View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Western Washington University

Western Washington University
Western CEDAR

Mathematics

College of Science and Engineering

2005

Connectivity of Random k-Nearest-Neighbor Graphs

Paul Balister
University of Memphis

Béla Bollobás
University of Memphis

Amites Sarkar
Western Washington University, amites.sarkar@wwu.edu

Mark Walters
Trinity College

Follow this and additional works at: https://cedar.wwu.edu/math_facpubs


Part of the Mathematics Commons

Recommended Citation

Balister, Paul; Bollobás, Béla; Sarkar, Amites; and Walters, Mark, "Connectivity of Random k-Nearest-Neighbor Graphs" (2005). *Mathematics*. 83.
https://cedar.wwu.edu/math_facpubs/83

This Article is brought to you for free and open access by the College of Science and Engineering at Western CEDAR. It has been accepted for inclusion in Mathematics by an authorized administrator of Western CEDAR. For more information, please contact westerncedar@wwu.edu.

Connectivity of random k-nearest neighbour graphs

Paul Balister[*†] Béla Bollobás[†‡§] Amites Sarkar[†¶]
Mark Walters[*§¶]

October 25, 2006

Abstract

Let $\mathcal{P}$ be a Poisson process of intensity one in a square S_n of area n. We construct a random geometric graph $G_{n,k}$ by joining each point of $\mathcal{P}$ to its $k = k(n)$ nearest neighbours. Recently, Xue and Kumar proved that if $k \leq 0.074 \log n$ then the probability that $G_{n,k}$ is connected tends to zero as $n \to \infty$, while if $k \geq 5.1774 \log n$ then the probability that $G_{n,k}$ is connected tends to one as $n \to \infty$. They conjectured that the threshold for connectivity is $k = (1 + o(1)) \log n$. In this paper we improve these lower and upper bounds to $0.3043 \log n$ and $0.5139 \log n$ respectively, disproving this conjecture. We also establish lower and upper bounds of $0.7209 \log n$ and $0.9967 \log n$ for the directed version of this problem.

A related question concerns coverage. With $G_{n,k}$ as above, surround each vertex by the smallest (closed) disc containing its k nearest neighbours. We prove that if $k \leq 0.7209 \log n$ then the probability that these discs cover S_n tends to zero as $n \to \infty$, while if $k \geq 0.9967 \log n$ then the probability that the discs cover S_n tends to one as $n \to \infty$.

1 Introduction

Suppose n radio transceivers are scattered at random over a desert. Each radio is able to establish a direct two-way connection with the k radios nearest to it. In addition,

[*]Research supported by NSF grant EIA-0130352

[†]University of Memphis, Department of Mathematics, Dunn Hall, 3725 Norriswood, Memphis, TN 38152, USA

[‡]Research supported by NSF grants DMS-9970404 and EIA-0130352 and DARPA grant F33615-01-C1900

[§]Trinity College, Cambridge CB2 1TQ, UK

[¶]Research supported by NSF grant ITR-0225610

messages can be routed via intermediate radios, so that a message can be sent indirectly from radio S to radio T through a series of radios $S = S_1, S_2, \ldots, S_n = T$, each one having a direct connection to its predecessor. How large does k have to be to ensure that any two radios can communicate (directly or indirectly) with each other?

To make this precise, we define a random geometric graph $G(A, \lambda, k)$ as follows. Let $\mathcal{P}$ be a Poisson process of intensity λ in a region A, and join every point of $\mathcal{P}$ to its k nearest neighbours. We would like to know the values of k for which the resulting graph $G(A, \lambda, k)$ is likely to be connected. Throughout this paper, distance is measured using the Euclidean l_2 norm, and is denoted by $\| \ \|$.

There are two equivalent ways of viewing the problem. The first is to fix the area A and let $\lambda \to \infty$. In the second formulation, we instead fix $\lambda = 1$ and grow the region A while keeping its shape fixed, so that the expected number of points in A again increases. As this is the formulation we shall use, we abbreviate $G(A, 1, k)$ to $G(A, k)$. We shall take $A = S_n$, the square of area n (*not* side length n), which ensures that the expected number of points in our region is n. (However, as it turns out, the shape is essentially irrelevant.) Thus we are interested in the values of $k = k(n)$ for which $G_{n,k} = G(S_n, k)$ is likely to be connected, as $n \to \infty$.

Much of the previous work on this problem has been done with the above application (namely, to wireless ad-hoc networks) in mind. In [6, 7, 8, 12, 16, 17] the network is modeled as a Poisson process in the plane, while in [9] the nodes (or transceivers) are located along a line.

Before we get to our main results, we observe that two essentially trivial arguments give the right order of magnitude for k: specifically, that there exist positive constants c_1 and c_2 so that if $k \leq c_1 \log n$ then the probability that $G_{n,k}$ is connected tends to zero as $n \to \infty$, and if $k \geq c_2 \log n$ then the probability that $G_{n,k}$ is connected tends to one as $n \to \infty$. (All logarithms in this paper are to base e). Throughout this paper, we shall say that an event occurs with high probability (**whp**) if it occurs with probability tending to one as $n \to \infty$. Thus, if $k \leq c_1 \log n$ then $G_{n,k}$ is disconnected **whp**, and if $k \geq c_2 \log n$ then $G_{n,k}$ is connected **whp**.

Let us tessellate the square S_n with small squares Q_i of area $\log n - O(1)$, where the (positive) $O(1)$ term is chosen so that the side length of Q_i exactly divides that of S_n. Then the probability that a small square contains no points of the process is $e^{-\log n + O(1)} = O(n^{-1}) = o(\frac{\log n}{n})$, so that **whp** every small square contains at least one point. Using the inequality $r! > (r/e)^r$, the probability that a disc of radius $\sqrt{5 \log n}$ (area $5\pi \log n$) contains more than $k = \lfloor 5\pi e \log n \rfloor < 42.7 \log n$ points is at most

$$e^{-5\pi \log n} \left(\frac{(5\pi \log n)^{k+1}}{(k+1)!} \right) \left(1 + \frac{5\pi \log n}{k+2} + \ldots \right) < e^{-5\pi \log n} (1 + e^{-1} + e^{-2} + \ldots) = o(n^{-1}),$$

so that **whp** every point has at most k points within distance $\sqrt{5\log n}$. Thus **whp** every point of $G_{n,k}$ contained in a square Q_i, is joined to every point in Q_i, and also to every point in every adjacent square. This is enough to make $G_{n,k}$ connected.

Further, if k is much smaller than $\log n$, then **whp** $G_{n,k}$ will not be connected. For consider a configuration of three concentric discs D_1, D_3 and D_5, of radii r, $3r$ and $5r$ respectively, where $\pi r^2 = k+1$. Call the configuration *bad* if (I) D_1 contains at least $k+1$ points, (II) the annulus $D_3 \setminus D_1$ contains no points, and (III) the intersection of $D_5 \setminus D_3$ with any disc of radius $2r$ centered at a point P on the boundary of D_3 contains at least $k+1$ points. Now if a bad configuration occurs anywhere in $G_{n,k}$, then $G_{n,k}$ will not be connected, because the k nearest neighbours of a point in D_1 all lie within D_1 and the k nearest neighbours of a point outside D_3 all lie outside D_3. Hence there will be no edge of $G_{n,k}$ connecting D_1 to $S_n \setminus D_3$. Condition (I) holds with probability approximately $1/2$, condition (II) holds with probability $e^{-8(k+1)}$, and condition (III) holds with probability $1-o(1)$ since a disc of radius $2r$ around a point on the boundary of D_3 is very likely to contain at least $2(k+1)$ points. Hence for $k \le (1-\varepsilon)(\log n)/8$, the probability of a configuration being bad is $p \ge (1/2 - o(1))n^{-1+\varepsilon}$. Since we can fit $\frac{Cn}{\log n}$ copies of D_5 in S_n, and each is bad independently with probability p, the probability that $G_{n,k}$ is connected is at most

$$(1-p)^{\frac{Cn}{\log n}} \le \exp(-C' n^{\varepsilon}/\log n) \to 0,$$

for $k \le (1-\varepsilon)(\log n)/8$.

