# LIMIT THEOREMS FOR BETTI NUMBERS OF RANDOM SIMPLICIAL COMPLEXES

MATTHEW KAHLE AND ELIZABETH MECKES

Abstract. There have been several recent articles studying homology of various types of random simplicial complexes. Several theorems have concerned thresholds for vanishing of homology, and in some cases expectations of the Betti numbers. However little seems known so far about limiting distributions of random Betti numbers.

In this article we establish Poisson and normal approximation theorems for Betti numbers of different kinds of random simplicial complex: Erdős-Rényi random clique complexes, random Vietoris-Rips complexes, and random Čech complexes. These results may be of practical interest in topological data analysis.

## 1. Introduction

Several papers have recently appeared concerning the topology of random simplicial complexes [11, 2, 10, 12, 13, 16, 9]. The results so far identify thresholds for vanishing of homology, or compute the expectation of the Betti numbers $\mathbb{E}[\beta_k]$ (i.e. the expected rank of these groups). In this article we prove Poisson and normal approximation theorems for $\beta_k$ for three models of random simplicial complex. The complexes themselves are defined precisely and given further motivation in the following sections but we first outline our results.

The first model considered is that of the Erdős-Rényi random clique complex $X(n,p)$, a higher dimensional analogue of the Erdős-Rényi random graph $G(n,p)$. It was shown in [11] that for each $k$ and a certain range of $p = p(n)$, $\beta_k \neq 0$ asymptotically almost surely (a.a.s), and in this regime, a formula for the asymptotic size of $\mathbb{E}[\beta_k]$in terms of $p$ is given. (Outside of this regime it is conjectured that $\beta_k = 0$ a.a.s. and some evidence for the conjecture is given in [11].) Here we prove a Central Limit Theorem for $\beta_k$. That is, we show that

$$\frac{\beta_k - \mathbb{E}[\beta_k]}{\sqrt{\mathrm{Var}[\beta_k]}} \Rightarrow \mathcal{N}(0,1),$$

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

The second model considered is the random Čech complex. This model is a higher-dimensional analog of the random geometric graph; the underlying graph is a random geometric graph and the presence of $(k-1)$-dimensional faces is determined by $k$-fold intersections of balls centered about the vertices. Čech complexes

Date: January 19, 2011.

M. Kahle's research was supported by the NSF Research Training Group grant in geometry and topology of Stanford University.

E. Meckes's research was supported by an American Institute of Mathematics Five-year Fellowship and NSF grant DMS-0852898.

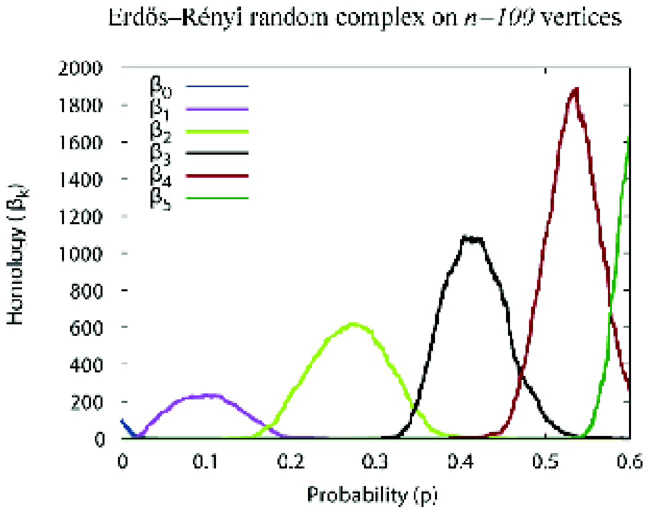

Figure 1. The Betti numbers of $X(n, p)$ plotted vertically against edge probability $p$; in this example $n = 100$. *Computation and graphic courtesy of Afra Zomorodian.*

are homotopy equivalent to Edelsbrunner and Mücke's *alpha shapes*, widely applied in computational geometry and topology [6]. The analysis needed to obtain limit theorems for the Betti numbers of random Čech complexes is more subtle that what is needed for the Erdös-Rényi model; to prove the normal and Poisson approximation theorems we must first establish limit theorems for certain hypergraph counts, extending some of Mathew Penrose's results for subgraph counts for geometric random graphs [15].

The final type of complex considered is the random Vietoris-Rips complex, denoted $VR(n, r)$. This is similar to the random Čech complex; the construction is to take the clique complex of a random geometric graph. (A useful reference for geometric random graphs is [15].) The topology is very different than for the clique complex of the Erdős-Rényi random graph; for the contrast between $X(n, p)$ and $VR(n, r)$ see Figures 1 and 2. The analysis needed to obtain limit theorems for the Betti numbers of $VR(n, r)$ is nevertheless essentially identical to that needed for the random Čech complex. A minor example of this fact is that in both cases, since $\beta_0$ counts the number of connected components for the Čech and Rips complexes, $\beta_0$ is actually the same in each of these cases and is equal to the number of components of the random geometric graph. This has already been treated in detail by Penrose [15], and so when convenient we will restrict attention to $\beta_k$ for $k \geq 1$.

The techniques throughout the paper are a combination of inequalities derived from combinatorial and topological considerations with Stein's method. (For an introduction to topological combinatorics see [4]; for a survey of Stein's method in proving Poisson approximation theorems see [5], and for an introduction to Stein's method for normal approximation, see [17].)

1.1. **Notation and conventions.** Throughout this article, we use Bachmann-Landau big-$O$, little-$O$, and related notations. In particular, for non-negative functions $g$ and $h$, we write the following.

- $g(n) = O(h(n))$ means that there exists $n_0$ and $k$ such that for $n > n_0$, we have that $g(n) \leq k \cdot h(n)$. (i.e. $g$ is asymptotically bounded above by $h$, up to a constant factor.)
- $g(n) = \Omega(h(n))$ means that there exists $n_0$ and $k$ such that for $n > n_0$, we have that $g(n) \geq k \cdot h(n)$. (i.e. $g$ is asymptotically bounded below by $h$, up to a constant factor.)
- $g(n) = \Theta(h(n))$ means that $g(n) = O(h(n))$ and $g(n) = \Omega(h(n))$. (i.e. $g$ is asymptotically bounded above and below by $h$, up to constant factors.)
- $g(n) = o(h(n))$ means that for every $\epsilon > 0$, there exists $n_0$ such that for $n > n_0$, we have that $g(n) \leq \epsilon \cdot h(n)$. (i.e. $g$ is dominated by $h$ asymptotically.)
- $g(n) = \omega(h(n))$ means that for every $k > 0$, there exists $n_0$ such that for $n > n_0$, we have that $g(n) \geq k \cdot h(n)$. (i.e. $g$ dominates $h$ asymptotically.)

We may also write $A_n \simeq B_n$ if $\lim_{n\to\infty} \frac{A_n}{B_n} = 1$, and $A_n \lesssim B_n$ if there is a constant $c$ such that $A_n \leq cB_n$ for all $n$.

A sequence $\{X_n\}_{n=1}^{\infty}$ of random variables is said to *converge weakly* to a limiting random variable $X$ (written $X_n \Rightarrow X$) if $\lim_{n\to\infty} \mathbb{E}[f(X_n)] = \mathbb{E}[f(X)]$ for all bounded continuous functions $f$ (there are several other equivalent definitions).

The *total variation distance* between random variables $X$ and $Y$ is defined by

$$d_{TV}(X,Y) := \sup_f \big|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big|,$$

with the supremum taken over all continuous functions bounded by one. Clearly, if $d_{TV}(X_n, X) \to 0$ as $n \to \infty$, then $X_n \Rightarrow X$; however, the topology induced by the total variation distance is stronger than the topology of weak convergence.

The $L_1$-*Wasserstein distance* or *Kantorovich-Rubenstein distance* between $X$ and $Y$ is defined by

$$d_1(X,Y) := \sup_f \big|\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]\big|,$$

where the supremum is over all functions $f$ with $\sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|} \leq 1$. This distance also induces a topology stronger than the topology of weak convergence.

Finally, the normal distribution with mean $\mu$ and variance $\sigma^2$ is denoted $\mathcal{N}(\mu, \sigma^2)$, and the distribution function of the standard normal distribution is denoted $\Phi(t)$.

## 2. Erdős-Rényi random clique complexes

Perhaps the first type of random simplicial complex studied was the 1-dimensional version studied by Erdős and Rényi [7].

**Definition 2.1.** The *Erdős-Rényi random graph* $G(n,p)$ is the probability space of all graphs on vertex set $[n] = \{1, 2, \ldots, n\}$ with each edge included independently with probability $p$.

The "clique complex" is used to generalize $G(n,p)$ from graphs to higher dimensional simplicial complexes.

**Definition 2.2.** The *clique complex* $X(H)$ of a graph $H$ is a the simplicial complex with vertex set $V(H)$ and a face for each set of vertices spanning a complete subgraph of $H$.

In other words, the clique complex $X(H)$ of a graph $H$ is the maximal simplicial complex with 1-skeleton $H$.

This section concerns the clique complex of the Erdős-Rényi random graph, i.e. $X(G(n,p))$. For simplicity in notation, this is denoted $X(n,p)$.

There are several motivations for using $X(n,p)$ as a model of a random simplicial complex. One motivation is that $X(n,p)$ provides a natural higher-dimensional generalization of $G(n,p)$, which has proved extremely useful in graph theory as well as in applications. (Other higher-dimensional generalizations are studied in [2, 12, 13].) Another motivation comes from the fact that every simplicial complex is homeomorphic to the clique complex of some graph (e.g. by barycentric subdivision) [8].

One interesting feature of $X(n,p)$ is that it provides homological analogues of the Erdős-Rényi theorem, but in a *non-monotone* setting: If edges are added at random to an empty graph, the Erdős-Rényi theorem characterizes the number of edges needed before the graph becomes connected. Connectivity is a monotone graph property – if one adds edges to a connected graph, it is still connected.

Topologically, connectivity is equivalent to a statement about zeroth homology $H_0(G(n,p))$ but if one asks about $H_k(X(n,p))$, $k > 0$, there is a problem – adding edges generates higher $k$-dimensional faces and $(k+1)$-dimensional faces at the same time. Since generators and relations are both being added, there is no reason that things have to behave in a monotone way. In fact, it is not just that things might not be monotone; they are non-monotone in an essential way. In particular, there seem to be two thresholds for higher homology – one where $H_k$ passes from vanishing to non-vanishing, and another where it passes back to vanishing.

The following theorem was proved in [11]. For any fixed $k > 0$, let $\beta_k$ denote the dimension of $k$th homology, i.e. $\beta_k = \dim[H_k(\Delta, \mathbb{Q})]$.

**Theorem 2.3.** *If $p = \omega(n^{-1/k})$ and $p = o(n^{-1/(k+1)})$ then*

$$\lim_{n\to\infty} \frac{\mathbb{E}[\beta_k(X(n,p))]}{n^k p^{\binom{k+1}{2}}} = \frac{1}{(k+1)!}.$$

(In [11] explicit nontrivial homology classes are exhibited, and several partial converses of Theorem 2.3 are proved; in particular it is shown that if $p = O(n^{-1/k-\epsilon})$ or $p = \Omega(n^{-1/(2k+1)+\epsilon})$ for some constant $\epsilon > 0$, then a.a.s. $\beta_k = 0$.)

The remainder of this section is devoted to showing that in the same regime, $\beta_k$ obeys a central limit theorem.

**Theorem 2.4.** *If $p = \omega(n^{-1/k})$ and $p = o(n^{-1/(k+1)})$ then*

$$\frac{\beta_k(X(n,p)) - \mathbb{E}[\beta_k(X(n,p))]}{\sqrt{\mathrm{Var}[\beta_k]}} \Rightarrow \mathcal{N}(0,1).$$

*Proof.* For a finite simplicial complex $\Delta$, let $f_i(\Delta)$ (or simply $f_i$ if context is clear) denote the number of $i$-dimensional faces of $\Delta$. A useful fact when proving Theorems 2.3 and 2.4 is that $\beta_k$ satisfies the following "Morse" inequalities:

$$-f_{k-1} + f_k - f_{k+1} \le \beta_k \le f_k, \tag{1}$$

for all $k$. These inequalities follow from the definition of simplicial homology and the rank-nullity law [8].

The next observation to make is that $X(n,p)$ is a clique complex, so $f_k$ counts the number of $(k+1)$-cliques. Since there are $\binom{n}{k+1}$ possible $(k+1)$-cliques and each appears with probability $p^{\binom{k+1}{2}}$,

$$\lim_{n\to\infty} \frac{\mathbb{E}[f_k]}{n^{k+1}p^{\binom{k+1}{2}}} = \frac{1}{(k+1)!}.$$

If $p = \omega(n^{-1/k})$ then

$$\frac{\mathbb{E}[f_{k-1}]}{\mathbb{E}[f_k]} = \frac{n^k p^{\binom{k}{2}}}{n^{k+1}p^{\binom{k+1}{2}}} = \frac{1}{np^k} = o(1),$$

and the same argument shows that if $p = o(n^{-1/(k+1)})$ then

$$\frac{\mathbb{E}[f_{k+1}]}{\mathbb{E}[f_k]} = o(1).$$

That is, in the regime of Theorems 2.3 and 2.4,

$$\lim_{n\to\infty} \frac{\mathbb{E}[f_k]}{\mathbb{E}[-f_{k-1}+f_k-f_{k+1}]} = 1,$$

which, in light of (1), reproves Theorem 2.3.

Let $\tilde{f}_k := -f_{k-1}+f_k-f_{k+1}$. The following claim together with (1) is used to show that $\beta_k$ satisfies a central limit theorem.

