*Adv. Appl. Prob. (SGSA)* **44,** 617–634 (2012)
*Printed in Northern Ireland*


# SHARPNESS IN THE $k$-NEAREST-NEIGHBOURS RANDOM GEOMETRIC GRAPH MODEL

VICTOR FALGAS-RAVRY * ** AND
MARK WALTERS,* *** *Queen Mary, University of London*

**Abstract**

Let $S_{n,k}$ denote the random graph obtained by placing points in a square box of area $n$ according to a Poisson process of intensity 1 and joining each point to its $k$ nearest neighbours. Balister, Bollobás, Sarkar and Walters (2005) conjectured that, for every $0 < \varepsilon < 1$ and all sufficiently large $n$, there exists $C = C(\varepsilon)$ such that, whenever the probability that $S_{n,k}$ is connected is at least $\varepsilon$, then the probability that $S_{n,k+C}$ is connected is at least $1 - \varepsilon$. In this paper we prove this conjecture. As a corollary, we prove that there exists a constant $C'$ such that, whenever $k(n)$ is a sequence of integers such that the probability $S_{n,k(n)}$ is connected tends to 1 as $n \to \infty$, then, for any integer sequence $s(n)$ with $s(n) = o(\log n)$, the probability $S_{n,k(n)+\lfloor C's \log\log n \rfloor}$ is $s$-connected (i.e. remains connected after the deletion of any $s - 1$ vertices) tends to 1 as $n \to \infty$. This proves another conjecture given in Balister, Bollobás, Sarkar and Walters (2009).

*Keywords:* Random geometric graph; connectivity; sharp transition

2010 Mathematics Subject Classification: Primary 60K35
Secondary 82B43

## 1. Introduction

Let $S_n$ be the square $[0, \sqrt{n}] \times [0, \sqrt{n}] \subset \mathbb{R}^2$, and let $k$ be an integer. Place points in $S_n$ according to a Poisson process of intensity 1, and put an undirected edge between each point and its $k$ nearest neighbours. Let $S_{n,k}$ be the resulting graph.

Several authors (see below) have considered the following question: for which $k$ is $S_{n,k}$ connected? Of course, it is always possible for $S_{n,k}$ to fail to be connected, no matter how large $k$ is; the best we can hope for is that $S_{n,k}$ is connected 'asymptotically'. Formally, given a function $k\colon \mathbb{N} \to \mathbb{N}$ and a property $\mathcal{Q}$ of geometric graphs, we say that $S_{n,k(n)}$ has a property $\mathcal{Q}$ *with high probability* (w.h.p.) if

$$\lim_{n\to\infty} \mathrm{P}(S_{n,k(n)} \text{ has property } \mathcal{Q}) = 1.$$

In words, this states that the probability that a random pointset gives rise to a graph with property $\mathcal{Q}$ tends to 1. Indeed, it will be convenient to distinguish between a *pointset* as an arbitrary set of points and a *Poisson pointset* as a random pointset chosen according to a Poisson process.

Elementary arguments indicate that there exist constants $c_l$ and $c_u$ such that, for every $c < c_l$, $S_{n,\lceil c \log n \rceil}$ is not connected w.h.p., while, for every $c > c_u$, $S_{n,\lfloor c \log n \rfloor}$ is connected w.h.p. Using a result of Penrose [6], Xue and Kumar [10] showed that $c_u \le 5.1774$. A bound

Received 18 January 2011; revision received 1 February 2012.
* Postal address: School of Mathematical Sciences, Queen Mary, University of London, London E1 4NS, UK.
** Email address: v.falgas-ravry@qmul.ac.uk
*** Email address: m.walters@qmul.ac.uk



https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

of $c_u \leq 2\log((4\pi/3 + \sqrt{3}/2)/(\pi + 3\sqrt{3}/4)) \approx 3.8597$ can also be read out of earlier work by Gonzáles-Barrios and Quiroz [5].

These results were significantly improved by Balister *et al.* [1], [2], who established the existence of a critical constant $c_* : 0.3043 < c_* < 1/\log 7 \approx 0.5139$ such that, for any $c < c_*$, $S_{n,\lceil c\log n\rceil}$ is not connected w.h.p. and, for any $c > c_*$, $S_{n,\lfloor c\log n\rfloor}$ is connected w.h.p. They also made the following conjecture about the sharpness of the transition.

**Conjecture 1.** (Conjecture 3 of [1].) *For any* $0 < \varepsilon < 1$*, there exists an integer constant* $C(\varepsilon)$ *such that, for all sufficiently large n, if*

$$\mathrm{P}(S_{n,k} \textit{ is connected}) \geq \varepsilon$$

*then*

$$\mathrm{P}(S_{n,k+C(\varepsilon)} \textit{ is connected}) > 1 - \varepsilon.$$

The main result of this paper is the following theorem which proves the conjecture for an explicit function $C(\varepsilon)$.

**Theorem 1.** *There exist absolute constants* $C > 0$ *and* $\gamma > 0$ *such that, for every* $0 < \varepsilon < 1$ *and all* $n > \varepsilon^{-\gamma}$*, if*

$$\mathrm{P}(S_{n,k} \textit{ is connected}) \geq \varepsilon$$

*then*

$$\mathrm{P}(S_{n,k+\lfloor C\log(1/\varepsilon)\rfloor} \textit{ is connected}) > 1 - \varepsilon.$$

Balister *et al.* [3] proved a weaker variant of Conjecture 1 which they used to prove the following result. Define a graph to be $s$-connected if it remains connected whenever at most $s-1$ vertices are removed. Then if $k = k(n)$ is such that $S_{n,k(n)}$ is connected w.h.p. then, for any $s = o(\log n)$, the graphs $S_{n,k'(n)}$, where $k'(n) = k(n) + \lfloor 6\sqrt{(s-1)\log n}\rfloor$, are $s$-connected w.h.p. in a technical sense of 'on average'. As an immediate corollary to Theorem 1, we remove the somewhat complicated hypothesis that Balister *et al.* needed in the statement of their Theorem 10 (admittedly with a weaker constant). Moreover, in the final section we strengthen this substantially, proving the following theorem.

**Theorem 2.** *Whenever* $k(n)$ *is an integer sequence such that* $S_{n,k(n)}$ *is connected w.h.p. and* $s(n)$ *is an integer sequence with* $s = s(n) = o(\log n)$*, then* $S_{n,k(n)+\lfloor 2Cs\log\log n\rfloor}$ *is* $s$*-connected w.h.p.*

This proves the main conjecture in [3].

Before we describe the structure of our paper, we briefly contrast the $k$-nearest-neighbours model with another classical random geometric graph model introduced by Gilbert [4]. As before, let $S_n$ be the square $[0, \sqrt{n}] \times [0, \sqrt{n}] \subset \mathbb{R}^2$. Let $r$ be a real number. Again, place points in $S_n$ according to a Poisson process of intensity 1, but this time put an undirected edge between any pair of points which lie at a distance of at most $r$ from one another. We denote by $G_{n,r}$ the resulting random geometric graph model; $G_{n,r}$ is often known as the Gilbert disc model. Penrose [6], [7], [8] proved very precise results on the connectivity of $G_{n,r}$. In particular, he showed that isolated vertices are the main obstacle to connectivity in the sense that, w.h.p.,

$$\inf\{r \geq 0\colon G_{n,r} \text{ is connected}\} = \inf\{r \geq 0\colon G_{n,r} \text{ has no isolated vertices}\}.$$

The situation is quite different for the $k$-nearest-neighbours model, which has no isolated vertices nor any immediately apparent analogous family of geometric obstructions to connectivity—indeed, the value of the critical constant $c_*$ is not known (although it may well be the lower bound of $0.3043\ldots$ proved in [1]).

One motivation for the study of $S_{n,k}$ (and the Gilbert disc model) comes from the theory of ad-hoc wireless networks. We imagine that we have various radio transmitters (nodes) that wish to communicate using multiple hops. The transmitters could have fixed range which naturally corresponds to the Gilbert disc model, or they could adjust their power so that each node has some fixed number of neighbours which is exactly the $k$-nearest-neighbour model. In this context Theorem 2 is a result about the fault tolerance of such a network: it states that we can have a fault tolerant network for very little additional cost over the minimum needed for communication.

### 1.1. Outline of the paper

In Section 2 we adapt techniques first introduced in [2] to relate the global property of connectivity to certain families of local 'disconnection' events: these will be events determined by the Poisson process inside a square of area of order $\log n$, which roughly say that the graph inside this square is not connected. We establish a correspondence between the probability of an individual disconnection event occurring and the probability that the graph as a whole is not connected (see Lemma 6).