These elementary arguments indicate that we should focus attention on the range $k = \Theta(\log n)$. Indeed, defining c_l and c_u by

$$c_l = \sup\{c : \mathbb{P}(G_{n,\lfloor c\log n\rfloor} \text{ is connected}) \to 0\},$$

and

$$c_u = \inf\{c : \mathbb{P}(G_{n,\lfloor c\log n\rfloor} \text{ is connected}) \to 1\},$$

we have just shown that

$$0.125 \le c_l \le c_u \le 42.7.$$

By making use of a substantial result of Penrose [13], Xue and Kumar [18] improved the upper bound to

$$c_u \le 5.1774,$$

although a bound of

$$c_u \le \left\{2\log\left(\tfrac{4\pi/3+\sqrt{3}/2}{\pi+3\sqrt{3}/4}\right)\right\}^{-1} \approx 3.8597$$

can be read out of earlier work of Gonzáles-Barrios and Quiroz [5].

It seems likely that $c_l = c_u = c$, and Xue and Kumar asked whether or not $c = 1$. In this paper we improve the above bounds considerably, disproving this conjecture.

The methods used in this paper are new and specific to this problem — however, it is interesting to compare our results with those relating to two similar problems. The first also concerns a Poisson process of intensity 1 in a region A. This time we join each point to all other points within a radius r, obtaining the graph $G_r(A)$: we shall refer to this as the *disc model*. This model originated in a paper of Gilbert [4]. He considered the model in the infinite plane, and was interested in the probability $P_r(\infty)$ that an arbitrary vertex of $G_r(\mathbb{R}^2)$ belongs to an infinite component. Define r_{crit} to be the supremum of the r for which $P_r(\infty) = 0$. Gilbert showed that

$$1.75 \leq \pi r_{\text{crit}}^2 \leq 17.4.$$

Simulations [1, 15] suggest $\pi r_{\text{crit}}^2 \approx 4.512$. The study of $G_r(\mathbb{R}^2)$ is known as continuum percolation, and is the subject of a monograph by Meester and Roy [11]. Many authors reserve the phrase "random geometric graphs" for the graphs $G_r(A)$: however we shall use it in a more general context, so that it includes the graphs $G_{n,k}$ as well.

Regarding connectivity, Penrose [13] showed that if $A = S_n$ and $\pi r^2 = c \log n$, so that each point has on average $c \log n$ neighbours, then there is a critical value of c, in the sense described above, and that it equals one. This is the result used by Xue and Kumar in the work cited above. There is an analogous result for classical random graphs: if in a random graph $G = G(n, p)$ the average degree is $c \log n$, then if $c < 1$, **whp** G is not connected, while if $c > 1$, **whp** G is connected. In both cases, the obstruction for connectivity is the existence of isolated vertices, in the sense that **whp** the graph becomes connected as soon as it has no isolated vertices.

In our problem we expressly forbid isolated vertices, indeed, each vertex has degree at least k. Thus the obstruction for connectivity must involve more complicated extremal configurations, making it harder to obtain precise results. Another complication is that the average vertex degree is not exactly k, but somewhere between k and $2k$. (In fact, it is easy to show that for $k \to \infty$, the average degree is $(1 + o(1))k$.) This motivates the study of the directed case, where, in a Poisson process of intensity 1 in a region A, we place directed edges pointing away from each point towards its k nearest neighbours. This ensures that in the resulting graph $\vec{G}(A, k)$, every vertex has out-degree exactly k. Again, we shall only consider the case $A = S_n$: we further let $k = \lfloor c \log n \rfloor$ and write $\vec{G}_{n,k} = \vec{G}(S_n, k)$. In this variant, we wish to know how large c should be to guarantee a directed path between any two vertices **whp**. Clearly the threshold value of c, if it exists, will be as least as large as in the undirected case. We provide upper and lower bounds for this problem as well.

At first sight it might seem that the following random graph problem might shed some light on the situation: in a graph on n vertices, join each vertex to k randomly chosen others. For what values of k is the resulting graph $G_{n,k\text{-out}}$ connected **whp**? Surprisingly, this question was posed by Ulam [10] in 1935 — see also page 40 of [2]. Here also we have expressly forbidden isolated vertices, however, it is easy to show that even $k = 2$ is enough to ensure connectivity **whp**. In contrast, for the directed version of the problem, where we send a directed edge from each vertex to k randomly chosen others, and ask for a directed path between any two vertices, we need $k \approx \log n$, the main obstruction to connectivity being vertices with zero in-degree.

All our results will apply not only for Poisson processes, but also for n points placed in a square of area n with the uniform distribution. Indeed, one can view our Poisson process as simply the result of placing X points in the square, where $X \sim \mathrm{Po}(n)$. For more details, see [13] and [18].

2 Results

Our main result concerns the undirected random geometric graph $G_{n,k}$.

Theorem 1. *If $c \leq 0.3043$ then $\mathbb{P}(G_{n,\lfloor c\log n\rfloor}$ is connected$) \to 0$ as $n \to \infty$. If $c > 1/\log 7 \approx 0.5139$ then $\mathbb{P}(G_{n,\lfloor c\log n\rfloor}$ is connected$) \to 1$ as $n \to \infty$. Thus*

$$0.3043 \leq c_l \leq c_u \leq 0.5139.$$

The lower bound appears as Theorem 5, while the upper bound is Theorem 13. The lower bound argument is essentially a modification of that given in the introduction, while the proof of the upper bound is more involved.

For the directed graph $\vec{G}_{n,k}$, we have the following result. (A directed graph is *connected* if, given any two vertices x and y, there is a directed path from x to y.)

Theorem 2. *If $c \leq 0.7209$ then $\mathbb{P}(\vec{G}_{n,\lfloor c\log n\rfloor}$ is connected$) \to 0$ as $n \to \infty$. If $c \geq 0.9967$ then $\mathbb{P}(\vec{G}_{n,\lfloor c\log n\rfloor}$ is connected$) \to 1$ as $n \to \infty$.*

Finally, let $\mathcal{P}_n$ be a Poisson process giving rise to the random geometric graph $G_{n,k}$. For a vertex in $x \in V(G_{n,k})$, we define the disc $B_k(x)$ to be the smallest closed disc containing the k nearest neighbours of x. Thus, in $G_{n,k}$, x is (almost surely) joined to every vertex in its disc $B_k(x)$. We say that $\mathcal{P}_n$ is a k-cover if the discs $B_k(x)$ cover S_n, and we prove the following result in Section 6.

Theorem 3. *If $c \leq 0.7209$ then $\mathbb{P}(\mathcal{P}_n$ is a $\lfloor c\log n\rfloor$-cover$) \to 0$ as $n \to \infty$. If $c \geq 0.9967$ then $\mathbb{P}(\mathcal{P}_n$ is a $\lfloor c\log n\rfloor$-cover$) \to 1$ as $n \to \infty$.*

3 Lower bounds

For any region $S \subseteq \mathbb{R}^2$, write $|S|$ for the Lebesgue measure of S. We start by proving a useful lemma.

Lemma 4. *Let $A_1, \dots, A_r$ be disjoint regions of $\mathbb{R}^2$ and $\rho_1, \dots, \rho_r \geq 0$ real numbers such that $\rho_i |A_i| \in \mathbb{Z}$. Then the probability that a Poisson process with intensity 1 has precisely $\rho_i |A_i|$ points in each region A_i is*

$$\exp\left\{\sum_{i=1}^{r} (\rho_i - 1 - \rho_i \log \rho_i)|A_i| + O(r \log_+ \textstyle\sum \rho_i |A_i|)\right\}$$

with the convention that $0 \log 0 = 0$, and $\log_+ x = \max(\log x, 1)$.

