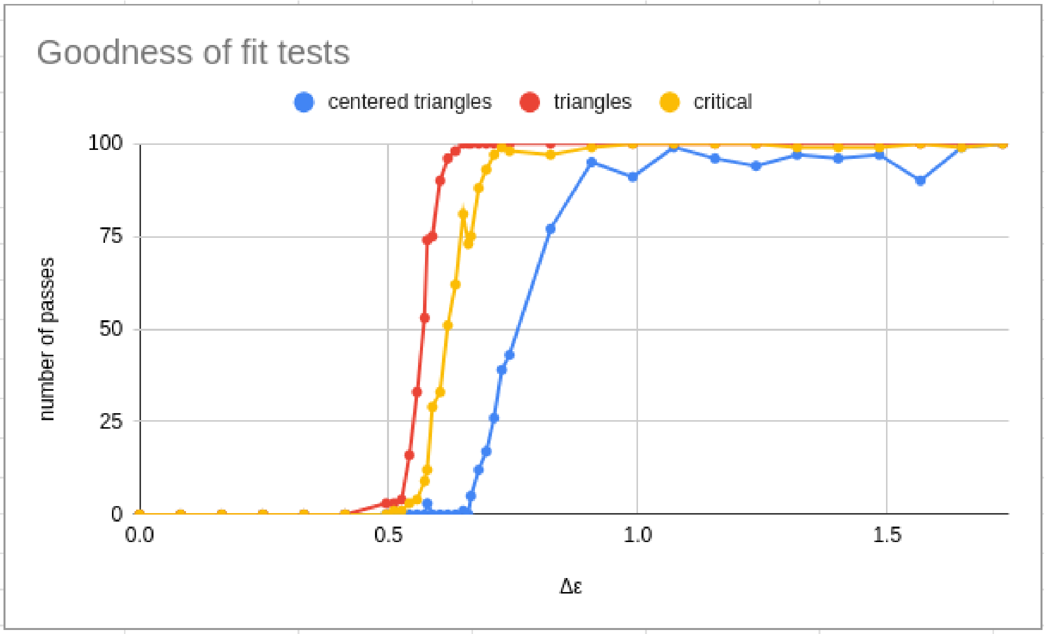

Figure 4: Goodness of fit testing results of the edge model. The x-axis corresponds to $\Delta\epsilon \coloneqq \epsilon_2 - \epsilon_1$.

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 [12, 38]. 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 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. 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 [49], 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

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

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

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

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 $\mathbb{X}^{(i)} \subseteq \mathbb{X}$.

Given a subset $s \subseteq [n]$ of size $m \geq k$ let $\binom{s}{i}$ be the set of i-subsets of s, which, if needed, we can turn into a sequence by ordering the subsets according to the lexicographical ordering.

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 4.1. 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

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

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 4.2. 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 [24, Section 10.7], [25], [22, 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

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

are all i.i.d. Bernoulli(p) but they are of different size. Because all vertices a 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 $Y_s = f_i(\mathcal{X}_\phi^{(1)}, \mathcal{X}_\phi^{(2)}, \dots, \mathcal{X}_\phi^{(k)})$ and

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

where $\mu_s = \mathbb{E}\left\{Y_s\right\}$ and $\sigma_i = \sqrt{\mathrm{Var}(S_{n,m_i}^{(k)}(f_i))}$, 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 $W = (W_1, W_2, \dots, W_d) \in \mathbb{R}^d$. 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

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

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 4.3. 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

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

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 4.4. *Let $Z \sim \mathrm{MVN}(0, \mathrm{Id}_{d\times d})$ and Σ be the covariance matrix of W, where W satisfies Assumption 4.3.*

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

$$\left|\mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}}Z)\right| \leq |h|_3\, B_{4.4} n^{-\frac{k'+1}{2}}.$$

2.

$$\sup_{A\in\mathcal{K}} \left|\mathbb{P}(W\in A) - \mathbb{P}(\Sigma^{\frac{1}{2}}Z \in A)\right| \leq 2^{\frac{7}{2}} 3^{-\frac{3}{4}} d^{\frac{3}{16}} B_{4.4}^{\frac{1}{4}} n^{-\frac{k'+1}{8}},$$

where

$$B_{4.4} = \frac{4}{3}\frac{\beta(k'+1)^{\frac{3}{2}(k'+1)}}{((k'+1)!)^2}(\sqrt{2}m)^{5m+2-3(k'+1)}\sum_{(i,j,k)}^{d}\frac{1}{\sqrt{\alpha_i\alpha_j\alpha_k}}$$

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 4.3, 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:

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

Using the language of Corollary B.3 and using the last part of Assumption 4.3 we can say that $\beta_{ijl} \leq \beta(\sigma_i \sigma_j \sigma_l)$. Moreover we can bound α_{ij} in Corollary B.3 by (4). The bound on the quantity $B_{B.3}$ from the corollary is:

$$\begin{aligned}
B_{B.3} &= \frac{1}{3} \sum_{(i,j,k)} |\mathbb{I}_i| \, \alpha_{ij} \left(\frac{3\alpha_{ik}}{2} + 2\alpha_{jk} \right) \beta_{ijk} \\
&\leq \frac{1}{3} \sum_{(i,j,k)} \frac{n^{m_i}}{m_i!} n^{m_j - 1 - k'} \frac{m_i^{m_i+1}}{(k'+1)!} \left(\frac{3}{2} n^{m_k - 1 - k'} \frac{m_i^{m_i+1}}{(k'+1)!} + 2n^{m_k - 1 - k'} \frac{m_j^{m_j+1}}{(k'+1)!} \right) \beta_{i,j,k}
\end{aligned}$$

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

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 $(2m_i^2 - m_i(1+k'))^{\frac{m_i}{2} - \frac{k'+1}{2}} \leq (2m^2)^{\frac{m}{2} - \frac{k'+1}{2}}$ and so

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

□

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 $\xi_\alpha^{(k+1)} \sim \text{Bernoulli}(p_k)$ 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 collection 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 Equation 3 is a sequence of the random elements, corresponding to the $(i+1)$-subsets of s, ordered lexicographically. 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

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

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

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

For example, if $\mathcal{L} = \{1, 2, 3, 4, 12, 13, 23, 24, 34, 123, 234\}$ (two 2-simplices with a common 1-simplex), then $\mathbb{E}\{T_{\mathcal{L}}\} = 3\binom{n}{4}p_1^5 p_2^2$. 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.

