Distribution of components in the k-nearest neighbour random geometric graph for k below the connectivity threshold[*]

Victor Falgas-Ravry[†]

Abstract

Let $S_{n,k}$ denote the random geometric graph obtained by placing points inside a square of area n according to a Poisson point process of intensity 1 and joining each such point to the $k = k(n)$ points of the process nearest to it.

In this paper we show that if $\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_1}$ then the probability that $S_{n,k}$ contains a pair of 'small' components 'close' to each other is $o(n^{-c_1})$ (in a precise sense of 'small' and 'close'), for some absolute constants $\gamma_1 > 0$ and $c_1 > 0$. This answers a question of Walters [13]. (A similar result was independently obtained by Balister.)

As an application of our result, we show that the distribution of the connected components of $S_{n,k}$ below the connectivity threshold is asymptotically Poisson.

1 Introduction

In this paper, we study the distribution of small connected components below the connectivity threshold in the k-nearest neighbour model. We prove that they lie far apart, and are approximately distributed like a Poisson process in both a numerical and a spatial sense.

1.1 Definitions

Let S_n denote the square $[0, \sqrt{n}]^2$ of area n, and let $|| \cdot ||$ denote the usual Euclidean norm in $\mathbb{R}^2$.

Let $k = k(n)$ be a nonnegative integer parameter. Scatter points at random inside S_n according to a Poisson process of intensity 1. (Or equivalently, scatter $N \sim \text{Poisson(n)}$

[†]Institutionen för matematik och matematisk statistik, Umeå Universitet, 901 87 Umeå, Sweden.
E-mail: victor.falgas-ravry@math.umu.se

points uniformly at random inside S_n.) This gives us a random geometric vertex set $\mathcal{P}$. The *k-nearest neighbour graph* on $\mathcal{P}$, $S_{n,k} = S_{n,k}(\mathcal{P})$, is obtained by setting an undirected edge between every vertex of $\mathcal{P}$ and the k vertices of $\mathcal{P}$ nearest to it.

The Poisson law of the random pointset $\mathcal{P}$ coupled with our deterministic definitions of $S_{n,k}$ gives rise to a probability measure on the space of k-nearest neighbour graphs with vertices in S_n, which we refer to as the *k-nearest neighbour model* $\mathcal{S}_{n,k}$.

As this model has some inherent randomness, rather than proving deterministic statements about $S_{n,k}$ we are mostly interested in establishing that properties hold for 'typical' k-nearest neighbour graphs. Formally, given a property $\mathcal{Q}$ of k-nearest neighbour graphs (i.e a sequence of subsets $\mathcal{Q}_n$ of the set of all geometric graphs on S_n) and a sequence $k = k(n)$ of nonnegative integers, we say that $S_{n,k(n)}$ has property $\mathcal{Q}$ *with high probability* (*whp*) if

$$\lim_{n \to \infty} \mathbb{P}(S_{n,k(n)} \in \mathcal{Q}_n) = 1.$$

1.2 Previous work on the k-nearest neighbour model

An important motivation for the study of the connectivity properties of the k-nearest neighbour model comes from the theory of ad-hoc wireless networks: suppose we have some radio transmitters (nodes) spread out over a large area and wishing to communicate using multiple hops, and that each transmitter can adjust its range so as to ensure two-way radio contact with the k nodes nearest to it. The connectivity of this radio network is modelled using $S_{n,k}$ in a natural way. As a result, the model has received a significant amount of attention, though many questions remain. (See e.g. the recent survey paper of Walters [12].)

Elementary arguments show that there exist constants $c_l \leq c_u$ such that for $k(n) \leq c_l \log n$ $S_{n,k}$ is whp not connected, while for $k(n) \geq c_u \log n$ $S_{n,k}$ is whp connected. A bound of $c_u \leq 5.1774$ was obtained by Xue and Kumar [14], using a substantial result of Penrose [10] for the related Gilbert disc model [8]. An earlier bound of $c_u \leq 3.5897$ could also be read out of earlier work of Gonzáles–Barrios and Quiroz [9]. These results were substantially improved by Balister, Bollobás, Sarkar and Walters in a series of papers [3, 5, 4] in which they established inter alia the existence of a critical constant $c_\star$: $0.3043 < c_\star < 0.5139$ such that for $c < c_\star$ and $k \leq c \log n$, $S_{n,k}$ is whp not connected while for $c > c_\star$ and $k \geq c \log n$, $S_{n,k}$ is whp connected. Building on their work, Walters and the author [7] recently proved that the transition from whp not connected to whp connected is sharp in k: there is an absolute constant $C > 0$ such that if $S_{n,k}$ is connected with probability at least $\varepsilon > 0$ and n is sufficiently large, then for $k' \geq k + C \log(1/\varepsilon)$, $S_{n,k'}$ is connected with probability at least $1 - \varepsilon$.

As part of their results, Balister, Bollobás, Sarkar and Walters [3] showed that when $0.3 \log n < k < 0.6 \log n$ whp all of the following hold:

 (i) all edges have length $O(\sqrt{\log n})$,

 (ii) there is a unique 'giant' connected component,

(iii) all other components have diameter $O(\sqrt{\log n})$.

(Strictly speaking, they only showed that $S_{n,k}$ has at most one giant component; that such a component exists follows from the study of percolation in the k-nearest neighbour graph: see e.g. [2].)

Refining their techniques, Walters [13] showed that around the connectivity threshold, there are no 'small' components (of diameter $O(\sqrt{\log n})$) lying 'close' to the boundary of S_n (within distance $O(\log n)$), and used this to improve the upper bound on $c_\star$ to 0.4125. Towards the end of his paper [13], Walters asked a number of questions

about the properties of small components of $S_{n,k}$ below the connectivity threshold, with the aim of completing the picture we currently have and perhaps improving the upper bound on $c_\star$.

1.3 Results of the paper

In this paper we contribute to this project by proving that the 'small' components are whp far apart (or more precisely, not 'close' together) and that they are distributed like a Poisson point process inside S_n. This answers one of Walters's questions.

To state our results formally, we must ascribe a precise meaning to 'small' and 'close'. We recall a result of Balister, Bollobás, Sarkar and Walters to do this.

Lemma 1.1 (Lemma 1 of [4]). *For any fixed α_1, α_2 with $0 < \alpha_1 < \alpha_2$ and any $\beta > 0$, there exists $\lambda = \lambda(\alpha_1, \alpha_2, \beta) > 0$, depending only on α_1, α_2 and β, such that for any k with $\alpha_1 \log n \leq k \leq \alpha_2 \log n$, the probability that $S_{n,k}$ contains two components each of diameter at least $\lambda\sqrt{\log n}$ or any edge of length at least $\lambda\sqrt{\log n}$ is $O(n^{-\beta})$.*

Definition 1.2. *Let $\lambda = \max(\lambda(0.3, 0.6, 2), e^2)$. A component shall be deemed* small *if it has diameter less than $\lambda\sqrt{\log n}$. Two small components shall be deemed* close *if they contain two points lying at a distance less than $8\lambda\sqrt{\log n}$ from one another.*

Our main result is the following:

Theorem 1.3. *There exist absolute constants $\gamma_1 > 0$ and $c_1 > 0$ such that if*

$$\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_1}$$

then

$$\mathbb{P}(S_{n,k} \text{ contains a pair of small, close components}) = o(n^{-c_1}).$$

This answers Question 1 in Walters [13]. Balister has independently obtained a similar result (personal communication).

Remark 1.4. *Theorem 1 of [7] showed that, for n large enough, we need to increase k by at most a constant times $\log(1/\varepsilon)$ for $S_{n,k}$ to go from having an ε chance of being connected to having a $1 - \varepsilon$ chance of being connected. Assuming there is a 'bluntness' converse to this result, i.e. that we need to increase k by at least a (smaller) constant times $\log(1/\varepsilon)$ for this transition to occur, then there must be some constant $\delta : 0 < \delta < c_\star$ such that for all $k = k(n)$ with $k(n) > (c_\star - \delta) \log n$ we have $\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_1}$. In particular Theorem 1.3 is not vacuous since there is a range of k for which*

$$n^{-\gamma_1} < \mathbb{P}(S_{n,k} \text{ connected}) << 1 - o(n^{-c_1}).$$

(For example, $k = \lfloor c \log n \rfloor$ for any c with $c_\star - \delta < c < c_\star$ will do.)

Next we turn to the distribution of the small components of $S_{n,k}$. Let $X = X_{n,k}$ denote the number of small connected components of $S_{n,k}$. (Since there is whp a unique non-small connected component [3], X is whp the number of components of $S_{n,k}$ minus 1.) Also, given $\nu \geq 0$ and $A \subseteq \mathbb{N} \cup \{0\}$, let $Po_\nu(A)$ denote the probability a Poisson random variable with parameter ν takes a value inside A.

As an application of Theorem 1.3, we prove:

Theorem 1.5. *There exist absolute constants γ_2 and $c_2 > 0$ such that if $k = k(n)$ is an integer sequence with $\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_2}$ for all n, then, writing $\nu = \nu(n)$ for $-\log(\mathbb{P}(S_{n,k} \text{ connected}))$, we have*

$$\sup_{A \subseteq \mathbb{N} \cup \{0\}} |\mathbb{P}(X \in A) - Po_\nu(A)| = o(n^{-c_2}).$$

Corollary 1.6. *Let $k(n)$ be an integer sequence. Suppose there is a subsequence $(k(n_i))_{i \in \mathbb{N}}$ such that*

$$\mathbb{P}(S_{n_i,k(n_i)} \text{ connected}) \to e^{-\nu}$$

for some constant $\nu \geq 0$. Then the law of $X_{n_i,k(n_i)}$ converges in distribution to Poisson with parameter ν:

$$\mathcal{L}(X_{n_i,k(n_i)}) \xrightarrow{d} \text{Poisson}(\nu).$$

We also prove a spatial analogue of Theorem 1.5: not only is the number of small components (approximately) Poisson distributed, but their spatial location is (approximately) distributed according to a Poisson point process inside S_n. We defer a precise statement of this result, Theorem 4.17, until Section 4.