In Section 3 we prove a geometric lemma which is crucial to our proof of Theorem 1, establishing that 'small' connected components in $S_{n,k}$ have a region of 'high point density'.

In Section 4 we show that removing points from such a dense region results in a much more likely configuration which still gives rise to a small connected component in the $k'$-nearest-neighbour graph for some $k'$ a little smaller than $k$. In other words, the disconnection event is much more likely to occur in the graph $S_{n,k'}$ than in the graph $S_{n,k}$. Then the correspondence between these local disconnection events and global connectivity shows that $S_{n,k'}$ is much more likely to be disconnected than $S_{n,k}$, which is exactly Theorem 1.

In the final section we prove Theorem 2.

## 2. Local obstacles to connectivity

Following [2], we shall relate the global connectivity of $S_{n,k}$ to certain families of local events. Let $M$ be an integer constant which we shall specify later on. Let $U_n$ be the square

$$U_n = \left[\frac{-M\sqrt{\log n}}{2}, \frac{M\sqrt{\log n}}{2}\right] \times \left[\frac{-M\sqrt{\log n}}{2}, \frac{M\sqrt{\log n}}{2}\right] \subset \mathbb{R}^2.$$

We shall refer to the subsquare $\frac{1}{2}U_n$ as the *central subsquare* of $U_n$. Place points in $U_n$ according to a Poisson process of intensity 1, and put an undirected edge between any point and the $k$ points nearest to it to obtain the graph $U_{n,k}$.

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

Note that our $A_k$ event is slightly different from the family of events defined in [2]: there the size of the box corresponding to $U_n$ varied with $k$ rather than $\log n$. One of the advantages of our definition of $U_n$ is that the $A_k$ events are nested: if $k \le k'$ then $A_{k'} \subseteq A_k$. We shall cover most of $S_n$ with copies of $U_n$ and show (approximately) that $S_{n,k}$ is disconnected if and only if the event $A_k$ occurs in one of these copies.

For this argument to work, we need to ensure that, w.h.p., $S_{n,k}$ contains no 'long' edges (relative to the size of $U_n$) and only one connected component of 'large' diameter. The following result is exactly what we want.

**Lemma 1.** (Lemma 1 of [2].) *For any fixed* $\alpha_1, \alpha_2$ *with* $0 < \alpha_1 < \alpha_2$ *and any* $\beta > 0$*, there exists* $c = c(\alpha_1, \alpha_2, \beta) > 0$*, depending only on* $\alpha_1$*,* $\alpha_2$*, and* $\beta$*, such that, for any* $k$ *with* $\alpha_1 \log n \le k \le \alpha_2 \log n$*, the probability that* $S_{n,k}$ *contains two components each of diameter at least* $c\sqrt{\log n}$ *or any edge of length at least* $c\sqrt{\log n}$ *is* $O(n^{-\beta})$*.*

**Remark.** In this paper we use the $O$ notation in a slightly nonstandard way. Most of our results depend on $n$ and $k$, where $k = k(n)$ is a function of $n$. When we say $f(n,k) = O(n)$, we mean 'uniformly in $k$', that is, there is a constant $B$ such that $f(n,k) \le Bn$ for all $n$ and $k$ (satisfying our other constraints).

Let $c_1 = c(0.3, 0.6, 2)$ in Lemma 1, and define a *small component* to be any component with diameter at most $c_1\sqrt{\log n}$. Let $M = \max(\lceil 16c_1 \rceil, 30)$. We shall also need the following lemma, which is an easy modification of Corollary 6 of [2].

**Lemma 2.** *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* $M\sqrt{\log n}/8$ *is* $O(n^{-6})$*.*

*Proof.* This is very similar to the proof of Corollary 6 of [2], but we have to make allowances for the slight difference in our definition of the event $A_k$.

Let $k < 0.6 \log n$. Suppose that some vertex $x \in U_n$ has its $k$th nearest neighbour lying at a distance of at least $M\sqrt{\log n}/8$. Then there must be fewer than $k < 0.6 \log n$ points within the intersection of the disk about $x$ of radius $M\sqrt{\log n}/8$ and the square $U_n$. Since at least one quarter of this disk lies within the square $U_n$, this intersection has area at least $\pi M^2 \log n/256$. (It may be only one quarter of the disk since $x$ may be close to a corner of $U_n$.) Since we picked $M \ge 30$, we have $\pi M^2 \log n/256 > 10 \log n$. Let $X \sim \text{Poisson}(10 \log n)$. Then

$$\begin{aligned}
\mathrm{P}(X < 0.6 \log n) &= \sum_{s < 0.6 \log n} \frac{(10 \log n)^s}{s!} \mathrm{e}^{-10 \log n} \\
&< (0.6 \log n) \left( \frac{10 \log n}{0.6 \log n/\mathrm{e}} \right)^{0.6 \log n} \mathrm{e}^{-10 \log n} \\
&< 0.6 (\log n) \mathrm{e}^{(0.6 \log(50\mathrm{e}/3) - 10) \log n} \\
&< \mathrm{e}^{-7 \log n} \quad \text{for sufficiently large } n.
\end{aligned}$$

Thus, the probability that any vertex $x \in U_n$ has its $k$th nearest neighbour lying at a distance at least $M\sqrt{\log n}/8$ away is at most

$$\mathrm{E}\{\text{number of vertices in } U_n\}\, \mathrm{P}(X < 0.6 \log n) < M^2 (\log n) \mathrm{e}^{-7 \log n} = O(n^{-6}),$$

as required.

We also need to define what we meant by '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 boundaries of $S_n$. The following lemma is a minor restatement of Theorem 1 of [9].

**Lemma 3.** *There is a positive constant* $0 < c_2 < 2$ *such that if* $k > 0.3 \log n$ *then the probability that* $S_{n,k}$ *contains any small component not wholly contained in* $T_n$ *is* $O(n^{-c_2})$*.*

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

We now define two covers of $T_n$ by translates of $U_n$. The *independent* cover $\mathcal{C}_1$ of $T_n$ is obtained by covering $T_n$ with translates 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 $V + (iM\sqrt{\log n}/4, jM\sqrt{\log n}/4)$, $i, j \in \{0, \pm 1, \pm 2\}$. By construction we have $\mathcal{C}_1 \subseteq \mathcal{C}_2$ and the translates of $\frac{1}{4}U_n$ corresponding to elements of $\mathcal{C}_2$ cover the whole of $T_n$. Also, $|\mathcal{C}_2| < 25n/M^2 \log n$.

We shall write '$A_k$ occurs in $\mathcal{C}_i$' as a convenient shorthand for 'there is a translate $V$ of $U_n$ in $\mathcal{C}_i$ for which the event corresponding to $A_k$ occurs'. We shall also write $V_k$ for the $k$-nearest-neighbour graph on $V$, and $\frac{1}{2}V$ for the centre subsquare of $V$.

Lemmas 1, 2, and 3 allow us to relate, up to some small error, the global connectivity to the local events $A_k$. Before we make this relationship precise we need a technical lemma.

**Lemma 4.** *Suppose that $S_{n,k}$ contains no edge of length greater than $M\sqrt{\log n}/16$ and that $V \in \mathcal{C}_2$ is a translate of $U_n$ such that $V_k$ contains no edge of length greater than $M\sqrt{\log n}/8$. Then $S_{n,k}$ has a connected component contained inside $\frac{1}{2}V$ whenever the event corresponding to $A_k$ occurs in $V$.*

*Proof.* Let $\Gamma_V$ denote the subgraph of $V_k$ consisting of all edges with at least one end in $\frac{1}{2}V$, and let $\Gamma_S$ be the subgraph of $S_{n,k}$ consisting of all edges with at least one end in $\frac{1}{2}V$. We aim to show that $\Gamma_V = \Gamma_S$. Obviously, this will imply the lemma.

Let $S_{n,k}[V]$ denote the induced subgraph of $S_{n,k}$ formed by the vertices contained in $V$. Trivially, $S_{n,k}[V]$ is a subgraph of $V_k$. What extra edges can there be in $V_k$? We assume that $S_{n,k}$ contains no edges of length greater than $M\sqrt{\log n}/16$. Thus, only the vertices within a distance $M\sqrt{\log n}/16$ of the boundary of $V$ may be joined in $S_{n,k}$ to points in $S_n \setminus V$. So every edge in $V_k \setminus S_{n,k}[V]$ (i.e. all extra edges) must meet one of these vertices.