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

□

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 $\sigma_{\mathcal{L}}^2 = \mathrm{Var}(T_{\mathcal{L}})$ and $\sigma_{\mathcal{M}}^2 = \mathrm{Var}(T_{\mathcal{M}})$. Then assuming that for all i the parameters p_i stay constant, for any fixed $m, l \in \mathbb{N}$ we have:*

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

Proof. Assume w.l.o.g. that $m \le l$. 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. Hence we see that

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

The limit of the ratio is determined by the ratio of the highest order term from the enumerator and the highest order term from the denominator. Those terms will correspond to $k = k'+2$ 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}, and the summands $\operatorname{Cov}(X_s^{\mathcal{L}}, X_u^{\mathcal{M}})$ do not depend on n. 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:

$$\begin{aligned}
&\lim_{n\to\infty} \sigma_{\mathcal{L}}^{-1}\sigma_{\mathcal{M}}^{-1}\operatorname{Cov}(T_{\mathcal{L}}, T_{\mathcal{M}}) = \\
&\lim_{n\to\infty} \binom{n}{l+m-2-k'}\binom{l+m-2-k'}{l}\binom{l}{k'+2} L_{k'+1}M_{k'+1}\left(p_{k'+1}^{-1}-1\right)\prod_{i=k'+1}^{l-1} p_i^{e_i(\mathcal{L})+e_i(\mathcal{M})} \\
&\left\{\binom{n}{2l-k'-2}\binom{2l-k'-2}{l}\binom{l}{k'+2} L_{k'+1}^2\left(p_{k'+1}^{-1}-1\right)\prod_{i=k'+1}^{l-1} p_i^{2e_i(\mathcal{L})}\right\}^{-\frac{1}{2}} \\
&\left\{\binom{n}{2m-k'-2}\binom{2m-k'-2}{m}\binom{m}{k'+2} M_{k'+1}^2\left(p_{k'+1}^{-1}-1\right)\prod_{i=k'+1}^{m-1} p_i^{2e_i(\mathcal{M})}\right\}^{-\frac{1}{2}} \\
&= \lim_{n\to\infty} L_{k'+1}M_{k'+1}\left(p_{k'+1}^{-1}-1\right)\prod_{i=k'+1}^{l-1} p_i^{e_i(\mathcal{L})+e_i(\mathcal{M})} \\
&\left\{L_{k'+1}^2\left(p_{k'+1}^{-1}-1\right)\prod_{i=k'+1}^{l-1} p_i^{2e_i(\mathcal{L})}\prod_{i=k'+1}^{m-1} M_{k'+1}^2\left(p_{k'+1}^{-1}-1\right)\prod_{i=k'+1}^{m-1} p_i^{2e_i(\mathcal{M})}\right\}^{-\frac{1}{2}} \\
&= 1.
\end{aligned}$$

□

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 Equation 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

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

where $\mu_s = \mathbb{E}\left\{Y_s\right\}$ 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 4.4, we need to show that Assumption 4.3 holds and find the quantities α_i and β from the assumption. The following proposition 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 4.4. 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 4.4, 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*

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

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

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

Proof. Take $s = (\phi, i), u = (\psi, i) \in \mathbb{I}_i$ such that $|\phi\cap\psi| \geq k'+2$. Using Proposition 5.3 we get that

$$\begin{aligned}\mathrm{Cov}(Y_s, Y_u) &\geq p_{k'+1}^{2e_{k'+1}(\mathcal{L}_i)-1}\prod_{j=k'+2}^{k_i} p_j^{2e_j(\mathcal{L}_i)} + (|[\mathcal{L}_i]|^2 - 1)\prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)} - |[\mathcal{L}_i]|^2 \prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)}\\ &= \prod_{j=k'+1}^{k_i} p_j^{2e_j(\mathcal{L}_i)}\left\{p_{k'+1}^{-1} - 1\right\} > 0.\end{aligned}$$

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 4.3.

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 $\mu_i = \prod_{k=k'+1}^{k_i} p_k^{e_k(\mathcal{L}_i)}$,

$$\begin{aligned}
&\mathbb{E}\left|\{Y_s - \mu_s\}\{Y_t - \mu_t\}\{Y_u - \mu_u\}\right| \leq \\
&\left\{\prod_{k=k'+1}^{k_i} p_k^{e_k(\mathcal{L}_i)} \prod_{k=k'+1}^{k_j} p_k^{e_k(\mathcal{L}_j)} \left(1 - \prod_{k=k'+1}^{k_i} p_k^{e_k(\mathcal{L}_i)}\right)\left(1 - \prod_{k=k'+1}^{k_j} p_k^{e_k(\mathcal{L}_j)}\right)\right\}^{\frac{1}{2}} \\
&\leq \prod_{k=k'+1}^{\min_{i\in[d]}(k_i)} p_k^{\min_{i\in[d]}(e_k(\mathcal{L}_i))} \left(1 - \prod_{k=k'+1}^{\max_{i\in[d]}(k_i)} p_k^{\max_{i\in[d]}(e_k(\mathcal{L}_i))}\right).
\end{aligned}$$

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 4.4 and bound the quantity $B_{4.4}$:

$$\begin{aligned}
B_{4.4} &\leq \frac{4}{3}\frac{(k'+2)^{\frac{3}{2}(k'+2)}}{((k'+2)!)^2}(\sqrt{2}m)^{5m+2-3(k'+2)} d^3 \beta \left\{(p_{k'+1}^{-1} - 1)\prod_{k=k'+1}^{\min_{i\in[d]}(k_i)} p_k^{\min_{i\in[d]}(e_k(\mathcal{L}_i))}\right\}^{-\frac{3}{2}} \\
&\leq \frac{4}{3}\frac{(k'+2)^{\frac{3}{2}(k'+2)}}{((k'+2)!)^2}(\sqrt{2}m)^{5m+2-3(k'+2)} d^3 \gamma.
\end{aligned}$$

□

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 and taken from [49].

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 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 [11, Section 2.2]. For any $k \in [n-1]$ and $\alpha \in \binom{[n]}{k+1}$ let $\xi_\alpha^{(k+1)} \sim \text{Bernoulli}(p_k)$.

Fix $s \in \binom{[n]}{k}$. Define the variables $X_s^+ = \mathbb{1}\left(s \text{ matches with its coface given it is a simplex}\right)$ as well as $X_s^- = \mathbb{1}\left(s \text{ matches with its face given it is a simplex}\right)$. The events that the two variables indicate are disjoint. By Lemma 6.1 we can see that $X_s^+ = 1 - \prod_{i=1}^{\min(s)-1}\left(1 - \prod_{j=k'+1}^{k}\prod_{\substack{\alpha\subseteq s\\|\alpha|=j}} \xi_{\alpha\cup\{i\}}^{(j+1)}\right)$ and the following: $X_s^- = \prod_{i=1}^{\min(s)-1}\left(1 - \prod_{j=k'+1}^{k-1}\prod_{\substack{\alpha\subseteq s_-\\|\alpha|=j}} \xi_{\alpha\cup\{i\}}^{(j+1)}\right)$, where $s_- := s \setminus \{\min(s)\}$. Hence,

$$\begin{aligned}
&\mathbb{1}\left(s \text{ is a critical simplex}\right) = \mathbb{1}\left(s \in \mathbf{X}(n,\mathbf{p})\right)\left(1 - (X_s^+ + X_s^-)\right)\\
&= \prod_{j=k'+1}^{k-1}\prod_{\substack{\alpha\subseteq s\\|\alpha|=j+1}} \xi_\alpha^{(j+1)}\left[\prod_{i=1}^{\min(s)-1}\left(1 - \prod_{j=k'+1}^{k}\prod_{\substack{\beta\subseteq s\\|\beta|=j}}\xi_{\beta\cup\{i\}}^{(j+1)}\right) - \prod_{i=1}^{\min(s)-1}\left(1 - \prod_{j=k'+1}^{k-1}\prod_{\substack{\beta\subseteq s_-\\|\beta|=j}}\xi_{\beta\cup\{i\}}^{(j+1)}\right)\right].
\end{aligned}$$

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

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

Note that this random variable does not fit into the framework of generalised U-statistics as in Definition 4.1 because the summands depend not only on the variables that are indexed by the subset s.