**Claim 2.5.**

(i)
$$\lim_{n\to\infty} \frac{\mathrm{Var}(f_k)}{\mathrm{Var}(\tilde{f}_k)} = 1.$$

(ii)
$$\frac{f_k - \mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}} \Rightarrow \mathcal{N}(0,1) \quad \text{as}\, n\to\infty.$$

(iii)
$$\frac{\tilde{f}_k - \mathbb{E}[\tilde{f}_k]}{\sqrt{\mathrm{Var}(\tilde{f}_k)}} \Rightarrow \mathcal{N}(0,1) \quad \text{as}\, n\to\infty.$$

For $t \in \mathbb{R}$, it follows from (1) that

$$\mathbb{P}\left[\frac{f_k - \mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}} \le t\right] \le \mathbb{P}\left[\frac{\beta_k - \mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}} \le t\right] \le \mathbb{P}\left[\frac{\tilde{f}_k - \mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}} \le t\right].$$

The left-hand side tends to $\Phi(t)$ as $n \to \infty$ by part (ii) of the claim. For the right-hand side, let $\epsilon > 0$ and observe that

(2)
$$\mathbb{P}\left[\frac{\tilde{f}_k - \mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}} \le t\right] \le \mathbb{P}\left[\frac{\tilde{f}_k - \mathbb{E}[\tilde{f}_k]}{\sqrt{\mathrm{Var}(\tilde{f}_k)}} \le t-\epsilon\right] + \mathbb{P}\left[\left|\frac{\tilde{f}_k - \mathbb{E}[\tilde{f}_k]}{\sqrt{\mathrm{Var}(\tilde{f}_k)}} - \frac{\tilde{f}_k - \mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}}\right| > \epsilon\right]$$
$$+ \mathbb{P}\left[\frac{\tilde{f}_k - \mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}} \le t, \left|\frac{\tilde{f}_k - \mathbb{E}[\tilde{f}_k]}{\sqrt{\mathrm{Var}(\tilde{f}_k)}} - t\right| \le \epsilon\right].$$

Now, it follows from part (iii) of the claim that the first term of the right-hand side of (2) tends to $\Phi(t-\epsilon)$ and that the last is asymptotically bounded above by $\Phi(t+\epsilon)-\Phi(t-\epsilon)$. For the second term, first require $n$ to be large enough that

$$\left|\frac{\mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}}-\frac{\mathbb{E}[\tilde{f}_k]}{\sqrt{\mathrm{Var}(\tilde{f}_k)}}\right|<\frac{\epsilon}{2}.$$

This condition together with Chebychev's inequality implies that

$$\begin{aligned}\mathbb{P}\left[\left|\frac{\tilde{f}_k-\mathbb{E}[\tilde{f}_k]}{\sqrt{\mathrm{Var}(\tilde{f}_k)}}-\frac{\tilde{f}_k-\mathbb{E}[f_k]}{\sqrt{\mathrm{Var}(f_k)}}\right|>\epsilon\right] &\le \mathbb{P}\left[\tilde{f}_k\left|\frac{1}{\sqrt{\mathrm{Var}(f_k)}}-\frac{1}{\sqrt{\mathrm{Var}(\tilde{f}_k)}}\right|>\frac{\epsilon}{2}\right]\\ &\le 4\epsilon^{-2}\left(\frac{\sqrt{\mathrm{Var}(\tilde{f}_k)}}{\sqrt{\mathrm{Var}(f_k)}}-1\right)^2,\end{aligned}$$

which tends to zero for fixed $\epsilon>0$ by part (i) of the claim. It thus follows that the right-hand side of (2) is asymptotically bounded above by $\Phi(t+\epsilon)$ as $n\to\infty$; as $\epsilon$ is arbitrary, this completes the proof of the central limit theorem for $\beta_k$, modulo proof of the claim.

To prove part (i) of the claim, first write

$$f_k=\sum_{\substack{A\subseteq\{1,\ldots,n\}\\|A|=k+1}}\xi_A,$$

where $\xi_A$ is the indicator that $A$ spans a face in $X(n,p)$; that is, that $A$ spans a complete graph in $G(n,p)$. Then, enumerating pairs of subsets of size $k+1$ of $\{1,\ldots,n\}$ by the size $r$ of their interesection,

$$\begin{aligned}\mathrm{Var}(f_k)&=\sum_{A,B}\mathbb{E}[\xi_A\xi_B]-\left[\binom{n}{k+1}p^{\binom{k+1}{2}}\right]^2\\ &=\binom{n}{k+1}\sum_{r=0}^{k+1}\binom{k+1}{r}\binom{n-k-1}{k+1-r}p^{2\binom{k+1}{2}-\binom{r}{2}}-\left[\binom{n}{k+1}p^{\binom{k+1}{2}}\right]^2.\end{aligned}$$

Now, it is not hard to see that in the range of $p$ considered here, only the $r=0,1,2$ terms contribute in the limit; there is cancellation of the terms of order $n^{k+1}$ and $n^k$, so that the main contribution is in fact from the $r=2$ term and

$$\lim_{n\to\infty}n^{-2k}p^{(-2\binom{k+1}{2}+1)}\mathrm{Var}(f_k)=c_k, \tag{3}$$

for some constant $c$ depending only on $k$. From this it follows immediately that

$$\frac{\mathrm{Var}(f_{k-1})}{\mathrm{Var}(f_k)}=o(1)\quad\text{and}\quad\frac{\mathrm{Var}(f_{k+1})}{\mathrm{Var}(f_k)}=o(1),$$

for $p$ in the range specified in the statement of the theorem.

Expanding the same way as above, it is clear that

$$\mathrm{Cov}(f_k,f_{k+1})=\binom{n}{k+1}p^{\binom{k+1}{2}+\binom{k+2}{2}}\left[\sum_{r=0}^{k+1}\binom{k+1}{r}\binom{n-k-1}{k+2-r}p^{-\binom{r}{2}}-\binom{n}{k+2}\right];$$

again there is cancellation of the terms of order $n^{k+2}$ and $n^{k+1}$ so that the leading contribution is from the $r = 2$ term and

$$\lim_{n\to\infty} n^{-2k-1} p^{-(\binom{k+1}{2}+\binom{k+2}{2}-1)} \mathrm{Cov}(f_k, f_{k+1}) = c_k$$

for a (different) constant $c_k$ depending only on $k$. Thus in the range of $p$ being considered,

$$\frac{\mathrm{Cov}(f_k, f_{k+1})}{\mathrm{Var}(f_k)} = o(1).$$

In exactly the same way, one can show that

$$\frac{\mathrm{Cov}(f_k, f_{k-1})}{\mathrm{Var}(f_k)} = o(1) \quad \text{and} \quad \frac{\mathrm{Cov}(f_{k-1}, f_{k+1})}{\mathrm{Var}(f_k)} = o(1),$$

completing the proof of part (i) of the claim.

The proofs of the second and third parts both follow from an abstract normal approximation theorem for dissociated random variables proved (via Stein's method) in [3]. Part (ii) is in fact proved there; the following is a a straightforward modification of their proof which obtains a central limit theorem for the lower bound $\tilde{f}_k$. One can also recover the proof of part (ii) from what is given below, simply by ignoring the extra terms present in $\tilde{f}_k$ beyond those coming from $f_k$.

A set $\{X_{\mathbf{j}} : \mathbf{j} = (j_1, \ldots, j_r) \in J\}$ for $J$ a set of $r$-tuples is *dissociated* if two subcollections of the random variables $\{X_{\mathbf{j}} : \mathbf{j} \in K\}$ and $\{X_{\mathbf{j}} : \mathbf{j} \in L\}$ are independent whenever $(\cup_{\mathbf{j}\in K}\{j_1, \ldots, j_r\}) \cap (\cup_{\mathbf{j}\in L}\{j_1, \ldots, j_r\}) = \emptyset$. Let $W := \sum_{\mathbf{j}\in J} X_{\mathbf{j}}$, and for each $\mathbf{j} \in J$, let $L_{\mathbf{j}} := \{\mathbf{k} \in J : \{k_1, \ldots, k_r\} \cap \{j_1, \ldots, j_r\} \neq \emptyset\}$. That is, $L_{\mathbf{j}}$ is a dependency neighborhood for $\mathbf{j}$. If $\mathbb{E}X_{\mathbf{j}} = 0$ and $\mathbb{E}W^2 = 1$, then it is shown in [3] that

$$d_1(W, Z) \le K \sum_{\mathbf{j}\in J} \sum_{\mathbf{k},\mathbf{l}\in L_{\mathbf{j}}} \Big[\mathbb{E}|X_{\mathbf{j}} X_{\mathbf{k}} X_{\mathbf{l}}| + \mathbb{E}|X_{\mathbf{j}} X_{\mathbf{k}}| \mathbb{E}|X_{\mathbf{l}}|\Big], \tag{4}$$

where $Z$ is a standard normal random variable.

To show that $\tilde{f}_k$ satisfies a central limit theorem, let the index set $J$ be the potential edge sets for complete graphs on $k + e$ ($e \in \{0, 1, 2\}$) vertices in $G(n, p)$; that is, an element of $J$ is a $\binom{k+e}{2}$-tuple of edges spanning a given set of $k + e$ vertices. Each $\mathbf{j} \in J$ can thus be associated with its spanning set $A_{\mathbf{j}}$ of vertices. If the random variables $X_{\mathbf{j}}$ are defined by

$$X_{\mathbf{j}} := \sigma^{-1}(\xi_{A_{\mathbf{j}}} - \mathbb{E}[\xi_{A_{\mathbf{j}}}]),$$

where $\sigma^2 = \mathrm{Var}(f_k)$, then $\{X_{\mathbf{j}}\}$ are evidently dissociated.

The second half of the sum from (4) is fairly straightforward to bound in this context. For each $\mathbf{j}$, partition $L_{\mathbf{j}}$ into the sets $L_{\mathbf{j}}^e$ of indices whose spanning sets have size $k + e$. Observe that for each $\mathbf{j}$, if $e_j = |L_{\mathbf{j}}| - k$, then

$$|L_{\mathbf{j}}^e| = \binom{n}{k+e} - \binom{n-k-e_j}{k+e} - (k+e_j)\binom{n-k-e_j}{k+e-1} = O(n^{k+e-2}).$$

Decomposing as in the variance estimate by the size $r$ of the intersection of $A_{\mathbf{j}}$ and $A_{\mathbf{k}}$ and using the bound above for $|L_{\mathbf{j}}^f|$ yields

$$\sum_{\mathbf{j}\in J}\sum_{\mathbf{k}\in L_{\mathbf{j}}^e}\sum_{\mathbf{l}\in L_{\mathbf{j}}^f}\mathbb{E}|X_{\mathbf{j}}X_{\mathbf{k}}|\mathbb{E}|X_{\mathbf{l}}|$$

$$\le \sigma^{-3}c_k n^{k+f-2}p^{\binom{k+f}{2}}\binom{n}{k+e_j}\sum_{r=2}^{k+(e_j\wedge e)}\binom{k+e_j}{r}\binom{n-k-e_j}{k+e-r}p^{\binom{k+e}{2}+\binom{k+e_j}{2}-\binom{r}{2}}$$
$$\le \sigma^{-3}c_k n^{3k+e_j+e+f-4}p^{\binom{k+e_j}{2}+\binom{k+e}{2}+\binom{k+f}{2}-1},$$

since the $r=2$ term yields the top-order contribution in the range of $p$ considered here. Moreover, it is easy to check that this expression is maximized for $e_j=e=f=1$. Combining this estimate with (3) shows that the contribution to the error from the second sum is bounded above by

$$\sigma^{-3}c_k n^{3k-1}p^{3\binom{k+1}{2}-1}\le\frac{c_k\sqrt{p}}{n},$$

which tends to zero as $n$ tends to infinity.

The first half of the sum is bounded similarly, although it requires that the intersections of three spanning sets of vertices be considered. Let $r$ denote the number of points common to $A_{\mathbf{j}}$ and $A_{\mathbf{k}}$. Let $p_1:=|A_{\mathbf{j}}\cap A_{\mathbf{l}}\cap A_{\mathbf{k}}^c|$, $p_2:=|A_{\mathbf{j}}\cap A_{\mathbf{l}}\cap A_{\mathbf{k}}|$ and $p_3:=|A_{\mathbf{j}}^c\cap A_{\mathbf{l}}\cap A_{\mathbf{k}}|$. Then

$$\mathbb{E}|X_{\mathbf{j}}X_{\mathbf{k}}X_{\mathbf{l}}|\le c\sigma^{-3}p^{\binom{k+e_j}{2}+\binom{k+e_k}{2}+\binom{k+e_l}{2}-\binom{p_1+p_2}{2}-\binom{p_2+p_3}{2}-\binom{r}{2}+\binom{p_2}{2}},$$

where the constant $c$ simply accounts for the fact that the $X_{\mathbf{j}}$ have been centered. The number of ways to choose $\mathbf{j}$, $\mathbf{k}$ and $\mathbf{l}$ is

$$\binom{n}{k+e_j}\binom{k+e_j}{r}\binom{n-k-e_j}{k+e_k-r}\binom{k+e_j-r}{p_1}$$
$$\times\binom{r}{p_2}\binom{k+e_k-r}{p_3}\binom{n-2k-e_j-e_k+r}{k+e_l-p_1-p_2-p_3}.$$

Combining these two facts, it is perhaps slightly unpleasant but not too hard to see that the main contribution to the error arises from the case that $r=2$, $p_1+p_2=2$ (in fact only when $p_1\neq 0$), and $e_j=e_k=e_l=1$. It follows that

$$\sum_{\mathbf{j}\in J}\sum_{\mathbf{k},\mathbf{l}\in L_{\mathbf{j}}}\mathbb{E}|X_{\mathbf{j}}X_{\mathbf{k}}X_{\mathbf{l}}|\le\sigma^{-3}c_k n^{3k-1}p^{3\binom{k+1}{2}-2}\le\frac{c_k}{n\sqrt{p}},$$

which also tends to zero as $n$ tends to infinity. This completes the proof of part (iii) of the claim, finishing the proof of Theorem 2.4.

□

## 3. Random Čech complexes

The second model of random simplicial complex considered is the random Čech complex. This is a higher-dimensional analog of a geometric random graph, constructed explicitly below. In order to analyze this model, we use the same techniques used by Penrose [15] in his study of subgraph counts of random geometric graph. The additional spacial dependence that is inherent in the random variables we consider presents an additional technical challenge, and means that Penrose's results cannot be applied directly to the problem.

Suppose that $\{X_i\}_{i=1}^\infty$ is an i.i.d. sequence of random vectors in $\mathbb{R}^d$, with bounded density $f$. Let $\{r_n\}_{n=1}^\infty \subseteq \mathbb{R}_+$, such that $nr_n^d \xrightarrow{n\to\infty} 0$ (the so-called "sparse" regime of geometric random graphs), and construct a random Čech complex $\mathcal{C}(X_1,\ldots,X_n)$ on $\{X_i\}_{i=1}^n$ as follows. If $|X_i - X_j| \le 2r_n$, put an edge between $X_i$ and $X_j$; that is, the 1-skeleton of the complex is a random geometric graph. More generally, make the convex hull of $\{X_{i_1}\ldots,X_{i_k}\}$ a face of the complex if the balls of radius $r_n$ about the points $\{X_{i_1}\ldots,X_{i_k}\}$ have non-trivial intersection.

**Definition 3.1.** The points $\{x_1,\ldots,x_k\} \subseteq \mathbb{R}^d$ form an *empty* $(k-1)$*-simplex* with respect to $r$ if for each $j_o \in \{1,\ldots,k\}$, the intersection $\bigcap_{\substack{1\le j\le k\\ j\neq j_o}} B_r(x_j)$ is non-empty, but the intersection $\bigcap_{1\le j\le k} B_r(x_j) = \emptyset$.