Now $V_k$ contains no edges of length greater than $M\sqrt{\log n}/8$, so all the vertices meeting an edge of $V_k \setminus S_{n,k}[V]$ must lie a distance at most

$$\frac{M\sqrt{\log n}}{8} + \frac{M\sqrt{\log n}}{16} < \frac{M\sqrt{\log n}}{4}$$

from the boundary of $V$. Since the vertices inside the central subsquare $\frac{1}{2}V$ all lie at a distance at least $M\sqrt{\log n}/4$ from the boundary of $V$, they do not meet any extra edges, and we have $\Gamma_V = \Gamma_S$ as claimed.

Next we relate the probability of $S_{n,k}$ being connected to the probability of an $A_k$ event occurring somewhere in $S_n$.

**Lemma 5.** *For all $n \in \mathbb{N}$ and all integers $k$ with $0.3\log n < k < 0.6\log n$, and $c_2$ as given by Lemma 3,*

$$\mathrm{P}(S_{n,k} \textit{ not connected}) = \mathrm{P}(A_k \textit{ occurs in } \mathcal{C}_2) + O(n^{-c_2}).$$

*Proof.* Suppose that $A_k$ occurs in $\mathcal{C}_2$. Then there is a translate $V$ of $U_n$ in $\mathcal{C}_2$ for which $A_k$ occurs; in other words, $V_k$ has a connected component $X$ wholly contained inside the central subsquare $\frac{1}{2}V$. By Lemma 1 and our choice of $M$, the probability that $S_{n,k}$ contains an edge of length at least $M\sqrt{\log n}/16$ is $O(n^{-2})$. Let us assume that this does not happen. Then there are no edges between $\frac{1}{2}V$ and $S_n \setminus V$ in $S_{n,k}$. It follows that $X$ is a connected component in $S_{n,k}$ as well as in $V_k$, so $S_{n,k}$ is disconnected. Thus,

$$\mathrm{P}(S_{n,k} \text{ not connected}) \geq \mathrm{P}(A_k \text{ occurs in } \mathcal{C}_2) + O(n^{-2}).$$

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

Conversely, suppose that $S_{n,k}$ is not connected. It must contain at least two connected components. By Lemma 1 and our choice of $M$, the probability that $S_{n,k}$ contains any edge of length at least $M\sqrt{\log n}/16$ or two components of diameter at least $M\sqrt{\log n}/16$ is at most $O(n^{-2})$. By Lemma 3, the probability that $S_{n,k}$ has a small component not contained entirely within $T_n$ is $O(n^{-c_2})$. Also, by Lemma 2, the probability that a particular translate of $U_n$ has any edge longer than $M\sqrt{\log n}/8$ is $O(n^{-6})$. Thus, the probability that $V_k$ has an edge longer than $M\sqrt{\log n}/8$ for *some* translate $V$ of $U_n$ in $\mathcal{C}_2$ is at most $|\mathcal{C}_2|O(n^{-6}) = O(n^{-5})$. Thus, the probability of any of the above occurring in $S_{n,k}$ is at most $O(n^{-c_2})$.

For the remainder of the proof, let us assume that none of the above occurs. Then at least one of the connected components of $S_{n,k}$ is contained in $T_n$ and has diameter less than $M\sqrt{\log n}/16$. Let $X$ be such a component, and let $x$ be a vertex of $X$. By our definition of $\mathcal{C}_2$, there exists a translate $V$ of $U_n$ such that $x \in \frac{1}{4}V$. For any point $x' \notin \frac{1}{2}V$, we have $d(x, x') > M\sqrt{\log n}/8$. By our assumption on the diameter of $X$, we have $x' \notin X$ and, hence, $X \subseteq \frac{1}{2}V$. We have shown that $S_{n,k}$ has a small component, namely $X$, contained entirely inside the central subsquare $\frac{1}{2}V$. Since $V_k$ and $S_{n,k}$ satisfy the hypotheses of Lemma 4, the event corresponding to $A_k$ occurs in $V$ and

$$\mathrm{P}(S_{n,k} \text{ not connected}) \le \mathrm{P}(A_k \text{ occurs in } \mathcal{C}_2) + O(n^{-c_2}).$$

The lemma follows.

Roughly speaking, $\mathrm{P}(A_k \text{ occurs in } C_2)$ is of order $(n/\log n)\,\mathrm{P}(A_k)$. Thus, from a heuristic perspective, Lemma 5 tells us that, as we increase $k$, the transition of $S_{n,k}$ from not connected w.h.p. to connected w.h.p. happens at the same time as the transition from $\mathrm{P}(A_k) \gg \log n/n$ to $\mathrm{P}(A_k) \ll \log n/n$. The following is a precise statement of this relationship.

**Lemma 6.** *There exists a constant $c_3 > 0$ such that, for all $\varepsilon$, $0 < \varepsilon \le \frac{1}{2}$, all integers $n > \varepsilon^{-c_3}$, and all integers $k$, $0.3\log n < k < 0.6\log n$, if*

$$\mathrm{P}(S_{n,k} \textit{ connected}) \ge \varepsilon$$

*holds then*

$$\mathrm{P}(A_k) \le \mathrm{e}\log\left(\frac{1}{\varepsilon}\right)\frac{M^2\log n}{n}.$$

*Conversely, if*

$$\mathrm{P}(A_k) \le \frac{\varepsilon}{\mathrm{e}^4}\frac{M^2\log n}{n}$$

*then*

$$\mathrm{P}(S_{n,k} \textit{ connected}) > 1-\varepsilon.$$

**Remark.** There is nothing special about the constants e and $\mathrm{e}^4$: we picked these values for later convenience, but all we needed was $\mathrm{e} > 2$ and $\mathrm{e}^4 > 25$.

*Proof of Lemma 6.* Suppose that $\mathrm{P}(S_{n,k} \text{ is connected}) \ge \varepsilon$. The translates of $U_n$ contained in $\mathcal{C}_1$ have disjoint interiors; hence, the event corresponding to $A_k$ occurs in each of them independently. Therefore,

$$\mathrm{P}(A_k \text{ occurs in } \mathcal{C}_1) = 1 - (1-\mathrm{P}(A_k))^{|\mathcal{C}_1|}.$$

Now,

$$\begin{aligned}\mathrm{P}(A_k \text{ occurs in } \mathcal{C}_1) &\le \mathrm{P}(A_k \text{ occurs in } \mathcal{C}_2) \quad (\text{since } \mathcal{C}_1 \subset \mathcal{C}_2)\\ &= \mathrm{P}(S_{n,k} \text{ not connected}) + O(n^{-c_2}) \quad (\text{by Lemma 5})\\ &\le 1-\varepsilon+O(n^{-c_2}).\end{aligned}$$

Thus,

$$(1 - \mathrm{P}(A_k))^{|\mathcal{C}_1|} \geq \varepsilon + O(n^{-c_2}).$$

Provided that we choose $c_3$ large enough, we see that, for all $n > \varepsilon^{-c_3}$, the right-hand side is at least $\varepsilon/2$. Taking logarithms of both sides and using the inequality $\log(1-x) \leq -x$ for $0 \leq x \leq 1$ yields

$$-|\mathcal{C}_1|\mathrm{P}(A_k) \geq \log\big(\tfrac{1}{2}\varepsilon\big),$$

so

$$\mathrm{P}(A_k) \leq \frac{1}{|\mathcal{C}_1|}\bigg(\log\frac{1}{\varepsilon/2}\bigg) = \frac{1}{|\mathcal{C}_1|}\bigg(\log\frac{1}{\varepsilon} + \log 2\bigg).$$

Now $\mathcal{C}_1$ contains $(n/M^2 \log n)(1 + O(\sqrt{\log n/n}))$ translates of $U_n$, $0 < \varepsilon \leq \frac{1}{2}$ and $\mathrm{e} > 2$. Hence, provided that we choose our constant $c_3$ sufficiently large, for all $n > \varepsilon^{-c_3}$, we have

$$\mathrm{P}(A_k) \leq \frac{\mathrm{e}M^2 \log n}{n} \log\frac{1}{\varepsilon}.$$

For the converse, suppose that $\mathrm{P}(A_k) \leq \varepsilon M^2 \log n/\mathrm{e}^4 n$. By Lemma 5 we have

$$\begin{aligned}
\mathrm{P}(S_{n,k} \text{ not connected}) &= \mathrm{P}(A_k \text{ occurs in } \mathcal{C}_2) + O(n^{-c_2}) \\
&\leq |\mathcal{C}_2|\mathrm{P}(A_k) + O(n^{-c_2}) \\
&\leq |\mathcal{C}_2|\varepsilon\frac{M^2 \log n}{\mathrm{e}^4 n} + O(n^{-c_2}) \\
&\leq \varepsilon\frac{25}{\mathrm{e}^4} + O(n^{-c_2}) \quad (\text{since } |\mathcal{C}_2| < 25n/M^2 \log n).
\end{aligned}$$

Since $0 < \varepsilon \leq \frac{1}{2}$ and $25/\mathrm{e}^4 < 1$, we have (again providing we choose $c_3$ sufficiently large), for all $n > \varepsilon^{-c_3}$,

$$\mathrm{P}(S_{n,k} \text{ not connected}) < \varepsilon.$$

## 3. Small components have high point density

Having made precise the relationship between $\mathrm{P}(A_k)$ and $\mathrm{P}(S_{n,k}$ connected), we turn our attention to the event $A_k$. Our aim in this section is to show that, provided $k > 0.3 \log n$, small connected components in $U_{n,k}$ witnessing $A_k$ must have a region with high point density.