Proof. Let $n_i = \rho_i |A_i|$. The probability in question is given exactly by

$$p = \prod_{i=1}^{r} \left(e^{-|A_i|} \frac{|A_i|^{n_i}}{n_i!} \right).$$

Taking logarithms and using Stirling's formula gives

$$\begin{aligned} \log p &= \sum_{i=1}^{r} \left(-|A_i| + n_i \log |A_i| - n_i \log n_i + n_i + O(\log_+ n_i) \right) \\ &= \sum_{i=1}^{r} (n_i - |A_i| - n_i \log \rho_i) + O(r \log_+ \max n_i) \\ &= \sum_{i=1}^{r} (\rho_i - 1 - \rho_i \log \rho_i)\, |A_i| + O(r \log_+ \textstyle\sum \rho_i |A_i|). \end{aligned}$$

□

Theorem 5. *If $c \leq 0.3043$ then $\mathbb{P}(G_{n, \lfloor c \log n \rfloor}$ is connected$) \to 0$ as $n \to \infty$.*

Proof. We first illustrate the proof with a simpler proof that $c < c_0 = 1/(\log \frac{50}{18} + 8 \log \frac{25}{18}) \approx 0.2739$ suffices. Let D be a disc with radius $5r_0$. Let A_1 be a concentric disc with radius r_0, A_2 a concentric annulus with radii r_0 and $3r_0$, and divide the remaining area A of D into $N - 2$ regions $A = \cup_{3 \leq i \leq N} A_i$, with each A_i of diameter at most εr_0 (see Figure 1). Define densities ρ_i by $\rho_1 = 2\rho = \frac{50}{18}$, $\rho_2 = 0$, and $\rho_i = \rho = \frac{25}{18}$ for $i \geq 3$. Suppose that $\rho_i |A_i| \in \mathbb{Z}$ and exactly $\rho_i |A_i|$ points lie in each A_i. (Note that $\sum \rho_i |A_i| = |D|$, so the

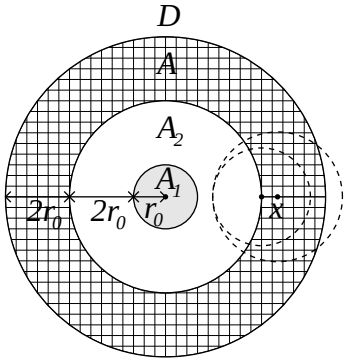

Figure 1: Lower bound, undirected case.

number of points in D is as expected.) Pick a point x at radius $r \geq 3r_0$ from the centre of D. Let D_x be the disc about x of radius $r - (1+\varepsilon)r_0$. Then x is at least εr_0 closer to all points in D_x than to any point in A_1. If $r = 3r_0$ and ε is sufficiently small, then $|D_x \cap A| \geq (1/2 + \delta)|D_x|$ for some $\delta > 0$, independent of ε. Hence for sufficiently small ε, $|D_x \cap A| \geq 2|A_1|$. If you move the point x radially outwards from the centre of D, the discs D_x form a nested family. Thus $|D_x \cap A| \geq 2|A_1|$ for all x. If some A_i, $i \geq 3$, intersects $D_x \cap A$, then all points in A_i are closer to x than any point of A_1. Hence the $2|A_1|\rho = \rho_1|A_i|$ points of the Poisson process closest to x all lie outside A_1. Clearly, if $x \in A_1$ then any point in A_1 is closer to x than any point outside A_1. Hence if we choose r_0 so that $\rho_1|A_1| = k+1 = \lfloor c \log n \rfloor + 1$, the points in A_1 form a component. If S_n contains such a configuration then $G_{n,k}$ is disconnected.

Now $\rho_1|A_1| = k+1$, $\rho_2|A_2| = 0$, and $\sum \rho_i|A_i| = 9\rho_1|A_1| = 9(k+1)$ are all integers. It is easy to see that if n (and hence k and r_0) are large enough, one can choose the regions A_i, $i \geq 3$, so that (i) $\rho_i|A_i| \in \mathbb{Z}$ for all i, (ii) the diameters of the A_i, $i \geq 3$, are at most εr_0, and (iii) the number of regions N is bounded above by some function of ε, independently of n. By Lemma 4, the probability of each A_i containing exactly $\rho_i|A_i|$ points is

$$p = \exp\left\{-\left(\log \tfrac{50}{18} + 8 \log \tfrac{25}{18}\right) \rho_1|A_1| + O(N \log |D|)\right\} = n^{-c/c_0 + o(1)}.$$

Since we can place $\Theta(n/\log n)$ disjoint regions D in S_n, the probability of at least one such configuration occurring in S_n tends to one as $n \to \infty$ when $c < c_0$.

To improve this bound, fix α with $0 < \alpha \leq \frac{1}{3}$. Let $\varepsilon \in (0, \alpha)$ and assume the circles in Figure 1 now have radii $(\alpha - \varepsilon)r_0$, r_0 and $(2-\alpha)r_0$ respectively. Let A_1 be the inner disc of radius $(\alpha - \varepsilon)r_0$, let A_2 be the surrounding annulus with outer radius r_0, and divide the remaining area A into regions A_i, $i = 3, \ldots, N$, each with diameter at most εr_0, and area at least 1. (Certainly possible if εr_0 is sufficiently large.) We shall define a function $\rho(r)$

that gives the approximate density of points in the regions A_i. Let B be the disc of radius αr_0 about O, so B is just a little larger than A_1. For $r \le \alpha r_0$, $\rho(r)$ will be a constant, and we shall require exactly $\rho_1|A_1| = \lfloor \rho(r)|B| \rfloor + 1$ points of $\mathcal{P}$ in A_1. For $\alpha r_0 < r < r_0$, $\rho(r) = 0$, and we shall require that A_2 have no points of the process. For $r \ge r_0$, $\rho(r)$ will be a continuous function, and the number of points in A_i will be $\rho_i|A_i| = \lfloor \int_{A_i} \rho(r)\, dA \rfloor + 1$, where r is the distance to the centre O of D. The function $\rho(r)$ will be determined later, but will be of the form $\rho(r) = \rho_0(r/r_0)$ where ρ_0 may depend on α, but will be independent of n, r_0 and ε. We shall also see that $|\log \rho(r)|$ is bounded on $B \cup A$. We now perform a similar calculation to above, requiring at least $k+1$ points in A_1 and for each point x at distance $r \ge r_0$ from O, at least $k+1$ points in A closer to x than any point of A_1. As before, the worst case is when x is at distance $r = r_0$ from O, and it is enough to ensure that there are at least $k+1$ points in sets A_i that intersect the disc $D_{(1-\alpha)r_0}(x)$ of radius $(1-\alpha)r_0$ about x. Thus it is enough if $\int_{D_{(1-\alpha)r_0}(x) \cap A} \rho\, dA \ge c \log n$. Define

$$g(r) = \tfrac{1}{\pi} \cos^{-1} \left(\tfrac{r^2 + r_0^2 - (1-\alpha)^2 r_0^2}{2 r_0 r} \right),$$

which is the proportion of the circle of radius r, centre O, that lies in $D_{(1-\alpha)r_0}(x)$. Hence

$$\int_{D_{(1-\alpha)r_0}(x) \cap A} \rho\, dA = \int_{r_0}^{(2-\alpha)r_0} \rho(r)\, 2\pi r g(r)\, dr = \int_A \rho g\, dA.$$

Thus it is enough to impose the following conditions on $\rho(r)$.

$$\int_B \rho\, dA = \int_A \rho g\, dA = c \log n. \tag{1}$$

Let δ_ε bound the variation of $\rho \log \rho$ across any of the sets A_i, $i \ge 3$. By the above assumptions, we can choose δ_ε independently of r_0 and n, with $\delta_\varepsilon \to 0$ as $\varepsilon \to 0$. Now by Lemma 4, the probability p of such a configuration occurring is given by

$$-\log p = \int_D (\rho - 1 - \rho \log \rho)\, dA + O(N \log |D| + N + \delta_\varepsilon |D| + \varepsilon c (\log n)/\alpha), \tag{2}$$

where the error terms include the error term of Lemma 4 plus $N-2$ error terms of magnitude $O(1+\delta_\varepsilon |A_i|)$ and one of magnitude $O(1+\varepsilon \rho_1 |A_1|/\alpha)$ arising from the differences between $\int_{A_i} (\rho - 1 - \rho \log \rho)\, dA$ and $(\rho_i - 1 - \rho_i \log \rho_i)|A_i|$ for $i = 1, \ldots, N$.