1.4 Structure of the paper

We follow the strategy introduced by Balister, Bollobás, Sarkar and Walters in [4] and developed in [7]: we prove Theorem 1.3 by looking at local events. In Section 2, we prove a local result, Theorem 2.2, which can be thought of as an analogue of the local sharpness result Lemma 12 in [7].

Theorem 2.2 is then used in Section 3 together with local-global correspondence results from [7] to prove the global result Theorem 1.3.

Finally in the last section we use a form of the Chen–Stein Method [6, 11] due to Arratia, Goldstein and Gordon [1] together with our results from the first two sections to prove Theorems 1.5 and 4.17 on the distribution of small components.

2 Proof of the local theorem

In this section we consider the connectivity of the k-nearest neighbour random geometric graph model on a local scale (i.e. within a region of area $O(\log n)$).

Pick n sufficiently large. Let $M = \max(160\lceil\lambda\rceil, 50)$. (We remark that this is similar to but slightly larger than the choice of M in [7].) We consider a Poisson point process of intensity 1 in the box

$$U_n = [-\frac{M}{2}\sqrt{\log n}, +\frac{M}{2}\sqrt{\log n}]^2.$$

We place an undirected edge between every point and its k nearest neighbours to obtain the graph $U_{n,k}$.

Let us define two families of events related to the connectivity of $U_{n,k}$:

Definition 2.1. *Let A_k be the event that $U_{n,k}$ has a connected component wholly contained inside the central subsquare $\frac{1}{2}U_n$. Let B_k be the event that $U_{n,k}$ has at least two connected components wholly contained inside the central subsquare $\frac{1}{2}U_n$.*

Our aim is to prove:

Theorem 2.2. *There exist constants $c_3, c_4 > 0$ such that for all integers k with $k \in (0.3\log n, 0.6\log n)$, we have*

$$\mathbb{P}(B_k) \leq c_3 n^{-c_4} \mathbb{P}(A_k) + o(n^{-2}).$$

This is saying that on a local scale it is far less likely that we have two small connected components close together than just one small connected component on its own. Our proof strategy is as follows: we show that whp if B_k occurs then there must be a large empty region inside U_n to which many points can be added without joining up all the components of $U_{n,k}$ which are contained inside $\frac{1}{2}U_n$. This ensures that A_k still
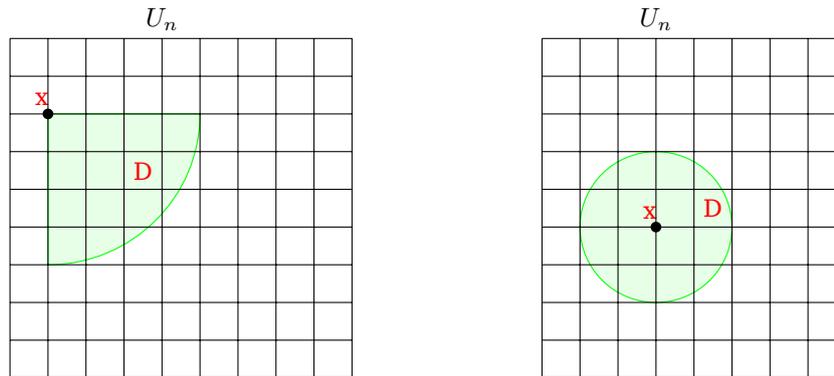
Figure 1: The square grid Γ, the point x and the quarter-disc and complete disc considered in the proof of Lemma 2.3.

occurs for the new pointset, and can be exploited to show A_k is much more likely than B_k.

To prove Theorem 2.2, we shall need a simple lemma on the concentration of Poisson random variables.

Lemma 2.3. *There exist constants λ_1 and λ_2, such that for all integers k with k between $0.3\log n$ and $0.6\log n$, the probability that there is any point $x \in U_n$ (not necessarily in the pointset arising from the Poisson point process) such that the ball of radius $\lambda_1\sqrt{\log n}$ about x contains at least k vertices of $U_{n,k}$ or that the ball of radius $\lambda_2\sqrt{\log n}$ about x contains fewer than k vertices of $U_{n,k}$ is $o(n^{-2})$.*

Moreover, λ_2 can be chosen to be less than λ.

Remark 2.4. *This says that the probability $U_{n,k}$ contains a pair of vertices not joined by an edge and lying within distance $\lambda_1\sqrt{\log n}$ of each other, or a pair of vertices joined by an edge and lying at distance at least $\lambda_2\sqrt{\log n}$ from one another is $o(n^{-2})$, i.e. a negligible quantity.*

Proof of Lemma 2.3. We first show that there is a constant $\lambda_2 \leq \lambda$ such that the probability that there is some point in U_n with fewer than $0.6\log n$ vertices of $U_{n,k}$ within distance $\lambda_2\sqrt{\log n}$ of itself is $o(n^{-2})$.

Let Γ denote the intersection of the integer grid $\mathbb{Z}^2$ with U_n. Let $\lambda_2' = 2\sqrt{\frac{e^3}{\pi}}$ and consider any $x \in \Gamma$. Since $M > 50$, at least one of the four standard quadrant quarter discs of radius $\lambda_2'\sqrt{\log n}$ about x is contained inside U_n; call this quarter disc D. The probability that D contains fewer than $0.6\log n$ vertices of $U_{n,k}$ is at most

$$
\begin{aligned}
&0.6(\log n)\exp\left(-\frac{\pi\lambda_2'^2}{4}\log n\right)\frac{(\pi\lambda_2'^2\log n/4)^{0.6\log n}}{(0.6\log n)!} \\
&< 0.6(\log n)\exp\left(-\frac{\pi\lambda_2'^2}{4}(\log n) + 0.6(\log n)\log(\frac{e\pi\lambda_2'^2}{0.6\times4})\right) \\
&< 0.6(\log n)\exp\left(-(e^3-3)\log n\right) \\
&< 0.6(\log n)n^{-4}.
\end{aligned}
$$

Thus the probability that Γ contains any point with fewer than $0.6\log n$ vertices of $U_{n,k}$ within distance $\lambda_2'\sqrt{\log n}$ is less than $|\Gamma| \times \frac{0.6\log n}{n^4} = o(n^{-2})$. Since every point in U_n lies

at distance at most $\sqrt{2}$ from a point in Γ, this proves our claim for $\lambda_2 = \lambda_2' + 1$, say. This is at most e^2, which is less than λ, as required.

Now let us show that there is a constant λ_1 such that the probability that there is some point in U_n having more than $0.3 \log n$ vertices of $U_{n,k}$ within distance $\lambda_1 \sqrt{\log n}$ of itself is $o(n^{-2})$.

Let Γ again denote the intersection of the integer grid $\mathbb{Z}^2$ with U_n. Let $\lambda_1' = \sqrt{\frac{\exp(-49/3)}{\pi}}$ and consider any $x \in \Gamma$. Let D denote the intersection of the disc of radius $\lambda_1' \sqrt{\log n}$ about x with U_n. The probability that D contains more than $0.3 \log n$ vertices of $U_{n,k}$ is at most

$$\frac{|D|^{\lceil 0.3 \log n \rceil}}{\lceil 0.3 \log n \rceil !} \exp\left(-|D|\right) + \frac{|D|^{\lceil 0.3 \log n \rceil + 1}}{(\lceil 0.3 \log n \rceil + 1)!} \exp\left(-|D|\right) + \cdots$$

$$< \frac{(\pi \lambda_1'^2 \log n)^{0.3 \log n}}{\lceil 0.3 \log n \rceil !} e^{-\pi \lambda_1'^2 \log n} \left(1 + \frac{\pi \lambda_1'^2}{0.3} + \left(\frac{\pi \lambda_1'^2}{0.3}\right)^2 + \cdots\right)$$

$$< \exp\left(0.3 \log n \left(\log\left(\pi \lambda_1'^2 \log n\right) - \log\left(0.3 \log n / e\right) - \frac{\pi \lambda_1'^2}{0.3}\right)\right) \frac{1}{1 - \pi \lambda_1'^2 / 0.3}$$

$$< \exp\left(0.3 \log n \left(-\frac{49}{3} + \log \frac{e}{0.3}\right) + O(1)\right)$$

$$< \exp\left(-4 \log n + O(1)\right)$$

$$= o(n^{-3}).$$

Thus the probability that Γ contains any point with more than $0.3 \log n$ vertices of $U_{n,k}$ within distance $\lambda_1' \sqrt{\log n}$ is less than $|\Gamma| \times \exp\left(-4 \log n + O(1)\right) = o(n^{-2})$. Now every point in U_n lies at distance at most $\sqrt{2}$ from a point in Γ, so this proves our claim for $\lambda_1 = \lambda_1'/2$, say. $\qquad \square$

Proof of Theorem 2.2. Let us assume that there is no point in U_n with more than k vertices of $U_{n,k}$ within distance $\lambda_1 \sqrt{\log n}$ of itself or fewer than k vertices of $U_{n,k}$ within distance $\lambda_2 \sqrt{\log n}$ of itself. We shall denote by $\mathcal{C}$ the set of pointsets we are thus excluding. By Lemma 2.3, $\mathbb{P}(\mathcal{C}) = o(n^{-2})$.