Let $N$ be an integer constant whose value we shall specify later. We consider a perfect tiling of $U_n$ by square tiles of area $\log n/N^2$. (Such a perfect tiling exists as $U_n$ has area $M^2 \log n$ and $M$, $N$ are integers.) The expected number of points of the Poisson point process on $U_n$ in each tile is $\log n/N^2$. Fix $0 < \eta \leq \frac{1}{2}$.

**Definition 2.** Given a tile $Q$ in $U_n$, we define three events:

1. $A_{k,Q}$ is the event that $A_k$ occurs *and* the tile $Q$ receives more than $(1+\eta)\log n/N^2$ points,
2. $A'_{k,Q}$ is the event that $A_k$ occurs *and* the tile $Q$ receives more than $(1+\eta/2)\log n/N^2$ points,
3. $A_{k,Q,L}$ is the event that if we remove any $L$ points of the process from $Q$ then $A'_{k,Q}$ still occurs.

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

**Lemma 7.** *Suppose that* $k \in [0.3 \log n, 0.6 \log n]$. *Then*

$$\mathrm{P}\Big(A_k \setminus \bigcup_Q A_{k,Q}\Big) = O(n^{-1.1}).$$

The main idea of the proof of this geometric lemma is the following: suppose that $X$ is a connected component of $U_{n,k}$ wholly contained inside $\frac{1}{2}U_n$, and suppose that $x$ is a vertex of $X$ which lies 'on the boundary' of $X$. Write $r$ for the distance between $x$ and its $k$th nearest neighbour.

If $U_{n,k}$ contains no tile with high density (i.e. no tile receiving more than $1+\eta$ times the expected number of points) then the intersection of the ball of radius $r$ centred at $x$ with the 'convex hull' of $X$ must have large area (at least $k/(1+\eta) - o(k)$). In particular, looking outwards from $X$ at $x$ there must be quite a few empty tiles. Doing the above in several different directions we find that $X$ is surrounded by a wide 'sea' of empty tiles of area at least $1.1 \log n$. Since the number of tiles $M^2N^2$ is a constant, the probability that such a collection of empty tiles exists is $O(n^{-1.1})$, yielding the desired result.

Before presenting the proof of Lemma 7, we need the following technical result.

**Lemma 8.** *Let* $\gamma\colon [0,1] \to U_n$ *be a closed continuously differentiable curve in* $U_n$. *Let* $l(\Gamma)$ *be the length of the curve* $\Gamma = \gamma([0,1])$, *and let* $D$ *be the number of tiles it meets. Then*

$$D \leq \frac{9l(\Gamma)}{\sqrt{\log n}/N}.$$

*Proof.* We define a graph $G$ on the set of tiles of $U_n$ by setting an edge between tiles $Q$ and $Q'$ if they meet in at least one point. ($G$ is just the usual square integer lattice on $\{1, 2, \ldots, MN\}^2$ with diagonal edges added.) Every tile has at most eight neighbours in this graph. Let $S$ be the set of tiles met by $\Gamma$. Greedily pick a maximal subset $S' \subseteq S$ which is independent in $G$: pick the tile $Q_1$ with $\gamma(0) \in Q_1$, then pick the first nonadjacent tile $Q_2$ which $\gamma(t)$ next meets, and so on. We have $D = |S| \leq 9|S'|$. Now $\Gamma$ is continuous and cycles through the tiles of $S'$ before coming back to $Q_1$. Since the minimum distance between points lying in nonadjacent tiles is at least one tile length (i.e. $\sqrt{\log n}/N$), it follows that the length of $\Gamma$ satisfies

$$l(\Gamma) \geq |S'| \frac{\sqrt{\log n}}{N}.$$

Substituting $D \leq 9|S'|$ and rearranging terms, we obtain the desired inequality

$$D \leq \frac{9l(\Gamma)}{\sqrt{\log n}/N}.$$

*Proof of Lemma 7.* Let $k$ be an integer with $0.3 \log n < k < 0.6 \log n$. By Lemma 2, the probability of $U_{n,k}$ containing any edge of length at least $M\sqrt{\log n}/8$ is $O(n^{-6})$. Since we are trying to show that $A_k \setminus \bigcup_Q A_{k,Q}$ has probability at most $O(n^{-1.1})$, we may assume in what follows that all edges in $U_{n,k}$ have length strictly less than $M\sqrt{\log n}/8$.

Suppose that $\mathcal{P}$ is a pointset for which $A_k$ occurs, but $A_{k,Q}$ does not occur for any tile $Q$—so, in particular, no tile of $A_k$ contains more than $(1+\eta)\log n/N^2$ points of $\mathcal{P}$. Write $U_{n,k}(\mathcal{P})$ for the $k$-nearest-neighbours graph on $U_n$ associated with the pointset $\mathcal{P}$. Let $X$ be

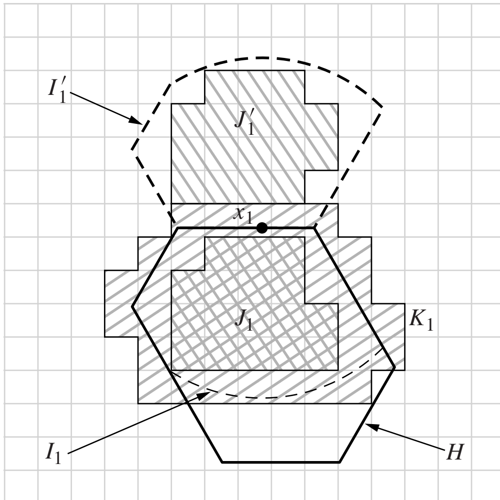


FIGURE 1: The hexagonal hull $H$ and regions $I_1$, $J_1$, $K_1$ and $I_1'$, $J_1'$.

the set of vertices of a connected component of $U_{n,k}(\mathcal{P})$ wholly contained in $\frac{1}{2}U_n$. Using an idea of Balister *et al.* [1], we shall consider the *hexagonal hull* of $X$, $H(X)$, which we now define.

We consider the six tangents to the convex hull of $X$ making angles of $0, \pi/3$, and $2\pi/3$ with the $x$-axis (two for each angle). Together, these define a hexagon $H(X)$ containing $X$ whose edges are segments of the tangents (some of which may have zero length). We shall call $H(X)$ the hexagonal hull of $X$, and label its edges $E_1, E_2, \ldots, E_6$ in cyclic clockwise order so that the top and bottom edges parallel to the $x$-axis are $E_1$ and $E_4$, respectively.

Consider $E_1$. There exists $x_1 \in E_1 \cap X$. Let $r_1$ be the distance between $x_1$ and its $k$th nearest neighbour. Note that, since $r_1$ is the length of an edge of $U_{n,k}$, $r_1 < M\sqrt{\log n}/8$. Let $I_1$ be the intersection of the ball of radius $r_1$ centred at $x_1$ with the hexagon $H(X)$. Let $I_1'$ be the reflection of $I_1$ with respect to $E_1$. Since $I_1' \cap H(X) = \varnothing$ and since every point of $I_1'$ lies at a distance at most $r_1$ from $x_1$, it follows that $I_1'$ contains no point of $\mathcal{P}$.