6.1 Moments and the limiting covariance

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

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

Proof.

$$\begin{aligned}
\mathbb{E}\left\{T_{k+1}\right\} &= \sum_{l=1}^{n-k} \sum_{\substack{s\in\binom{[n]}{k+1}\\ \min(s)=l}} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}} \left[\left(1-\prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}}\right)^{l-1} - \left(1-\prod_{i=k'+1}^{k} p_i^{\binom{k}{i}}\right)^{l-1}\right] \\
&= \sum_{l=1}^{n-k} \binom{n-l}{k} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}} \left[\left(1-\prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}}\right)^{l-1} - \left(1-\prod_{i=k'+1}^{k} p_i^{\binom{k}{i}}\right)^{l-1}\right] \\
&\leq \binom{n-1}{k} \prod_{i=k'+1}^{k} p_i^{\binom{k+1}{i+1}} \sum_{l=0}^{\infty} \left[\left(1-\prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i}}\right)^{l} - \left(1-\prod_{i=k'+1}^{k} p_i^{\binom{k}{i}}\right)^{l}\right] \\
&= \binom{n-1}{k} \prod_{i=k'+1}^{k+1} p_i^{\binom{k+1}{i+1}-\binom{k+1}{i}} \left(1-\prod_{i=k'+1}^{k} p_i^{\binom{k}{i-1}}\right).
\end{aligned}$$

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

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

□

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

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

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

Let $W_i = \sum_{s \in \mathbb{I}_i} X_s$ and $W = (W_{k'+1}, W_{k'+2}, \ldots, W_{k'+d}) \in \mathbb{R}^d$. 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*
$$\left| \mathbb{E}h(W) - \mathbb{E}h(\Sigma^{\frac{1}{2}} Z) \right| \leq B_{6.6.1} \left|h\right|_3 n^{-1-\frac{k'}{2}}.$$

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. We write C for an unspecified positive constant that does not depend on n. For the bounds on the quantity $B_{B.2}$ we use Lemma B.4 and Lemma 6.3 as well as the fact that for fixed $b \in [n], i, j \in$

$[k'+1, k'+d], s \in \mathbb{I}_i$ we have $|\mathbb{D}_j(s)| \le Cn^{j-k'}$ and $|\{(\psi, j) \in \mathbb{D}_j(s) \mid \min(\psi) = b\}| \le Cn^{j-k'-1}$. Also, we assume here that n is large enough for the bound in Lemma 6.3 to apply. 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.

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

□

Acknowledgements

TT acknowledges funding from EPSRC studentship 2275810. VN and GR are supported by the EPSRC grant EP/R018472/1. GR is also funded in part by the EPSRC grant EP/T018445/1. The authors would like to thank Christina Goldschmidt, Heather Harrington, Adrian Röllin and Fang Xiao for helpful discussions.

References

[1] Ayat Ababneh and Matthew Kahle. "Maximal persistence in random clique complexes". In: *Journal of Applied and Computational Topology* (2023), pp. 1–15.

[2] Karim A Adiprasito, Bruno Benedetti, and Frank H Lutz. "Extremal Examples of Collapsible Complexes and Random Discrete Morse Theory". In: *Discrete & Computational Geometry* 57 (2017), pp. 824–853.

[3] Andrew D Barbour, Michal Karoński, and Andrzej Ruciński. "A central limit theorem for decomposable random variables with applications to random graphs". In: *Journal of Combinatorial Theory, Series B* 47.2 (1989), pp. 125–145.

[4] Marc Barthelemy. "Class of models for random hypergraphs". In: *Physical Review E* 106.6 (2022), p. 064310.

[5] Ulrich Bauer. "Ripser: efficient computation of Vietoris–Rips persistence barcodes". In: *Journal of Applied and Computational Topology* 5 (2021), pp. 391–423.

[6] Ulrich Bauer and Abhishek Rathod. “Parameterized inapproximability of Morse matching”. In: *arXiv:2109.04529* (2021).

[7] Ginestra Bianconi. *Higher-Order Networks*. Cambridge University Press, 2021.

[8] Christian Bick, Elizabeth Gross, Heather A Harrington, and Michael T Schaub. “What are higher-order networks?” In: *SIAM Review* 65.3 (2023), pp. 686–731.

[9] Christophe A. N. Biscio, Nicolas Chenavier, Christian Hirsch, and Anne Marie Svane. “Testing goodness of fit for point processes via topological data analysis”. In: *Electronic Journal of Statistics* 14.1 (2020), pp. 1024 –1074.

[10] Omer Bobrowski and Matthew Kahle. “Topology of random geometric complexes: a survey”. In: *Journal of Applied and Computational Topology* 1.3 (2018), pp. 331–364.

[11] Omer Bobrowski and Dmitri Krioukov. “Random simplicial complexes: models and phenomena”. In: *Higher-Order Systems*. Springer, 2022, pp. 59–96.

[12] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. “Testing for high-dimensional geometry in random graphs”. In: *Random Structures & Algorithms* 49.3 (2016), pp. 503–532.

[13] Alessandra Cipriani, Christian Hirsch, and Martina Vittorietti. “Topology-based goodness-of-fit tests for sliced spatial data”. In: *Computational Statistics & Data Analysis* 179 (2023), p. 107655.

[14] Armindo Costa and Michael Farber. “Large random simplicial complexes, I”. In: *Journal of Topology and Analysis* 8.03 (2016), pp. 399–429.

[15] Armindo Costa and Michael Farber. “Large random simplicial complexes, III the critical dimension”. In: *Journal of Knot Theory and Its Ramifications* 26.02 (2017), p. 1740010.

[16] Justin Curry, Robert Ghrist, and Vidit Nanda. “Discrete Morse theory for computing cellular sheaf cohomology”. In: *Foundations of Computational Mathematics* 16 (4 2016), pp. 875 –897.

[17] Peter Eichelsbacher, Benedikt Rednoß, Christoph Thäle, and Guangqu Zheng. “A simplified second-order Gaussian Poincaré inequality in discrete setting with applications”. In: *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*. Vol. 59. 1. Institut Henri Poincaré. 2023, pp. 271–302.