The function $\rho(r)$ is chosen to maximize the above integral subject to (1). Using the method of Lagrange multipliers, we maximize

$$\int_D (\rho - 1 - \rho \log \rho)\, dA - \mu \int_B \rho\, dA - \nu \int_A \rho g\, dA. \tag{3}$$

By applying the calculus of variations, we obtain

$$\rho(r) = \begin{cases} \exp(\mu) & \text{if } r \le \alpha r_0; \\ 0 & \text{if } r \in (\alpha r_0, r_0); \\ \exp(\nu g(r)) & \text{if } r \ge r_0, \end{cases} \tag{4}$$

where the constants μ and ν are chosen so that

$$\int_B \rho \, dA = \int_A \rho g \, dA \qquad \text{and} \qquad \int_D (\rho - 1) \, dA = 0.$$

(The second condition comes from varying the scale r_0, which implies that the expression (3) should equal zero.) It is easy to check that each value of α gives a unique value of μ and ν, and the conditions assumed for $\rho(r)$ above do indeed hold. Also, $|D| = O(\log n)$ and $N = O(\varepsilon^{-2})$, so by taking, say, $\varepsilon \sim (\log n)^{-1/3}$, $\varepsilon r_0 \to \infty$ and the error term in (2) is $o(\log n)$. Substituting into (2) we get $-\log p = (c(\mu + \nu) + o(1)) \log n$. Since we can place $\Theta(n/\log n)$ disjoint copies of D inside S_n, $G_{n,k}$ is disconnected **whp** whenever $c < (\mu + \nu)^{-1}$. Finally, optimizing over α gives a value of $(\mu + \nu)^{-1}$ just larger than 0.3043 when $\alpha = 0.3302$. □

Note that we were lucky that the optimum value of α was less than $\frac{1}{3}$. For $\alpha > \frac{1}{3}$ the distances between points in A_1 could be larger than the distance from A_1 to A. Hence we would need more points in A_1, and we would need to cut A_1 into smaller regions with varying densities in a similar manner to that done with A.

Theorem 6. *If $c \le 0.7209$ then $\mathbb{P}(\vec{G}_{n,\lfloor c \log n \rfloor}$ is connected$) \to 0$ as $n \to \infty$.*

Proof. We first illustrate the proof with a simpler proof that $c < c_1 = 1/(6 \log \frac{4}{3}) \approx 0.5793$ suffices. Let D be a disc with radius $2r_0$ and centre O. Set A_1 to be a disc about O with radius εr_0, A_2 an annulus with centre O and radii εr_0 and r_0, and divide the remaining annulus A of D into regions $A_3, \dots, A_N$, each with diameter at most εr_0 (see Figure 2). Define densities ρ_i by $\rho_2 = 0$, and $\rho_i = \rho = \frac{4}{3}$ for $i \ge 3$. Suppose that there is one point of the Poisson process in A_1 and $\rho_i |A_i|$ points of the Poisson process lie in each A_i for $i \ge 2$. Pick a point x at distance $r \ge r_0$ from O and let D_x be the disc about x of radius $r - 2\varepsilon r_0$. Then x is at least εr_0 closer to every point in D_x than to A_1. As r moves radially outwards, $D_x \cap A$ increases, so $|D_x \cap A|$ is at least as large as when $r = r_0$. In this case $|D_x \cap A| > \pi r_0^2/2$ for sufficiently small ε. If some A_i, $i \ge 3$, intersects $D_x \cap A$ then all points in A_i are closer to x than O, so the $\rho \pi r_0^2/2$ closest points to x lie outside A_1. Choose r_0 so that $\rho \pi r_0^2/2 = k + 1 = \lfloor c \log n \rfloor + 1$. Then the unique point in A_1 has zero in-degree, so if S_n contains such a configuration then $\vec{G}_{n,k}$ is disconnected. As before, fixing $\varepsilon > 0$

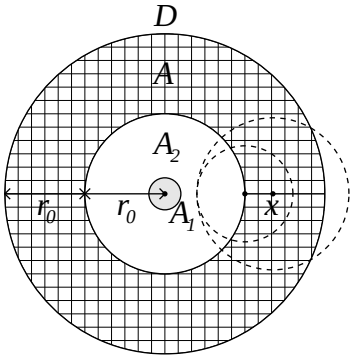

Figure 2: Lower bound, directed case.

and assuming n is sufficiently large, once can choose the A_i so that $\rho_i|A_i| \in \mathbb{Z}$, and N is bounded by a function of ε, independently of n. Now by Lemma 4, the probability of such a configuration is

$$p = \exp\left\{-4\pi r_0^2 \log\tfrac{4}{3} + O((\log|A_1|)/|A_1|) + O(N\log|D|)\right\} = n^{-c/c_1+o(1)}.$$

Since we can find $\Theta(n/\log n)$ disjoint copies of D in S_n, the probability of at least one such configuration occurring tends to 1 as $n \to \infty$ provided $c < c_1$.

To improve this bound, we follow the proof of Theorem 5 and make the assumption that the ρ_i are given by a function $\rho(r)$ of the distance r to the centre of D. We shall define the A_i exactly as in Theorem 5 with a small $\alpha > 0$, but insist now that A_1 contains precisely one point of $\mathcal{P}$, and $\rho(r) = 0$ for all $r < r_0$. We obtain (2) again (with the last term in the error estimate replaced with $\log|A_1|$), which we wish to maximize subject to the conditions $\rho(r) = 0$ for $r \le r_0$ and $\int_A \rho g\,dA = c\log n$. To do this we maximize (3) without the $\mu\int_B \rho\,dA$ term. After optimizing we obtain

$$\rho(r) = \begin{cases} 0 & \text{if } r \le r_0; \\ \exp(\nu g(r)) & \text{if } r > r_0, \end{cases}$$

where $\nu = \nu(\alpha)$ is chosen so that $\int_D(\rho - 1)\,dA = 0$. On substituting back into (2) and choosing $\varepsilon \sim (\log n)^{-1/3}$, this gives $-\log p = (c\nu + o(1))\log n$. As before, we can find $\Theta(n/\log n)$ disjoint discs D. Hence provided $c < \nu^{-1}$, $\vec{G}_{n,k}$ is disconnected **whp**, with an isolated point as an in-component. Finally, for sufficiently small α, ν^{-1} is just larger than 0.7209. □

4 Upper bounds

In this section we shall establish upper bounds for the directed and undirected cases. The basic arguments are simple, but, in both cases, the situation is complicated by points near the boundary. In principle these should be less of a problem than in the disc model: unfortunately, for both problems the most natural arguments run into trouble at the boundary. For the moment we shall ignore boundary effects, and assume that all points are *normal*: a point P is normal if the smallest circle containing its k nearest neighbours does not intersect the boundary. This excludes $O(\sqrt{n \log n})$ points from consideration, and enables us to give the following "one line" argument.

Theorem 7. *Let $c > \frac{1}{\log 2} \approx 1.4427$. Then the probability that $G_{n,\lfloor c \log n \rfloor}$ contains a component consisting entirely of normal points tends to zero as $n \to \infty$.*

Proof. Suppose that $G_{n,\lfloor c \log n \rfloor}$ has a component G' containing only normal points. Let P be a northernmost point of G'. Then P is "extreme" in the sense that its $k = \lfloor c \log n \rfloor$ nearest neighbours all lie below it. The probability that a normal point is extreme is 2^{-k}, and so the expected number of extreme normal points is at most $n2^{-k} = o(1)$. Thus the probability of such a G' arising tends to zero as $n \to \infty$. □

As an aside, we can consider the analogous problem on the *torus*, rather than the square S_n. Unfortunately, the above proof does not show that the corresponding graph on the torus is connected **whp** for $c > \frac{1}{\log 2}$, since a component on the torus need not have any extreme points.

Next we establish an upper bound. The proof splits into two parts. In the first (Lemma 12) we show that there do not exist two "large" components; indeed we show that even if k is far smaller than $\log n$ then these components do not exist. Secondly we show that there are no small components.

We shall use the following simple lemma that bounds the edge lengths. There are many results in the literature bounding the Poisson distribution; we give a simple bound in a form convenient for our needs.