Let $h_r(x_1,\ldots,x_k)$ be the indicator that $\{x_1,\ldots,x_k\}$ form an empty $(k-1)$-simplex with respect to $r$, and for a multiindex $\mathbf{i} = (i_1,\ldots,i_k)$ with $1 \le i_1 < \cdots < i_k \le n$, let $\xi_{\mathbf{i}} = h_{r_n}(X_{i_1},\ldots,X_{i_k})$. Let

$$S_{n,k} := \sum_{\substack{\mathbf{i}=(i_1,\ldots,i_k)\\ 1\le i_1<\cdots<i_k\le n}} \xi_{\mathbf{i}};$$

that is, $S_{n,k}$ is the number of empty $(k-1)$-simplices in $\mathcal{C}(X_1,\ldots,X_n)$. Another object of equal importance in what follows is $\widetilde{S}_{n,k}$, the number of *isolated* empty $k$-simples. That is, if $\zeta_{(i_1,\ldots,i_k)}$ is the indicator that $\{X_{i_1},\ldots,X_{i_k}\}$ form an empty $(k-1)$-simplex with respect to $r_n$ and that there are no edges between $\{X_j\}_{j\in\{i_1,\ldots,i_k\}}$ and $\{X_j\}_{j\notin\{i_1,\ldots,i_k\}}$, then

$$\widetilde{S}_{n,k} = \sum_{\substack{\mathbf{i}=(i_1,\ldots,i_k)\\ 1\le i_1<\cdots<i_k\le n}} \zeta_{\mathbf{i}}.$$

The random variables $S_{n,k}$ and $\widetilde{S}_{n,k}$ are related to $\beta_{k-1}$ as follows. Firstly, $\beta_{k-1}$ is bounded below by the number of isolated empty $k$-simplices; that is, $\beta_{k-1}(\mathcal{C}(X_1,\ldots,X_n)) \ge \widetilde{S}_{n,k}$. Furthermore, any contribution to $\beta_{k-1}$ not coming from an isolated empty $(k-1)$-simplex comes from a component in $\mathcal{C}(X_1,\ldots,X_n)$ on at least $k+1$ vertices. In order for such a component to contribute to $\beta_{k-1}$, $(k-2)$-dimensional faces. Such faces are necessarily triangulated (by the construction of $\mathcal{C}(X_1,\ldots,X_n)$), and so any further contribution to $\beta_{k-1}$ contains at least one simplex on $k-1$ vertices, with either an extra edge attached to each of two different vertices (terminating in different places), or else an extra path of length two attached to one vertex. Let $Y_{n,k}$ denote the number of simplices in $\mathcal{C}(X_1,\ldots,X_n)$ on $k-1$ vertices with two extra edges attached, counted once for each simplex on $k-1$ vertices which occurs and for each distinct pair of simplex vertices with an extra edge. Similarly, let $Z_{n,k}$ denote the number of simplices in $\mathcal{C}(X_1,\ldots,X_n)$ on $k-1$ vertices with at least one extra path of length 2 attached, counted once for each simplex which occurs and for each vertex with a path of length two attached. The argument above shows that

$$\widetilde{S}_{n,k} \le \beta_{k-2}(\mathcal{C}(X_1,\ldots,X_n)) \le S_{n,k} + Y_{n,k} + Z_{n,k}, \tag{5}$$

where the trivial bound $\widetilde{S}_{n,k} \le S_{n,k}$ has also been used.

The limiting distribution of $\beta_{k-1}$ will follow as in the previous section by proving the same limit theorems for the upper and lower bounds of (5). The theorem is the following.

**Theorem 3.2.**

(i) *If* $n^k r_n^{d(k-1)} \to 0$ *as* $n \to \infty$, *then*

$$\beta_k(\mathcal{C}(X_1, \ldots, X_n)) \to 0 \quad a.a.s. \text{ as } n \to \infty.$$

(ii) *If* $n^k r_n^{d(k-1)} \to \alpha \in (0, \infty)$ *as* $n \to \infty$, *then*

$$d_{TV}(\beta_k(\mathcal{C}(X_1, \ldots, X_n)), Y) \le cnr_n^d, ,$$

*where* $Y$ *is a Poisson random variable with* $\mathbb{E}[Y] = \mathbb{E}[\beta_k]$ *and* $c$ *is a constant depending only on* $d$, $k$, *and* $f$.

(iii) *If* $n^k r_n^{d(k-1)} \to \infty$ *as* $n \to \infty$ *and* $nr_n^d \to 0$ *as* $n \to \infty$, *then*

$$\frac{\beta(\mathcal{C}(X_1, \ldots, X_n)) - \mathbb{E}[\beta(\mathcal{C}(X_1, \ldots, X_n))]}{\sqrt{\mathrm{Var}(\beta(\mathcal{C}(X_1, \ldots, X_n)))}} \Rightarrow \mathcal{N}(0, 1).$$

The first step in proving Theorem 3.2 is to determine the order in $n$ and $r_n$ of $\mathbb{E}[\widetilde{S}_{n,k}]$ and $\mathbb{E}[S_{n,k} + Y_{n,k} + Z_{n,k}]$. In fact, slightly more is needed. Let $A$ be an open subset of $\mathbb{R}^d$ such that $vol(\partial A) = 0$. Let $\mathcal{X}$ be a finite subset of $\mathbb{R}^d$, and call $x \in \mathcal{X}$ the "left-most" point of $\mathcal{X}$ (denoted $LMP(\mathcal{X})$) if $x$ is the first element of $\mathcal{X}$ when $\mathcal{X}$ is ordered lexicographically. Now, define $S_{n,k,A}$ to be the number of empty $(k-1)$-simplices formed from $X_1, \ldots, X_n$, such that the left-most point of the $k$-simplex is in $A$. Define $\widetilde{S}_{n,k,A}$ in the analogous way.

**Lemma 3.3.** *For* $k > 1$, *let*

$$\mu_A := \left( \int_A f(x)^k dx \right) \int_{(\mathbb{R}^d)^{k-1}} h_1(0, y_2, \ldots, y_k) d(y_2, \ldots, y_k).$$

*Then*

$$\lim_{n \to \infty} n^{-k} r_n^{-d(k-1)} \mathbb{E}\left[S_{n,k,A}\right] = \lim_{n \to \infty} n^{-k} r_n^{-d(k-1)} \mathbb{E}[\widetilde{S}_{n,k,A}] = \frac{\mu_A}{k!}.$$

Observe that $\mu_A$ depends only on $f$ and $A$ and can be trivially bounded by $\|f\|_\infty^{k-1}(2^d\theta_d)^{k-1}$, where $\theta_d$ is the volume of the unit ball in $\mathbb{R}^d$.

**Lemma 3.4.** *Let*

$$\mu' := \left( \int_{\mathbb{R}^d} f(x)^{k+1} dx \right) \int_{(\mathbb{R}^d)^k} g_1^{1,2}(0, y_1, \ldots, y_k) dy_1 \cdots dy_k,$$

*where* $g_1^{1,2}(x_0, \ldots, x_k)$ *is the indicator that* $\{x_0, \ldots, x_{k-2}\}$ *form a simplex (where a complex is built as described on* $x_0, \ldots, x_k$ *with threshhold radius* 1*) and that* $\{x_0, x_{k-1}\}$ *and* $\{x_1, x_k\}$ *are edges. Let*

$$\mu'' := \left( \int_{\mathbb{R}^d} f(x)^{k+1} dx \right) \int_{(\mathbb{R}^d)^k} k_1^1(0, y_1, \ldots, y_k) dy_1 \cdots dy_k.$$

*Let* $k_1^1(x_0, \ldots, x_k)$ *be the indicator that* $\{x_0, \ldots, x_{k-2}\}$ *form a simplex and that* $\{x_0, x_{k-1}\}$ *and* $\{x_{k-1}, x_k\}$ *are edges. Then*

$$\lim_{n \to \infty} n^{-(k+1)} r_n^{-dk} \mathbb{E}[Y_{n,k}] = \frac{\mu'}{2(k-3)!},$$

*and*

$$\lim_{n\to\infty} n^{-(k+1)} r_n^{-dk} \mathbb{E}[Z_{n,k}] = \frac{\mu''}{(k-2)!}.$$

**Corollary 3.5.** *For $S_{n,k}, Y_{n,k}, Z_{n,k}$ as above,*

$$\mathbb{E}[S_{n,k} + Y_{n,k} + Z_{n,k}] \simeq \mathbb{E}[\widetilde{S}_{n,k}].$$

The proofs of these facts are identical to the proofs of the corresponsing facts for subgraph counts of random geometric graphs given in Chapter 3 of [15].

This last corollary is already enough to prove part (i) of Theorem 3.2: if $n^k r_n^{d(k-1)} \to 0$ as $n \to \infty$, then

$$\mathbb{P}\big[\beta_k(\mathcal{C}(X_1,\ldots,X_n) \ge 1\big] \le \mathbb{E}\big[\beta_k(\mathcal{C}(X_1,\ldots,X_n)\big] \le \mathbb{E}\big[S_{n,k} + Y_{n,k} + Z_{n,k}\big] \xrightarrow{n\to\infty} 0.$$

In order to prove part (ii), the following abstract approximation theorem of Arratia, Goldstein, and Gordon is needed.

**Theorem 3.6** ([1])**.** *Let $(\xi_i, i \in I)$ be a finite collection of Bernoulli random variables with dependency graph $(I, \sim)$. Let $p_i := \mathbb{E}[\xi_i]$ and $p_{ij} := \mathbb{E}[\xi_i \xi_j]$. Let $\lambda := \sum_{i\in I} p_i$, and let $W := \sum_{i \in I} \xi_i$. Then*

$$d_{TV}(W, Poi(\lambda)) \le \min(3, \lambda^{-1}) \left( \sum_{i\in I} \sum_{\substack{j\sim i\\ j\neq i}} p_{ij} + \sum_{i\in I}\sum_{j\sim i} p_i p_j \right).$$

Penrose [15] used this theorem to prove Poisson approximation results for subgraph counts of random geometric graphs; one can follow this approach essentially without change to prove the following result, which holds in the entire sparse regime.

**Theorem 3.7.** *With definitions as above,*

$$d_{TV}\big(S_{n,k}, Poi(\mathbb{E}[S_{n,k}])\big) \le c_{k,d,f}\big[nr_n^d\big],$$

*for a constant $c_{d,k,f}$ depending only on $d$, $k$, and $\|f\|_\infty$.*

**Corollary 3.8.** *If $n^k r_n^{d(k-1)} \to \alpha \in (0,\infty)$ as $n \to \infty$, then*

$$d_{TV}\big(\widetilde{S}_{n,k}, Poi(\mathbb{E}[\widetilde{S}_{n,k}])\big) \le \tilde{c}_{d,k,f}\,\alpha(nr_n^d).$$

That is, in the regime of part (ii) of the theorem, the lower bound for $\beta_k$ given in (5) is approximately Poisson.

*Proof.* Note that $S_{n,k} - \widetilde{S}_{n,k}$ is the number of empty $(k-1)$-simplices among $\{X, \ldots, X_n\}$ which are not isolated, and is thus bounded above by the number of connected subsets of $\{X_{,} \ldots, X_n\}$ with $k+1$ points, $k$ of which form an empty $k$-simplex. The expected number of such sets is bounded by

$$\binom{n}{k+1} k \|f\|_\infty^{k+1} \theta_d^{k+1} (2r_n)^{d(k-1)} (4r_n)^d \simeq \left( \frac{k\|f\|_\infty^{k+1} \theta_d^{k+1} 2^{d(k+1)}}{(k+1)!} \right) n^{k+1} r_n^{dk},$$

so that

$$\begin{aligned} d_{TV}(S_{n,k},\widetilde{S}_{n,k}) &= \left|\mathbb{P}[S_{n,k}\in A]-\mathbb{P}[\widetilde{S}_{n,k}\in A]\right| \\ &= \left|\mathbb{P}[S_{n,k}\in A, S_{n,k}\neq \widetilde{S}_{n,k}]-\mathbb{P}[\widetilde{S}_{n,k}\in A, S_{n,k}\neq\widetilde{S}_{n,k}]\right| \\ &\le c_{d,k,f}n^{k+1}r_n^{dk} \\ &\le \tilde{c}_{d,k,f}\alpha n r_n^d. \end{aligned}$$

Moreover, it is easy to see in general that if $Y_\alpha$ and $Y_\beta$ have Poisson distributions with means $\alpha$ and $\beta$, respectively, then $d_{TV}(Y_\alpha,Y_\beta)\le|\alpha-\beta|$, and so

$$d_{TV}(Poi(\mathbb{E}[S_{n,k}]),Poi(\mathbb{E}[\widetilde{S}_{n,k}]))\le c_{d,k,f}\alpha n r_n^d$$

as well.

□

The following result, proved below using Theorem 3.6, holds throughout the sparse regime.

**Theorem 3.9.** *There is a constant $c_{d,k,f}$ depending on $d$, $k$, and $f$ only, so that with $S_{n,k},Y_{n,k},Z_{n,k}$ as above,*

$$d_{TV}(S_{n,k}+Y_{n,k}+Z_{n,k},Poi(\mathbb{E}[\widetilde{S}_{n,k}]))\le c_{d,k,f}nr_n^d.$$

The inequalities in (5) together with Corollary 3.8 and Theorem 3.9 yield part (ii) almost immediately.

*Proof of part (ii) of Theorem 3.2.* By the left-hand inequality in (5) and Corollary 3.8,

$$\mathbb{P}[\beta_{k-1}\le m]\le\mathbb{P}[\widetilde{S}_{n,k}\le m]\le\mathbb{P}[Y\le m]+c_{d,k,f}nr_n^d,$$

where $Y$ is a Poisson random variable with mean $\mathbb{E}[\widetilde{S}_{n,k}]$.