[18] Michael Farber and Lewis Mead. “Random simplicial complexes in the medial regime”. In: *Topology and its Applications* 272 (2020), p. 107065.

[19] Robin Forman. “A user's guide to discrete Morse theory”. In: *Séminaire Lotharingien de Combinatoire* 48 (2002), B48c.

[20] Shaun Harker, Konstantin Mischaikow, Marian Mrozek, and Vidit Nanda. "Discrete Morse theoretic algorithms for computing homology of complexes and maps". In: *Foundations of Computational Mathematics* 14.1 (2014), pp. 151 –184.

[21] Svante Janson. "Poisson approximation for large deviations". In: *Random Structures & Algorithms* 1.2 (1990), pp. 221–229.

[22] Svante Janson. *Gaussian Hilbert spaces*. 129. Cambridge University Press, 1997.

[23] Svante Janson. "Large deviations for sums of partly dependent random variables". In: *Random Structures & Algorithms* 24.3 (2004), pp. 234–248.

[24] Svante Janson and Krzysztof Nowicki. "The asymptotic distributions of generalized U-statistics with applications to random graphs". In: *Probability Theory and Related Fields* 90.3 (1991), pp. 341–375.

[25] Peter de Jong. "A central limit theorem with applications to random hypergraphs". In: *Random Structures & Algorithms* 8.2 (1996), pp. 105–120.

[26] Michael Joswig and Marc E Pfetsch. "Computing optimal discrete Morse functions". In: *SIAM J. Discrete Math.* 20.1 (2004), pp. 11–25.

[27] Matthew Kahle. "Topology of random clique complexes". In: *Discrete Mathematics* 309.6 (2009), pp. 1658–1671.

[28] Matthew Kahle. "Random geometric complexes". In: *Discrete & Computational Geometry* 45.3 (2011), pp. 553–573.

[29] Matthew Kahle. "Sharp vanishing thresholds for cohomology of random flag complexes". In: *Annals of Mathematics* (2014), pp. 1085–1107.

[30] Matthew Kahle. "Topology of random simplicial complexes: a survey". In: *AMS Contemp. Math* 620 (2014), pp. 201–222.

[31] Matthew Kahle and Elizabeth Meckes. "Limit theorems for Betti numbers of random simplicial complexes". In: *Homology, Homotopy and Applications* 15.1 (2013), pp. 343–374.

[32] Gursharn Kaur and Adrian Röllin. "Higher-order fluctuations in dense random graph models". In: *Electronic Journal of Probability* 26 (2021), pp. 1–36.

[33] Johannes Krebs and Christian Hirsch. "Functional central limit theorems for persistent Betti numbers on cylindrical networks". In: *Scandinavian Journal of Statistics* 49.1 (2022), pp. 427–454.

[34] Jing Lei. "A goodness-of-fit test for stochastic block models". In: *The Annals of Statistics* 44.1 (2016), pp. 401–424.

[35] Christophe Ley, Gesine Reinert, and Yvik Swan. “Stein’s method for comparison of univariate distributions”. In: *Probability Surveys* 14 (2017), pp. 1 –52.

[36] Nathan Linial and Roy Meshulam. “Homological connectivity of random 2-complexes”. In: *Combinatorica* 26.4 (2006), pp. 475–487.

[37] Nathan Linial and Yuval Peled. “On the phase transition in random simplicial complexes”. In: *Annals of Mathematics* (2016), pp. 745–773.

[38] Suqi Liu and Miklós Z Rácz. “Phase transition in noisy high-dimensional random geometric graphs”. In: *arXiv preprint arXiv:2103.15249* (2021).

[39] WG McGinley and Robin Sibson. “Dissociated random variables”. In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 77. 1. Cambridge University Press. 1975, pp. 185–188.

[40] Konstantin Mischaikow and Vidit Nanda. “Morse theory for filtrations and efficient computation of persistent homology”. In: *Discrete and Computational Geometry* 50.2 (2013), pp. 330–353.

[41] Andrew Newman. “One-sided sharp thresholds for homology of random flag complexes”. In: *arXiv preprint arXiv:2108.04299* (2021).

[42] Mark Newman. *Networks*. Oxford University Press, 2018.

[43] Sarah Ouadah, Stéphane Robin, and Pierre Latouche. “Degree-based goodness-of-fit tests for heterogeneous random graph models: Independent and exchangeable cases”. In: *Scandinavian Journal of Statistics* 47.1 (2020), pp. 156–181.

[44] Agniva Roy and D Yogeshwaran. “Random clique complex process inside the critical window”. In: *arXiv preprint arXiv:2303.17535* (2023).

[45] Andrzej Ruciński. “When are small subgraphs of a random graph normally distributed?” In: *Probability Theory and Related Fields* 78.1 (1988), pp. 1–10.

[46] Gennady Samorodnitsky and Takashi Owada. “Large deviations for subcomplex counts and Betti numbers in multiparameter simplicial complexes”. In: *Random Structures & Algorithms* (2023).

[47] Jona Schulz. “The optimal Berry-Esseen constant in the binomial case”. PhD thesis. University of Trier, 2016.

[48] Edwin Spanier. *Algebraic Topology*. McGraw-Hill, 1966.

[49] Tadas Temčinas, Vidit Nanda, and Gesine Reinert. “Multivariate Central Limit Theorems for Random Clique Complexes”. In: *Journal of Applied and Computational Topology* (2023+).

[50] Naya Yerolemou and Vidit Nanda. “Morse theory for complexes of groups”. In: *arXiv:2203.00539 [math.GR]* (2022).

[51] Zhuo-Song Zhang. “Berry–Esseen bounds for generalized U-statistics”. In: *Electronic Journal of Probability* 27 (2022), pp. 1–36.

[52] Konstantin Zuev, Or Eisenberg, and Dmitri Krioukov. “Exponential random simplicial complexes”. In: *Journal of Physics A: Mathematical and Theoretical* 48.46 (2015), p. 465002.