Next we show that $I_1'$ covers many tiles. Since $r_1 < M\sqrt{\log n}/8$ and $H \subset \frac{1}{2}U_n$, we see that $I_1' \subset U_n$. Let $J_1$ be the union of all of the tiles wholly contained inside $I_1$, and let $J_1'$ be the union of all of the tiles wholly contained inside $I_1'$. Let $K_1$ be the union of all of the tiles meeting $I_1$, and let $K_1'$ be the union of all of the tiles meeting $I_1'$. (See Figure 1.) Since no tile in $U_n$ contains more than $(1+\eta)\log n/N^2$ points of $\mathcal{P}$, it follows that $K_1$ is the union of at least $k/((1+\eta)\log n/N^2)$ tiles.

A tile is contained in $K_1 \setminus J_1$ only if it meets the boundary of $I_1$. Now, since $I_1$ is a convex subset of a disc of radius $r_1$, the boundary of $I_1$ has length less than $2\pi r_1$, so, by Lemma 8, $K_1 \setminus J_1$ is the union of at most $18\pi r_1/(\sqrt{\log n}/N)$ tiles. By the same argument, $K_1' \setminus J_1'$ is the union of at most $18\pi r_1/(\sqrt{\log n}/N)$ tiles. Denote by $|I_1|$ the area of $I_1$; $|I_1'|, |J_1|, \ldots, |K_1'|$ similarly. We have

$$|J_1'| \geq |I_1'| - |K_1' \setminus J_1'| \geq |I_1| - |K_1' \setminus J_1'| \geq |K_1| - |K_1 \setminus J_1| - |K_1' \setminus J_1'|.$$

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

Since each tile has area $\log n/N^2$, our bounds on the number of tiles in $K_1$, $K_1 \setminus J_1$, and $K_1' \setminus J_1'$ imply that

$$|J_1'| \geq \frac{\log n}{N^2}\left(\frac{k}{(1+\eta)\log n/N^2} - \frac{36\pi r_1}{\sqrt{\log n}/N}\right)$$
$$\geq \frac{k}{1+\eta} - \frac{36\pi r_1\sqrt{\log n}}{N}$$
$$\geq \frac{k}{1+\eta} - \frac{9M\pi \log n}{2N}, \tag{1}$$

where the last inequality follows since $r_1$, the radius of the $k$-nearest-neighbour disc about $x_1$, is the length of an edge, so is at most $M\sqrt{\log n}/8$.

We turn at last to the choice of $N$: let $N = 10\lceil 27M\pi \rceil$. For $k > 0.3\log n$ and $\eta \leq \frac{1}{2}$, (1) becomes

$$|J_1'| > \tfrac{11}{60}\log n.$$

For $i = 2, 3, \dots, 6$, we may define $I_i$, $I_i'$, etc. as above. It is easy to see that the $J_i'$ are disjoint: each $J_i'$ lies between the bisectors of two adjacent angles of the convex hexagon $H(X)$. Repeating the argument above to bound below $|J_2'|, \dots, |J_6'|$, we obtain

$$\left|\bigcup_{i=1}^{6} J_i'\right| = \sum_{i=1}^{6} |J_i'| > \frac{11}{10}\log n.$$

Thus, there are at least

$$\frac{11}{10}\frac{\log n}{\log n/N^2} = 110(\lceil 27M\pi \rceil)^2$$

tiles which receive no points. There are at most

$$\binom{M^2N^2}{110(\lceil 27M\pi \rceil)^2}$$

ways of choosing this many tiles. Since $M$ and $N$ are constants, this is just a (large) constant. The probability that there exist $110(\lceil 27M\pi \rceil)^2$ empty tiles (i.e. empty tiles with total area $\frac{11}{10}\log n$) in $U_n$ is therefore

$$O\big(\exp(-\tfrac{11}{10}\log n)\big) = O(n^{-1.1}).$$

Thus,

$$\mathrm{P}\bigg(A_k \setminus \bigcup_{Q} A_{k,Q}\bigg) = O(n^{-1.1}),$$

as claimed.

## 4. The sharp connectivity threshold for $S_{n,k}$

In Lemma 7 we proved that small components witnessing $A_k$ have high point density. We use this fact to prove a sharpness result for $\mathrm{P}(A_k)$, which, by Lemma 6, implies in turn a sharp threshold for the connectivity of $S_{n,k}$ (i.e. Theorem 1). We shall do this by showing that, for all $k' > k$, most pointsets in $A_{k'}$ may be obtained by adding points to already dense parts of $A_k$ pointsets.

We shall need the following lemma, which is a convenient restatement of Theorem 5 of [1].

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

**Lemma 9.** *There exists a positive constant $c_4 > 0$ such that, for every $\varepsilon$ with $0 < \varepsilon \le \frac{1}{2}$ and all $n > \varepsilon^{-c_4}$,*

- *if $k \le 0.3 \log n$ then $\mathrm{P}(S_{n,k}$ connected$) < \varepsilon$, and*
- *if $k \ge 0.6 \log n$ then $\mathrm{P}(S_{n,k}$ connected$) > 1 - \varepsilon$.*

Recall that in the previous section we fixed constants $0 < \eta \le \frac{1}{2}$ and $N \in \mathbb{N}$, and introduced a tiling of $U_n$ into $M^2N^2$ small square tiles as well as the families of events $A_{k,Q}$ and $A'_{k,Q}$. Lemma 7 says that, provided $\mathrm{P}(A_k) = \Omega(n^{-1})$, we have $\mathrm{P}(\bigcup_Q A_{k,Q}) = (1 - O(n^{-0.1}))\, \mathrm{P}(A_k)$. Thus, if a small $A_k$ connected component occurs then, w.h.p., some tile $Q$ receives far more points than expected. We show that if $k' > k$ then most $A_{k'}$ pointsets can be obtained by adding points to an overpopulated tile of an $A_k$ pointset.

Recall our definition of $A_{k,Q,L}$ (Definition 2): it is the event that if we remove any $L$ points of the process from $Q$ then $A'_{k,Q}$ occurs.

**Lemma 10.** *For any tile $Q$ and positive integer $L < \eta \log n/2N^2$, we have*

$$A_{k+L,Q} \subseteq A_{k,Q,L}.$$

*Proof.* Suppose that $\mathcal{P} \subset U_n$ is a pointset for which the event $A_{k+L,Q}$ occurs. It is enough to show that the removal of any $L$ points from $\mathcal{P} \cap Q$ yields a pointset $\mathcal{P}'$ for which the event $A'_{k,Q}$ occurs.

As in Lemma 7, write $U_{n,k}(\mathcal{P})$ for the $k$-nearest-neighbours graph on $U_n$ associated with the pointset $\mathcal{P}$. Since we remove at most $L$ vertices from $\mathcal{P}$, every vertex in $\mathcal{P}$ loses at most $L$ of its $k+L$ nearest neighbours; the set of its $k$ nearest neighbours in $\mathcal{P}'$ is thus a subset of the set of its $k+L$ nearest neighbours in $\mathcal{P}$. It follows that $U_{n,k}(\mathcal{P}')$ is a subgraph of $U_{n,k+L}(\mathcal{P})$.

Since we assume that $A_{k+L,Q}$ occurs, $U_{n,k+L}(\mathcal{P})$ has a connected component wholly contained inside $\frac{1}{2}U_n$. This component must contain at least $k+L+1 > L$ vertices and, since we have removed only $L$ vertices from $\mathcal{P}$ to obtain $\mathcal{P}'$, some vertices of this component remain, that is, $U_{n,k}(\mathcal{P}')$ must also have a component wholly contained inside $\frac{1}{2}U_n$. Thus, $\mathcal{P}' \in A_k$.

Moreover, the number of points in $\mathcal{P}' \cap Q$ is exactly

$$|\mathcal{P} \cap Q| - L > (1+\eta)\frac{\log n}{N^2} - \frac{\eta \log n}{2N^2} = \left(1 + \frac{\eta}{2}\right)\frac{\log n}{N^2},$$

and, hence, $\mathcal{P}' \in A'_{k,Q}$, as claimed.

**Lemma 11.** *Let $L < \eta \log n/2N^2$ be a positive integer, and let $Q$ be a tile. Then*

$$\mathrm{P}(A_{k+L,Q}) < \left(1 + \frac{\eta}{2}\right)^{-L} \mathrm{P}(A'_{k,Q}).$$

*Proof.* First, note that we may consider the Poisson process on $U_n$ as the union of a Poisson process on $Q$ and an independent Poisson process on the disjoint set $U_n \setminus Q$. Now a Poisson point process on $Q$ is just a uniform point process placing

$$Z \sim \mathrm{Poisson}\left(\frac{\log n}{N^2}\right)$$

points in $Q$.