By the right-hand inequality in (5) and Theorem 3.9,

$$\mathbb{P}[\beta_{k-1}\le m]\ge\mathbb{P}[S_{n,k}+Y_{n,k}+Z_{n,k}\le m]\ge\mathbb{P}[Y\le m]-c_{d,k,f}nr_n^d.$$

As in the previous proof, $Y$ can be replaced by a Poisson random variable with mean $\mathbb{E}[\beta_k(\mathcal{C}(X_1,\dots,X_n))]$ with only a change of constant in the error term. □

*Proof of Theorem 3.9.* For notational convenience, let $W_{n,k}:=S_{n,k}+Y_{n,k}+Z_{n,k}$. For $1\le p<q\le k-1$, let $g_{r_n}^{p,q}(x_1,\dots,x_{k+1})$ be the indicator that $\{x_1,\dots,x_{k-1}\}$ form a simplex (where a complex is built as described on $x_1,\dots,x_{k+1}$ with threshhold radius $r_n$) and that $\{x_p,x_k\}$ and $\{x_q,x_{k+1}\}$ are edges. Let $k_{r_n}^p(x_1,\dots,x_{k+1})$ be the indicator that $\{x_1,\dots,x_{k-1}\}$ form a simplex and that $\{x_p,x_k\}$ and $\{x_k,x_{k+1}\}$ are edges. For $\mathbf{j}=(j_1,\dots,j_{k+1})$, let $\gamma_{\mathbf{j}}^{p,q}=g_{r_n}^{p,q}(X_{j_1},\dots,X_{j_{k+1}})$ and let $\eta_{\mathbf{j}}^p=k_{r_n}^p(X_{j_1},\dots,X_{j_{k+1}})$. Then

$$\begin{aligned} W_{n,k}=\sum_{1\le i_1<\cdots<i_k\le n}\xi_{\mathbf{i}}+\sum_{\substack{1\le j_1<\cdots<j_{k-1}\le n\\ j_k,j_{k+1}\notin\{j_1,\dots,j_{k-1}\}\\ j_k\neq j_{k+1}}}\ \sum_{1\le p<q\le k-1}\gamma_{\mathbf{j}}^{p,q} \\ +\sum_{\substack{1\le j_1<\cdots<j_{k-1}\le n\\ j_k,j_{k+1}\notin\{j_1,\dots,j_{k-1}\}\\ j_k\neq j_{k+1}}}\ \sum_{1\le p\le k-1}\eta_{\mathbf{j}}^p. \end{aligned}$$

The proof that $W_{n,k}$ has an approximate Poisson distribution proceeds along the same lines as the proof given by Penrose for subgraph counts. For the Bernoulli random variables in the sum above, one can take a dependency graph to be $\mathbf{i} \sim \mathbf{j}$ if $\mathbf{i} \cap \mathbf{j} \neq \emptyset$. (Abusing notation, $\mathbf{i}$ is also used here to denote the set of indices from the multiindex $\mathbf{i}$.) Note that it is not important that $\mathbf{i}$ and $\mathbf{j}$ be the same size.

Now, $\mathbb{E}[\xi_{\mathbf{i}}] \leq [(2r_n)^d \theta_d \|f\|_\infty]^{k-1}$ and if $|\mathbf{i} \cap \mathbf{i}'| = \ell$, then

$$\mathbb{E}[\xi_{\mathbf{i}} \xi_{\mathbf{i}'}] \leq [(2r_n)^d \theta_d \|f\|_\infty]^{2k-\ell-1},$$

since if set of $k$ points forms a simplex, they must all be in the ball of radius $2r_n$ about the first point. Given $\mathbf{i} = (i_1, \ldots, i_k)$, the number of $\mathbf{i}' = (i_1', \ldots, i_k')$ with $\mathbf{i} \sim \mathbf{i}'$ (including $\mathbf{i}$ itself) is

$$\binom{n}{k} - \binom{n-k}{k} = \frac{k^2 n^{k-1}}{k!} + O\left(n^{k-2}\right);$$

for $\mathbf{i}$ as above, the number of $\mathbf{i} = (i_1', \ldots, i_k')$ with $\left|\mathbf{i} \cap \mathbf{i}'\right| = \ell$ is

$$\binom{k}{\ell}\binom{n-k}{k-\ell} = \binom{k}{\ell}\frac{1}{(k-\ell)!} n^{k-\ell} + O\left(n^{k-\ell-1}\right).$$

This means that the contribution to the error term (without the $\min(3, \lambda^{-1})$ factor in front) from Theorem 3.6 of the form $p_{\mathbf{i}} p_{\mathbf{i}'}$ for $\mathbf{i} \sim \mathbf{i}'$ is, to top-order in $n$,

$$\frac{k n^{2k-1}}{k!(k-1)!} \left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k-2},$$

and the contribution from terms of the form $p_{\mathbf{i}\mathbf{i}'}$ is (to top order)

$$\binom{n}{k} \sum_{\ell=1}^{k-1} \binom{k}{\ell} \frac{1}{(k-\ell)!} n^{k-\ell} [(2r_n)^d \theta_d \|f\|_\infty]^{2k-\ell-1} \lesssim n^{k+1} r_n^{dk}.$$

Similar to above, $\mathbb{E}[\gamma_{\mathbf{j}}^{p,q}] \leq 2^d \left[(2r_n)^d \theta_d \|f\|_\infty\right]^k$ and if $|\mathbf{j} \cap \mathbf{j}'| = \ell$, then

$$\mathbb{E}[\gamma_{\mathbf{j}}^{p,q} \gamma_{\mathbf{j}'}^{p',q'}] \leq 2^{3d} \left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k+1-\ell}.$$

Given $\mathbf{j} = (j_1, \ldots, j_{k+1})$, the number of $\mathbf{j}' = (j_1', \ldots, j_{k+1}')$ with $\mathbf{j} \sim \mathbf{j}'$ is

$$\frac{(k+1)^2 n^k}{(k+1)!} + O(n^{k-1})$$

and the number of $\mathbf{j}'$ with $|\mathbf{j} \cap \mathbf{j}'| = \ell$ is

$$\binom{k+1}{\ell} \frac{n^{k+1-\ell}}{(k+1-\ell)!} + O(n^{k-\ell}).$$

This yields a top-order contribution to the error from Theorem 3.6 from the $\mathbb{E}[\gamma_{\mathbf{j}}]\mathbb{E}[\gamma_{\mathbf{j}'}]$ and $\mathbb{E}[\gamma_{\mathbf{j}} \gamma_{\mathbf{j}'}]$ terms of order

$$\begin{aligned}
&\frac{(k+1)^2 n^{2k+1}}{[(k+1)!]^2} \binom{k-1}{2}^2 2^{2d} \left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k} \\
&\qquad + \binom{n}{k+1} \sum_{\ell=1}^{k+1} \binom{k-1}{2}^2 \binom{k+1}{\ell} \frac{n^{k+1-\ell}}{(k+1-\ell)!} 2^{3d} \left[(2r_n)^d \theta_d \|f\|_\infty\right]^{2k+1-\ell} \\
&\qquad \lesssim n^{k+1} r_n^{dk}.
\end{aligned}$$

In the same way, $\mathbb{E}[\eta_{\mathbf{j}}^p] \le 2^d \left[(2r_n)^d\theta_d\|f\|_\infty\right]^k$, and if $|\mathbf{j}\cap\mathbf{j'}| = \ell$, then

$$\mathbb{E}[\eta_{\mathbf{j}}^p\eta_{\mathbf{j'}}^{p'}] \le 2^{3d}\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k+1-\ell},$$

thus the contribution from the terms of the form $\mathbb{E}[\eta_{\mathbf{j}}]\mathbb{E}[\eta_{\mathbf{j'}}]$ and of the form $\mathbb{E}[\eta_{\mathbf{j}}\eta_{\mathbf{j'}}]$ is of the same order as the contribution above from the corresponding $\gamma$ terms.

The cross terms are essentially the same: if $|\mathbf{i}\cap\mathbf{j}| = \ell$, then

$$\mathbb{E}[\xi_{\mathbf{i}}\gamma_{\mathbf{j}}^{p,q}] \le 2^{2d}\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k-\ell} \qquad \mathbb{E}[\xi_{\mathbf{i}}\eta_{\mathbf{j}}^{p}] \le 2^{3d}\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k-\ell}$$
$$\mathbb{E}[\gamma_{\mathbf{i}}^{p,q}\eta_{\mathbf{j}}^{r}] \le 2^{4d}\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k+1-\ell}.$$

The number of $\mathbf{j} = (j_1, \ldots, j_{k+1})$ with $\mathbf{i} \sim \mathbf{j}$ is

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

and the number of such $\mathbf{j}$ with $|\mathbf{i}\cap\mathbf{j}| = \ell$ is

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

This yields a contribution from the $\xi$-$\gamma$ cross-terms of

$$\begin{aligned}&\frac{n^{2k}}{k!(k-1)!}\binom{k-1}{2}2^d\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k-1}\\ &\qquad\qquad + \binom{n}{k}\sum_{\ell=0}^{k}\binom{k-1}{2}\binom{k}{\ell}\frac{n^{k+1-\ell}}{(k+1-\ell)!}2^{2d}\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k-\ell}\\ &\qquad\qquad \lesssim n^{k+1}r_n^{dk}.\end{aligned}$$

The contribution from the $\xi$-$\eta$ cross terms is the same up to constants depending only on $k$ and $d$, and the contribution from the $\gamma$-$\eta$ cross terms is

$$\begin{aligned}&\frac{(k+1)^2n^{2k+1}}{[(k+1)!]^2}(k-1)\binom{k-1}{2}2^{2d}\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k}\\ +\ &\binom{n}{k+1}\sum_{\ell=0}^{k+1}(k-1)\binom{k-1}{2}\binom{k+1}{\ell}\frac{n^{k+1-\ell}}{(k+1-\ell)!}2^{4d}\left[(2r_n)^d\theta_d\|f\|_\infty\right]^{2k+1-\ell}\\ \lesssim\ & n^{k+1}r_n^{dk}.\end{aligned}$$

Collecting terms and using that $\lambda = \mathbb{E}[W_{n,k}] \simeq n^kr_n^{d(k-1)}\left(\frac{\mu}{k!}\right)$, Theorem 3.6 yields

$$d_{TV}(W, Poi(\lambda)) \le c_{d,k,f}nr_n^d.$$

Again, one can replace $\lambda$ with $\mathbb{E}[\widetilde{S}_{n,k}]$ with only a loss in the value of the constant $c_{d,k,f}$.

□

The remainder of the section is devoted to the proof of part (iii) of Theorem 3.2. A central limit theorem for the recentered, renormalized upper bound of $\beta_k$ given in (5) follows immediately from Theorem 3.9 in this range of $r_n$, by the classical result that a Poisson random variable with mean tending to infinity tends to a Gaussian random variable when recentered and renormalized.

**Theorem 3.10.** *If $nr_n^d \xrightarrow{n\to\infty} 0$ and $n^k r_n^{d(k-1)} \xrightarrow{\infty}$, then*

$$\frac{S_{n,k} + Y_{n,k} + Z_{n,k} - \mathbb{E}[\widetilde{S}_{n,k}]}{\sqrt{\mathbb{E}[\widetilde{S}_{n,k}]}} \Longrightarrow \mathcal{N}(0,1)$$

*as $n$ tends to infinity.*

Clearly the approach to the lower bound of (5) taken in the regime in which $n^k r_n^{d(k-1)} \to \alpha \in (0,\infty)$ also works in the case that $n^k r_n^{d(k-1)}$ tends to infinity but $n^{k+1} r_n^{dk}$ tends to zero to show that $\widetilde{S}_{n,k}$ is approximately Gaussian in that regime as well. However, to deal with the regime in which $r_n = o(n^{-1/d})$ but $n^{k+1} r_n^{dk}$ is bounded away from zero, a different argument is needed for the lower bound of (5). Following Penrose, the approach taken here is to consider the Poissonized version of the problem (the vertices distributed as a Poisson process of intensity $nf(\cdot)$ instead of i.i.d. with density $f$), and then to recover the i.i.d. case.

Let $N_n$ be a Poisson random variable with mean $n$, and let $\mathcal{P}_n = \{X_1, \ldots, X_{N_n}\}$, where $\{X_i\}_{i=1}^{\infty}$ is an i.i.d. sequence of random points in $\mathbb{R}^d$ with density $f$. Then $\mathcal{P}_n$ is a Poisson process with intensity $nf(\cdot)$, and one can define $S_{n,k}^P$ and $\widetilde{S}_{n,k}^P$ for the random points $\mathcal{P}_n$ analogously to the earlier definitions. In what follows, assume that $k \geq 3$; that is, the empty $(k-1)$-simplices are at least empty triangles. Empty 1-simplices are simply pairs of vertices which are not connected, and different arguments are needed in that case.

In order to compute expectations for the expressions which arise in the Poissonized case, the following results are useful.

**Theorem 3.11** (See [15]). *Let $\lambda > 0$ and let $\mathcal{P}_\lambda$ be a Poisson process with intensity $\lambda f(\cdot)$. Let $j \in \mathbb{N}$, and suppose that $h(\mathcal{Y}, \mathcal{X})$ is a bounded measurable function on pairs $(\mathcal{Y}, \mathcal{X})$ with $\mathcal{X}$ a finite subset of $\mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathcal{X}$, such that $h(\mathcal{Y}, \mathcal{X}) = 0$ unless $|\mathcal{Y}| = j$. Then*

$$\mathbb{E}\left[\sum_{\mathcal{Y} \subseteq \mathcal{P}_\lambda} h(\mathcal{Y}, \mathcal{P}_\lambda)\right] = \frac{\lambda^j}{j!} \mathbb{E} h(\mathcal{X}_j', \mathcal{X}_j' \cup \mathcal{P}_\lambda),$$

*where $\mathcal{X}_j'$ is a set of $j$ i.i.d. points in $\mathbb{R}^d$ with density $f$, independent of $\mathcal{P}_\lambda$.*

From this, one can prove the following.

**Theorem 3.12.** *Let $\lambda > 0$ and $k, j_1, \ldots, j_k \in \mathbb{N}$; define $j := \sum_{i=1}^k j_i$. For $1 \leq i \leq k$, suppose $h_i(\mathcal{Y}, \mathcal{X})$ is a bounded measurable function of pairs $(\mathcal{Y}, \mathcal{X})$ of finite subsets of $\mathbb{R}^d$ with $\mathcal{Y} \subseteq \mathcal{X}$, such that $h_i(\mathcal{Y}, \mathcal{X}) = 0$ if $|\mathcal{Y}| \neq j_i$. Then*

$$\mathbb{E}\left[\sum_{\mathcal{Y}_1, \subseteq \mathcal{P}_\lambda} \cdots \sum_{\mathcal{Y}_k \subseteq \mathcal{P}_\lambda} \left(\prod_{i=1}^k h_i(\mathcal{Y}_i)\right) \mathbb{1}_{\{\mathcal{Y}_i \cap \mathcal{Y}_j = \emptyset \text{ for } i \neq j\}}\right] = \mathbb{E}\left[\prod_{i=1}^k \left(\frac{\lambda^{j_i}}{j_i!}\right) h_i(\mathcal{X}_{j_i}', \mathcal{X}_j' \cup \mathcal{P}_n)\right],$$

where $\mathcal{X}_j'$ are $j$ i.i.d points in $\mathbb{R}^d$ with density $f$, $\mathcal{P}_\lambda$ is a Poisson process with intensity $\lambda f(\cdot)$, and $\mathcal{X}_j'$ and $\mathcal{P}_\lambda$ are independent.