We consider a perfect tiling of U_n into tiles of area $\frac{\log n}{N^2}$, for some (large) constant N. Explicitly, we shall choose

$$N = \max(N_1, N_2, N_3),$$

where N_1, N_2 and N_3 are the constants

$$N_1 = \left\lceil \sqrt{5}/\lambda_1 \right\rceil + 1,$$

$$N_2 = \left\lceil \frac{2}{\lambda_1} + \frac{4\sqrt{5}\lambda_2}{\lambda_1^2} \right\rceil, \text{ and}$$

$$N_3 = \left\lceil \frac{1}{\lambda_1^2}\left((1+\sqrt{5})\lambda_1 + \lambda_2 + \sqrt{\left((1+\sqrt{5})\lambda_1 + \lambda_2\right)^2 - (5 + 2\sqrt{5})\lambda_1^2}\right)\right\rceil + 1,$$

each of which will appear at one of the stages of our argument. The choice of $N \geq N_2$ ensures that the inequality

$$\frac{1}{N} + \left(\frac{4\sqrt{5}\lambda_2}{N} + \frac{1}{N^2}\right)^{1/2} \leq \lambda_1$$

holds, while the choice of $N \geq N_3$ ensures that

$$\frac{1}{N^2} + \frac{2\lambda_2}{N} < \left(\lambda_1 - \frac{(1 + \sqrt{5})}{N}\right)^2$$

holds. (Both inequalities clearly hold for all sufficiently large N, and it is easily checked by solving two quadratic equations that the values of N_2 and N_3 given above will do.)

Given a pointset $\mathcal{P} \subset U_n$, write $U_{n,k}(\mathcal{P})$ for the k-nearest neighbour graph on $\mathcal{P}$.

Definition 2.5. *For each tile Q, let $B_k(Q)$ be the event that the pointset $\mathcal{P}$ resulting from the Poisson process on U_n has the following properties:*

(i) *$\mathcal{P} \in B_k$ (i.e. $U_{n,k}(\mathcal{P})$ has at least two connected components contained inside $\frac{1}{2}U_n$),*

(ii) *Q contains no point of $\mathcal{P}$, and*

(iii) *for any set of points $\mathcal{B} \subset Q$, the pointset $\mathcal{P} \cup \mathcal{B}$ lies in A_k (i.e. $U_{n,k}(\mathcal{P} \cup \mathcal{B})$ has at least one connected component wholly contained inside $\frac{1}{2}U_n$).*

The key step in the proof of Theorem 2.2 is to show:

Theorem 2.6.
$$B_k \setminus \mathcal{C} \subseteq \bigcup_Q B_k(Q).$$

In other words, if B_k occurs we can (except in a negligible proportion of cases) find an empty tile to which we can add many points and still have A_k occurring in the resulting pointset. This can be thought of as an analogue of the key Lemma 7 in [7].

Proof of Theorem 2.6. Let $\mathcal{P}$ be a pointset for which $B_k \setminus \mathcal{C}$ occurs. Say that a tile is *empty* if it contains no point of $\mathcal{P}$. Let X, Y be the vertex sets of two small connected components of $U_{n,k}(\mathcal{P})$ witnessing B_k. (At least two such components must exist, though there could potentially be more.)

Let a be a vertex in $X \cup Y$ nearest to the bottom side of U_n. Without loss of generality, we may assume that $a \in X$. Let E be the horizontal line through a. Then all points of X and Y must lie on or above E. Now a lies in some tile, Q_a say.

We consider the tiles directly below Q_a. Since $N \geq N_1 > \frac{\sqrt{5}}{\lambda_1}$, the topmost of these tiles must be empty. There are two cases to consider. Either all tiles directly below Q_a are empty, in which case we let Q denote the one among them which is incident with the boundary of U_n; or there is some tile directly below Q_a, which is nonempty. Then let Q' be the topmost of these nonempty tiles, and let Q denote the (empty) tile directly above it.

Lemma 2.7.
$$\mathcal{P} \in B_k(Q).$$

Proof. Let $\mathcal{B} \subset Q$ be a nonempty set of points in Q, and let $\mathcal{P}' = \mathcal{P} \cup \mathcal{B}$. Our claim is that $\mathcal{P}' \in A_k$. To establish this, it is enough to show that there are no edges from Y to Y^c in $U_{n,k}(\mathcal{P}')$ (as then Y will be a connected component of $U_{n,k}(\mathcal{P}')$ contained inside $\frac{1}{2}U_n$).

Since the only edges in $U_{n,k}(\mathcal{P}')$ that are not also edges of $U_{n,k}(\mathcal{P})$ have at least one end in Q, it suffices to show no vertex of Y is joined by an edge of $U_{n,k}(\mathcal{P}')$ to a point $b \in \mathcal{B}$. We split into two cases.

First of all, suppose Q is incident with the boundary of U_n. Since $\mathcal{P} \notin \mathcal{C}$ and $\lambda_2 < \frac{M}{4}$ (since $\lambda_2 \leq \lambda \leq \frac{M}{160}$), we know that no point in Q can have any of its k nearest neighbours inside $\frac{1}{2}U_n$ and vice-versa. Thus there are no edges between Q and $Y \subseteq \frac{1}{2}U_n$ in $U_{n,k}(\mathcal{P}')$, and $\mathcal{P}' \in A_k$ as required.
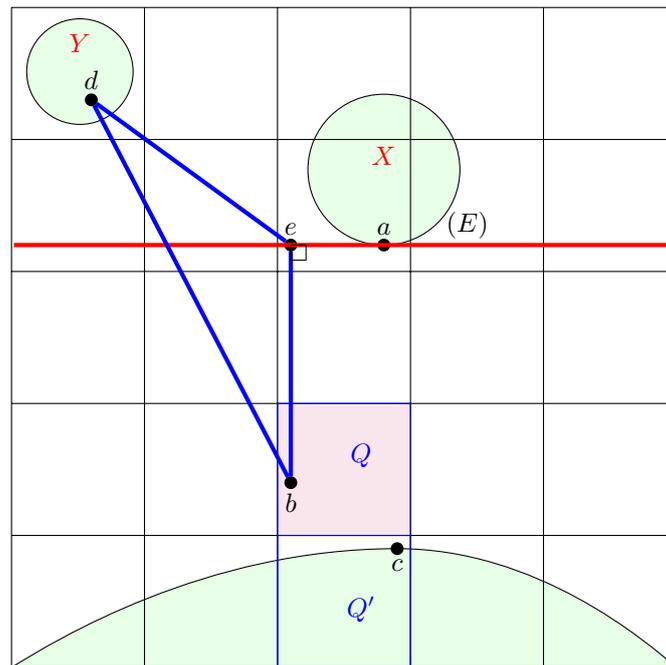
Figure 2: The points a, b, c, d and e, the line E, the tile Q and the components X and Y.

We now turn to the less trivial case where Q is not incident with the boundary of U_n. Then the tile Q' directly below Q is nonempty: there exists $c \in \mathcal{P} \cap Q'$.

Let R denote the distance between c and its kth nearest neighbour in $\mathcal{P}$. For any $b \in \mathcal{B}$, $\|c - b\| \leq \frac{\sqrt{5}}{N} \sqrt{\log n}$. Thus the distance between b and its kth nearest neighbour in $\mathcal{P}'$ will be at most $R + \frac{\sqrt{5}}{N} \sqrt{\log n}$. Also, ac is not an edge of the k-nearest neighbour graph on $\mathcal{P}$, whence $\|a - c\| \geq R$ and $\|a - b\| \geq R - \frac{\sqrt{5}}{N} \sqrt{\log n}$ for all $b \in \mathcal{B}$.

Now, let $b \in \mathcal{B} \subseteq Q$ and suppose d is a point lying above E such that d is one of the k nearest neighbours of b in $\mathcal{P}'$. Our earlier choice of $N \geq N_2$ ensures the following holds:

Claim 2.8.

$$\|a - d\| \leq \lambda_1 \sqrt{\log n}.$$

Remark 2.9. *By our assumption on $\mathcal{P}$ (namely our assumption that $\mathcal{P} \notin \mathcal{C}$), this implies that ad is an edge in $U_{n,k}(\mathcal{P})$. Since $a \in X$ it follows that $d \notin Y$.*

Proof of Claim 2.8. This is an exercise in Euclidean geometry. (See Figure 2.) We know that $\|b - d\| \leq R + \frac{\sqrt{5}}{N} \sqrt{\log n}$. Let e be the foot of the perpendicular to E which goes through b. As b lies in a tile directly below a's tile, we have $\|a - e\| \leq \frac{1}{N} \sqrt{\log n}$. It follows by Pythagoras's Theorem that

$$\|b - e\|^2 = \|b - a\|^2 - \|a - e\|^2$$
$$\geq (R - \frac{\sqrt{5}}{N} \sqrt{\log n})^2 - \frac{1}{N^2} \log n.$$

Now the angle bed is obtuse. Hence,

$$
\begin{aligned}
||e-d||^2 &\leq ||b-d||^2 - ||b-e||^2 \\
&\leq (R + \frac{\sqrt{5}}{N}\sqrt{\log n})^2 - (R - \frac{\sqrt{5}}{N}\sqrt{\log n})^2 + \frac{1}{N^2}\log n \\
&= \frac{4\sqrt{5}}{N}R\sqrt{\log n} + \frac{1}{N^2}\log n.
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
||a-d|| &\leq ||a-e|| + ||e-d|| \\
&\leq \frac{1}{N}\sqrt{\log n} + \left(\frac{4\sqrt{5}}{N}R\sqrt{\log n} + \frac{1}{N^2}\log n\right)^{\frac{1}{2}}.
\end{aligned}
$$

Now $R \leq \lambda_2\sqrt{\log n}$ (since $\mathcal{P} \notin \mathcal{C}$). Substituting this into the above yields

$$
||a-d|| \leq \left(\frac{1}{N} + \left(\frac{4\sqrt{5}\lambda_2}{N} + \frac{1}{N^2}\right)^{1/2}\right)\sqrt{\log n},
$$

which by our choice of $N \geq N_2$ is less than $\lambda_1\sqrt{\log n}$. $\qquad\square$