Lemma 8. *Fix $c > 0$, and set*

$$c_- = ce^{-1-1/c} \qquad \text{and} \qquad c_+ = 4e(1+c).$$

If r and R are such that $\pi r^2 = c_- \log n$ and $\pi R^2 = c_+ \log n$, then **whp** *every vertex in $G_{n,\lfloor c \log n \rfloor}$ is joined to every vertex within distance r, and no vertex is joined to a vertex at distance more than R. The same is true for the directed model $\vec{G}_{n,\lfloor c \log n \rfloor}$.*

Proof. This lemma will follow from simple properties of the Poisson distribution. Write $D_\rho(P)$ for the open disc of radius ρ centred at P. Fix $k = \lfloor c \log n \rfloor$, and suppose that a vertex P of $G_{n,k}$ is not joined to every other vertex of $G_{n,k}$ in $D_r(P) \cap S_n$, where $\pi r^2 = c_- \log n = \lambda$. Then $D_r(P) \cap S_n$, which has area at most λ, contains at least k additional vertices of $G_{n,k}$. The probability p of this happening can be bounded as follows (by comparison with a geometric series):

$$p = e^{-\lambda} \sum_{l=k}^{\infty} \frac{\lambda^l}{l!} < e^{-\lambda} \frac{k}{k-\lambda} \frac{\lambda^k}{k!} < e^{-\lambda} \frac{k}{k-\lambda} \left(\frac{\lambda e}{k}\right)^k = \frac{c}{c-c_-} n^{c(\log(c_-/c)+1)-c_-}(1+o(1)),$$

which is $o(n^{-1})$ provided

$$c_- < c \qquad \text{and} \qquad c \log(c_-/c) + c - c_- < -1,$$

which is true for c_- as in the statement of the theorem.

Since the expected number of vertices in S_n is n, the expected number of vertices P such that $D_r(P) \cap S_n$ contains at least k additional vertices is $o(1)$, and hence the probability that there is any such vertex P in $G_{n,k}$ is $o(1)$ as claimed.

The proof of the upper bound is almost the same. Let R satisfy $\pi R^2 = c_+ \log n$. If a vertex is joined to another at distance at least R then the circle of radius R about one of the two, P say, contains at most k additional vertices of $G_{n,k}$. The area of $D_R(P) \cap S_n$ is at least $\pi R^2/4 = (c_+/4) \log n = \lambda$, so the probability p that this occurs for a particular vertex can be bounded by

$$p = e^{-\lambda} \sum_{l=0}^{k} \frac{\lambda^l}{l!} < e^{-\lambda} \frac{\lambda}{\lambda-k} \frac{\lambda^k}{k!} < e^{-\lambda} \frac{\lambda}{\lambda-k} \left(\frac{\lambda e}{k}\right)^k = \frac{c_+}{c_+-4c} n^{c(\log(c_+/4c)+1)-c_+/4}(1+o(1)),$$

which is $o(n^{-1})$ provided

$$c_+ > 4c \qquad \text{and} \qquad c \log(c_+/4c) + c - c_+/4 < -1,$$

which is true for c_+ as in the statement of the theorem (using the inequality $\log((c+1)/c) \leq 1/c$). Hence, the probability we have any such vertex P is $o(1)$. □

Remark. Although we only claim that the above result holds **whp**, much more is true: indeed, for any fixed constant K, we can find c_- and c_+ such that it holds with probability $1 - O(n^{-K})$.

The next two lemmas state simple facts about the components of $G_{n,k}$.

Lemma 9. *No two edges belonging to different components of $G_{n,k}$ may cross.*

Proof. Let $G_1, G_2, \ldots, G_N$ be the components of $G_{n,k}$. Suppose that $i_1i_2 = e_i \in E(G_i)$ and $j_1j_2 = e_j \in E(G_j)$, for $i \neq j$, and that e_i and e_j cross. Then, considering e_i, if i_2 is one of the kth nearest neighbours of i_1, then $\|j_1 - i_1\| > \|i_1 - i_2\|$, while if i_1 is one of the kth nearest neighbours of i_2, then $\|j_1 - i_2\| > \|i_1 - i_2\|$. Therefore, in either case, e_i is not the longest edge of the triangle $i_1i_2j_1$, and so the angle $i_1j_1i_2$ is less than $\frac{\pi}{2}$. But this applies to all four angles of the quadrilateral $i_1j_1i_2j_2$, which gives a contradiction. □

Lemma 10. *With r as in Lemma 8,* **whp** *the distance between any two edges belonging to different components of $G_{n,k}$ is at least $r/2$.*

Proof. As before, let $G_1, G_2, \ldots, G_N$ be the components of $G_{n,k}$, and let $i_1i_2 = e_i \in E(G_i)$ and $j_1j_2 = e_j \in E(G_j)$, for $i \neq j$. Since e_i and e_j do not cross, the distance between them is attained at a vertex of one of them, say j_1, and thus, we need only show that j_1 is not within distance $r/2$ of e_i.

Suppose otherwise. Let z be the foot of the perpendicular from j_1 onto the line through i_1i_2, so that $\|j_1 - z\| \leq r/2$. If z does not lie between i_1 and i_2 then the minimum distance between e_i and j_1 is attained at one of the endpoints of the edge, say i_1, and thus $\|i_1 - j_1\| \leq r/2$, so that the edge i_1j_1 is in $G_{n,k}$, by Lemma 8. Now suppose z does lie between i_1 and i_2, and assume that the edge e_i is present because i_2 is one of the k nearest neighbours of i_1. Suppose that z lies within distance $r/2$ of i_2. Then

$$\|i_2 - j_1\| \leq \|i_2 - z\| + \|z - j_1\| \leq \tfrac{r}{2} + \tfrac{r}{2} = r,$$

and thus, by Lemma 8, the edge i_2j_1 is contained in G. Otherwise,

$$\|z - i_2\| > \tfrac{r}{2} \geq \|z - j_1\|,$$

and so

$$\|i_1 - j_1\| \leq \|i_1 - z\| + \|z - j_1\| = (\|i_1 - i_2\| - \|i_2 - z\|) + \|z - j_1\| < \|i_1 - i_2\|$$

so that, since i_1i_2 is an edge, so is i_1j_1. In each case j_1 is in the same component as e_i. □

Next we need a geometric lemma.

Lemma 11. *Let Λ_l be the graph of the $l \times l$ square integer grid $\{1, \ldots, l\}^2 \subset \mathbb{R}^2$ with all the unit length edges. Suppose that $A \subset V(\Lambda_l)$ with both A and $A^c = V(\Lambda_l) \setminus A$ connected in Λ_l. Let ∂A denote the set of vertices of A^c that are adjacent to vertices of A. Then the set ∂A is diagonally connected, i.e, connected if we include all edges of length $\leq \sqrt{2}$.*

Proof. Let B be the set of edges from an element of A to an element of A^c and let B' be the corresponding edges in the dual lattice. If we consider B' as a subgraph of the dual lattice then every vertex has even degree except those vertices corresponding to the boundary of Λ_l. Thus we can decompose B' into edge disjoint subgraphs each of which is either a cycle, or a path starting and ending at the boundary. Any such cycle or path splits Λ_l into two components. Since all of any connected set must lie in the same component, we see that all of A lies in the same component and all of A^c lies in the same component. This implies that the cycle or path partitions Λ_l into exactly A and A^c, and hence is all of B'. Thus ∂A is diagonally connected and the result follows. □

The following lemma asserts that there are no two large components.

Lemma 12. *Fix $c > 0$. Then, there exists a constant c' such that the probability that $G_{n,\lfloor c \log n \rfloor}$ contains two components each of (Euclidean) diameter at least $c'\sqrt{\log n}$ tends to zero as $n \to \infty$.*

Proof. Fix c' to be chosen later, and let $D = c'\sqrt{\log n}$. Let c_- be as in Lemma 8 and r satisfy $\pi r^2 = c_- \log n$. By Lemma 8 **whp** every vertex is joined to every other vertex within distance r. Thus, we may ignore all configurations for which this does not hold. Also by assumption and the definition of D there exist two components, G_1 and G_2 of $G = G_{n,\lfloor c \log n \rfloor}$, each of diameter at least D. Let G_3 be the rest of the vertices.

We tessellate the square S_n with squares of side $r/\sqrt{20}$; letting $l = \sqrt{20n}/r$, we identify the squares with the square grid $\Lambda_l = \mathbb{Z}_l^2$. (Here, and in the proof of Lemma 14, we assume for convenience that $r/\sqrt{20}$ divides $\sqrt{n}$.) We colour the squares as follows. Colour red any square containing a vertex of G_1 or intersecting an edge of G_1. Colour blue any square containing a vertex of G_2 or intersecting an edge of G_2. Colour black the remaining squares containing a vertex. All other squares we call empty and colour white. This colouring is well defined by Lemma 10. The same lemma also shows that a red square can only be adjacent to another red square or an empty square, since any two points in adjacent squares must be within distance $\sqrt{5}(r/\sqrt{20}) = r/2$. In addition, the set of red squares and the set of blue squares each forms a connected component in Λ_l.