A Proof of Lemmas 6.3 and 6.4

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

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

where

$$\begin{aligned}
V_1 = {} & 2\sum_{l<m}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}\binom{n-m}{2k+1-j-q}\binom{2k+1-j-q}{k}\binom{k}{j-1}\binom{m-l+1}{q-1}\\
& \Big\{\rho^+(k+1)^2\rho^+(j)^{-1}\tau(l,m,q,k,j-1)\\
& \Big[\big(1-2\rho(k)+\rho(k)^2\rho(j-1)^{-1}\big)^{l-1}-\big(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j-1)^{-1}\big)^{l-1}\Big]\\
& +\rho^+(k+1)^2\rho^+(j)^{-1}\tau(l,m,q,k+1,j)\\
& \Big[\big(1-2\rho(k+1)+\rho(k+1)^2\rho(j)^{-1}\big)^{l-1}-\big(1-\rho(k+1)-\rho(k)+\rho(k)\rho(k+1)\rho(j)^{-1}\big)^{l-1}\Big]-\mu(l)\mu(m)\Big\};\\
V_2 = {} & 2\sum_{l<m}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}\binom{n-m}{2k+1-j-q}\binom{2k+1-j-q}{k}\binom{k}{j}\binom{m-l+1}{q-1}\\
& \Big\{\rho^+(k+1)^2\rho^+(j)^{-1}\tau(l,m,q,k,j)\\
& \Big[\big(1-2\rho(k)+\rho(k)^2\rho(j)^{-1}\big)^{l-1}-\big(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\big)^{l-1}\Big]\\
& +\rho^+(k+1)^2\rho^+(j)^{-1}\tau(l,m,q,k+1,j)\\
& \Big[\big(1-2\rho(k+1)+\rho(k+1)^2\rho(j)^{-1}\big)^{l-1}-\big(1-\rho(k+1)-\rho(k)+\rho(k)\rho(k+1)\rho(j)^{-1}\big)^{l-1}\Big]-\mu(l)\mu(m)\Big\};\\
V_3 = {} & \sum_{l=1}^{n-k}\sum_{j=k'+1}^{k}\binom{n-l}{2k+1-j}\binom{2k+1-j}{k}\binom{k}{j-1}\Big\{\rho^+(k+1)^2\rho^+(j)^{-1}\\
& \Big[\big(1-2\rho(k+1)+\rho(k+1)^2\rho(j)^{-1}\big)^{l-1}+\big(1-2\rho(k)+\rho(k)^2\rho(j-1)^{-1}\big)^{l-1}\\
& -2\big(1-\rho(k)-\rho(k+1)+\rho(k)\rho(k+1)\rho(j-1)^{-1}\big)^{l-1}\Big]-\mu(l)^2\Big\};\\
V_4 = {} & \sum_{l=1}^{n-k}\binom{n-l}{k}\left\{\mu(l)-\mu(l)^2\right\}.
\end{aligned}$$

Here we have used the following abbreviations:

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

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

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 [49, Lemma 34]. For $s\in\binom{[n]}{k+1}$ recall that $s_-=s\setminus\{\min(s)\}$. We write:

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

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

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

$$\mathrm{Var}(T_{k+1})=\sum_{s\in\binom{[n]}{k+1}}\mathrm{Var}(Z_sY_s)+\sum_{\substack{s\neq t\in\binom{[n]}{k+1}\\\min(s)\neq\min(t)}}\mathrm{Cov}(Z_sY_s,Z_tY_t)+\sum_{\substack{s\neq t\in\binom{[n]}{k+1}\\\min(s)=\min(t)}}\mathrm{Cov}(Z_sY_s,Z_tY_t).\tag{9}$$

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

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

Now we see:

$$\begin{aligned}\sum_{s\in\binom{[n]}{k+1}}\mathrm{Var}(Z_sY_s)&=\sum_{l=1}^{n}\sum_{\substack{s\in\binom{[n]}{k+1}\\\min(s)=l}}\left(\mathbb{E}\left\{(Z_sY_s)^2\right\}-\mathbb{E}\left\{Z_sY_s\right\}^2\right)\\&=\sum_{l=1}^{n-k}\binom{n-l}{k}\left(\mathbb{P}(Z_sY_s=1)-\mathbb{P}(Z_sY_s=1)^2\right)=\sum_{l=1}^{n-k}\binom{n-l}{k}\left\{\mu(l)-\mu(l)^2\right\}=V_4.\end{aligned}$$

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_sY_s and Z_tY_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 $\mathbb{P}(Z_sZ_t=1)=\prod_{i=k'+1}^{k}p_i^{2\binom{k+1}{i+1}-\binom{|s\cap t|}{i}}$. Also, $Y_sY_t=Y_s^+Y_t^++Y_s^-Y_t^--Y_s^+Y_t^--Y_s^-Y_t^+$. 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 variable $\prod_{\substack{i=\min(s)\\i\notin s}}^{\min(t)-1}\left(1-\prod_{j=k'+1}^{k+1}\prod_{\substack{\alpha\subseteq t\\|\alpha|=j}}\xi_{\alpha\cup\{i\}}^{(j+1)}\right)$ are independent of Z_sZ_t. Recall the notation $[a,b]=\{a,a+1,\dots,b\}$ for two positive integers $a\le b$. 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:

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

Using the described strategy we get:

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

Now we are ready to calculate the covariance:

$$\begin{aligned}
&\mathrm{Cov}(Z_sY_s, Z_tY_t) = \mathbb{E}\left\{Z_sZ_tY_s^+Y_t^+\right\} + \mathbb{E}\left\{Z_sZ_tY_s^-Y_t^-\right\} - \mathbb{E}\left\{Z_sZ_tY_s^+Y_t^-\right\} - \mathbb{E}\left\{Z_sZ_tY_s^-Y_t^+\right\} \\
&- \mathbb{E}\left\{Z_sY_s\right\}\mathbb{E}\left\{Z_tY_t\right\} \\
&= \mathbb{P}(Z_sZ_t = 1)\Big\{\mathbb{P}\left(Y_s^+Y_t^+ = 1|Z_sZ_t = 1\right) + \mathbb{P}\left(Y_s^-Y_t^- = 1|Z_sZ_t = 1\right) \\
&- \mathbb{P}\left(Y_s^+Y_t^- = 1|Z_sZ_t = 1\right) - \mathbb{P}\left(Y_s^-Y_t^+ = 1|Z_sZ_t = 1\right)\Big\} - \mathbb{P}(Z_sY_s = 1)\mathbb{P}(Z_tY_t = 1) \\
&= \prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1}-\binom{|s\cap t|}{i+1}}(\pi(l,m,k+1,k+1,|s\cap t|,|s\cap t|,q) + \pi(l,m,k,k,|s_-\cap t_-|,|s\cap t_-|,q) \\
&- \pi(l,m,k,k+1,|s\cap t_-|,|s\cap t_-|,q) - \pi(l,m,k+1,k,|s_-\cap t|,|s\cap t|,q)) - \mu(l)\mu(m).
\end{aligned}$$