*Proof.* Consider the case $k = 2$ for simplicity (the case of larger $k$ is the same with more notation). Define $h(\mathcal{Y}, \mathcal{X})$ on subsets $\mathcal{Y}$ of $\mathcal{X}$ of size $j_1 + j_2$ by

$$h(\mathcal{Y}, \mathcal{X}) := \sum_{\substack{\mathcal{Y}_1 \subseteq \mathcal{Y} \\ |\mathcal{Y}_1| = j_1}} h_1(\mathcal{Y}_1, \mathcal{X}) h_2(\mathcal{Y} \setminus \mathcal{Y}_1, \mathcal{X}).$$

Then by Theorem 3.11,

$$\begin{aligned}\mathbb{E}&\left[\sum_{\mathcal{Y}_1,\subseteq\mathcal{P}_\lambda}\sum_{\mathcal{Y}_2,\subseteq\mathcal{P}_\lambda}h_1(\mathcal{Y}_1,\mathcal{P}_n)h_2(\mathcal{Y}_2,\mathcal{P}_n)\mathbb{1}_{\{\mathcal{Y}_1\cap\mathcal{Y}_2=\emptyset\}}\right]\\&=\mathbb{E}\left[\sum_{\mathcal{Y}\subseteq\mathcal{P}_n}h(\mathcal{Y},\mathcal{P}_n)\right]\\&=\frac{\lambda^{j_1+j_2}}{(j_1+j_2)!}\mathbb{E}h(\mathcal{X}_j',\mathcal{X}_j'\cup\mathcal{P}_n)\\&=\frac{\lambda^{j_1+j_2}}{j_1!j_2!}\mathbb{E}\left[h_1(\mathcal{X}_{j_1}',\mathcal{X}_j'\cup\mathcal{P}_n)h_2(\mathcal{X}_j'\setminus\mathcal{X}_{j_1}',\mathcal{X}_j'\cup\mathcal{P}_n)\right].\end{aligned}$$

□

One can apply these results to compute the mean and variance of $\widetilde{S}^P_{n,k,A}$, the number of isolated empty $k$-simplices in $\mathcal{P}_n$ whose left-most vertex is in the set $A$. Recall that $A$ is assumed to be open with $\operatorname{vol}(\partial A)=0$.

**Lemma 3.13.** *For $\mu_A$ as in Lemma 3.3,*

$$\lim_{n\to\infty}n^{-k}r_n^{-d(k-1)}\mathbb{E}\left[\widetilde{S}^P_{n,k}\right]=\lim_{n\to\infty}n^{-k}r_n^{-d(k-1)}\operatorname{Var}\left[\widetilde{S}^P_{n,k}\right]=\frac{\mu_A}{k!}.$$

*Proof.* Let $\tilde{h}_{r_n,A}(\{x_1,\ldots,x_k\},\mathcal{X})$ be the indicator that $\{x_1,\ldots,x_k\}\subseteq\mathcal{X}$ form an isolated empty $(k-1)$-simplex in $\mathcal{X}$, whose left-most point is in $A$. Then

$$\mathbb{E}[\widetilde{S}^P_{n,k,A}]=\mathbb{E}\left[\sum_{\mathcal{Y}\subseteq\mathcal{P}_\lambda}\tilde{h}_{r_n,A}(\mathcal{Y},\mathcal{P}_n)\right]=\frac{n^k}{k!}\mathbb{E}\left[\tilde{h}_{r_n,A}(\mathcal{X}_k',\mathcal{X}_k'\cup\mathcal{P}_n)\right].\tag{6}$$

Now, $\mathbb{E}\left[\tilde{h}_{r_n,A}(\mathcal{X}_k',\mathcal{X}_k'\cup\mathcal{P}_n)\right]\le\mathbb{E}\left[h_{r_n,A}(\mathcal{X}_k')\right]\simeq r_n^{d(k-1)}\mu_A$. Note that the conditional probability that $\mathcal{X}_k'$ is isolated from $\mathcal{P}_n$ given that $\mathcal{X}_k'$ forms an empty $(k-1)$-simplex with left-most vertex in $A$ is bounded below by the probability that there are no points of $\mathcal{P}_n$ in the ball of radius $4r_n$ about $X_1$, which is given by $e^{-n\operatorname{vol}_f(B_{4r_n}(X_1))}\ge e^{-n\|f\|_\infty\theta_d(4r_n)^d}$, since $\mathcal{P}_n$ is a Poisson process with intensity $nf(\cdot)$. It thus follows that

$$\mathbb{E}\left[\tilde{h}_{r_n,A}(\mathcal{X}_k',\mathcal{X}_k'\cup\mathcal{P}_n)\right]\ge e^{-n\|f\|_\infty\theta_d(4r_n)^d}\mathbb{E}[h_{r_n,A}(\mathcal{X}_k')]\simeq e^{-n\|f\|_\infty\theta_d(4r_n)^d}r_n^{d(k-1)}\mu_A.$$

Since $nr_n^d\to0$, this shows that

$$\mathbb{E}[\widetilde{S}^P_{n,k}]\simeq\frac{n^kr_n^{d(k-1)}\mu_A}{k!}.$$

A similar approach is taken to compute the variance:

$$\begin{aligned}\mathbb{E}\left[(\widetilde{S}^P_{n,k,A})^2\right]&=\mathbb{E}\left[\sum_{\mathcal{Y}\subseteq\mathcal{P}_n}\tilde{h}_{r_n,A}(\mathcal{Y},\mathcal{P}_n)\right]\\&\quad+\mathbb{E}\left[\sum_{j=0}^{k-1}\sum_{\mathcal{Y},\mathcal{Y}'\subseteq\mathcal{P}_n}\tilde{h}_{r_n,A}(\mathcal{Y},\mathcal{P}_n)\tilde{h}_{r_n,A}(\mathcal{Y}',\mathcal{P}_n)\mathbb{1}_{\{|\mathcal{Y}\cap\mathcal{Y}'|=j\}}\right].\end{aligned}$$

The first summand has already been analyzed: $\mathbb{E}\left[\widetilde{S}^P_{n,k,A}\right] \simeq \frac{n^k r_n^{d(k-1)}\mu_A}{k!}$. For the second, observe first that the terms corresponding to $j \neq 0$ vanish: $\tilde{h}_{r_n,A}(\mathcal{Y}, \mathcal{P}_n)\tilde{h}_{r_n,A}(\mathcal{Y}', \mathcal{P}_n) \equiv 0$ if $|\mathcal{Y} \cap \mathcal{Y}'| = j$, because if $\mathcal{Y}$ and $\mathcal{Y}'$ both form empty $k$-simplices, then neither is isolated. When $j = 0$, applying Theorem 3.12 yields

$$\mathbb{E}\left[\sum_{\mathcal{Y},\mathcal{Y}'\subseteq\mathcal{P}_n} \tilde{h}_{r_n,A}(\mathcal{Y}, \mathcal{P}_n)\tilde{h}_{r_n,A}(\mathcal{Y}', \mathcal{P}_n)\mathbb{1}_{\{\mathcal{Y}\cap\mathcal{Y}'=\emptyset\}}\right]$$
$$= \frac{n^{2k}}{(k!)^2}\mathbb{E}\left[\tilde{h}_{r_n,A}(\mathcal{X}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\tilde{h}_{r_n,A}(\mathcal{X}'_{2k} \setminus \mathcal{X}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\right],$$

and thus (making use of (6)),

$$\mathrm{Var}\left[\widetilde{S}^P_{n,k,A}\right] = \mathbb{E}\left[\widetilde{S}^P_{n,k,A}\right] + \frac{n^{2k}}{(k!)^2}\Big(\mathbb{E}\left[\tilde{h}_{r_n,A}(\mathcal{X}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\tilde{h}_{r_n,A}(\mathcal{X}'_{2k} \setminus \mathcal{X}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\right]$$
$$- \left(\mathbb{E}\left[\tilde{h}_{r_n,A}(\mathcal{X}'_k, \mathcal{X}'_k \cup \mathcal{P}_n)\right]\right)^2\Big),$$

Now, let $\mathcal{P}'_n$ be an independent copy of $\mathcal{P}_n$. For notational convenience, denote $\mathcal{X}'_{2k} \setminus \mathcal{X}'_k$ by $\mathcal{Y}'_k$ and abbreviate $\tilde{h}_{r_n,A}$ by $\tilde{h}$. Then

$$\begin{aligned}
&\mathbb{E}\left[\tilde{h}(\mathcal{X}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\tilde{h}(\mathcal{Y}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\right] - \left(\mathbb{E}\left[\tilde{h}(\mathcal{X}'_k, \mathcal{X}'_k \cup \mathcal{P}_n)\right]\right)^2 \\
&\quad= \mathbb{E}\left[\tilde{h}(\mathcal{X}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\tilde{h}(\mathcal{Y}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n) - \tilde{h}(\mathcal{X}'_k, \mathcal{X}'_k \cup \mathcal{P}_n)\tilde{h}(\mathcal{Y}'_k, \mathcal{Y}'_k \cup \mathcal{P}'_n)\right] \\
&\quad= \mathbb{E}\left[\left(\tilde{h}(\mathcal{X}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n) - \tilde{h}(\mathcal{X}'_k, \mathcal{X}'_k \cup \mathcal{P}_n)\right)\tilde{h}(\mathcal{Y}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n)\right] \\
&\qquad+ \mathbb{E}\left[\tilde{h}(\mathcal{X}'_k, \mathcal{X}'_k \cup \mathcal{P}_n)\left(\tilde{h}(\mathcal{Y}'_k, \mathcal{X}'_{2k} \cup \mathcal{P}_n) - \tilde{h}(\mathcal{Y}'_k, \mathcal{Y}'_k \cup \mathcal{P}_n)\right)\right] \\
&\qquad+ \mathbb{E}\left[\tilde{h}(\mathcal{X}'_k, \mathcal{X}'_k \cup \mathcal{P}_n)\left(\tilde{h}(\mathcal{Y}'_k, \mathcal{Y}'_k \cup \mathcal{P}_n) - \tilde{h}(\mathcal{Y}'_k, \mathcal{Y}'_k \cup \mathcal{P}'_n)\right)\right] \\
&\quad= E_1 + E_2 + E_3.
\end{aligned}$$

Now, observe that in fact $E_1 = 0$: the difference is non-zero if and only if $\mathcal{X}'_k$ and $\mathcal{Y}'_k$ are connected by an edge, in which case the second factor is zero.

Observe that the difference in $E_2$ is non-positive. Furthermore, it is non-zero if and only if $\mathcal{X}'_k$ and $\mathcal{Y}'_k$ are connected by an edge, and both $\mathcal{X}'_k$ and $\mathcal{Y}'_k$ form empty $k$-simplices. This probability is bounded above by $\|f\|_\infty^{2k-1}\theta_d^{2k-1}(2r_n)^{2d(k-1)}(8r_n)^d$.

Finally, if $\left[\cup_{i=1}^k B_{2r_n}(X'_i)\right] \cap \left[\cup_{i=k+1}^{2k} B_{2r_n}(X'_i)\right] = \emptyset$, then the two terms of $E_3$ have the same distribution by the spacial independence property of the Poisson process. A contribution from $E_3$ therefore only arises if in particular $|X_1 - X_j| \leq 2r_n$ for each $2 \leq j \leq k$, if $|X_{k+1} - X_j| \leq 2r_n$ for $k+2 \leq j \leq 2k$, and $|X_1 - X_{k+1}| \leq 8r_n$. The probability of this event is bounded above by $\|f\|_\infty^{2k-1}\theta_d^{2k-1}(2r_n)^{2d(k-1)}(8r_n)^d$. It follows that

$$\mathrm{Var}\left[\widetilde{S}^P_{n,k,A}\right] = \mathbb{E}\left[\widetilde{S}^P_{n,k,A}\right] + E,$$

and

$$|E| \leq \frac{n^{2k}(2r_n)^{2dk-d}}{(k!)^2} 2\|f\|_\infty^{2k-1}\theta_d^{2k-1}4^d = C(f, d, k)(nr_n^d)^k(n^k r_n^{d(k-1)}),$$

where $C(f, k, d)$ is a constant depending on $f$, $d$, and $k$. This completes the proof. □

The following abstract normal approximation theorem is another version of the dependency graph approach to Stein's method. It is used in what follows to prove a central limit theorem for $\widetilde{S}^P_{n,k}$.

**Theorem 3.14** (Penrose)**.** *Suppose $\{\xi_i\}_{i\in I}$ is a finite collection of random variables with dependency graph $(I,\sim)$ with maximum degree $D-1$, with $\mathbb{E}[\xi_i]=0$ for each $i$. Set $W:=\sum_{i\in I}\xi_i$; suppose $\mathbb{E}[W^2]=1$. Let $Z$ be a standard normal random variable. Then for all $t\in\mathbb{R}$,*

$$\big|\mathbb{P}[W\le t]-\mathbb{P}[Z\le t]\big|\le\frac{2}{\sqrt[4]{2\pi}}\sqrt{D^2\sum_{i\in I}\mathbb{E}|\xi_i|^3}+6\sqrt{D^3\sum_{i\in I}\mathbb{E}|\xi_i|^4}.$$

Making use of this result, we prove the following.

**Theorem 3.15.** *With notation as above, and for $n^k r_n^{d(k-1)}\to\infty$ and $nr_n^d\to 0$,*

$$\frac{\widetilde{S}^P_{n,k}-\mathbb{E}\left[\widetilde{S}^P_{n,k}\right]}{\sqrt{\operatorname{Var}\left[\widetilde{S}^P_{n,k}\right]}}\Rightarrow\mathcal{N}(0,1).$$

*Proof.* To define a dependency graph for the summands of $\widetilde{S}^P_{n,k}$, the independence properties of the Poisson process are exploited. Let $\{Q_{i,n}\}_{i\in\mathbb{N}}$ be a partition of $\mathbb{R}^d$ into cubes of side length $r_n$. For the moment, assume that $A$ is a bounded set, and let $I_A$ be the set of indices $i$ such that $diam(A\cap Q_{i,n})>2r_n$. Write

$$\widetilde{S}^P_{n,k,A}=\sum_{i\in I_A}\sum_{\mathcal{Y}\subseteq\mathcal{P}_n}\tilde{h}_{r_n,A\cap Q_{i,n}}(\mathcal{Y},\mathcal{P}_n). \tag{7}$$

Observe that if one defines a relation $\sim$ on $I_A$ by $i\sim j$ if and only if the Euclidean distance from $Q_{i,n}$ to $Q_{j,n}$ is less than $8r_n$, then $(I_A,\sim)$ is a dependency graph for the summands in (7). The degree of vertices in this dependency graph is then bounded by $17^d$.