Since G_1 and G_2 have diameter at least D, the squares have diameter $\sqrt{2}r/\sqrt{20} < r$, and the set of red squares and the set of blue squares are each connected there must be at least D/r red squares and D/r blue squares.

Let U be the set of red squares and let $V = U^c$ be the complement of U. V splits into components $V_1, V_2, \ldots, V_s$ for some $s \geq 1$. Since the blue squares are connected, at most one of these components, say V_1, can contain blue squares.

Let $U_1 = V_1^c$; i.e., U and all the components of U^c that do not contain any blue squares. Note that both U_1 and U_1^c are connected, and each contains at least D/r squares, since all

the red squares lie in U_1 and all the blue squares lie in $V_1 = U_1^c$.

Let ∂U_1 be the set of squares not in U_1, but adjacent to at least one square in U_1. Each square in ∂U_1 is empty, and the set ∂U_1 is a diagonally connected component of squares, since both U_1 and $U_1^c = V_1$ are connected.

By the vertex isoperimetric inequality in the grid [3],

$$|\partial U_1| \geq \min\{\sqrt{2|U_1|}, \sqrt{2|U_1^c|}\} \geq (D/r)^{1/2}.$$

Hence, if we have G_1, G_2 both with diameter at least D we can find a set connected in Λ_l of size $K = (D/r)^{1/2} = \sqrt[4]{\pi c'^2/c_-}$ consisting entirely of empty squares. To complete the proof we just need to show that such a set is unlikely to exist.

We use the following graph theoretic lemma. For any graph G with maximum degree Δ, the number of connected subsets of size n containing a particular vertex v_0 is at most $(e\Delta)^n$.

Define Λ_l^* as the graph with vertex set Λ_l and edges joining diagonally connected vertices. The graph Λ_l^* has maximum degree 8, so the number of connected sets of K squares in Λ_l^* containing a particular square is at most $(8e)^K$. There are $l^2 \leq n$ squares in Λ_l so the total number of connected sets of size K is at most $n(8e)^K$. Therefore the probability p that any connected set K consists entirely of empty squares satisfies

$$\begin{aligned} p &\leq n(8e)^K e^{-Kr^2/20} \\ &\leq n \exp\left(K(\log(8e) - r^2/20)\right) \\ &\leq n^{1-Kc_-/20\pi+o(1)} \end{aligned}$$

which tends to zero provided we chose c' and thus K large enough. Hence the probability that there are two components with diameter at least D tends to zero as n tends to infinity. □

Theorem 13. *If $c > \frac{1}{\log 7} \approx 0.5139$, then $\mathbb{P}(G_{n,\lfloor c \log n \rfloor}$ is connected$) \to 1$ as $n \to \infty$.*

Proof. Let $k = \lfloor c \log n \rfloor$. We shall show that for any fixed $c' > 0$ there is no component G' of $G = G_{n,k}$ with diameter less than $c'\sqrt{\log n}$ **whp**. This, together with Lemma 12, will prove the result. By Lemma 8 we may assume that the k nearest neighbours of any point all lie within distance R, where $\pi R^2 = c_+ \log n$.

Firstly let us assume such a small component G' exists and that G' contains only normal points. Consider the six tangents to the convex hull of G' which are inclined at angles 0, $\frac{\pi}{3}$, and $\frac{2\pi}{3}$ to the horizontal. These tangents form a hexagon H containing G', as shown in Figure 3, and each tangent t_i intersects G' in a point $P_i \in V(G')$ (some of the P_i may coincide). The exterior angle bisectors of H divide the exterior of H into six regions H_i, each of which is bounded by two bisectors and t_i. Consider the smallest disc D_i centered

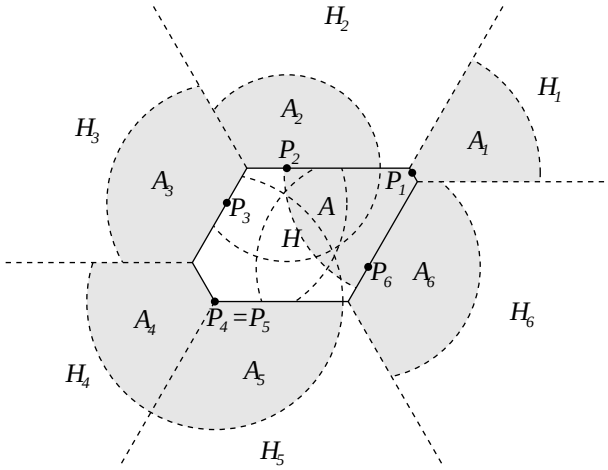

Figure 3: The hexagon H

at P_i and containing its k nearest neighbours. By assumption, all the D_i are contained in S_n. Write $A_i = H_i \cap D_i$. Without loss of generality, $|A_1| \leq |A_i|$ for all i, so that, writing $A = H \cap D_1$ and noting that $|A| \leq |A_1|$ (since A_1 does not meet the boundary of S_n), we obtain $|A| \leq \frac{1}{7}|A \cup (\cup_i A_i)|$. Now we require that there are exactly k points in the region $A \cup (\cup_i A_i)$, and that they all lie within A. The probability of this happening is at most 7^{-k}. However, the number of choices for the regions A, A_i, can be estimated as follows. There are $O(n)$ choices for the point P_1 (**whp**), and, fixing P_1, there are **whp** $O(\log n)$ choices for each $P_2, \ldots, P_6$ (since they lie within $c'\sqrt{\log n}$ of P_1), and $O((\log n)^6)$ choices for the six radii of the D_i, since they are determined by a point within distance R of P_i. Thus the number of choices for the A and A_i is $O(n(\log n)^{11})$ which is $n^{1+o(1)}$. Thus, the probability that we have a G' of diameter at most $c' \log n$ is at most $n^{1+o(1)}7^{-k}$, which is $o(1)$ for $c > \frac{1}{\log 7}$.

The above argument applies if G' is not too close to the boundary of S_n. Suppose now that G' is within distance R of the boundary, but further than R from a corner of S_n. In this case we ignore the two tangents t_i whose normal vectors point out of S_n, and define H and the relevant H_i and A_i as the intersections of the previously defined H, H_i and A_i with S_n (see Figure 4). (For the horizontal boundaries, rotate the tangents by 90 degrees.) Now, supposing that again $|A_1| \leq |A_i|$ for all i, and writing $A = H \cap D_1$ as before, we obtain $|A| \leq \frac{1}{5}|A \cup (\cup_i A_i)|$. Therefore the probability that all k points in $A \cup (\cup_i A_i)$ are in fact contained in A is at most 5^{-k}. Thus the probability of obtaining such a small component lying near the boundary is $n^{\frac{1}{2}+o(1)}5^{-k}$, which is $o(1)$ for $c > \frac{1}{\log 7} > \frac{1}{2\log 5}$. (Note

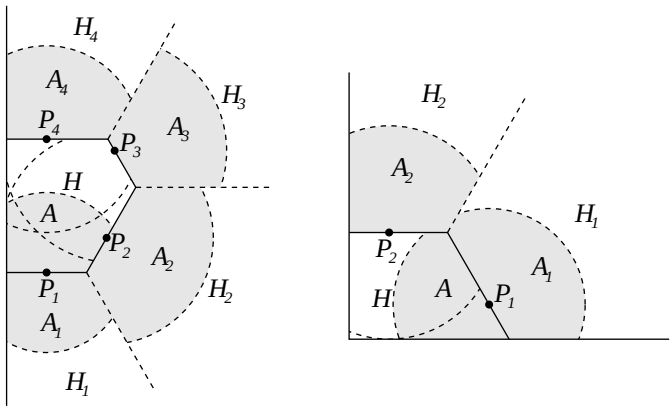

Figure 4: G' lies near an edge or corner

that there are now only $O(\sqrt{n \log n})$ choices for P_1.)