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

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

and

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

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 [49] 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

$$\begin{aligned}
&\sum_{\substack{s\neq t\in\binom{[n]}{k+1}\\ \min(s)\neq\min(t)}} \mathrm{Cov}(Z_sY_s, Z_tY_t)\\
=&\sum_{l=1}^{n-k}\sum_{m=i+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}\sum_{(s,t)\in\Gamma^+_{k+1}(l,m,j,q)}\mathrm{Cov}(Z_sY_s, Z_tY_t)\\
&+\sum_{l=1}^{n-k}\sum_{m=i+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}\sum_{(s,t)\in\Gamma^-_{k+1}(l,m,j,q)}\mathrm{Cov}(Z_sY_s, Z_tY_t)\\
&+\sum_{m=1}^{n-k}\sum_{l=j+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,l-m)}\sum_{(s,t)\in\Gamma^+_{k+1}(m,l,j,q)}\mathrm{Cov}(Z_sY_s, Z_tY_t)\\
&+\sum_{m=1}^{n-k}\sum_{l=j+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,l-m)}\sum_{(s,t)\in\Gamma^-_{k+1}(m,l,j,q)}\mathrm{Cov}(Z_sY_s, Z_tY_t)\\
=&\sum_{l=1}^{n-k}\sum_{m=i+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}|\Gamma^+_{k+1}(l,m,j,q)|\Big\{\prod_{i=k'+1}^{k}p_i^{2\binom{k+1}{i+1}-\binom{j}{i+1}}(\pi(l,m,k+1,k+1,j,j,q)\\
&+\pi(l,m,k,k,j-1,j-1,q)-\pi(l,m,k,k+1,j-1,j-1,q)-\pi(l,m,k+1,k,j,j,q))-\mu(l)\mu(m)\Big\}\\
&+\sum_{l=1}^{n-k}\sum_{m=i+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}|\Gamma^-_{k+1}(l,m,j,q)|\Big\{\prod_{i=k'+1}^{k}p_i^{2\binom{k+1}{i+1}-\binom{j}{i+1}}(\pi(l,m,k+1,k+1,j,j,q)\\
&+\pi(l,m,k,k,j,j,q)-\pi(l,m,k,k+1,j,j,q)-\pi(l,m,k+1,k,j,j,q))-\mu(l)\mu(m)\Big\}\\
&+\sum_{m=1}^{n-k}\sum_{l=j+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,l-m)}|\Gamma^+_{k+1}(m,l,j,q)|\Big\{\prod_{i=k'+1}^{k}p_i^{2\binom{k+1}{i+1}-\binom{j}{i+1}}(\pi(m,l,k+1,k+1,j,j,q)\\
&+\pi(m,l,k,k,j-1,j-1,q)-\pi(m,l,k,k+1,j-1,j-1,q)-\pi(m,l,k+1,k,j,j,q))-\mu(l)\mu(m)\Big\}\\
&+\sum_{m=1}^{n-k}\sum_{l=j+1}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,l-m)}|\Gamma^-_{k+1}(m,l,j,q)|\Big\{\prod_{i=k'+1}^{k}p_i^{2\binom{k+1}{i+1}-\binom{j}{i+1}}(\pi(m,l,k+1,k+1,j,j,q)\\
&+\pi(m,l,k,k,j,j,q)-\pi(m,l,k,k+1,j,j,q)-\pi(m,l,k+1,k,j,j,q))-\mu(l)\mu(m)\Big\} = V_1+V_2.
\end{aligned}$$

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

$$\sum_{\substack{s\neq t\in\binom{[n]}{k+1}\\ \min(s)=\min(t)}} \mathrm{Cov}(Z_sY_s, Z_tY_t) = \sum_{l=1}^{n-k}\sum_{j=k'+1}^{k}\sum_{(s,t)\in\Gamma_{k+1}(l,l,j,0)} \mathrm{Cov}(Z_sY_s, Z_tY_t)$$
$$= \sum_{l=1}^{n-k}\sum_{j=k'+1}^{k}\binom{n-l}{2k+1-j}\binom{2k+1-j}{k}\binom{k}{j-1}\Big\{\prod_{i=k'+1}^{k} p_i^{2\binom{k+1}{i+1}-\binom{j}{i+1}}$$
$$[\pi(l,l,k+1,k+1,j,0,0) + \pi(l,l,k,k,j-1,0,0) - 2\pi(l,l,k+1,k,j-1,0,0)] - \mu(l)^2\Big\} = V_3.$$

□

Proof for Lemma 6.3. Fix $k'+1 \le k \le n-1$ 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:

$$V_1 \ge -2\rho^+(k+1)^2\sum_{l<m}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}\binom{n-m}{2k+1-j-q}\binom{2k+1-j-q}{k}\binom{k}{j-1}\binom{m-l+1}{q-1}$$
$$\Big\{\rho^+(j)^{-1}\tau(l,m,q,k,j-1)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j-1)^{-1}\right)^{l-1}$$
$$+\rho^+(j)^{-1}\tau(l,m,q,k+1,j)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1}$$
$$+(1-\rho(k+1))^{l+m-2} + (1-\rho(k))^{l+m-2}\Big\};$$
$$V_2 \ge -2\rho^+(k+1)^2\sum_{l<m}^{n-k}\sum_{j=k'+1}^{k}\sum_{q=1}^{\min(k+1,m-l)}\binom{n-m}{2k+1-j-q}\binom{2k+1-j-q}{k}\binom{k}{j}\binom{m-l+1}{q-1}$$
$$\Big\{\rho^+(j)^{-1}\tau(l,m,q,k,j)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1}$$
$$+\rho^+(j)^{-1}\tau(l,m,q,k+1,j)\left(1-\rho(k+1)-\rho(k)+\rho(k+1)\rho(k)\rho(j)^{-1}\right)^{l-1}$$
$$+(1-\rho(k+1))^{l+m-2} + (1-\rho(k))^{l+m-2}\Big\}.$$

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:

$$
\begin{aligned}
V_3 &\geq \rho^+(k+1)^2 \sum_{l=1}^{n-k} \binom{n-l}{2k-k'} \binom{2k-k'}{k} \binom{k}{k'} \{(1-2\rho(k+1)+\rho(k+1)^2\rho(k'+1)^{-1})^{l-1} - (1-\rho(k+1))^{2l-2}\} \\
&- 2\rho^+(k+1)^2 \sum_{l=1}^{n-k} \sum_{j=k'+2}^{k} \binom{n-l}{2k+1-j} \binom{2k+1-j}{k} \binom{k}{j-1} \\
&\left\{\rho^+(j)^{-1}\left(1-\rho(k)-\rho(k+1)+\rho(k)\rho(k+1)\rho(j-1)^{-1}\right)^{l-1} + (1-\rho(k+1))^{2l-2}\right\} \\
&= \rho^+(k+1)^2 R_4 - 2\rho^+(k+1)^2 R_3;
\end{aligned}
$$

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 $V_4 \geq 0$. Hence, we have:

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

Let us now upper bound R_1:

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

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:

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

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:

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

$$
\begin{aligned}
&\le \sum_{l=1}^{n-k} \sum_{j=k'+2}^{k} \frac{n^{2k+1-j}}{(2k+1-j)!} \frac{(2k+1-j)^k}{k!} \frac{k^{j-1}}{(j-1)!} (\rho^+(k)^{-1}+1)(1-\rho(k+1))^{l-1} \\
&\le (k-k'-1) \frac{n^{2k-k'-1}}{(k+1)!} \frac{(2k-k'-1)^k}{k!} \frac{k^k}{(k'+1)!} (\rho^+(k)^{-1}+1) \sum_{l=1}^{\infty} (1-\rho(k+1))^{l-1} \\
&= n^{2k-k'-1} \frac{k-k'-1}{(k+1)!} \frac{(2k-k'-1)^k}{k!} \frac{k^k}{(k'+1)!} (\rho^+(k)^{-1}+1)\rho(k+1)^{-1}.
\end{aligned}
$$