On the other hand, suppose d is a point lying above the line E such that b is one of the k nearest neighbours of d in $\mathcal{P}'$. Then our earlier choice of $N \geq N_3$ ensures the following holds:

Claim 2.10.
$$
||a-d|| < ||b-d||.
$$

Remark 2.11. *This implies that a is one of the k nearest neighbours of d in $U_{n,k}(\mathcal{P}')$, and hence also in $U_{n,k}(\mathcal{P})$. As $a \in X$ it follows that $d \notin Y$.*

Proof of Claim 2.10. This is again an exercise in Euclidean geometry. (See Figure 2.) Since b is among the k nearest neighbours of d, it follows that $||b-d|| \leq \lambda_2\sqrt{\log n}$ (since $d \in \mathcal{P}$ and $\mathcal{P} \notin \mathcal{C}$). Similarly, as ac is not an edge of $U_{n,k}(\mathcal{P})$, we must have that $||a-c|| \geq \lambda_1\sqrt{\log n}$. Let e be the foot of the perpendicular to the line E which goes through b. Since the triangle bed is obtuse, we have $||e-d|| \leq ||b-d||$. As b lies in a tile directly below a's tile, we have $||a-e|| \leq \frac{1}{N}\sqrt{\log n}$. Also, as c lies in a tile directly below b's tile we have $||b-c|| \leq \frac{\sqrt{5}}{N}\sqrt{\log n}$. Thus

$$
\begin{aligned}
||b-e|| &\geq ||a-c|| - ||a-e|| - ||b-c|| \\
&\geq \left(\lambda_1 - \frac{(1+\sqrt{5})}{N}\right)\sqrt{\log n}.
\end{aligned}
$$

Using again the fact that the triangle bed is obtuse, we have

$$
\begin{aligned}
||b-d||^2 &\geq ||b-e||^2 + ||e-d||^2. \\
&\geq \left(\lambda_1 - \frac{(1+\sqrt{5})}{N}\right)^2\log n + ||e-d||^2. \qquad\qquad(1)
\end{aligned}
$$

Finally we have

$$
\begin{aligned}
||a - d||^2 &\leq (||a - e|| + ||e - d||)^2 \\
&= ||a - e||^2 + 2||a - e|| \times ||e - d|| + ||e - d||^2 \\
&\leq \frac{\log n}{N^2} + \frac{2\sqrt{\log n}}{N}||e - d|| + ||e - d||^2 \\
&\leq \frac{\log n}{N^2} + \frac{2\lambda_2 \log n}{N} + ||e - d||^2,
\end{aligned}
\tag{2}
$$

where the last line follows from the fact that $||e - d|| \leq ||b - d|| \leq \lambda_2 \sqrt{\log n}$.

Now our choice of N (more specifically our choice of $N \geq N_3$) guarantees that

$$
\frac{1}{N^2} + \frac{2\lambda_2}{N} < (\lambda_1 - \frac{(1 + \sqrt{5})}{N})^2
$$

so that by comparing (1) and (2) we have

$$
||a - d|| < ||b - d||
$$

as claimed. $\qquad\square$

As remarked Claim 2.8 tells us that if d is one of b's k nearest neighbours in $\mathcal{P}'$ then d is one of a's k nearest neighbours in $\mathcal{P}$ (since $\mathcal{P} \notin \mathcal{C}$) — and in particular $d \notin Y$. On the other hand Claim 2.10 tells us that if b is one of d's k nearest neighbours in $\mathcal{P}'$ then a is one of d's k nearest neighbours in $\mathcal{P}$ — so that again $d \notin Y$.

Combining these two claims, we see that there are no edges between $\mathcal{B}$ and Y in $U_{n,k}(\mathcal{P}')$, and hence that Y is a connected component of $U_{n,k}(\mathcal{P}')$ contained inside $\frac{1}{2}U_n$. We thus have $\mathcal{P}' \in A_k$, as claimed. $\qquad\square$

For every $\mathcal{P} \in B_k \setminus \mathcal{C}$, we have thus shown there exists a tile Q such that $\mathcal{P} \in B_k(Q)$, proving Theorem 2.6. $\qquad\square$

Having established Theorem 2.6, the rest of the proof of Theorem 2.2 is straightforward. We can consider a Poisson process of intensity 1 on U_n as the union of two independent Poisson processes on $U_n \setminus Q$ and Q respectively. Call the corresponding random pointsets $\mathcal{P}$ and $\mathcal{B}$ respectively. The event $B_k(Q)$ can then be considered as a product event

$$
B_k(Q) = \{\mathcal{P} \in \tilde{B}_k(Q)\} \times \{\mathcal{B} = \emptyset\},
$$

where $\tilde{B}_k(Q)$ is an event depending only on the points inside $U_n \setminus Q$.

We then have

$$
\begin{aligned}
\mathbb{P}(B_k(Q)) &= \mathbb{P}(\tilde{B}_k(Q))\mathbb{P}(\mathcal{B} = \emptyset) \\
&= \mathbb{P}(\tilde{B}_k(Q)) \exp\left(-\frac{\log n}{N^2}\right) \\
&\leq \mathbb{P}(A_k) \exp\left(-\frac{\log n}{N^2}\right)
\end{aligned}
\tag{3}
$$

where the last line follows from property (iii) of $B_k(Q)$ (which stated that if $B_k(Q)$ occurs then no matter what points we add to Q, the event A_k will occur in the resulting modified pointset).

Now

$$\mathbb{P}(B_k) \leq \mathbb{P}(B_k \setminus \mathcal{C}) + \mathbb{P}(\mathcal{C})$$

$$= \mathbb{P}\left(\bigcup_Q B_k(Q)\right) + \mathbb{P}(\mathcal{C}) \qquad \text{by Theorem 2.6}$$

$$\leq \sum_Q \mathbb{P}(B_k(Q)) + \mathbb{P}(\mathcal{C})$$

$$\leq (MN)^2 \mathbb{P}(A_k) \exp\left(-\frac{\log n}{N^2}\right) + \mathbb{P}(\mathcal{C}) \qquad \text{by (3)}$$

$$= (MN)^2 n^{-1/N^2} \mathbb{P}(A_k) + o(n^{-2}) \qquad \text{by Lemma 2.3}$$

This concludes the proof of Theorem 2.2 with $c_3 = (MN)^2$ and $c_4 = \frac{1}{N^2}$. $\qquad \square$

3 Proof of the global theorem

In this section we prove our global result, Theorem 1.3. It will be convenient to prove something slightly stronger than Theorem 1.3 (albeit with a more cumbersome statement), namely:

Theorem 3.1. *There exist constants γ_1' and $c_1' > 0$ such that for every $\varepsilon : 0 < \varepsilon \leq \frac{1}{2}$, all $n > \varepsilon^{-1/\gamma_1'}$ and all integers $k \in (0.3 \log n, 0.6 \log n)$, if*

$$\mathbb{P}(S_{n,k} \text{ connected}) \geq \varepsilon$$

holds, then

$$\mathbb{P}(S_{n,k} \text{ contains a pair of small, close components}) < \left(\log \frac{1}{\varepsilon}\right) n^{-c_1'}.$$

Proof of Theorem 1.3 from Theorem 3.1. The proof of Theorem 5 of [3] establishes that there exists a constant $\gamma > 0$ such that for any $k \leq 0.3 \log n$, the probability that $S_{n,k}$ is connected is $o(n^{-\gamma})$. Also, Theorem 13 of [3] shows that for $k \geq 0.6 \log n$, the probability that $S_{n,k}$ contains any small component is $o(n^{-(0.6 \log 7 - 1)})$. Thus Theorem 1.3 is immediate from Theorem 3.1 together with an appropriate choice of the constants γ_1 and c_1. $\qquad \square$

Proof of Theorem 3.1 from Theorem 2.2. Let B be the event that $S_{n,k}$ contains a pair of small, close components. We need a few results from [7] relating local connectivity to global connectivity.

Lemma 3.2 (Lemma 2 of [7])**.** *For any n and any integer k with $0.3 \log n < k < 0.6 \log n$, the probability that $U_{n,k}$ contains an edge of length at least $\frac{M\sqrt{\log n}}{8}$ is $O(n^{-6})$.*

(Note that our choice of M is slightly larger than in [7]; however the proof of Lemma 2 in that paper only used the fact that $M > 30$, and so holds in the present setting also.)

To do away with boundary effects, we shall restrict our attention to 'most' of S_n. Let

$$T_n = \left[M\sqrt{\log n}, \left(\left\lfloor \frac{\sqrt{n}}{M\sqrt{\log n}} \right\rfloor - 1\right) M\sqrt{\log n}\right]^2.$$

The nice feature of T_n is that it is not very close to any of the boundary of S_n. The following is an easy Corollary of Theorem 1 of [13]:

Lemma 3.3. *There is a positive constant $0 < c_5 < 2$ such that if $k > 0.3 \log n$ then the probability that $S_{n,k}$ contains any small component not wholly contained inside T_n is $O(n^{-c_5})$.*

As in [4, 7], we now define two covers of T_n by copies of U_n. The *independent* cover $\mathcal{C}_1$ of T_n is obtained by covering T_n with copies of U_n with disjoint interiors. The *dominating* cover $\mathcal{C}_2$ of T_n is obtained from $\mathcal{C}_1$ by replacing each square $V \in \mathcal{C}_1$ by the twenty-five translates $V + (i\frac{M\sqrt{\log n}}{4}, j\frac{M\sqrt{\log n}}{4})$, $i, j \in \{0, \pm 1, \pm 2\}$. By construction, we have $\mathcal{C}_1 \subseteq \mathcal{C}_2$ and the copies of $\frac{1}{4}U_n$ corresponding to elements of $\mathcal{C}_2$ cover the whole of T_n. Also $|\mathcal{C}_2| < \frac{25n}{M^2 \log n}$.