Finally, if some point of G' is within R of a corner of S_n, we now have $|A| \leq \frac{1}{3}|A \cup (\cup_i A_i)|$ (see Figure 4), and thus the probability of all k points in $A \cup (\cup_i A_i)$ lying in A is at most 3^{-k}. Here, the shape of the region H is not critical — we only need to ensure that the reflections of H in the tangents t_i are disjoint and lie within S_n. Hence the probability of obtaining a small component lying at a corner is $n^{o(1)}3^{-k} = o(1)$, there now being only $O(\log n)$ choices for P_1. □

4.1 The directed case

As in the undirected case we first show that **whp** there do not exist two large components. The proof is very similar to that of the undirected case, so we sketch the parts that are the same and concentrate on the differences. The first key difference is that in a directed graph there is no clear idea of component. We define two such notions which will satisfy our needs. A set C is a *out-component* if, for some x_0, it is of the form $\{\, y :$ there exists a directed path from x_0 to $y \,\}$. It is an *in-component* if it is of the form $\{\, y :$ there exists a directed path from y to $x_0 \,\}$. If the graph is undirected then both of these reduce to the normal definition of component. The following lemma is analogous to Lemma 12.

Lemma 14. *Fix $c > 0$ and let $k = \lfloor c \log n \rfloor$. Then there exists c' such that the probability that $\vec{G}_{n,k}$ contains an in-component and an out-component that are disjoint and both of diameter at least $c'\sqrt{\log n}$ tends to zero as $n \to \infty$.*

Proof. As before fix c' to be chosen later and let $D = c'\sqrt{\log n}$. This time, since we shall also need an upper bound on the edge length, let c_- and c_+ be as in Lemma 8 and

let r and R satisfy $\pi r^2 = c_- \log n$ and $\pi R^2 = c_+ \log n$. We may ignore all configurations which have two points at distance at most r that are not joined, or have two points at distance at least R that are joined.

Let G_1 be an out-component and G_2 an in-component, both of diameter at least D. Let G_3 be the rest of the vertices. This time edges of G_i and G_j may cross for $i \neq j$. However, it is still true that no vertex not in G_1 may lie within distance $r/2$ of an edge of G_1. Indeed the proof of Lemma 10 shows that (with notation as in that proof) in this case either $i_1\vec{j}$ or $i_2\vec{j}$ is an edge. Thus, since G_1 is an out-component, $j \in G_1$. (Note that it is important that G_1 is an out-component: it would not be true for an in-component.)

Again, we tessellate the square with squares of side $r/\sqrt{20}$; letting $l = \sqrt{20n}/r$, we identify the squares with the square grid Λ_l. We colour the squares almost exactly as before: colour the squares containing a vertex of G_1 or intersecting an edge of G_1 red, colour the squares containing a vertex of G_2 blue (note we do *not* colour the squares intersecting an edge of G_2 as that might conflict with the squares already coloured), colour the remaining squares containing a vertex black, and finally colour the empty squares white. As before, the colouring is well defined and also we see that a red square can only be adjacent to another red square or an empty square. In addition, the set of red squares forms a connected component of squares.

This time, since no point is joined to another at distance greater than R, there must be at least D/R red squares, and at least D/R blue squares.

Let U be the set of red squares and let $V = U^c$ be the complement of U. V splits into components $V_1, V_2, \ldots, V_s$ for some $s \geq 1$. This time the blue squares need not be connected and so need not all be in the same set V_i. Suppose that the components that contain blue squares are $V_1, V_2, \ldots, V_t$.

Let $U_1 = U \cup \bigcup_{i=t+1}^{s} V_i$; i.e., U and all the components of U^c that do not contain any blue squares. U_1 and U_1^c each contain at least D/R squares, since all the red squares lie in U_1 and all the blue squares lie in U_1^c.

Let ∂U_1 be the set of squares not in U_1, but adjacent to at least one square in U_1. Each square in ∂U_1 lies in ∂U, so is empty. The set ∂U_1 is not necessarily a connected component of squares in Λ_l, however, we show that, for some d, it is connected in $\Lambda_{l,d}$, the dth power of the lattice Λ_l, where we join vertices if their distance in the lattice (i.e., their l_1 distance) is at most d.

Let $d = 2\lceil\sqrt{20}R/r\rceil$. Then the blue squares are joined in $\Lambda_{l,d}$. Suppose that ∂U_1 is not connected in $\Lambda_{l,d}$; i.e., we can partition ∂U_1 into two non-empty sets A and B with no square in A within d of any square in B. For $i \leq t$ write ∂V_i for $\partial U_1 \cap V_i$. Since V_i and V_i^c are both connected in Λ_l, ∂V_i is connected in $\Lambda_{l,2}$, and hence A and B are both the union of such ∂V_i. Every V_i with $i \leq t$ contains a blue square so there must be a pair $i, j \leq t$ with $\partial V_i \subseteq A$, $\partial V_j \subseteq B$ and blue squares b_i, b_j with $b_i \in V_i$, $b_j \in V_j$ and l_1 distance

$d(b_i, b_j) \leq d$. The shortest path from b_i to b_j in Λ_l passes through ∂V_i and ∂V_j and has length at most d, so $d(\partial V_i, \partial V_j) < d$, contradicting the assumption that ∂V_i and ∂V_j were in different components in $\Lambda_{l,d}$.

As before, by the vertex isoperimetric inequality in the grid [3],

$$|\partial U_1| \geq \min\{\sqrt{2|U_1|}, \sqrt{2|U_1^c|}\} \geq (D/R)^{1/2}.$$

Hence, if we have G_1, G_2 both with diameter at least D, we can find a set connected in $\Lambda_{l,d}$ of size $K = (D/R)^{1/2} = \sqrt[4]{\pi c'^2/c_+}$ consisting entirely of empty squares. Once again we show that it is unlikely that such a set exists.

$\Lambda_{l,d}$ has maximum degree $2d^2 + 2d$. Thus, applying the lemma stated in the undirected case, the number of connected sets of K squares in $\Lambda_{l,d}$ containing a particular square is at most $(e(2d^2 + 2d))^k \leq (4ed^2)^k$. Since there are $l^2 \leq n$ squares in Λ_l, the probability p that there exists a set connected in $\Lambda_{l,d}$ of empty squares satisfies

$$\begin{aligned} p &\leq n(4ed^2)^K e^{-Kr^2/20} \\ &\leq n \exp\left(K(\log(4ed^2) - r^2/20)\right) \\ &\leq n^{1-Kc_-/20\pi + o(1)} \end{aligned}$$

which, again, tends to zero provided we chose c' and thus K large enough. Hence the probability that we have an in-component and an out-component each of size at least D tends to zero. □

Theorem 15. *If $c \geq 0.9967$ then $\mathbb{P}(\vec{G}_{n,\lfloor c \log n \rfloor}$ is connected$) \to 1$ as $n \to \infty$.*

Proof. Suppose that $k = \lfloor c \log n \rfloor$ and $\vec{G} = \vec{G}_{n,k}$ is not connected. Then there will be two points $x, y \in V(\vec{G})$ such that there is no directed path from x to y. We consider two subsets of $V(\vec{G})$, C_x and C_y, defined as follows:

$$C_x = \{x\} \cup \{\, x' : \text{ there is a directed path from } x \text{ to } x' \,\},$$

and

$$C_y = \{y\} \cup \{\, y' : \text{ there is a directed path from } y' \text{ to } y \,\}.$$

C_x and C_y are disjoint, since if we had $z \in C_x \cap C_y$, there would be a directed path from x to z and another directed path from z to y, giving us a directed path from x to y.

Lemma 14 shows that there exists a $c' > 0$ such that the probability that both C_x and C_y have diameter more than $c'\sqrt{\log n}$ tends to zero. The proof of Theorem 13 shows that the probability that an out-component C_x exists with diameter less than $c'\sqrt{\log n}$ tends to

zero since $c > \frac{1}{\log 7}$. We complete the proof by showing that for all $c' > 0$, the probability that an in-component C_y exists with diameter less than $c'\sqrt{\log n}$ also tends to zero.

We first illustrate the proof with a simpler proof that $c \geq 1.0293 > \frac{1}{\log \gamma}$ is sufficient, where $\gamma = (\frac{4\pi}{3} + \frac{\sqrt{3}}{2})/(\frac{\pi}{3} + \frac{\sqrt{3}}{2})$.