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

We may think of this uniform point process as adding points one by one. If $A_{k,Q,L}$ occurs then, in particular, $A'_{k,Q}$ occurs if we remove the last $L$ points added by the point process. It follows that

$$\begin{aligned}\mathrm{P}(A_{k+L,Q}) \leq \mathrm{P}(A_{k,Q,L}) \quad &(\text{by Lemma 10}) \\ &= \sum_m \mathrm{P}(A_{k,Q,L} \mid Z = m + L)\,\mathrm{P}(Z = m + L) \\ &\leq \sum_m \mathrm{P}(A'_{k,Q} \mid Z = m)\,\mathrm{P}(Z = m + L) \quad (\text{by the definition of } A_{k,Q,L}) \\ &= \sum_m \mathrm{P}(A'_{k,Q} \mid Z = m)\,\mathrm{P}(Z = m) \prod_{i=1}^{L} \frac{N^{-2}\log n}{m+i}. \end{aligned} \tag{2}$$

By the definition of $A'_{k,Q}$,

$$\mathrm{P}(A'_{k,Q} \mid Z = m) = 0 \quad \text{for all } m < \left(1 + \frac{\eta}{2}\right)\frac{\log n}{N^2}.$$

For $m \geq (1 + \eta/2)\log n / N^2$, we have

$$\prod_{i=1}^{L} \frac{N^{-2}\log n}{m+i} < \left(\frac{N^{-2}\log n}{m}\right)^{L} \leq \left(1 + \frac{\eta}{2}\right)^{-L}.$$

Substituting this into (2) gives

$$\mathrm{P}(A_{k+L,Q}) < \left(1 + \frac{\eta}{2}\right)^{-L} \mathrm{P}(A'_{k,Q}),$$

as claimed.

**Lemma 12.** *There exist constants $c_5$ and $L \in \mathbb{N}$ such that, for all $n > c_5$ and all $k$ satisfying*

$$0.3\log n \leq k \leq 0.6\log n - L \quad \textit{and} \quad \mathrm{P}(A_k) \geq n^{-1.05},$$

*we have*

$$\mathrm{P}(A_{k+L}) < \mathrm{e}^{-1}\,\mathrm{P}(A_k).$$

*Proof.* Let $L$ be an integer constant which we shall specify later on. As $\eta$, $L$, and $N$ are all constants, provided that we choose the constant $c_5 > 0$ sufficiently large, then, for all $n > c_5$, we have $L < \eta \log n / 2N^2$—so, in particular, the hypothesis of Lemma 11 is satisfied. Also, $k + L \in [0.3\log n, 0.6\log n]$, so the hypothesis of Lemma 7 is satisfied for $k + L$. Applying the two lemmas successively, we obtain

$$\begin{aligned}\mathrm{P}(A_{k+L}) &= \mathrm{P}\left(\bigcup_Q A_{k+L,Q}\right) + O(n^{-1.1}) \quad (\text{by Lemma 7}) \\ &\leq \sum_Q \mathrm{P}(A_{k+L,Q}) + O(n^{-1.1}) \\ &\leq \sum_Q \left(1 + \frac{\eta}{2}\right)^{-L} \mathrm{P}(A'_{k,Q}) + O(n^{-1.1}) \quad (\text{by Lemma 11}) \\ &\leq M^2 N^2 \left(1 + \frac{\eta}{2}\right)^{-L} \mathrm{P}(A_k) + O(n^{-1.1}), \end{aligned} \tag{3}$$

where the final line follows since $\mathrm{P}(A'_{k,Q}) \leq \mathrm{P}(A_k)$.

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

We now choose $L$: let

$$L = \left\lceil \frac{\log(M^2 N^2 \mathrm{e}^2)}{\log(1+\eta/2)} \right\rceil,$$

so that

$$M^2 N^2 \left(1 + \frac{\eta}{2}\right)^{-L} \leq \mathrm{e}^{-2}.$$

Thus, (3) becomes

$$\mathrm{P}(A_{k+L}) \leq \mathrm{e}^{-2}\, \mathrm{P}(A_k) + O(n^{-1.1}).$$

By assumption, $\mathrm{P}(A_k) \geq n^{-1.05}$, so, again provided that we choose $c_5$ large enough, for all $n > c_5$, we have

$$\mathrm{P}(A_{k+L}) < \mathrm{e}^{-1}\, \mathrm{P}(A_k),$$

as claimed. (Note that the choice of our constant $L$ depended only on the constants $M$, $N$, and $\eta$.)

*Proof of Theorem 1.* In essence, we just iterate Lemma 12. However, we have to choose the right parameters and make sure the conditions hold at each stage.

We choose $\gamma > 0$ such that $\gamma > \max(c_3, c_4, \log_2 c_5, 20)$, where $c_3$, $c_4$, and $c_5$ are the constants in Lemma 6, Lemma 9, and Lemma 12, respectively. Note that, since we defined $M \geq 30$, we have $\mathrm{e}^4/M^2 \log n \leq \mathrm{e}^4/900 \log 2 < 0.09 < 1$ for all $n \geq 2$, so

$$n^{-1/\gamma} > \frac{\mathrm{e}^4}{M^2 \log n} n^{-0.05} \tag{4}$$

for all $n \geq 2$.

Suppose that $n$ and $k$ are such that $\mathrm{P}(S_{n,k}$ is connected$) > \varepsilon$ and $n > \varepsilon^{-\gamma}$. We may assume that $\varepsilon \leq \frac{1}{2}$ and $\mathrm{P}(S_{n,k}$ connected$) \leq 1 - \varepsilon$, for otherwise we have nothing to prove. Since $n > \varepsilon^{-\gamma} > \varepsilon^{-c_4}$ and $\varepsilon < \mathrm{P}(S_{n,k}$ connected$) < 1 - \varepsilon$, Lemma 9 implies that

$$0.3 \log n < k < 0.6 \log n. \tag{5}$$

In particular, for $n > \varepsilon^{-\gamma}$, the assumptions of Lemma 6 are satisfied.

Let $C$ be a strictly positive real constant which we shall specify later on. There are three cases to consider, the first two of which are essentially trivial.

Suppose first of all that

$$k + \left\lfloor C \log \frac{1}{\varepsilon} \right\rfloor \geq 0.6 \log n.$$

Then by Lemma 9 we have $\mathrm{P}(S_{n,k+\lfloor C \log(1/\varepsilon)\rfloor}$ connected$) > 1 - \varepsilon$, and we are done.

Second, suppose that $k + \lfloor C \log(1/\varepsilon) \rfloor < 0.6 \log n$ and

$$\mathrm{P}(A_{k+\lfloor C \log(1/\varepsilon)\rfloor}) < n^{-1.05}.$$

Since $n > \varepsilon^{-\gamma}$, by (4) we have

$$n^{-1.05} < n^{-1/\gamma} \frac{M^2 \log n}{\mathrm{e}^4 n} < \varepsilon \frac{M^2 \log n}{\mathrm{e}^4 n},$$

so by Lemma 6 we have $\mathrm{P}(S_{n,k+\lfloor C \log(1/\varepsilon)\rfloor}$ connected$) > 1 - \varepsilon$, and we are done.

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

Finally, we turn to the case when neither of the above occurs, that is, when

$$k + \left\lfloor C \log \frac{1}{\varepsilon} \right\rfloor < 0.6 \log n$$

and

$$\mathrm{P}(A_{k+\lfloor C \log(1/\varepsilon)\rfloor}) \geq n^{-1.05}.$$

Since $\mathrm{P}(A_{k'})$ monotonically decreases as $k'$ increases, we have

$$\mathrm{P}(A_{k'}) \geq n^{-1.05}$$

for every $k'$, $k \leq k' \leq k + \lfloor C \log(1/\varepsilon)\rfloor - L$. Trivially, for any $k'$ in this range, we have $k' + L < 0.6 \log n$ and, by (5), $0.3 \log n < k'$. Thus, applying Lemma 12 we have, for all $k'$, $k \leq k' \leq k + \lfloor C \log(1/\varepsilon)\rfloor - L$,

$$\mathrm{P}(A_{k'+L}) < \mathrm{e}^{-1}\, \mathrm{P}(A_{k'}). \tag{6}$$

Since $\mathrm{P}(S_{n,k}$ connected$) \geq \varepsilon$, Lemma 6 implies that

$$\mathrm{P}(A_k) \leq \frac{\mathrm{e} M^2 \log n}{n} \log \frac{1}{\varepsilon}.$$