We shall write 'A_k occurs in $\mathcal{C}_i$' as a convenient shorthand for 'there is a copy V of U_n in $\mathcal{C}_i$ for which the event corresponding to A_k occurs', and similarly for B_k. We shall need the following results from [7]:

Lemma 3.4 (Lemma 5 in [7]). *For all $n \in \mathbb{N}$ and all integers k with $0.3 \log n < k < 0.6 \log n$, and c_5 as given by Lemma 3.3,*

$$\mathbb{P}(S_{n,k} \text{ not connected}) = \mathbb{P}(A_k \text{ occurs in } \mathcal{C}_2) + O(n^{-c_5}).$$

Lemma 3.5 (Lemma 6 in [7]). *There exists a constant $\gamma_1' > 0$ such that for all $\varepsilon : 0 < \varepsilon \leq \frac{1}{2}$, all integers $n > \varepsilon^{-1/\gamma_1'}$ and all integers k with $k \in (0.3 \log n, 0.6 \log n)$, if*

$$\mathbb{P}(S_{n,k} \text{ connected}) \geq \varepsilon$$

holds then

$$\mathbb{P}(A_k) \leq \frac{eM^2 \log n}{n} \log\left(\frac{1}{\varepsilon}\right).$$

Similarly to Lemma 3.4, we have

Lemma 3.6. *For all $n \in \mathbb{N}$ and all integers k with $0.3 \log n < k < 0.6 \log n$, and c_5 as given by Lemma 3.3,*

$$\mathbb{P}(B) = \mathbb{P}(B_k \text{ occurs in } \mathcal{C}_2) + O(n^{-c_5}).$$

Proof. This is an easy modification of the proof of Lemma 3.4. . $\qquad\square$

Now, fix $\varepsilon : 0 < \varepsilon \leq \frac{1}{2}$. Suppose $\mathbb{P}(S_{n,k} \text{ connected}) \geq \varepsilon$. Provided $n > \varepsilon^{-1/\gamma_1'}$, we have by Lemma 3.5 that

$$\mathbb{P}(A_k) < \frac{eM^2 \log n}{n} \log\left(\frac{1}{\varepsilon}\right).$$

Now,

$$
\begin{aligned}
\mathbb{P}(B) &= \mathbb{P}(B_k \text{ occurs in } \mathcal{C}_2) + O(n^{-c_5}) && \text{by Lemma 3.6} \\
&\leq \frac{25n}{M^2 \log n}\mathbb{P}(B_k) + O(n^{-c_5}) && \\
&\leq \frac{25c_3}{M^2}\frac{n^{1-c_4}}{\log n}\mathbb{P}(A_k) + O\left(\max\left(n^{-c_5}, \frac{n^{-1}}{\log n}\right)\right) && \text{by Theorem 2.2} \\
&\leq 25ec_3 n^{-c_4}\log\left(\frac{1}{\varepsilon}\right) + O\left(\max\left(n^{-c_5}, \frac{n^{-1}}{\log n}\right)\right) && \text{by our bound on } \mathbb{P}(A_k) \\
&\leq \log\left(\frac{1}{\varepsilon}\right)n^{-c_1'} &&
\end{aligned}
$$

for all $n > \varepsilon^{-1/\gamma_1'}$ and sufficiently small choices of $\gamma_1' > 0$ and $c_1' > 0$. This is the claimed inequality. $\qquad\square$

4 The distribution of the small connected components

In this section, we use Theorems 1.3 and 2.2 together with a form of the Chen–Stein Method due to Arratia, Goldstein and Gordon [1] to show that the small components in $S_{n,k}$ are asymptotically Poisson distributed in a spatial as well as a numerical sense.

The Arratia, Goldstein and Gordon result essentially tells us that if the presence of one small component in a subregion of area $O(\log n)$ does not greatly increase the chance of having other small components in the same subregion, then the number of small components is Poisson distributed (just as we would expect it to be if small components were rare events occurring independently at random inside S_n).

We thus proceed in two stages. First of all we find a good approximation to the distribution of the small components of $S_{n,k}$ using local events. This requires us to use Theorem 1.3, amongst other things. Then we adapt our local result Theorem 2.2 to show that the local events we define are negatively correlated — to be more precise, they are independent if sufficiently far apart, and negatively dependent otherwise. An application of the theorem of Arratia, Goldstein and Gordon concludes the proof.

4.1 Local approximation

We set up some counting functions for the small components. Given a point $x \in S_n$, let $V_n(x)$ be the square of area $\left(4\lambda\sqrt{\log n}\right)^2$ centred at x.

(Recall that λ is the constant given by Definition 1.2 in the introduction; with probability $1 - o(n^{-2})$ there are no edges of length greater or equal to $\lambda\sqrt{\log n}$ and not more than one component of diameter greater than $\lambda\sqrt{\log n}$ in $S_{n,k}$.)

Also let $V_{n,k}(x)$ be the k-nearest neighbour graph on the set of points placed inside $V_n(x)$ by the Poisson point process on S_n.

Definition 4.1. *Let Γ be the* grid $\Gamma = \left\{x \in \mathbb{Z}^2 : V_n(x) \subseteq S_n\right\}$. *Given $x \in \Gamma$, let the* local counting function $Y(x) = Y_{n,k}(x)$ *be the random variable taking the value 1 if there is a connected component H in $V_{n,k}(x)$ such that H has diameter less than $\lambda\sqrt{\log n}$ and x is the (almost surely unique) member of $\mathbb{Z}^2$ closest to the (almost surely unique) bottom-most vertex of H.*

Pick $x \in \Gamma$ and set
$$p = p(n,k) := \mathbb{P}(Y(x) = 1).$$
Note that $\mathbb{P}(Y(x) = 1) = \mathbb{P}(Y(x') = 1)$ for all $x, x' \in \Gamma$, so that the definition of p is independent of x.

Definition 4.2. *Given $x \in \Gamma$, let the* global counting function $X(x) = X_{n,k}(x)$ *be the random variable taking the value 1 if there is a connected component H in $S_{n,k}$ such that H has diameter less than $\lambda\sqrt{\log n}$ and x is the (almost surely unique) member of $\mathbb{Z}^2$ closest to the (almost surely unique) bottom-most vertex of H.*

We shall show that whp the small components of $S_{n,k}$ are counted exactly by $\sum_x X(x)$, and then that whp $X(x) = Y(x)$ for all $x \in \Gamma$. For this we need some easy lemmas.

Lemma 4.3. *Suppose $S_{n,k}$ contains no edge of length greater than $\lambda\sqrt{\log n}$ and that $x \in \Gamma$ is such that $V_{n,k}(x)$ contains no edge of length greater than $\lambda\sqrt{\log n}$. Then $X(x) = Y(x)$.*

Proof. This is a minor modification of Lemma 4 of [7]. □

Lemma 4.4. *Suppose $S_{n,k}$ has at most one non-small component, no two small components close together and no small component close to the boundary of S_n. Then there is (almost surely) a one-to-one correspondence between the small components of $S_{n,k}$ and the $x \in \Gamma$ for which $X(x) = 1$.*

Proof. Almost surely every connected component of $S_{n,k}$ is counted by at most one $X(x)$. Since there are no small components close to the boundary, it follows that every small component is counted by at least one $X(x)$. Finally since there are no small components close together, every $X(x)$ counts at most one component. $\quad\square$

Definition 4.5. *We set $\mathcal{D}$ to be a collection of* bad events. *Let $\mathcal{D}$ be the event that any of the following occur:*

(i) $S_{n,k}$ *contains an edge of length at least* $\lambda\sqrt{\log n}$

(ii) *there is some $x \in \Gamma$ for which $V_{n,k}(x)$ contains an edge of length at least* $\lambda\sqrt{\log n}$

(iii) $S_{n,k}$ *contains at least two components of diameter at least* $\lambda\sqrt{\log n}$

(iv) $S_{n,k}$ *contains a small component H such that the point of $\mathbb{Z}^2$ closest to the bottom-most vertex of H does not lie in Γ*

(v) $S_{n,k}$ *contains at least two components of diameter less than $\lambda\sqrt{\log n}$ lying within distance less than $8\lambda\sqrt{\log n}$ of each other*

(vi) $S_{n,k}$ *contains a small component H such that there is more than one element of $\mathbb{Z}^2$ closest to a bottom-most vertex of H*

(vii) *there is some $x \in \Gamma$ for which $V_{n,k}(x)$ contains a small component H such that there is more than one element of $\mathbb{Z}^2$ closest to a bottom-most vertex of H.*

Our two previous lemmas have the following corollary:

Corollary 4.6. *Suppose that $\mathcal{D}$ does not occur. Then there is a one-to-one correspondence between the small components of $S_{n,k}$ and the x for which $Y(x) = 1$.*

$\quad\square$

How large can this bad set $\mathcal{D}$ be? All but (v) were shown whp not to occur in [7] (up to some trivial changes of constants); so all we need to do is apply Theorem 1.3.