Let $\xi_i:=\sum_{\mathcal{Y}\subseteq\mathcal{P}_n}\tilde{h}_{r_n,A\cap Q_{i,n}}(\mathcal{Y},\mathcal{P}_n)$; to apply Theorem 3.14, bounds are needed for $\mathbb{E}|\xi_i-\mathbb{E}\xi_i|^p$ for $p=3,4$, for which it suffices to have bounds on $\mathbb{E}|\xi|^p$ for $p=3,4$. Observe that if $Z_i$ is the number of points within $2r_n$ of $Q_{i,n}$, then $Z_{i,n}$ is distributed as a Poisson random variable with mean $n\operatorname{vol}_f((Q_{i,n})_{2r_n})$, and

$$|\xi_i|\le(Z_i)(Z_i-1)\cdots(Z_i-k+1)=:(Z_i)_k.$$

It follows that there is a constant $c$ depending only on $d$ and $f$, such that for $\rho_n:=nr_n^d$,

$$\mathbb{E}|\xi_i|^p\le\mathbb{E}(Z_i)_k^p\le\sum_{m=k}^{\infty}(m)_k^p\frac{e^{-c\rho_n}(c\rho_n)^m}{m!}\le c'\rho_n^k$$

for some new constant $c'$ depending only on $d$, $f$, and $k$.

Note that since $A$ is bounded, $|I_A|$ is at worst of the order $r_n^{-d}$, with coefficient depending on $A$. Applying Theorem 3.14 to $\frac{\xi_i-\mathbb{E}\xi_i}{\sqrt{\operatorname{Var}(\widetilde{S}_{n,k,A})}}$ gives

$$\left|\mathbb{P}\left[\frac{\widetilde{S}^P_{n,k,A}-\mathbb{E}\widetilde{S}^P_{n,k,A}}{\sqrt{\operatorname{Var}(\widetilde{S}^P_{n,k,A})}}\le t\right]-\mathbb{P}[Z\le t]\right|\le c''[n^k r_n^{d(k-1)}]^{-1/4},$$

which tends to zero as $n$ tends to infinity.

To move to $A = \mathbb{R}_d$, let $\zeta_{n,k}(A) := \frac{\widetilde{S}^P_{n,k,A} - \mathbb{E}[\widetilde{S}^P_{n,k,A}]}{\sqrt{n^k r_n^{d(k-1)}}}$ and consider $A_K := (-K, K)^d$ and $A^K := \mathbb{R}^d \setminus [-K, K]^d$. Given $t \in \mathbb{R}$ and $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}[\zeta_{n,k}(\mathbb{R}^d) \leq t] = \mathbb{P}[\zeta_{n,k}(A_K) \leq t - \epsilon] &- \mathbb{P}[\{\zeta_{n,k}(A_K) \leq t - \epsilon\} \cap \{\zeta_{n,k}(\mathbb{R}^d) > t\}] \\ &+ \mathbb{P}[\{|\zeta_{n,k}(A_K) - t| < \epsilon\} \cap \{\zeta_{n,k}(\mathbb{R}^d) \leq t\}] \\ &+ \mathbb{P}[\{\zeta_{n,k}(A_K) \geq t + \epsilon\} \cap \{\zeta_{n,k}(\mathbb{R}^d) \leq t\}].\end{aligned}$$

Now, $\zeta_{n,k}(\mathbb{R}^d) = \zeta_{n,k}(A_K) + \zeta_{n,k}(A^K)$ almost surely since $vol(A_K^c \cup (A^K)^c) = 0$, so $\big|\mathbb{P}[\zeta_{n,k}(\mathbb{R}^d) \leq t] - \mathbb{P}[\zeta_{n,k}(A_K) \leq t - \epsilon]\big| \leq \mathbb{P}[|\zeta_{n,k}(A^K)| \geq \epsilon] + \mathbb{P}[|\zeta_{n,k}(A_K) - t| < \epsilon]$. By Chebychev's inequality and the central limit theorem already established for bounded sets, this last expression is bounded above by

$$\begin{aligned}&\frac{1}{\epsilon^2}\mathrm{Var}(\zeta_{n,k}(A^K)) + \mathbb{P}\left[\left|\sqrt{\frac{\mathrm{Var}(\widetilde{S}^P_{n,k,A_K})}{n^k r_n^{d(k-1)}}}Z - t\right| < \epsilon\right] + c_K\left[(n^k r_n^{d(k-1)})^{-1/4}\right] \\ &\qquad \leq \frac{1}{\epsilon^2}\mathrm{Var}(\zeta_{n,k}(A^K)) + \frac{2\epsilon\sqrt{n^k r_n^{d(k-1)}}}{\sqrt{2\pi \mathrm{Var}(\widetilde{S}^P_{n,k,A_K})}} + c_K\left[(n^k r_n^{d(k-1)})^{-1/4}\right] \\ &\qquad \simeq \frac{1}{\epsilon^2}\frac{\mu_{A^K}}{k!} + \frac{2\epsilon\sqrt{k!}}{\sqrt{2\pi\mu_{A_K}}} + c_K\left[(n^k r_n^{d(k-1)})^{-1/4}\right],\end{aligned}$$

for a constant $c_K$ depending on $K$. Taking $n$ to infinity for $K$ and $\epsilon$ fixed yields

$$\limsup_{n\to\infty}\big|\mathbb{P}[\zeta_{n,k}(\mathbb{R}^d) \leq t] - \mathbb{P}[\zeta_{n,k}(A_K) \leq t - \epsilon]\big| \leq \frac{1}{\epsilon^2}\frac{\mu_{A^K}}{k!} + \frac{2\epsilon\sqrt{k!}}{\sqrt{2\pi\mu_{A_K}}},$$

which, together with the central limit theorem for $\zeta_{n,k}(A_K)$, implies that

$$\limsup_{n\to\infty}\left|\mathbb{P}[\zeta_{n,k}(\mathbb{R}^d) \leq t] - \mathbb{P}\left[\sqrt{\frac{\mathrm{Var}(\widetilde{S}^P_{n,k,A_K})}{n^k r_n^{d(k-1)}}}Z \leq t - \epsilon\right]\right| \leq \frac{1}{\epsilon^2}\frac{\mu_{A^K}}{k!} + \frac{2\epsilon\sqrt{k!}}{\sqrt{2\pi\mu_{A_K}}}.$$

Now,

$$\begin{aligned}&\mathbb{P}\left[\sqrt{\frac{\mathrm{Var}(\widetilde{S}^P_{n,k,A_K})}{n^k r_n^{d(k-1)}}}Z \leq t - \epsilon\right] \\ &\quad = \quad \Phi\left(\sqrt{\frac{n^k r_n^{d(k-1)}}{\mathrm{Var}(\widetilde{S}^P_{n,k,A_K})}}(t-\epsilon)\right) \xrightarrow{n\to\infty} \Phi\left(\sqrt{\frac{k!}{\mu_{A_K}}}(t-\epsilon)\right);\end{aligned}$$

that is,

$$\limsup_{n\to\infty}\left|\mathbb{P}[\zeta_{n,k}(\mathbb{R}^d) \leq t] - \Phi\left(\sqrt{\frac{k!}{\mu_{A_K}}}(t-\epsilon)\right)\right| \leq \frac{1}{\epsilon^2}\frac{\mu_{A^K}}{\mu} + \frac{2\epsilon\sqrt{k!}}{\sqrt{2\pi\mu_{A_K}}}.$$

Recall that $\lim_{K\to\infty}\mu_{A_K} = \mu$ and $\lim_{K\to\infty}\mu_{A^K} = 0$. Thus for $n$ and $K$ large enough,

$$\left|\mathbb{P}[\zeta_{n,k}(\mathbb{R}^d) \leq t] - \Phi\left(\sqrt{\frac{k!}{\mu}}(t-\epsilon)\right)\right| \leq \frac{2\epsilon\sqrt{k!}}{\sqrt{2\pi\mu}} + \epsilon.$$

Since $\Phi\left(\sqrt{\frac{k!}{\mu}}(t-\epsilon)\right) \xrightarrow{\epsilon\to 0} \Phi\left(\sqrt{\frac{k!}{\mu}}t\right)$ and $\epsilon$ was arbitrary, this finally shows that

$$\lim_{n\to\infty}\left|\mathbb{P}[\widetilde{S}^P_{n,k}\le t] - \Phi\left(\sqrt{\frac{k!}{\mu}}t\right)\right| = 0.$$

□

The remaining work is to use this result to obtain the same result for $\widetilde{S}_{n,k}$ itself. To do so, the following "de-Poissonization result" is used.

**Theorem 3.16** (See [15]). *Suppose that for each $n\in\mathbb{N}$, $H_n(\mathcal{X})$ is a real-valued functional on finite sets $\mathcal{X}\subseteq\mathbb{R}^d$. Suppose that for some $\sigma^2\ge 0$,*

(i) $\frac{1}{n}\mathrm{Var}(H_n(\mathcal{P}_n)) \longrightarrow \sigma^2$, *and*

(ii) $\frac{1}{\sqrt{n}}\big[H_n(\mathcal{P}_n) - \mathbb{E}H_n(\mathcal{P}_n)\big] \Longrightarrow \sigma^2 Z$, *for $Z$ a standard normal random variable.*

*Suppose that there are constants $\alpha\in\mathbb{R}$ and $\gamma>\frac{1}{2}$ such that the increments $R_{m,n} = H_n(\mathcal{X}_{m+1}) - H_n(\mathcal{X}_m)$ satisfy*

$$\lim_{n\to\infty}\left(\sup_{n-n^\gamma\le m\le n+n^\gamma}|\mathbb{E}[R_{m,n}]-\alpha|\right) = 0, \tag{8}$$

$$\lim_{n\to\infty}\left(\sup_{n-n^\gamma\le m<m'\le n+n^\gamma}|\mathbb{E}[R_{m,n}R_{m',n}]-\alpha^2|\right) = 0, \tag{9}$$

*and*

$$\lim_{n\to\infty}\left(\frac{1}{\sqrt{n}}\sup_{n-n^\gamma\le m\le n+n^\gamma}\mathbb{E}[R^2_{m,n}]\right) = 0. \tag{10}$$

*Finally, assume that there is a constant $\beta>0$ such that, with probability one,*

$$|H_n(\mathcal{X}_m)|\le\beta(n+m)^\beta.$$

*Then $\alpha^2\le\sigma^2$ and as $n\to\infty$, $\frac{1}{n}\mathrm{Var}(H_n(\mathcal{X}_n))\to\sigma^2-\alpha^2$ and*

$$\frac{1}{\sqrt{n}}\big[H_n(\mathcal{X}_n) - \mathbb{E}H_n(\mathcal{X}_n)\big] \Longrightarrow \sqrt{\sigma^2-\alpha^2}Z.$$

In conjunction with Theorem 3.15, this yields the following.

**Theorem 3.17.** *With notation as above, and for $n^k r_n^{d(k-1)}\to\infty$ and $nr_n^d\to 0$,*

$$\frac{\widetilde{S}_{n,k} - \mathbb{E}\left[\widetilde{S}_{n,k}\right]}{\sqrt{\mathrm{Var}\left[\widetilde{S}_{n,k}\right]}} \Rightarrow \mathcal{N}(0,1).$$

*Proof.* Theorem 3.16 is applied to the functional

$$H_n(\mathcal{X}) := \frac{1}{\sqrt{(nr_n^d)^{k-1}}}\sum_{\mathcal{Y}\subseteq\mathcal{X}}\tilde{h}_{r_n}(\mathcal{Y},\mathcal{X});$$

$\sigma^2 = \frac{\mu}{k!}$ and the central limit theorem holds for $H_n(\mathcal{P}_n)$ by Theorem 3.15.

Let $D_{m,n} := \sum_{\mathcal{Y}\subseteq\mathcal{X}_{m+1}}\tilde{h}_{r_n}(\mathcal{Y},\mathcal{X}_{m+1}) - \sum_{\mathcal{Y}\subseteq\mathcal{X}_m}\tilde{h}_{r_n}(\mathcal{Y},\mathcal{X}_m)$, and observe that $D_{m,n}$ is the number of isolated empty $(k-1)$-simplices in $\mathcal{X}_{m+1}$ with $X_{m+1}$ as a vertex, minus the number of empty $(k-1)$-simplices in $\mathcal{X}_m$ which are isolated in $\mathcal{X}_m$ but connected to $X_{m+1}$. Thus

$$\begin{aligned}\text{(11)} \qquad \mathbb{E}[D_{m,n}] &= \binom{m}{k-1}\mathbb{E}[\tilde{h}_{r_n}(\mathcal{X}_k, \mathcal{X}_{m+1})] \\ &\quad -\binom{m}{k}\mathbb{E}[\tilde{h}_{r_n}(\mathcal{X}_k, \mathcal{X}_m)]\mathbb{P}\left[X_{m+1} \in \cup_{i=1}^k B_{2r_n}(X_i)\right].\end{aligned}$$

It is clear that

$$(1-\|f\|_\infty \theta_d (4r_n)^d)^{m+1-k} r_n^{d(k-1)}\mu \le \mathbb{E}[\tilde{h}_{r_n}(\mathcal{X}_k, \mathcal{X}_{m+1})] \le r_n^{d(k-1)}\mu,$$

with the upper bound arising from removing the condition that $\mathcal{X}_k$ be a component in $\mathcal{C}(\mathcal{X}_{m+1})$ and the lower bound arising by bounding below the conditional probability that $\mathcal{X}_k$ is a component, given that it forms an empty $(k-1)$-simplex. If $\gamma < 1$, then $\lim_{n\to\infty}(1-\|f\|_\infty\theta_d(4r_n)^d)^{m+1-k} = 1$, uniformly in $m \in [n-n^\gamma, n+n^\gamma]$, thus $\mathbb{E}[\tilde{h}_{r_n}(\mathcal{X}_k, \mathcal{X}_{m+1})] \simeq r_n^{d(k-1)}\mu$ uniformly in $m \in [n-n^\gamma, n+n^\gamma]$, and the same is true for $\mathbb{E}[\tilde{h}_{r_n}(\mathcal{X}_k, \mathcal{X}_m)]$.