Thus, by repeatedly applying (6), we see that

$$\begin{aligned}
\mathrm{P}(A_{k+\lfloor C \log(1/\varepsilon)\rfloor}) &\leq \mathrm{P}(A_k) \exp\left(-\left\lfloor \frac{\lfloor C \log(1/\varepsilon)\rfloor}{L} \right\rfloor\right) \\
&\leq \frac{\mathrm{e} M^2 \log n}{n} \log \frac{1}{\varepsilon} \cdot \exp\left(-\left\lfloor \frac{C \log(1/\varepsilon)}{L} \right\rfloor\right) \\
&\leq \frac{M^2 \log n}{n} \exp\left(-\left\lfloor \frac{C \log(1/\varepsilon)}{L} \right\rfloor + 1 + \log\log \frac{1}{\varepsilon}\right).
\end{aligned} \tag{7}$$

We now choose $C$: let

$$C = \left(2 + \frac{6}{\log 2}\right) L,$$

where $L$ is the constant in Lemma 12. Since $\varepsilon \leq \frac{1}{2}$, we have $\log(1/\varepsilon)/\log 2 \geq 1$. Thus, for this choice of $C$, we have

$$\begin{aligned}
&-\left\lfloor \frac{C}{L} \log \frac{1}{\varepsilon} \right\rfloor + 1 + \log\log \frac{1}{\varepsilon} \\
&\quad \leq 2 + \log\log \frac{1}{\varepsilon} - \frac{C}{L} \log \frac{1}{\varepsilon} \\
&\quad = \left(2 - 2\frac{\log(1/\varepsilon)}{\log 2}\right) + \left(\log\log \frac{1}{\varepsilon} - \log \frac{1}{\varepsilon}\right) - \frac{4 \log(1/\varepsilon)}{\log 2} - \log \frac{1}{\varepsilon} \\
&\quad \leq -4 - \log \frac{1}{\varepsilon}.
\end{aligned}$$

Substituting this into (7) we obtain

$$\mathrm{P}(A_{k+\lfloor C \log(1/\varepsilon)\rfloor}) \leq \varepsilon \frac{M^2 \log n}{\mathrm{e}^4 n}.$$

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

By Lemma 6, this implies that

$$\mathrm{P}(S_{n,k+\lfloor C\log(1/\varepsilon)\rfloor} \text{ connected}) > 1-\varepsilon,$$

proving the theorem.

## 5. Higher connectivity

In this section we shall apply our sharpness result, Theorem 1, to show Theorem 2, proving a conjecture of Balister *et al.* [3]. Suppose that $\mathcal{P}$ is any pointset in the square $S_n = [0, \sqrt{n}]^2$. Let $G_k(\mathcal{P})$ denote the $k$-nearest-neighbour graph on $\mathcal{P}$.

**Lemma 13.** *Suppose that* $s < 0.9\log n/\log\log n$ *and* $0.3\log n + s < k < 0.6\log n$. *Then there exists a constant* $c_6$ *such that*

$$\mathrm{P}(S_{n,k} \textit{ not } s\textit{-connected}) \le c_6(\log n)\,\mathrm{P}(S_{n,k-1} \textit{ not } (s-1)\textit{-connected}) + O(n^{-3}).$$

*Moreover,*

$$\mathrm{P}(S_{n,k} \textit{ not } s\textit{-connected}) \le (c_6\log n)^{s-1}\,\mathrm{P}(S_{n,k-s+1} \textit{ not connected}) + o(n^{-2}).$$

We shall need the following technical result to prove Lemma 13.

**Lemma 14.** *Suppose that* $0.3\log n < k < 0.6\log n$. *Then there exists* $c_7$ *such that the collection* $\mathcal{C}$ *of pointsets* $\mathcal{P}$ *from which we may delete a set* $T$ *of at most* $0.9\log n/\log\log n$ *points so that either*

- *there exists any point* $x \in S_n$ *(not necessarily in* $\mathcal{P}$*) such that the disc of radius* $c_7\sqrt{\log n}$ *centred at* $x$ *contains less than* $0.6\log n$ *points of* $\mathcal{P}\setminus T$, *or*
- $G_k(\mathcal{P})\setminus T$ *contains at least two components of diameter at least* $c_7\sqrt{\log n}$,

*holds, satisfies* $\mathrm{P}(\mathcal{C}) = O(n^{-3})$.

*Proof.* This is an easy modification of Lemmas 2 and 6 of [1].

*Proof of Lemma 13.* We can view a Poisson pointset as follows. Suppose that $X_1, X_2, X_3, \dots$ is an infinite sequence of uniformly distributed random variables in $S_n$, and let $Z \sim \mathrm{Poisson}(n)$. Then let the points in $\mathcal{P}$ be given by $(X_i)_{i=1}^{Z}$. Let $\mathcal{P}_m$ denote the collection of pointsets with exactly $m$ points that we give the conditional measure, which we will sometimes denote by $\mathrm{P}_m$. From this point of view, it is easy to see that we have $m$ measure preserving maps $\phi_i$ for $1 \le i \le m$ from $\mathcal{P}_m$ to $\mathcal{P}_{m-1}$ given by deleting the point $X_i$. We shall usually abbreviate $\phi_1$ to $\phi$.

Let $\mathcal{A}_s$ denote the collection of pointsets $\mathcal{P}$ for which $G_k(\mathcal{P})$ is not $s$-connected but $G_{k-1}(\mathcal{P})$ is $(s-1)$-connected. Let $\mathcal{B}_s$ denote those pointsets $\mathcal{P}$ for which $G_{k-1}(\mathcal{P})$ is not $(s-1)$-connected. Finally, let $\mathcal{C}$ denote the collection of pointsets $\mathcal{P}$ for which either of the conditions in Lemma 14 holds, which we shall think of as the 'bad' pointsets. By Lemma 14, $\mathrm{P}(\mathcal{C}) = O(n^{-3})$.

For any pointset $\mathcal{P}$ in $\mathcal{A}_s$, it is clear that (at least) one of the functions $\phi_i$ maps $\mathcal{P}$ into $\mathcal{B}_s$. Indeed, since $G_k(\mathcal{P})$ is not $s$-connected, there is a point $X_i$ which we can delete to make the graph not $(s-1)$-connected. Since $G_{k-1}(\mathcal{P}\setminus X_i)$ is a subgraph of $G_k(\mathcal{P})\setminus X_i$, the map $\phi_i$ is one such function. Thus, $\mathcal{A}_s \subseteq \bigcup_{i=1}^{m}\phi_i^{-1}(\mathcal{B}_s)$.

https://doi.org/10.1239/aap/1346955257 Published online by Cambridge University Press

Note that $\mathrm{P}(|Z-n|>n/2)=o(n^{-3})$. We have

$$\begin{aligned}\mathrm{P}(\mathcal{A}_s) &= \sum_{m=0}^{\infty} \mathrm{P}(\mathcal{A}_s \mid Z=m)\,\mathrm{P}(Z=m) \\ &= \sum_{m=n/2}^{3n/2} \mathrm{P}(\mathcal{A}_s \mid Z=m)\,\mathrm{P}(Z=m)+o(n^{-3}) \\ &= \sum_{m=n/2}^{3n/2} \mathrm{P}_m(\mathcal{A}_s \setminus \mathcal{C})\,\mathrm{P}(Z=m)+O(n^{-3}) \\ &= \sum_{m=n/2}^{3n/2} \mathrm{P}_m(\mathcal{P} \in \mathcal{A}_s \setminus \mathcal{C} \text{ and there exists } i\colon \phi_i(\mathcal{P}) \in \mathcal{B}_s)\,\mathrm{P}(Z=m)+O(n^{-3}) \quad (8) \\ &\le \sum_{m=n/2}^{3n/2} \sum_{i=1}^{m} \mathrm{P}_m(\mathcal{P} \in \mathcal{A}_s \setminus \mathcal{C} \text{ and } \phi_i(\mathcal{P}) \in \mathcal{B}_s)\,\mathrm{P}(Z=m)+O(n^{-3}) \\ &= \sum_{m=n/2}^{3n/2} m\,\mathrm{P}_m(\mathcal{P} \in \mathcal{A}_s \setminus \mathcal{C} \text{ and } \phi(\mathcal{P}) \in \mathcal{B}_s)\,\mathrm{P}(Z=m)+O(n^{-3}). \quad (9)\end{aligned}$$

Next we bound $\mathrm{P}_m(\mathcal{P} \in \mathcal{A}_s \setminus \mathcal{C}$ and $\phi(\mathcal{P}) \in \mathcal{B}_s)$. For each $\mathcal{P} \in \mathcal{A}_s \setminus \mathcal{C}$ with $\phi(\mathcal{P}) \in \mathcal{B}_s$, we see that $G_{k-1}(\mathcal{P})$ is $(s-1)$-connected but $G_{k-1}(\phi(\mathcal{P}))$ is not $(s-1)$-connected.