Lemma 4.7. *There exists a constant $c_6 > 0$ such that if k is an integer with $k \in (0.3\log n, 0.6\log n)$ and $\mathbb{P}(S_{n,k}$ connected$) \geq n^{-\gamma_1}$, then*

$$\mathbb{P}(\mathcal{D}) = o\left(n^{-c_6}\right).$$

Proof. Provided we choose c_6 small enough, this is immediate from the properties of λ given in Definition 1.2 (properties (i) and (iii)), Lemma 3.2 applied n times (property (ii) – note we use the fact $\lambda_2 \leq \lambda$ here), Lemma 3.3 (property (iv)), Theorem 1.3 (property (v)) and the fact that almost surely no point of the Poisson process falls on the midpoints of two members of Γ and no two points of the Poisson process fall on the same horizontal line (properties (vi) and (vii)). $\quad\square$

4.2 Global approximation

We now study the distribution of $Y := \sum_{x\in\Gamma} Y(x)$. Our aim is to show its law is Poisson-like. To achieve this we use the Chen–Stein Method [6, 11] in a form due to Arratia, Goldstein and Gordon [1].

Informally this says that provided that mutually dependent pairs of random variables $Y(x)$ and $Y(x')$ are not likely to both equal 1 and that there are not too many such pairs, then the distribution of Y is approximately Poisson. To state Arratia, Goldstein and Gordon's theorem precisely, we need some definitions and notation.

Let $x \in \Gamma$. We can consider a Poisson point process on S_n as the union of independent Poisson point processes on $V_n(x)$ and $S_n \setminus V_n(x)$. By definition $Y(x)$ is independent of the point process on $S_n \setminus V_n(x)$.

Set Γ_x to be the set of $y \in \Gamma$ for which $V_n(x) \cap V_n(y) \neq \emptyset$; this can be thought of as the set of possible dependencies for $Y(x)$.

Define

$$b_1 = \sum_{x \in \Gamma} \sum_{y \in \Gamma_x} \mathbb{P}(Y(x) = 1)\mathbb{P}(Y(y) = 1)$$

$$b_2 = \sum_{x \in \Gamma} \sum_{y \in \Gamma_x \setminus \{x\}} \mathbb{P}(Y(x) = 1, Y(y) = 1).$$

Now let μ be the mean of Y,

$$\mu = \mathbb{E}Y = \sum_x Y(x)$$

$$= |\Gamma|p,$$

and recall from the introduction that $Po_\mu(A)$ is the probability that a Poisson random variable with parameter μ takes a value inside the set A.

Then the following holds:

Theorem 4.8 (Arratia, Goldstein and Gordon [1])**.**

$$\sup_{A \subseteq \mathbb{N} \cup \{0\}} |\mathbb{P}(Y \in A) - Po_\mu(A)| \leq b_1 + b_2.$$

Thus provided we can obtain good upper bounds on b_1 and b_2, we will be close to done.

That b_1 is small will follow from our assumption that the probability of $S_{n,k}$ being connected is not too small. Let $c_8 > 0$ and $\gamma_2 > 0$ be strictly positive constants chosen sufficiently small to satisfy $c_8 < \min(1, c_4, c_6)$ and $\gamma_2 \leq \min(\gamma_1, \frac{c_8}{2})$ respectively.

Lemma 4.9. *There is a constant $c_7 > 0$ such that if $\gamma \leq \min(\gamma_1, c_6/2)$, $k \in (0.3 \log n, 0.6 \log n)$ and $\mathbb{P}(S_{n,k}$ connected$) > n^{-\gamma}$, then*

$$p < \frac{c_7(\log n)^2}{n} \text{ and } \mu < c_7(\log n)^2.$$

Lemma 4.10. *Suppose $k \in (0.3 \log n, 0.6 \log n)$ and*

$$\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_2}.$$

Then

$$b_1 = o(n^{-c_8}).$$

Proof of Lemma 4.10 from Lemma 4.9. We have $|\Gamma| \leq 2n$. Also, as $Y(x)$ is independent of $Y(y)$ for all y with $||x - y|| > 4\sqrt{2}\lambda\sqrt{\log n}$, we have that $|\Gamma_x| < 256\lambda^2 \log n$.

Now suppose $\mathbb{P}(S_{n,k}$ connected$) > n^{-\gamma_2}$. We have

$$b_1 = \sum_{x \in \Gamma} \sum_{y \in \Gamma_x} \mathbb{P}(Y(x) = 1)\mathbb{P}(Y(y) = 1)$$

$$= \sum_{x \in \Gamma} |\Gamma_x|p^2$$

$$\leq 2n(256\lambda^2 \log n)p^2$$

$$< 512{c_7}^2 \lambda^2 \frac{(\log n)^5}{n} \qquad\qquad \text{by Lemma 4.9}$$

$$= o(n^{-c_8}) \qquad\qquad\qquad \text{for } c_8 < 1.$$

$\square$

We now prove Lemma 4.9.

Proof of Lemma 4.9. Suppose that $\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma}$ with $\gamma \leq \min(\gamma_1, \frac{c_6}{2})$. Now

$$
\begin{aligned}
\mathbb{P}(S_{n,k} \text{ connected}) &\leq \mathbb{P}(\{S_{n,k} \text{ connected}\} \cap \mathcal{D}^c) + \mathbb{P}(\mathcal{D}) \\
&= \mathbb{P}(\{Y = 0\} \cap \mathcal{D}^c) + \mathbb{P}(\mathcal{D}) \qquad \text{by Corollary 4.6} \\
&\leq \mathbb{P}(Y = 0) + \mathbb{P}(\mathcal{D}).
\end{aligned}
$$

We have $\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma}$ by hypothesis and $\mathbb{P}(\mathcal{D}) = o(n^{-c_6})$ by Lemma 4.7, so that

$$\mathbb{P}(Y = 0) \geq n^{-\gamma} + o(n^{-c_6}).$$

Now by definition $Y(x)$ and $Y(x')$ are independent random variables for any $x, x' \in \Gamma$ with $\|x - x'\| > 4\sqrt{2}\lambda\sqrt{\log n}$. Since there is a set of at least $\frac{n}{1000\lambda^2 \log n}$ members of Γ which are $4\sqrt{2}\lambda\sqrt{\log n}$ separated, we have

$$\mathbb{P}(Y = 0) \leq (1 - p)^{n/1000\lambda^2 \log n}.$$

Combining this with the lower bound for $\mathbb{P}(Y = 0)$ obtained above, we get

$$(1 - p)^{n/1000\lambda^2 \log n} \geq n^{-\gamma} + o(n^{-c_6}) = n^{-\gamma}\left(1 + o(n^{\gamma - c_6})\right),$$

so that

$$
\begin{aligned}
p &\leq -\log(1 - p) \\
&\leq \frac{1000\lambda^2 \gamma (\log n)^2}{n} + o\left(\frac{\log n}{n} n^{\gamma - c_6}\right) \\
&< \frac{c_7 (\log n)^2}{n}
\end{aligned}
$$

for some suitable constant $c_7 > 0$ (since $\gamma \leq \frac{c_6}{2}$). In particular, $\mu = |\Gamma| p < c_7 (\log n)^2$. $\qquad \square$

We now turn our attention to b_2. Here we will need to use a variant of Theorem 2.2:

Lemma 4.11. *There exists a constant $c_3' > 0$ such that if $k \in (0.3\log n, 0.6\log n)$ and x, x' are points in Γ with $x' \in \Gamma_x$, then*

$$\mathbb{P}(\{Y(x) = 1\} \cap \{Y(x') = 1\}) \leq c_3' n^{-c_4} \mathbb{P}(\{Y(x) = 1\}) + o(n^{-2}).$$

Proof. Lemma 4.11 does not follow directly from Theorem 2.2 since instead of working with a nice square U_n, we begin instead with the union of two intersecting squares $V_n(x)$ and $V_n(x')$. However only a slight modification of our argument in the proof of Theorem 2.2 is needed to establish Lemma 4.11.

We consider a translate U_n' of U_n such that the centre subsquare $\frac{1}{2}U_n'$ contains $V_n(x) \cup V_n(x')$. (This is possible since $M \geq 160\lambda\sqrt{\log n}$ while $V_n(x), V_n(x')$ have side-length $4\lambda\sqrt{\log n}$ and $\|x - x'\| \leq 4\sqrt{2}\lambda\sqrt{\log n}$.) We consider the k-nearest neighbour graph $U_{n,k}'$ on the points placed inside U_n' by our Poisson process on S_n.

We can define events $A_k(x)$ and $B_k(x, x')$ corresponding to $\{Y(x) = 1\}$ and $\{Y(x) = Y(x') = 1\}$ in a natural way: for $y = x, x'$ let $A_k(y)$ be the event that $U_{n,k}'$ contains a small connected component H such that y is the unique element of $\mathbb{Z}^2$ closest to a bottom-most point of H, and let $B_k(x, x')$ be the event that both of these happen, $B_k(x, x') = A_k(x) \cap A_k(x')$.

By following exactly the proof of Theorem 2.2, we get

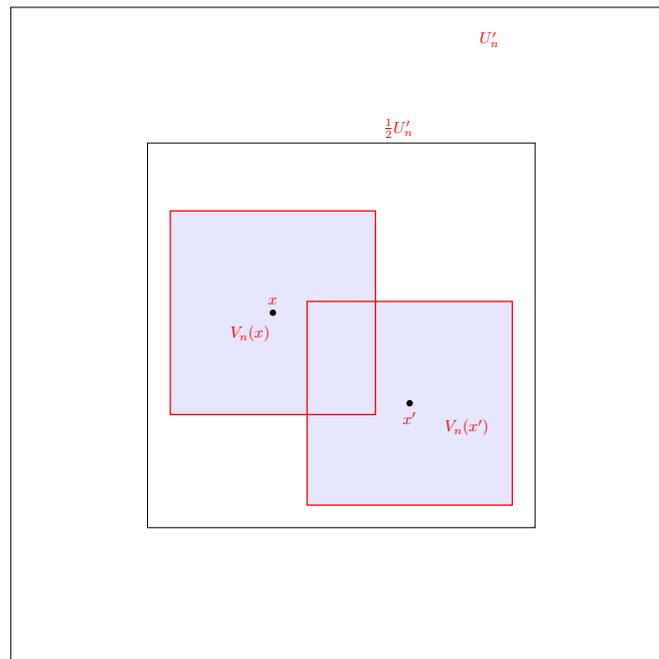$$\mathbb{P}(B_k(x, x')) \leq c_3 n^{-c_4} \mathbb{P}\left(A_k(x) \cup A_k(x')\right) + o(n^{-2}).$$

Figure 3: The intersecting squares V_n and $V_n(x')$, and the translate U_n' with its centre subsquare $\frac{1}{2}U_n'$ featuring in the proof of Lemma 4.11.