For the second term of (11), observe that

$$\frac{\binom{m}{k}}{\binom{m}{k-1}}\mathbb{P}\left[X_{m+1} \in \cup_{i=1}^k B_{2r_n}(X_i)\right] \lesssim \frac{m}{k}\|f\|_\infty \theta_d (4r_n)^d,$$

and $\lim_{n\to\infty} m r_n^d = 0$, uniformly in $m \in [n-n^\gamma, n+n^\gamma]$. That is, the second term is of strictly smaller order than the first. Thus

$$\lim_{n\to\infty} \sup_{n-n^\gamma \le m \le n+n^\gamma} \left| (nr_n^d)^{1-k}\mathbb{E}[D_{m,n}] - \frac{1}{(k-1)!}\mu \right| = 0.$$

This implies that

$$\lim_{n\to\infty} \sup_{n-n^\gamma \le m \le n+n^\gamma} \left| (nr_n^d)^{(1-k)/2}\mathbb{E}[D_{m,n}] \right| = 0,$$

since $nr_n^d \to 0$ as $n \to \infty$, and so the first increment condition of the theorem is satisfied with $\alpha = 0$ and any choice of $\gamma \in (\frac{1}{2}, 1)$.

Next, consider the quantity $\mathbb{E}[D_{m,n}D_{m',n}]$ for $m \le m'$. Recall that

$$D_{m,n} = \sum_{\substack{\mathcal{Y}\subseteq\mathcal{X}_m \\ |\mathcal{Y}|=k-1}} \tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\}, \mathcal{X}_{m+1}) - \sum_{\substack{\mathcal{Y}\subseteq\mathcal{X}_m \\ |\mathcal{Y}|=k}} \tilde{h}_{r_n}(\mathcal{Y}, \mathcal{X}_m)\mathbb{1}_{\left\{X_{m+1}\in\bigcup_{y\in\mathcal{Y}} B_{2r_n}(y)\right\}}.$$

First consider the contribution to $\mathbb{E}[D_{m,n}D_{m',n}]$ from terms of the form

$$\mathbb{E}\left[\tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\}, \mathcal{X}_{m+1})\tilde{h}_{r_n}(\mathcal{Y}'\cup\{X_{m'+1}\}, \mathcal{X}_{m'+1})\right]$$

for $\mathcal{Y}, \mathcal{Y}'$ such that $\left(\mathcal{Y}\cup\{X_{m+1}\}\right)\cap\mathcal{Y}' = \emptyset$. By conditioning on the event $\tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\}, \mathcal{X}_{m+1}) = 1$, it follows that

$$\mathbb{E}[\tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\}, \mathcal{X}_{m+1})\tilde{h}_{r_n}(\mathcal{Y}'\cup\{X_{m'+1}\}, \mathcal{X}_{m'+1})] \simeq r_n^{2d(k-1)}\mu^2\zeta,$$

where $\zeta$ is the conditional probability that $\mathcal{Y}'\cup X_{m'+1}$ is a component in $\mathcal{X}_{m'+1}$, given that it forms an empty $(k-1)$-simplex, and that $\mathcal{Y}\cup X_{m+1}$ forms an empty $(k-1)$-simplex which is not connected to any other points of $\mathcal{X}_{m+1}$. Note that if $m = m'$ then $\zeta = 0$. Otherwise, simply bound $\zeta \le 1$, so that these terms have asymptotic order bounded above by $r_n^{2d(k-1)}\mu^2$, uniformly in $m$. The number of such terms is bounded by $\frac{(n+n^\gamma)^{2k-2}}{[(k-1)!]^2}$.

Note that if $\big(\mathcal{Y}\cup\{X_{m+1}\}\big)\cap\mathcal{Y}'\neq\emptyset$, and $m\neq m'$, then $\tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\},\mathcal{X}_{m+1})\tilde{h}_{r_n}(\mathcal{Y}'\cup\{X_{m'+1}\},\mathcal{X}_{m'+1})\equiv 0$. If $m=m'$ and then it must be that $\mathcal{Y}=\mathcal{Y}'$ to get a non-zero contribution. In this case, one gains a contribution to $\mathbb{E}[D_{m,n}^2]$ of

$$\binom{m}{k-1}r_n^{d(k-1)}\mu\leq\frac{(n+n^\gamma)^{k-1}r_n^{d(k-1)}\mu}{(k-1)!}.$$

Moving on to the cross terms, if $m'=m$ then

$$\tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\},\mathcal{X}_{m+1})\tilde{h}_{r_n}(\mathcal{Y}',\mathcal{X}_m)\mathbb{1}_{\left\{X_{m+1}\in\bigcup_{y\in\mathcal{Y}'}B_{2r_n}(y)\right\}}\equiv 0.$$

If $m<m'$ (or $m>m'$), then

$$\begin{aligned}\mathbb{E}&\left[\tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\},\mathcal{X}_{m+1})\tilde{h}_{r_n}(\mathcal{Y}',\mathcal{X}_{m'})\mathbb{1}_{\left\{X_{m'+1}\in\bigcup_{y\in\mathcal{Y}'}B_{2r_n}(y)\right\}}\right]\\&\leq\mathbb{E}\left[\tilde{h}_{r_n}(\mathcal{Y}\cup\{X_{m+1}\},\mathcal{X}_{m+1})\tilde{h}_{r_n}(\mathcal{Y}',\mathcal{X}_{m'})\right]\|f\|_\infty\theta_d(4r_n)^d.\end{aligned}$$

Again, to get a non-zero contribution, it must be that $(\mathcal{Y}\cup\{X_{m+1}\})\cap\mathcal{Y}'=\emptyset$. In this case, the expression above is bounded above by

$$(r_n^{d(k-1)}\mu)^2\|f\|_\infty\theta_d(4r_n)^d.$$

The number of such terms is bounded by $\binom{m}{k-1}\binom{m}{k}\leq\frac{(n+n^\gamma)^{2k-1}}{k!(k-1)!}$.

For the product of the second sums from $D_{m,n}$ and $D_{m',n}$, we have already seen that the conditional probability that $X_{m+1}\in\bigcup_{y\in\mathcal{Y}}B_{2r_n}(y)$ given $\mathcal{Y}$ is bounded above by $\|f\|_\infty\theta_d(4r_n)^d$, and so if $m=m'$,

$$\mathbb{E}\left[\sum_{\mathcal{Y}\subseteq\mathcal{X}_{m'}}\left(\tilde{h}_{r_n}(\mathcal{Y},\mathcal{X}_{m'})\mathbb{1}_{\left\{X_{m'+1}\in\bigcup_{y\in\mathcal{Y}}B_{2r_n}(y)\right\}}\right)^2\right]\leq\\\frac{(n+n^\gamma)^k}{k!}r_n^{d(k-1)}\mu\|f\|_\infty\theta_d(4r_n)^d,$$

while if $\mathcal{Y}\neq\mathcal{Y}'$,

$$\left(\tilde{h}_{r_n}(\mathcal{Y},\mathcal{X}_{m'})\mathbb{1}_{\left\{X_{m'+1}\in\bigcup_{y\in\mathcal{Y}}B_{2r_n}(y)\right\}}\right)\left(\tilde{h}_{r_n}(\mathcal{Y}',\mathcal{X}_{m'})\mathbb{1}_{\left\{X_{m'+1}\in\bigcup_{y\in\mathcal{Y}'}B_{2r_n}(y)\right\}}\right)\equiv 0.$$

For $m\neq m'$, $\mathcal{Y}\subseteq\mathcal{X}_m$ and $\mathcal{Y}\subseteq\mathcal{X}_{m'}$, let $\xi$ be the indicator that $\mathcal{Y}$ forms an empty $(k-1)$-simplex and $\eta$ the indicator that it is a component in $\mathcal{X}_m$. Let $\xi'$ and $\eta'$ be the corresponding indicators that $\mathcal{Y}'$ is an empty $(k-1)$-simplex and that it is a component in $\mathcal{X}_{m'}$. Let $\zeta$ and $\zeta'$ be the indicators that $X_{m+1}$ is connected to $\mathcal{Y}$ and that $X_{m'+1}$ is connected to $\mathcal{Y}'$, respectively. Then what is needed is

$$\mathbb{E}[\xi\eta\zeta\xi'\eta'\zeta'].$$

Note that for the product to be non-zero, it must be that $(\mathcal{Y}\cup\{X_{m+1}\})\cap\mathcal{Y}'=\emptyset$. Now,

$$\mathbb{P}\left[\zeta\zeta'=1\middle|\xi\eta\xi'\eta'=1\right]\leq\frac{\|f\|_\infty^2\theta_d^2(4r_n)^{2d}}{\mathrm{vol}_f(\cap_{y\in\mathcal{Y}'}B_{2r_n}(y)^c)}\leq\frac{\|f\|_\infty^2\theta_d^2(4r_n)^{2d}}{1-\|f\|_\infty\theta_d(4r_n)^d},$$

since if $\xi\eta\xi'\eta'=1$, then $\mathcal{Y}$ and $\mathcal{Y}'$ make up empty $(k-1)$-simplices; and morover, while nothing at all is known about $X_{m'+1}$, it is known that $X_{m+1}$ is not connected to $\mathcal{Y}'$. Trivially, $\mathbb{P}\left[\eta\eta' = 1 \middle| \xi\xi' = 1\right] \leq 1$, and $\mathbb{P}[\xi\xi' = 1] = \mathbb{P}[\xi = 1]\mathbb{P}[\xi' = 1] \simeq r_n^{2d(k-1)}\mu^2$, since $\mathcal{Y} \cap \mathcal{Y}' = \emptyset$. Thus

$$\mathbb{E}\left[\sum_{\mathcal{Y}\subseteq\mathcal{X}_m} \sum_{\substack{\mathcal{Y}\subseteq\mathcal{X}_{m'}\\ \mathcal{Y}'\neq\mathcal{Y}}} \tilde{h}_{r_n}(\mathcal{Y}, \mathcal{X}_m)\mathbb{1}_{\left\{X_{m+1}\in\bigcup_{y\in\mathcal{Y}} B_{2r_n}(y)\right\}} \tilde{h}_{r_n}(\mathcal{Y}, \mathcal{X}_{m'})\mathbb{1}_{\left\{X_{m'+1}\in\bigcup_{y\in\mathcal{Y}'} B_{2r_n}(y)\right\}}\right]$$
$$\lesssim \frac{c_{d,f}(nr_n^d)^{2k}\mu^2}{(k!)^2}.$$

It now follows that $\mathbb{E}[D_{m,n}D_{m',n}] \lesssim c_{d,f,k}(nr_n^d)^k$ for all $m, m' \in [n - n^\gamma, n + n^\gamma]$ with $m \neq m'$, and so

$$\lim_{n\to\infty} \sup_{n-n^\gamma\leq m<m'\leq n+n^\gamma} (nr_n^d)^{1-k}\mathbb{E}[D_{m,n}D_{m',n}] = 0.$$

If $m = m'$, then $\mathbb{E}[D_{m,n}^2] \lesssim c_{d,f}(nr_n^d)^{k-1}$, and so

$$\lim_{n\to\infty} \sup_{n-n^\gamma\leq m\leq n+n^\gamma} \frac{1}{\sqrt{n}}(nr_n^d)^{1-k}\mathbb{E}[D_{m,n}^2] = 0.$$

Thus the increment conditions of the theorem are satisfied with $\alpha = 0$.

Finally, observe that

$$H_n(\mathcal{X}_m) \leq \frac{\sqrt{n}m}{n^k r_n^{d(k-1)}k} \leq \frac{(\sqrt{n} + m)^2}{n^k r_n^{d(k-1)}k};$$

since $n^k r_n^{d(k-1)}$ is assumed to go to infinity as $n \to \infty$, the polynomial boundedness condition of Theorem 3.16 is satisfied and the central limit theorem for $\widetilde{S}_{n,k}$ is proved.

□

As was previously noted, that the same central limit theorem holds for upper and lower bounds for $\beta_k$ given in (5) immediately yields part (iii) of Theorem 3.2.

**Theorem 3.18.**

$$\frac{\beta_{k-1} - \mathbb{E}[\widetilde{S}_{n,k}]}{\sqrt{\mathbb{E}[\widetilde{S}_{n,k}]}} \Longrightarrow \mathfrak{N}(0, 1).$$

## 4. Vietoris-Rips complexes

Vietoris-Rips complexes were introduced by Leopold Vietoris in the context of algebraic topology, and independently by Eliyahu Rips in the context of geometric group theory. These complexes continue to be a useful construction in both fields, and are also useful in computational topology – although they do not carry the same homotopy information that the Čech complex does, the fact that they are determined by their underlying graph makes them much smaller in memory and more amenable to certain kinds of calculation.

Let $f : \mathbb{R}^d \to \mathbb{R}^{\geq 0}$ be a bounded measurable density function and et $\mathcal{X}_n$ denote a set of $n$ points drawn independently from this distribution. For any $r > 0$ define a (random geometric) graph $G(n, r)$ on $\mathcal{X}_n$ by inserting an edge $\{x, y\}$ whenever $d(x, y) < 2r$. Usually $r = r(n)$ and we consider the limit as $n$ tends to infinity.

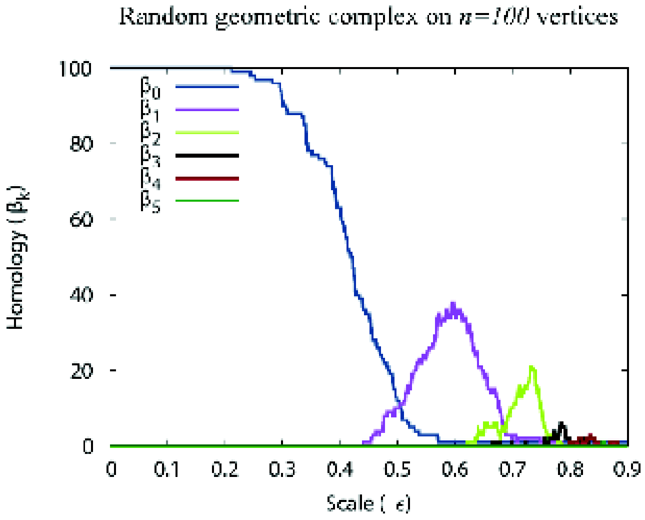

FIGURE 2. The Betti numbers of $VR(n,r)$ plotted vertically against $r$ horizontally; $n = 100$. *Computation and graphic courtesy of Afra Zomorodian.*

The *random Vietoris-Rips complex* $VR(n,r)$ is the clique complex of this random geometric graph; that is, the maximal simplicial complex with 1-skeleton $G(n,r)$. To see the contrast with $X(n,p)$, Figure 4 has a picture of the Betti numbers of a random Rips complex $VR(n,r)$ on 100 uniform points in a 6-dimensional cube, with $n = 100$ and $0 \leq r \leq 1$; compare with Figure 1.