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

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

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 $k \ge k'+1$ 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 $n \ge N$:

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

$\square$

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

$$
\begin{aligned}
\sigma_k^2 = \sum_{l=1}^{n-k} & \binom{n-l}{2k-k'} \binom{2k-k'}{k} \binom{k}{k'} \rho^+(k+1)^2 \\
& \left[(1 - 2\rho(k+1) + \rho(k+1)^2 \rho(k'+1)^{-1})^{l-1} - (1 - 2\rho(k+1) + \rho(k+1)^2)^{l-1} \right] + o(n^{2k-k'}).
\end{aligned}
$$

Analogously to the calculations in Lemmas A.1 and 6.3, one can perform calculations to find the highest order term of $\mathrm{Cov}(T_{k+1}, T_{r+1})$. 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

$$\mathrm{Cov}(T_{k+1},T_{r+1}) = \sum_{\substack{s\in\binom{[n]}{k+1}, t\in\binom{[n]}{r+1}\\ \min(s)=\min(t)\\ |s\cap t|=k'+1}} \mathrm{Cov}(Z_sY_s, Z_tY_t) + o(n^{k+r-k'})$$

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

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

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:*

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

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.

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

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

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

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

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

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

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

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

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

□

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

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

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 [49, 3, 39].

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 [49])**.** *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.

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

2.

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

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

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

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

Corollary B.3 (Corollaries 8 and 11 in [49]). *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.

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

2.

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

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 $|\mathbb{D}_j(s)|$ over $s \in \mathbb{I}_i$, and

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

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 [49]). *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*

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

C Simulation Study Results

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.

Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

Title of Paper	Goodness-of-fit via Count Statistics in Dense Random Simplicial Complexes
Publication Status	□Published □Accepted for Publication X Submitted for Publication □Unpublished and unsubmitted work written in a manuscript style
Publication Details	T. Temčinas, V. Nanda, G. Reinert "Goodness-of-fit via Count Statistics in Dense Random Simplicial Complexes", submitted to the Electronic Journal of Probability.

Student Confirmation

Student Name:	Tadas Temčinas		
Contribution to the Paper	Under the guidance of my supervisors, who are the co-authors, I have originated the idea for the paper, proved the main results, and wrote most of the manuscript (excluding only parts of the introduction).		
Signature		Date	14/09/2023

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title: Prof Gesine Reinert		
Supervisor comments The description is accurate.		
Gesine Reinert	Date	14/09/2023

This completed form should be included in the thesis, at the end of the relevant chapter.

4 Conclusions and Future Work

The results presented in this thesis contribute to developing probabilistic and statistical tools that allow rigorous analysis of simplicial complex data - a structure used in topological data analysis (TDA). The central limit theorems (CLTs) proved in this thesis both establish the limiting distribution of statistics of interest as well as quantify the convergence non-asymptotically. As a result, the normal approximations can be used in statistical applications in a quantitative fashion when the observed simplicial complexes are finite, which is the practical setting.

A particular application of the CLTs that the thesis develops is the goodness-of-fit tests based on the number of critical simplices with respect to a lexicographical acyclic partial matching. In Chapter 3 it is shown that the goodness-of-fit tests based on this statistic outperform state of the art tests and can distinguish between the $\mathbf{X}(n, \mathbf{p})$ model and a wide range of models that depend even in a small way on some latent geometric space.

Here we detail some possible directions for future work.

4.1 Goodness-of-fit tests via centered counts

The task of detecting latent geometry in network data has attracted significant attention in recent years [3, 13, 14]. When the data are simplicial complexes, however, this task of detecting latent geometry had not been considered before this work. Naturally when a simplicial complex depends only on the latent geometry (for example, when we consider the so-called Čech or Vietoris-Rips complexes of a finite point cloud in a metric space), one could consider its 1-skeleton as a graph and use the existing methods to detect the dependence on geometry. But when a simplicial complex has a non-geometric 1-skeleton and higher dimensional geometric structure, such an approach would not be fruitful.

In the parameter regimes where it is possible, an alternative statistic that distinguishes between the Erdős–Rényi random graph $\mathbf{G}(n, p)$ and a geometric random graph on a sphere is the centered triangle count. Writing $X_{i,j}$ for the indicator of the edge $\{i, j\}$ in a random graph model, the statistic is defined as

$$T = \sum_{i<j<k} (X_{i,j} - \mathbb{E}\{X_{i,j}\})(X_{i,k} - \mathbb{E}\{X_{i,k}\})(X_{k,j} - \mathbb{E}\{X_{k,j}\}).$$

It is analogous to the number of triangles but with the edge indicators being replaced with their centered versions. Recently the centered subgraph count statistics, not only the centered triangle count, have been studied from a distributional approximation perspective in [11]. There the authors prove a multivariate CLT via Stein's method for the centered subgraph counts in a graphon based random graph model and show that in the limit the centered counts are uncorrelated, which is in contrast with perfect correlation in the non-centered case. Therefore, these centered statistics theoretically seem much more convenient for goodness-of-fit tests as different centered subgraphs capture different aspects on the null model. However, as discussed in Chapter 3, the centered triangle count could be more sensitive to parameter estimation error and therefore less robust, especially on smaller datasets. From a theoretical perspective, though, a natural next step in our research of goodness-of-fit tests for random simplicial complexes would be to understand how centered subcomplex counts behave in the $\mathbf{X}(n, \mathbf{p})$ model or more general models that still preserve the conditional independence of simplex indicators. Once the limiting distributions are established, it would be interesting to see how these statistics perform in goodness-of-fit tests. Moreover, a more thorough quantification of robustness of these tests with respect to re-sampling and parameter estimation from both theoretical and practical points of view would be desirable. We conjecture that goodness-of-fit tests based on uncentered counts are more robust but have less statistical power. Replacing simplex indicators with their centered versions in the critical simplex counts and investigating if the statistic becomes more useful for testing purposed would also be a very interesting project.

4.2 Multivariate and functional distributions of algebraic invariants

Another possible research theme is to study distributions of algebraic invariants (e.g. Euler characteristic or Betti numbers) of random simplicial complexes beyond the univariate setting that has been studied so far [8]. One way is to investigate using the multivariate results on combinatorial properties of simplicial complexes that are proved in this work in order to study multivariate distributions of the algebraic invariants. In certain parameter regimes for models like the clique complex of $\mathbf{G}(n, p)$ random graph, it is proven that asymptotically with high probability only one Betti number is non-trivial [7]. Therefore, while extending the multivariate results one would need to carefully choose a regime where the multivariate distribution is not trivial. It is also important to note that new ideas are required in order to extend the existing results to multivariate ones. As noted in [9], simply using that for any simplicial complex $\mathcal{L}$ we have the inequality

$$e_k(\mathcal{L}) - e_{k+1}(\mathcal{L}) - e_{k-1}(\mathcal{L}) \leq \beta_k(\mathcal{L}) \leq e_k(\mathcal{L})$$

does not translate into a distributional approximation result in a straightforward manner, even in the univariate case. In general, if we have an inequality between random variables $X \leq Y \leq Z$, proving a CLT for X and Z does not guarantee a CLT for Y.