Fix a separating set $T$ of $s-2$ vertices for $G_{k-1}(\phi(\mathcal{P}))$. Since $\mathcal{P} \notin \mathcal{C}$, all but (at most) one of the components in the separated graph $G_{k-1}(\phi(\mathcal{P})) \setminus T$ are small: less than $c_7\sqrt{\log n}$ in diameter. Let $C$ be one such small component.

Since $G_{k-1}(\mathcal{P})$ is $(s-1)$-connected, we see that $G_{k-1}(\mathcal{P}) \setminus T$ is connected. However, $G_{k-1}(\mathcal{P}) \setminus T \setminus \{X_1\}$ is not connected (since it is a subgraph of $G_{k-1}(\mathcal{P} \setminus \{X_1\}) \setminus T = G_{k-1}(\phi(\mathcal{P})) \setminus T$). Thus, $X_1$ must be joined to $C$ in $G_{k-1}(\mathcal{P})$. Since we assume that $\mathcal{P} \notin \mathcal{C}$, the bound on the edge length from Lemma 14 holds and we see that $X_1$ lies within a distance $c_7\sqrt{\log n}$ of $C$.

Combining this with the bound on the diameter of $C$ we see that $X_1$ lies within a set of measure less than $4\pi c_7^2 \log n$ which is determined by $\mathcal{P} \setminus X_1$. This event has probability less than $4\pi c_7^2 \log n/n$. Thus, since $\phi$ is a measure preserving transformation from $\mathrm{P}_m$ to $\mathrm{P}_{m-1}$,

$$\begin{aligned}\mathrm{P}_m(\mathcal{P} \in \mathcal{A}_s \setminus \mathcal{C} \text{ and } \phi(\mathcal{P}) \in \mathcal{B}_s) &\le \frac{4\pi c_7^2 \log n}{n}\,\mathrm{P}_m(\phi(\mathcal{P}) \in \mathcal{B}_s) \\ &= \frac{4\pi c_7^2 \log n}{n}\,\mathrm{P}_{m-1}(\mathcal{P} \in \mathcal{B}_s). \quad (10)\end{aligned}$$

To complete the proof, note that $\mathrm{P}(Z=m) \le 2\,\mathrm{P}(Z=m-1)$ for all $m>n/2$. Thus, substituting (10) into (8),

$$\begin{aligned}\mathrm{P}(\mathcal{A}_s) &\le \sum_{m=n/2}^{3n/2} m\,\mathrm{P}_m(\mathcal{P} \in \mathcal{A}_s \setminus \mathcal{C} \text{ and } \phi(\mathcal{P}) \in \mathcal{B}_s)\,\mathrm{P}(Z=m)+O(n^{-3}) \\ &\le \sum_{m=n/2}^{3n/2} m\frac{4\pi c_7^2 \log n}{n}\,\mathrm{P}_{m-1}(\mathcal{P} \in \mathcal{B}_s)\,\mathrm{P}(Z=m)+O(n^{-3})\end{aligned}$$

$$\leq \sum_{m=n/2}^{3n/2} (12\pi c_7^2 \log n)\, \mathrm{P}_{m-1}(\mathcal{P} \in \mathcal{B}_s)\, \mathrm{P}(Z = m-1) + O(n^{-3})$$
$$\leq (12\pi c_7^2 \log n)\, \mathrm{P}(\mathcal{B}_s) + O(n^{-3}).$$

Finally, observe that

$$\{\mathcal{P} : S_{n,k} \text{ not } s\text{-connected}\} \subseteq \mathcal{A}_s \cup \mathcal{B}_s,$$

so the first part of the lemma holds with $c_6 = 12\pi c_7^2 + 1$, i.e.

$$\mathrm{P}(S_{n,k} \text{ not } s\text{-connected}) \leq c_6 \log n\, \mathrm{P}(S_{n,k-1} \text{ not } (s-1)\text{-connected}) + O(n^{-3}).$$

Iterating this $s - 1 = O(\log n)$ times we obtain the second part of our claim.

We can now finally turn to the proof of Theorem 2.

*Proof of Theorem 2.* By Theorem 2 of [3] we may restrict ourselves to the case where $s(n)$ is an integer sequence with $s(n) \leq \min(\log n/(2\gamma \log\log n), 0.9 \log n/\log\log n)$. Suppose that $k = k(n)$ is such that $S_{n,k}$ is connected w.h.p., so that

$$\mathrm{P}(S_{n,k} \text{ is not connected}) \to 0.$$

Now, letting $\varepsilon = (c_6 \log n)^{-s}$ and applying Theorem 1,

$$\mathrm{P}(S_{n,k+\lfloor C \log(1/\varepsilon)\rfloor} \text{ is not connected}) < \varepsilon = (c_6 \log n)^{-s}$$

for all sufficiently large $n$. (Explicitly, this is for all $n$ with $n > \varepsilon^{-\gamma}$. Given our choice of $\varepsilon$ and the restriction on $s$, $\varepsilon^{-\gamma}$ is at most $\exp(\frac{1}{2}\log n + O(\log n/\log\log n))$, so this is indeed satisfiable for large enough $n$.) Now

$$C \log\frac{1}{\varepsilon} + s - 1 < 2Cs \log\log n$$

for all sufficiently large $n$. If $k + \lfloor 2Cs \log\log n\rfloor < 0.6 \log n$, we have, by Lemma 13,

$$\begin{aligned}
&\mathrm{P}(S_{n,k+\lfloor 2Cs \log\log n\rfloor} \text{ not } s\text{-connected}) \\
&\quad\leq (c_6 \log n)^{s-1}\, \mathrm{P}(S_{n,k-s+1+\lfloor 2Cs \log\log n\rfloor} \text{ not connected}) + o(n^{-2}) \\
&\quad\leq (c_6 \log n)^{s-1}\, \mathrm{P}(S_{n,k+\lfloor C \log 1/\varepsilon\rfloor} \text{ not connected}) + o(n^{-2}) \\
&\quad< (c_6 \log n)^{s-1}\varepsilon + o(n^{-2}) \\
&\quad= O\left(\frac{1}{\log n}\right) \\
&\quad= o(1),
\end{aligned}$$

as required. If, on the other hand, $k + \lfloor 2Cs \log\log n\rfloor \geq 0.6 \log n$, we have

$$\mathrm{P}(S_{n,k+\lfloor 2Cs \log\log n\rfloor} \text{ is not } s\text{-connected}) = o(1)$$

by Theorem 2 of [3]. The result follows.

## References


[1] Balister, P., Bollobás, B., Sarkar, A. and Walters, M. (2005). Connectivity of random $k$-nearest-neighbour graphs. *Adv. Appl. Prob.* **37,** 1–24.

[2] Balister, P., Bollobás, B., Sarkar, A. and Walters, M. (2009). A critical constant for the $k$-nearest-neighbour model. *Adv. Appl. Prob.* **41,** 1–12.

[3] Balister, P., Bollobás, B., Sarkar, A. and Walters, M. (2009). Highly connected random geometric graphs. *Discrete Appl. Math.* **157,** 309–320.

[4] Gilbert, E. N. (1961). Random plane networks. *J. Soc. Indust. Appl. Math.* **9,** 533–543.

[5] González-Barrios, J. M. and Quiroz, A. J. (2003). A clustering procedure based on the comparison between the $k$ nearest neighbors graph and the minimal spanning tree. *Statist. Prob. Lett.* **62,** 23–34.

[6] Penrose, M. D. (1997). The longest edge of the random minimal spanning tree. *Ann. Appl. Prob.* **7,** 340–361.

[7] Penrose, M. D. (1999). On $k$-connectivity for a geometric random graph. *Random Structures Algorithms* **15,** 145–164.

[8] Penrose, M. (2003). *Random Geometric Graphs* (Oxford Stud. Prob. **5**). Oxford University Press.

[9] Walters, M. (2012). Small components in k-nearest neighbour graphs. *Discrete Appl. Math.* **160,** 2037–2047.

[10] Xue, F. and Kumar, P. R. (2004). The number of neighbors needed for connectivity of wireless networks. *Wireless Networks* **10,** 169–181.