Now all we have to show is that $A_k(x)$ and $A_k(x')$ are essentially the same events as $\{Y(x) = 1\}$ and $\{Y(x') = 1\}$ respectively. This is straightforward from Lemma 4 of [7], Lemma 4.3 and Lemma 3.2, which taken together show

$$\mathbb{P}(A_k(x)) = \mathbb{P}(Y(x) = 1) + o(n^{-2}),$$
$$\mathbb{P}(A_k(x')) = \mathbb{P}(Y(x') = 1) + o(n^{-2}).$$

Finally

$$\mathbb{P}\left(A_k(x) \cup A_k(x')\right) \leq \mathbb{P}(A_k(x)) + \mathbb{P}(A_k(x'))$$

so that Lemma 4.11 follows with $c_3' = 2c_3$. □

We are now able to bound b_2 from above:

Lemma 4.12. *Suppose* $k \in (0.3 \log n, 0.6 \log n)$ *and*

$$\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_2}.$$

Then

$$b_2 = o(n^{-c_8}).$$

As $c_8 < \min(1, c_4)$ and $\gamma_2 \leq \min(\gamma_1, c_6/2)$, Lemma 4.12 comes as a straightforward consequence of Lemma 4.11.

Proof. Suppose that $\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma}$ with $\gamma \leq \min(\gamma_1, \frac{c_6}{2})$. Then by Lemma 4.9

$$p < c_7 \frac{(\log n)^2}{n}.$$

Now $|\Gamma| < 2n$ and (as $Y(x)$ is independent of $Y(x')$ for all x' with $||x - x'|| \geq 4\sqrt{2}\lambda\sqrt{\log n}$), $|\Gamma_x| < 256\lambda^2 \log n$ for all $x \in \Gamma$.

Thus

$$
\begin{aligned}
b_2 &= \sum_{x \in \Gamma} \sum_{y \in \Gamma_x \setminus \{x\}} \mathbb{P}(Y(x) = 1, Y(y) = 1) \\
&\leq \sum_{x \in \Gamma} |\Gamma_x| \left(c_3' n^{-c_4} p + o(n^{-2}) \right) && \text{by Lemma 4.11} \\
&\leq 512\lambda^2 c_3' (\log n) n^{1-c_4'} p + o\left(\frac{\log n}{n} \right) && \text{by our bounds on } |\Gamma|, |\Gamma_x| \\
&< 512\lambda^2 c_3' c_7 \frac{(\log n)^3}{n^{c_4}} + o\left(\frac{\log n}{n} \right) && \text{by our bound on } p \\
&= o(n^{-c_8}) && \text{for } c_8 < \min(1, c_4).
\end{aligned}
$$

$\square$

Corollary 4.13. *Suppose $k \in (0.3 \log n, 0.6 \log n)$ and $\mathbb{P}(S_{n,k}$ connected$) > n^{-\gamma_2}$. Then*

$$
\sup_{A \subseteq \mathbb{N} \cup \{0\}} |\mathbb{P}(Y \in A) - Po_\mu(A)| = o(n^{-c_8}).
$$

Proof. The hypotheses of Lemma 4.10 and Lemma 4.12 are satisfied, so that $b_1 + b_2 = o(n^{-c_8})$. Thus by Theorem 4.8,

$$
\sup_{A \subseteq \mathbb{N} \cup \{0\}} |\mathbb{P}(Y \in A) - Po_\mu(A)| \leq b_1 + b_2 = o(n^{-c_8}).
$$

$\square$

4.3 Putting it together

We now put together the results of the previous sections to prove Theorem 1.5.

Corollary 4.13 tells us that Y is approximately Poisson with parameter $\mu = \mathbb{E}Y$, while Corollary 4.6 tells us that whp Y counts exactly the small components of $S_{n,k}$. Our aim is to show that the number of small components X of $S_{n,k}$ is approximately Poisson with parameter $\nu = \nu(n, k)$, where

$$
\nu(n, k) := -\log \mathbb{P}(S_{n,k} \text{ connected}).
$$

Thus what we have left to show is that Po_μ and Po_ν are approximately the same probability measure. We do this in two stages: first we prove μ and ν are almost equal, and then use that to show Po_μ and Po_ν are approximately the same.

Let $c_8 > 0$ and $\gamma_2 > 0$ be the constants defined in the previous subsection. Recall that they satisfy $c_8 < \min(1, c_4, c_6)$ and $\gamma_2 \leq \min(\gamma_1, \frac{c_8}{2})$ respectively.

Lemma 4.14. *Suppose $k \in (0.3 \log n, 0.6 \log n)$ and $\mathbb{P}(S_{n,k}$ connected$) > n^{-\gamma_2}$. Then,*

$$
\nu = \mu + o(n^{-c_8/2}).
$$

Proof. We have

$$
\begin{aligned}
e^{-\nu} &= \mathbb{P}(S_{n,k} \text{ connected}) \\
&= \mathbb{P}\left(\{S_{n,k} \text{ connected}\} \cap \mathcal{D}^c \right) + \mathbb{P}\left(\{S_{n,k} \text{ connected}\} \cap \mathcal{D} \right) \\
&= \mathbb{P}\left(\{Y = 0\} \cap \mathcal{D}^c \right) + \mathbb{P}\left(\{S_{n,k} \text{ connected}\} \cap \mathcal{D} \right) && \text{by Corollary 4.6} \\
&= \mathbb{P}(Y = 0) + o\left(n^{-c_6} \right) && \text{by Lemma 4.7} \\
&= e^{-\mu} + o\left(\max(n^{-c_8}, n^{-c_6}) \right) && \text{by Corollary 4.13} \\
&= e^{-\mu} + o(n^{-c_8}) && \text{since } c_8 < c_6.
\end{aligned}
$$

Now $e^{-\nu} > n^{-\gamma_2}$ and $\gamma_2 < \frac{c_8}{2}$ by assumption, so that

$$
\begin{aligned}
\mu &= -\log\left(e^{-\nu} + o\left(n^{-c_8}\right)\right) \\
&= \nu - \log\left(1 + o(n^{-(c_8-\gamma_2)})\right) \\
&= \nu + o\left(n^{-(c_8-\gamma_2)}\right) \\
&= \nu + o(n^{-c_8/2})
\end{aligned}
$$

as claimed. $\qquad\square$

It readily follows that Po_μ and Po_ν are close to being the same measure.

Corollary 4.15. *If $k \in (0.3\log n, 0.6\log n)$ and*

$$
\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_2}
$$

both hold, then

$$
\sup_{A \subset \mathbb{N} \cup \{0\}} |Po_\nu(A) - Po_\mu(A)| = o\left(n^{-c_8/2}\right).
$$

Proof. Assume without loss of generality that $\nu \geq \mu$. Then we can consider a Poisson random variable with mean ν as the sum of two independent Poisson random variables with means μ and $\nu - \mu$ respectively. Thus for any set $A \subseteq \mathbb{N} \cup \{0\}$,

$$
\begin{aligned}
\mathrm{Po}_\nu(A) &\geq \mathrm{Po}_\mu(A)\mathrm{Po}_{\nu-\mu}(0) \\
&= \mathrm{Po}_\mu(A) \exp(-(\nu - \mu)) \\
&= \mathrm{Po}_\mu(A) + o(n^{-c_8/2}) \qquad\qquad \text{by Lemma 4.14.}
\end{aligned}
$$

In the other direction,

$$
\begin{aligned}
\mathrm{Po}_\nu(A) &\leq \mathrm{Po}_\mu(A)\mathrm{Po}_{\nu-\mu}(0) + \mathrm{Po}_{\nu-\mu}\left([1,\infty)\right) \\
&= \mathrm{Po}_\mu(A) \exp(-(\nu - \mu)) + (1 - \exp(-(\nu - \mu))) \\
&= \mathrm{Po}_\mu(A) + o(n^{-c_8/2}) \qquad\qquad\quad \text{by Lemma 4.14.}
\end{aligned}
$$

Thus for all $A \subseteq \mathbb{N} \cup \{0\}$ we have

$$
|\mathrm{Po}_\mu(A) - \mathrm{Po}_\nu(A)| = o(n^{-c_8/2}),
$$

as claimed.

$\qquad\square$

Theorem 1.5 then follows from the triangle inequality and appropriately small choices of the constants $c_2 > 0$ and $\gamma_2 > 0$.

Proof of Theorem 1.5. Let $X = X_{n,k}$ denote the number of small components in $S_{n,k}$. Assume $\mathbb{P}(S_{n,k} \text{ connected}) > n^{-\gamma_2}$.

The proof of Theorem 5 of [3] establishes the existence of a constant $\gamma > 0$ such that for $k \leq 0.3\log n$ we have

$$
\mathbb{P}(S_{n,k} \text{ connected}) = o(n^{-\gamma}).
$$

Provided $\gamma_2 < \gamma$ and n is sufficiently large this together with our assumption on the connectivity of $S_{n,k}$ guarantees $k > 0.3\log n$.