In the sparse range of parameter, $r = o(n^{-1/d})$, a formula for the asymptotic expectation of $\beta_k$ was given in [9].

**Theorem 4.1.** *For $d \geq 2$, $k \geq 1$, $\epsilon > 0$, and $r_n = O(n^{-1/d-\epsilon})$, the expectation of the kth Betti number $\mathbb{E}[\beta_k]$ of the random Vietoris-Rips complex $VR(X_n; r_n)$ satisfies*

$$\frac{\mathbb{E}[\beta_k]}{n^{2k+2} r_n^{d(2k+1)}} \to C_k,$$

*as $n \to \infty$, where $C_k$ is a constant that depends only on $k$ and the underlying density function $f$.*

In the same regime we prove limit theorems for $\beta_k$.

**Theorem 4.2.** *With the same hypothesis as in Theorem 4.1,*

(i) *if $n^{2k+2} r_n^{d(2k+1)} \to 0$ as $n \to \infty$, then*

$$\beta_k(VR(X_n; r_n)) \to 0 \qquad a.a.s.;$$

(ii) *if $n^{2k+2} r_n^{d(2k+1)} \to \alpha \in (0, \infty)$ as $n \to \infty$, then*

$$d_{TV}(\beta_k(VR(X_n; r_n)), Y) \leq c\alpha n r_n^d,$$

*where $Y$ is a Poisson random variable with $\mathbb{E}[Y] = \mathbb{E}[\beta_k]$ and $c$ is a constant depending only on $d$, $k$, and $f$;*

(iii) *if $n^{2k+2}r_n^{d(2k+1)} \to \infty$, then*
$$\frac{\beta_k - \mathbb{E}[\beta_k]}{\sqrt{\mathrm{Var}[\beta_k]}} \to \mathcal{N}(0,1).$$

(The case $k = 0$ is handled in detail by Penrose [15].)

The main idea of the proof of Theorem 4.2 is again to bound $\beta_k$ between two random variables which satisfy the same central limit theorem. The intuition behind the bounds is that almost all of the homology of $VR(n, r)$ is contributed from a single source: the octahedral components.This is essentially because they are the smallest possible support of homology (smallest in the sense of vertex support), in the same way that empty $(k - 1)$-simplices were the smallest possible support of homology in the previous section.

**Definition 4.3.** The $(k+1)$-dimensional *cross-polytope* is defined to be the convex hull of the $2k+2$ points $\{\pm e_i\}$, where $e_1, e_2, \dots, e_{k+1}$ are the standard basis vectors of $\mathbb{R}^{k+1}$. The boundary of this polytope is a $k$-dimensional simplicial complex, denoted $O_k$.

Simplicial complexes which arise as clique complexes of graphs are sometimes called *flag complexes.* A useful fact in combinatorial topology is the following; for a proof see [11].

**Lemma 4.4.** *If $\Delta$ is a flag complex, then any nontrivial element of $k$-dimensional homology $H_k(\Delta)$ is supported on a subcomplex $S$ with at least $2k + 2$ vertices. Moreover, if $S$ has exactly $2k + 2$ vertices, then $S$ is isomorphic to $O_k$.*

**Definition 4.5.** Let $o_k(\Delta)$ (or $o_k$ if context is clear) denote the number of induced subgraphs of $\Delta$ combinatorially isomorphic to the 1-skeleton of the cross-polytope $O_k$, and let $\tilde{o}_k(\Delta)$ denote the number of components of $\Delta$ combinatorially isomorphic to the 1-skeleton of the cross-polytope $O_k$.

**Definition 4.6.** Let $f_k^{=i}(\Delta)$ denote the number of $k$-dimensional faces on connected components containing with exactly $i$ vertices. Similarly, let $f_k^{\geq i}(\Delta)$ denote the number of $k$-dimensional faces on connected components containing at least $i$ vertices.

In [15], Penrose proved the following limit theorems for subgraph counts of random geometric graphs.

**Theorem 4.7** (Penrose)**.** *Let $\Gamma_1, \dots, \Gamma_m$ be graphs on $v \geq 2$ vertices, such that $\mathbb{P}[G(v, r) \cong \Gamma_j] > 0$ for each $j$. Let $G_n(\Gamma)$ denote the number of induced subgraphs of $G(n, r_n)$ isomorphic to $\Gamma$. Then with $r_n$ as in the statement of Theorem 4.2,*

(i) *There is a constant $\mu_j$ depending only on $\Gamma_j$ and $v$ such that*
$$\lim_{n\to\infty} r_n^{-d(v-1)} n^{-v} \mathbb{E}[G_n(\Gamma_j)] = \mu_j.$$

(ii) *Let $Z_1, \dots, Z_m$ be indpendent Poisson random variables with $\mathbb{E}Z_j = \mathbb{E}[G_n(\Gamma_j)]$. There is a constant $c$ depending only on $m$ such that*
$$d_{TV}\left[(G_n(\Gamma_1), \dots, G_n(\Gamma_m)), (Z_1, \dots, Z_m)\right] \leq cn^{v+1}r_n^{dv}.$$

(iii) *Suppose that $n^v r_n^{d(v-1)} \to \infty$ as $n \to \infty$. Let $\tau = \sqrt{n^v r_n^{d(v-1)}}$. Then the joint distribution of the random variables $\{G_n(\Gamma_j)\}_{j=1}^m$ converges to a centered Gaussian distribution with covariance matrix $\Sigma = diag(\mu_1, \dots, \mu_m)$, for $\mu_j$ as in part (i)*

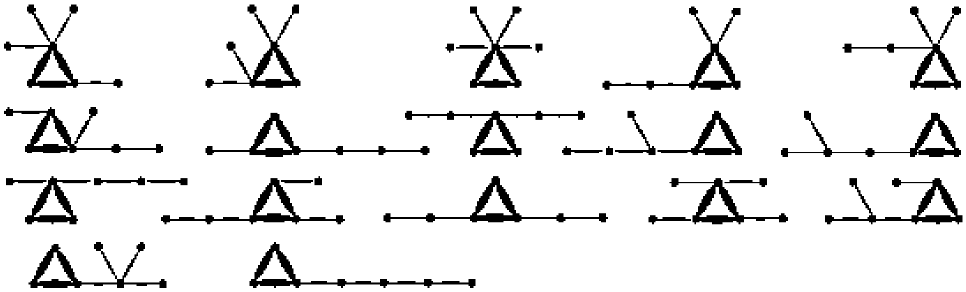

FIGURE 3. The case $k = 2$: the seventeen isomorphism types of subgraphs which arise when extending a 3-clique to a connected graph on 7 vertices with 7 edges. Each subgraph isomorphic to one of these can contribute at most 1 to the sum bounding the error term $f_2^{\geq 7}$.

A dimension bound paired with Lemma 4.4 yields

$$\tilde{o}_k \leq \beta_k \leq \tilde{o}_k + f_k^{\geq 2k+3}, \tag{12}$$

in analogy to the Morse inequalities used in the first section.

One could work with $f_k^{\geq 2k+3}$ directly, but it turns out to be sufficient to overestimate $f_k^{\geq 2k+3}$ as follows. For each $k$-dimensional face, consider the underlying $(k+1)$-clique; if it is in a component with at least $2k+3$ vertices, extend the clique to a connected subgraph with exactly $2k+3$ vertices and $\binom{k+1}{2} + k + 2$ edges, by the following algorithm.

(i) Set $G$ to be the 1-skeleton of the complex, and initialize $H$ to be the $(k+1)$-clique.
(ii) Find some edge connecting $V(H)$ to $V(G) - V(H)$. Add this edge (and its endpoint) to $H$. This is always possible since by assumption $H$ is contained in a component with at least $2k+3$ vertices.
(iii) Repeat step 2 until $H$ has exactly $2k+3$ vertices.

For example, let $k = 2$; then

$$\tilde{o}_2 \leq \beta_2 \leq \tilde{o}_2 + f_2^{\geq 7}.$$

Up to isomorphism, the seventeen graphs that arise when extending a 2-dimensional face (i.e. a 3-clique) to a minimal connected graph on 7 vertices are exhibited in Figure 3.

In particular,$f_2^{\geq 7} \leq \sum_{i=1}^{17} s_i$, where $s_i$ counts the number of subgraphs isomorphic to graph $i$ for some indexing of the seventeen graphs in Figure 3.

In general, one can express the number of graphs on $2k + 3$ vertices that can arise from the algorithm above as a function of $k$. Moreover, as is noted in [15], the number of occurances of a given graph $\Gamma$ on $v$ vertices (that is, the subgraph count corresponding to $\Gamma$) can be written as a linear combination of the induced subgraph counts for those graphs on $v$ vertices which have $\Gamma$ as a subgraph. That is,

$$\tilde{o}_k \leq \beta_k \leq o_k + g_{2k+3}, \tag{13}$$

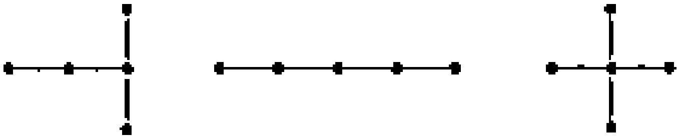

FIGURE 4. The case $k = 1$: the three isomorphism types of trees on five vertices. Each subgraph isomorphic to one of these can contribute at most 4 to the sum bounding the error term $f_1^{\geq 5}$.

where $g_{2k+3}$ is a linear combination of the induced subgraph counts of graphs on $2k+3$ vertices, the number of which depends only on $k$, and the trivial bound $\tilde{o}_k \leq o_k$ has been used on the right-hand side.

The induced subgraph counts appearing on the right-hand side of (13) are among the components of a random vector whose joint distribution is identified in Theorem 4.7 (for two different values of $v$), and thus limiting distributions for $o_k$ and $g_{2k+3}$ are known in those regimes. Moreover, it is easy to modify Penrose's proofs (just as in the previous section) to show that

$$d_{TV}(o_k + g_{2k+3}, Y) \leq c\alpha n r_n^d,$$

where $Y$ is a Poisson random variable with $\mathbb{E}[Y] = \mathbb{E}[o_k + g_{2k+3}]$, which in particular yields a central limit theorem if $n^{2k+2} r_n^{d(2k+1)} \to \infty$ as $n \to \infty$.

To obtain the limiting distribution for the lower bound of (13) is also just as in the previous section; all the proofs go through in exactly the same way, and will therefore not be repeated.

For $k = 1$ there are several ways of extending a 2-clique (i.e. an edge) to a connected graph on 5 vertices and 4 edges. In this case the graph must be a tree, and it is no longer possible to recover the clique from the connected graph. However, there are only three isomorphism types of trees on five vertices, shown in Figure 4. Counting these types of subgraphs may therefore result in an underestimate for $f_1^{\geq 5}$ because some edges might get extended to the same tree. However, each tree has only four edges, and so one can obtain the bound

$$f_1^{\geq 5} \leq 4(t_1 + t_2 + t_3),$$

where $t_1, t_2, t_3$ count the number of subgraphs isomorphic to the three trees in Figure 4. The proof is then the same as in the case $k \geq 2$.

## 5. Comments

We studied here three different kinds of random simplicial complex in order to work as generally as possible; however there are various ways in which we believe it may be possible to extend our results.

1. The random Vietoris-Rips and Čech complexes studied here are on Euclidean space, but this is mostly a matter of convenience. It would seem that the same proofs work, mutatis mutandis, for arbitrary Riemannian manifolds. This may be of interest in topological data analysis, as in earlier work of Niyogi, Smale, and Weinberger [14].

2. It may be possible to extend the central limit theorems for the random Vietoris-Rips and Čech complexes into denser regimes, at least into the thermodynamic limit. We expect, for example, that there exists some $c > 0$ such that CLT's hold for all Betti numbers $\beta_k$ simultaneously, whenever $r \geq cn^{-1/d}$.

3. An easier argument than those presented here should yield central limit theorems for Euler characteristic $\chi$ of geometric random complexes, in the sparse range. Again it would be nice to know this this in denser regimes, and we would guess that it holds at least partway into the thermodynamic limit.

**Acknowledgements:** The authors met and began discussing this project at the Workshop on Topological Complexity of Random Sets held at the American Institute of Mathematics in August, 2009; many thanks to AIM and to the organizers of the workshop. The authors also thank Omer Bobrowski for pointing out a mistake in the original version of the paper.

## References

[1] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen-Stein method. *Ann. Probab.*, 17(1):9–25, 1989.

[2] Eric Babson, Chris Hoffman, and Matthew Kahle. The fundamental group of random 2-complexes. To appear in *J. Amer. Math. Soc.*

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

[4] A. Björner. Topological methods. In *Handbook of combinatorics, Vol. 1, 2*, pages 1819–1872. Elsevier, Amsterdam, 1995.

[5] Sourav Chatterjee, Persi Diaconis, and Elizabeth Meckes. Exchangeable pairs and Poisson approximation. *Probab. Surv.*, 2:64–106 (electronic), 2005.

[6] Herbert Edelsbrunner and Ernst Peter Mücke. Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. In *Proceedings of the Fourth Annual Symposium on Computational Geometry (Urbana, IL, 1988)*, pages 118–133, New York, 1988. ACM.

[7] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[8] Allen Hatcher. *Algebraic topology.* Cambridge University Press, Cambridge, 2002.

[9] Matthew Kahle. Random geometric complexes. Preprint, arXiv:0910.1649.

[10] Matthew Kahle. The neighborhood complex of a random graph. *J. Combin. Theory Ser. A*, 114(2):380–387, 2007.

[11] Matthew Kahle. Topology of random clique complexes. *Discrete Math.*, 309(6):1658–1671, 2009.

[12] Nathan Linial and Roy Meshulam. Homological connectivity of random 2-complexes. *Combinatorica*, 26(4):475–487, 2006.

[13] R. Meshulam and N. Wallach. Homological connectivity of random $k$-dimensional complexes. *Random Structures Algorithms*, 34(3):408–417, 2009.

[14] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.*, 39(1-3):419–441, 2008.

[15] Mathew Penrose. *Random geometric graphs*, volume 5 of *Oxford Studies in Probability.* Oxford University Press, Oxford, 2003.

[16] Nicholas Pippenger and Kristin Schleich. Topological characteristics of random triangulated surfaces. *Random Structures Algorithms*, 28(3):247–288, 2006.

[17] Yosef Rinott and Vladimir Rotar. Normal approximations by Stein's method. *Decis. Econ. Finance*, 23(1):15–29, 2000.

School of Mathematics, Institute for Advanced Study, Einstein Drive, Princeton NJ 08540, U.S.A.
*E-mail address*: mkahle@math.stanford.edu

Case Western Reserve University, Cleveland, OH 44106, U.S.A.
*E-mail address*: elizabeth.meckes@case.edu