Another idea, perhaps more relevant in applications from a persistent homology point of view, is to study functional distributions of algebraic invariants with respect to some filtration of a simplicial complex. Stein's method for functional distributions has been developed for many types of random variables [1, 6, 5] and it has already seen some applications in the study of random simplicial complexes [16]. Hence it could provide a useful tool for extending our results into the functional setting.

4.3 Simplicial complex distributions and conditional central limit theorems

Stein's method is also used to study probability distributions on the space of graphs (i.e. random graphs) directly rather than via graph statistics like subgraph counts [15]. So far this approach has not been extended to the study of probability distributions on the space of simplicial complexes, which could be a fruitful direction for future research. Stein operators for the graph distributions have also been leveraged in order to develop goodness-of-fit tests [17]. Once the theory of Stein's method on the space of simplicial complexes has been established, similar arguments should carry over to random simplicial complexes.

A classical CLT is a result about the asymptotic distribution of a sum of i.i.d. variables $X_1, \ldots, X_n$. That is, the distribution of a random variable $W = \sum_{i=1}^{n} X_i$ is asymptotically normal, under suitable assumptions about the underlying distribution of the summands. A classical conditional CLT would be a result that the asymptotic distribution of W conditional on the event $\{Y = y\}$, where Y is a related random variable, is normal under suitable assumptions. Such conditional CLTs seem to be more challenging to prove since conditioning on such an event, the summands could become dependent on each other in non-trivial ways.

Another paradigm where Stein's method seems to be not fully developed yet are conditional CLTs.

Recently an exchangeable pair approach to a conditional CLT has been established [4]. However, a conditional CLT for vectors of dissociated sums, which are studied in Chapter 2, has not yet been proved. Working in this direction would not only contribute to the development of Stein's method but it would also allow the extension of the results presented in this thesis to models where the number of simplices is fixed.

Thinking about the graph case, when a graph is observed as a datapoint, the number of edges present is known. Therefore, the $\mathbf{G}(n, m)$ model, which uniformly picks a graph on n vertices with m edges, seems to be more appropriate compared to $\mathbf{G}(n, p)$. Studying statistics of $\mathbf{G}(n, m)$ proves to be more challenging because the edge indicators are no longer independent. However, conditioning on having m edges in the $\mathbf{G}(n, p)$ random graph results in a $\mathbf{G}(n, m)$ model. In this context, conditional CLTs can be used in order to establish normal approximation results for popular statistics like subgraph counts in the more realistic $\mathbf{G}(n, m)$ model.

The same thinking can be carried over to conditional random simplicial complex models. Once a simplicial complex is observed, the number of simplices is known and in the models where this is a sufficient statistic, one can condition on it to obtain a new conditional model, which in practice would likely be more realistic. For example, conditioning on having m vertices, the $\mathbf{X}(n, p)$ model becomes $\mathbf{X}(n, m)$, which is the clique complex of the $\mathbf{G}(n, m)$ random graph. As explained in [2], in the case of the multi-parameter model $\mathbf{X}(n, \mathbf{p})$, one can condition on having s_i i-simplices and h_i hollow i-simplices for every $1 \leq i \leq n-1$ and obtain a random simplicial complex model that uniformly picks a simplicial complex having the prescribed number of i-simplices and hollow i-simplices for every i. Using the Kruskal-Katona theorem [10, 12] one can verify that the sequence of the number of simplices is viable and such simplicial complexes with the prescribed number of i-simplices for each i exist. Therefore, having conditional CLT for vectors of dissociated sums via Stein's method would allow the extension of CLTs presented in this thesis to such conditional random simplicial complex models.

References

[1] Andrew D Barbour. "Stein's method for diffusion approximations". In: *Probability Theory and Related Fields* 84.3 (1990), pp. 297–322.

[2] Omer Bobrowski and Dmitri Krioukov. “Random simplicial complexes: models and phenomena”. In: *Higher-Order Systems*. Springer, 2022, pp. 59–96.

[3] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. “Testing for high-dimensional geometry in random graphs”. In: *Random Structures & Algorithms* 49.3 (2016), pp. 503–532.

[4] Partha S Dey and Grigory Terlov. “Stein’s method for conditional central limit theorem”. In: *The Annals of Probability* 51.2 (2023), pp. 723–773.

[5] Christian Döbler, Mikołaj Kasprzak, and Giovanni Peccati. “The multivariate functional de Jong CLT”. In: *Probability Theory and Related Fields* 184.1-2 (2022), pp. 367–399.

[6] Christian Döbler and Mikołaj J Kasprzak. “Stein’s method of exchangeable pairs in multivariate functional approximations”. In: *Electronic Journal of Probability* 25 (2021), pp. 1–50.

[7] Matthew Kahle. “Sharp vanishing thresholds for cohomology of random flag complexes”. In: *Annals of Mathematics* (2014), pp. 1085–1107.

[8] Matthew Kahle and Elizabeth Meckes. “Limit theorems for Betti numbers of random simplicial complexes”. In: *Homology, Homotopy and Applications* 15.1 (2013), pp. 343–374.

[9] Matthew Kahle and Elizabeth Meckes. “Erratum: Limit theorems for Betti numbers of random simplicial complexes”. In: *arXiv preprint arXiv:1501.03759* (2015).

[10] Gyula O H Katona. “Intersection theorems for systems of finite sets”. In: *Quart. J. Math. Oxford Ser.(2)* 12 (1961), pp. 313–320.

[11] Gursharn Kaur and Adrian Röllin. “Higher-order fluctuations in dense random graph models”. In: *Electronic Journal of Probability* 26 (2021), pp. 1–36.

[12] Joseph B Kruskal. “The number of simplices in a complex”. In: *Mathematical optimization techniques* 10 (1963), pp. 251–278.

[13] Siqi Liu, Sidhanth Mohanty, Tselil Schramm, and Elizabeth Yang. “Testing thresholds for high-dimensional sparse random geometric graphs”. In: *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*. 2022, pp. 672–677.

[14] Suqi Liu and Miklós Z Rácz. “Phase transition in noisy high-dimensional random geometric graphs”. In: *arXiv preprint arXiv:2103.15249* (2021).

[15] Gesine Reinert and Nathan Ross. "Approximating stationary distributions of fast mixing Glauber dynamics, with applications to exponential random graphs". In: *The Annals of Applied Probability* 29.5 (2019), pp. 3201 –3229.

[16] Andrew M Thomas and Takashi Owada. "Functional limit theorems for the Euler characteristic process in the critical regime". In: *Advances in Applied Probability* 53.1 (2021), pp. 57–80.

[17] Wenkai Xu and Gesine Reinert. "A Stein goodness-of-test for exponential random graph models". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 415–423.