Also Theorem 13 of [3] shows that for $k \geq 0.6\log n$ the probability that $S_{n,k}$ contains *any* small component is $o\left(n^{-(0.6\log 7-1)}\right)$. Thus if $c_2 < 0.6\log 7 - 1 = 0.16\ldots$ and $k \geq 0.6\log n$ we have

$$
\mathbb{P}(X = 0) = 1 - o(n^{-c_2})
$$

and
$$\nu = -\log\left(\mathbb{P}(S_{n,k} \text{ connected})\right) = o(n^{-c_2})$$

so that
$$\text{Po}_\nu(\{0\}) = 1 - o(n^{-c_2})$$

and
$$\text{Po}_\nu((1,\infty)) = o(n^{-c_2}).$$

Thus in this case
$$\sup_{A \in \mathbb{N} \cup \{0\}} |\mathbb{P}(X \in A) - \text{Po}_\nu(A)| \leq |\mathbb{P}(X = 0) - \text{Po}_\nu(\{0\})| + \mathbb{P}(X \neq 0) + \text{Po}_\nu((1,\infty))$$
$$= o(n^{-c_2}),$$

so that the conclusion of Theorem 1.5 holds in this case.

Now supppose $k \in (0.3 \log n, 0.6 \log n)$. Then,

$$\sup_{A \in \mathbb{N} \cup \{0\}} |\mathbb{P}(X \in A) - \text{Po}_\nu(A)|$$
$$\leq \sup_{A \in \mathbb{N} \cup \{0\}} |\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)| + \sup_{A \in \mathbb{N} \cup \{0\}} |\mathbb{P}(Y \in A) - \text{Po}_\mu(A)|$$
$$+ \sup_{A \in \mathbb{N} \cup \{0\}} |\text{Po}_\mu(A) - \text{Po}_\nu(A)|$$
$$\leq 2\mathbb{P}(\mathcal{D}) + o(n^{-c_8/2}) \quad \text{by Corollary 4.6, Corollary 4.13 and Corollary 4.15}$$
$$= o(n^{-c_2}) \quad \text{by Lemma 4.7, provided } c_2 < \min(c_6, c_8/2).$$

Thus picking c_2 and γ_2 to satisfy

$$0 < c_2 < \min\left(c_6, 0.6 \log 7 - 1, \frac{c_8}{2}\right)$$

and
$$0 < \gamma_2 < \min\left(\gamma, \gamma_1, \frac{c_8}{2}\right)$$

we are done.

$\square$

4.4 Process version

Let us conclude this paper by showing as promised that the locations of the small components are approximately distributed like a Poisson process.

Let $\mathbf{X}$ and $\mathbf{Y}$ be the $|\Gamma|$-dimensional vectors $\mathbf{X} = (X(x))_{x \in \Gamma}$ and $\mathbf{Y} = (Y(x))_{x \in \Gamma}$ respectively. We define two Poisson processes on Γ.

Let $(Z(x))_{x \in \Gamma}$ be a set of $|\Gamma|$ independent, identically distributed random variables, with $Z(x) \sim \text{Poisson}(p)$ for every x. (Recall $p = p(n,k) = \mathbb{P}(Y(x) = 1)$.) Set $\mathbf{Z}$ to be the resulting Poisson process on Γ, $\mathbf{Z} = (Z(x))_{x \in \Gamma}$.

Similary set $p' = \nu/|\Gamma|$ and let $(Z'(x))_{x \in \Gamma}$ be a set of $|\Gamma|$ independent, identically distributed random variables, with $Z'(x) \sim \text{Poisson}(p')$ for every x. Set $\mathbf{Z}'$ to be the resulting Poisson process on Γ, $\mathbf{Z}' = (Z'(x))_{x \in \Gamma}$.

Arratia, Goldstein and Gordon [1] gave the following process version of their Poisson approximation theorem:

Theorem 4.16 (Arratia, Goldstein and Gordon [1])**.** *Let b_1 and b_2 be as in Theorem 4.8. Then*
$$\sup\left\{|\mathbb{P}(\mathbf{Y} \in \mathcal{A}) - \mathbb{P}(\mathbf{Z} \in \mathcal{A})| : \mathcal{A} \subseteq (\mathbb{N} \cup \{0\})^{|\Gamma|}\right\} \leq 2(b_1 + b_2).$$

We shall write $D(\mathbf{Y}, \mathbf{Z})$ for the difference

$$D(\mathbf{Y}, \mathbf{Z}) := \sup \left\{ |\mathbb{P}(\mathbf{Y} \in \mathcal{A}) - \mathbb{P}(\mathbf{Z} \in \mathcal{A})| : \ \mathcal{A} \subseteq (\mathbb{N} \cup \{0\})^{|\Gamma|} \right\},$$

which is known as the *total variation distance* between the distributions of $\mathbf{Y}$ and $\mathbf{Z}$.

Let us now apply Theorem 4.16 to prove a process version of Theorem 1.5:

Theorem 4.17. *Suppose* $\mathbb{P}(S_{n,k}$ *connected*$) > n^{-\gamma_2}$. *Then*

$$D(\mathbf{X}, \mathbf{Z}') = o(n^{-c_2}).$$

Proof. Suppose $\mathbb{P}(S_{n,k}$ connected$) > n^{-\gamma_2}$. As in the proof of Theorem 1.5, we may assume $0.3 \log n < k < 0.6 \log n$ provided n is sufficiently large.

We know by Corollary 4.6 that if $\mathcal{D}^c$ does not occur then $\mathbf{X} = \mathbf{Y}$. Theorem 4.16 and the bounds on b_1 and b_2 we obtained in Lemmas 4.10 and 4.12 tell us $D(\mathbf{Y}, \mathbf{Z})$ is at most $o(n^{-c_2})$. We know by Lemma 4.14 that $\nu = \mu + o(n^{-c_8/2})$. Dividing through by $|\Gamma|$ we get that

$$p' = p + o(n^{-(1+c_8/2)}), \tag{5}$$

from which it is easy to show that $D(\mathbf{Z}, \mathbf{Z}') = o(n^{-c_2})$: running exactly the same argument as in Corollary 4.15 but with p and p' instead of μ and ν, and using (5) instead of Lemma 4.14, we get that

$$\sup_{A \subset \mathbb{N} \cup \{0\}} |\mathbb{P}(Z(x) \in A) - \mathbb{P}(Z'(x) \in A)| = o(n^{-1-c_8/2})$$

for all $x \in \Gamma$. Thus

$$\begin{aligned} D(\mathbf{Z}, \mathbf{Z}') &\leq \sum_{x \in \Gamma} \sup_{A \subset \mathbb{N} \cup \{0\}} |\mathbb{P}(Z(x) \in A) - \mathbb{P}(Z'(x) \in A)| \\ &= o(|\Gamma| n^{-1-c_8/2}) \\ &= o(n^{-c_2}), \end{aligned}$$

as required.

Putting it all together, we have

$$\begin{aligned} D(\mathbf{X}, \mathbf{Z}') &\leq D(\mathbf{X}, \mathbf{Y}) + D(\mathbf{Y}, \mathbf{Z}) + D(\mathbf{Z}, \mathbf{Z}') \\ &\leq 2\mathbb{P}(\mathcal{D}) + o(n^{-c_2}) \\ &= o(n^{-c_2}) \qquad\qquad \text{by Lemma 4.7, since } c_2 < c_6. \end{aligned}$$

$\square$

What Theorem 4.17 effectively says is that the location of the small components of $S_{n,k}$ inside S_n is approximately a Poisson point process. Thus the distribution of the small components is approximately Poisson in both a numerical and in a spatial sense.

References

[1] R. Arratia, L. Goldstein, and L. Gordon, *Two moments suffice for Poisson approximations: the Chen–Stein method*, The Annals of Probability **17** (1989), no. 1, 9–25. MR-0972770

[2] P. Balister and B. Bollobás, *Percolation in the k-nearest neighbor graph*, 2009.

[3] P. Balister, B. Bollobás, A. Sarkar, and M. Walters, *Connectivity of random k-nearest-neighbour graphs*, Advances in Applied Probability **37** (2005), no. 1, 1–24. MR-2135151

[4] _____ , *A critical constant for the k nearest-neighbour model*, Advances in Applied Probability **41** (2009), no. 1, 001–012. MR-2514943

[5] _____ , *Highly connected random geometric graphs*, Discrete Applied Mathematics **157** (2009), no. 2, 309–320. MR-2479805

[6] L.H.Y. Chen, *Poisson approximation for dependent trials*, The Annals of Probability (1975), 534–545. MR-0428387

[7] V. Falgas-Ravry and M. Walters, *Sharpness in the k-nearest neighbour random geometric graph model*, Advances in Applied Probability **44** (2012), 617–634. MR-3024602

[8] E.N. Gilbert, *Random plane networks*, Journal of the Society for Industrial and Applied Mathematics **9** (1961), no. 4, 533–543. MR-0132566

[9] J.M. González-Barrios and A.J. Quiroz, *A clustering procedure based on the comparison between the k-nearest neighbors graph and the minimal spanning tree*, Statistics and probability letters **62** (2003), no. 1, 23–34. MR-1965368

[10] M.D. Penrose, *The longest edge of the random minimal spanning tree*, The annals of applied probability **7** (1997), no. 2, 340–361. MR-1442317

[11] C. Stein, *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*, Proc. Sixth Berkeley Symp. Math. Stat. Prob., 1972, pp. 583–602. MR-0402873

[12] M. Walters, *Random geometric graphs*, Surveys in Combinatorics 2011, Cambridge University Press, 2011, pp. 365–402. MR-2866737

[13] _____ , *Small components in k-nearest neighbour graphs*, Discrete Applied Mathematics **160** (2012), 2037–2047. MR-2927533

[14] F. Xue and P.R. Kumar, *The number of neighbors needed for connectivity of wireless networks*, Wireless networks **10** (2004), no. 2, 169–181.