Suppose first that no point of C_y lies within a distance R of the boundary of S_n, where R is as in Lemma 8. Let $z \notin C_y$ be the closest point of $V(\vec{G}) \setminus C_y$ to C_y and y_z its nearest neighbour in C_y. Write $\rho = \|z - y_z\|$ for the distance between them, and, for an arbitrary point P, write $D_\rho(P)$ for the open disc of radius ρ, centered at P. Consider the leftmost point y_l and the rightmost point y_r of C_y. There can be no points in $B = D^l_\rho(y_l) \cup D^r_\rho(y_r)$, the left half of $D_\rho(y_l)$ or the right half of $D_\rho(y_r)$. By the proof of Lemma 8, we may assume $D^l_R(y_l)$ contains at least k points. Hence $\rho < R$, B is contained within S_n, and $|B| = |D_\rho(x)| = \pi\rho^2$. On the other hand, there are at least k points in $A = D_\rho(z) \setminus D_\rho(y_z)$, since otherwise z would send a directed edge to either y_z, or to a point $y' \in D_\rho(z) \cap D_\rho(y_z)$. The first possibility contradicts the hypothesis $z \notin C_y$, and for the second possibility, we must have $y' \notin C_y$ to ensure $z \notin C_y$, but then $y' \notin C_y$ is closer to C_y than is z, contradicting the choice of z. Therefore, as shown in Figure 5, there must be at least k points in $A \cup B$, which must all lie in $A \setminus B$. The probability of this happening is at most $\left(\frac{|A\setminus B|}{|A\cup B|}\right)^k \leq \left(\frac{|A|}{|A|+|B|}\right)^k = \gamma^{-k}$. The number of choices for z, y_z, y_l, and y_r is $O(n(\log n)^3)$, so the probability such a configuration occurs anywhere is at most $n^{1+o(1)}\gamma^{-k}$, which is $o(1)$ for $c > \frac{1}{\log \gamma}$.

If some point of C_y is close to an edge or corner of S_n we use a single half disc or quarter disc for B, and a similar argument to the one used to complete the proof of Theorem 13 shows that the probability of obtaining a small C_y near the boundary is also $o(1)$.

With a little more work, we can obtain a slight improvement by showing there is a region $C \subseteq A$ containing no points in its interior.

Suppose that $w \in D_\rho(z)$. Write $\rho' = \|w - y_z\|$ and set

$$\begin{aligned} A_1 &= (A \setminus D_{\rho'}(w)) \setminus B, \\ A_2 &= (A \cap D_{\rho'}(w)) \setminus B, \\ A_3 &= (D_{\rho'}(w) \setminus (D_\rho(z) \cup D_\rho(y_z))) \setminus B, \\ A_4 &= B \end{aligned}$$

as illustrated in Figure 5 (for simplicity, the set B is not shown). Writing n_i for the number of points (other than y_z, z, or w) in regions A_i, we see that the following must hold:

$$n_1 + n_2 \geq k - 1, \qquad n_3 + n_2 \geq k - 1, \qquad n_4 = 0. \tag{5}$$

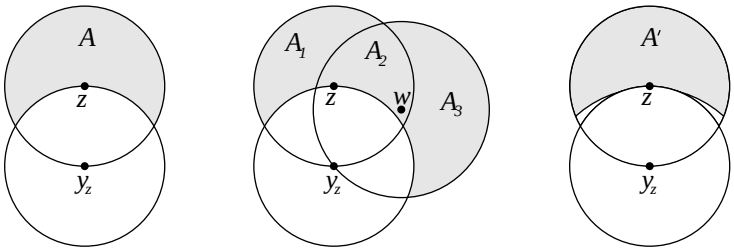

Figure 5: Upper bound, directed case (B not shown)

We need to show that for some w, the probability p of such an arrangement is small. By Lemma 4, we have

$$\log p = \sum_i \left(n_i - |A_i| - n_i \log \tfrac{n_i}{|A_i|} \right) + O(\log \textstyle\sum n_i). \tag{6}$$

We now maximize the right hand side of (6). Since (5) becomes more likely if $|A_1|$, $|A_2|$, or $|A_3|$ is increased, we may assume B is disjoint from $A \cup D_{\rho'}(w)$. Also, as we shall only be interested in ratios of areas, we first maximize (6) under uniform scaling of areas, giving

$$n_1 + n_2 + n_3 = |A_1| + |A_2| + |A_3| + |A_4|.$$

Now vary the n_i subject to $n_1 + n_2$ and $n_3 + n_2$ being fixed. This gives

$$\eta = \frac{n_2}{|A_2|} = \frac{n_1}{|A_1|} \frac{n_3}{|A_3|}.$$

Also, by varying just n_1, we see that either $n_1 + n_2 = k - 1$ or $n_1 = |A_1|$. Similarly, either $n_3 + n_2 = k - 1$ or $n_3 = |A_3|$. Hence

$$\begin{aligned}
\log p &= \sum -n_i \log \tfrac{n_i}{|A_i|} + O(\log \textstyle\sum n_i) \\
&= -n_1 \log \tfrac{n_1}{|A_1|} - n_3 \log \tfrac{n_3}{|A_3|} - n_2 \log(\tfrac{n_1}{|A_1|} \tfrac{n_3}{|A_3|}) + O(\log \textstyle\sum n_i) \\
&= -(n_1 + n_2) \log \tfrac{n_1}{|A_1|} - (n_3 + n_2) \log \tfrac{n_3}{|A_3|} + O(\log \textstyle\sum n_i) \\
&= -(k-1) \log(\tfrac{n_1}{|A_1|} \tfrac{n_3}{|A_3|}) + O(\log \textstyle\sum n_i).
\end{aligned}$$

Therefore,

$$p = \eta^{-(k-1)} n^{o(1)}.$$

Define γ' by $(\log \gamma')^{-1} = 0.9967$ and let C be the set of points $w \in A$ such that

$$\sum_i |A_i| > \gamma' |A_2| + \sqrt{4\gamma' |A_1||A_3|} \qquad \text{and} \qquad |A_3| < 2|A_1|.$$

We shall show that with the above constraints

$$\eta = \tfrac{n_2}{|A_2|} = \tfrac{n_1\ n_3}{|A_1||A_3|} > \gamma'.$$

If $n_3 + n_2 > k - 1 = n_1 + n_2$, then $n_3 = |A_3|$ and so $2|A_1| > |A_3| = n_3 > n_1 = \eta |A_1|$. But then $\eta < 2$ and $|A_1| + |A_2| + |A_4| = n_1 + n_2 < 2(|A_1| + |A_2|)$, contradicting the fact that $|A_1| + |A_2| < |A_4|$. On the other hand, if $n_1 + n_2 > k - 1 = n_3 + n_2$ then $|A_1| = n_1 > n_3 = \eta |A_3|$. But $|A_3| \geq |A_1|$, so $\eta \leq 1$. But then $n_1 + n_2 + n_3 \leq |A_1| + |A_2| + |A_3|$ and so $|A_4| \leq 0$, a contradiction. Similarly, if $n_1 + n_2 > k - 1$ and $n_3 + n_2 > k - 1$ then $\eta = 1$ and $|A_4| \leq 0$ again. Hence we may assume $n_1 + n_2 = n_2 + n_3 = k - 1$, $n_1 = n_3$ and so $\sum_i |A_i| = n_2 + (n_1 + n_3) = n_2 + \sqrt{4n_1 n_3} = \eta |A_2| + \sqrt{4\eta |A_1||A_3|}$. But this then implies $\eta > \gamma'$ as required.

Computer calculations show that

$$\tfrac{|B| + |A \setminus C|}{|A \setminus C|} > \gamma'.$$

Now suppose that the region C contains no points in its interior. Then we have at least k points in the region $(A \setminus C) \cup B$, all of which are constrained to lie in $A' = A \setminus (C \cup B)$ (see Figure 5). This event has probability at most $\gamma'^{-k} n^{o(1)} = o(n^{-1})$. On the other hand, the probability that a configuration exists with a point $w \in C$ is also at most $\gamma'^{-k} n^{o(1)} = o(n^{-1})$. Therefore, **whp** $\vec{G}$ is connected. □

5 Sharp threshold

Theorems 5 and 13 show that if $n = n(k) \leq e^{k/0.5139}$ then $\lim_{k\to\infty} \mathbb{P}(G_{n,k} \text{ is connected}) = 1$ and if $n = n(k) \geq e^{k/0.3043}$ then $\lim_{k\to\infty} \mathbb{P}(G_{n,k} \text{ is connected}) = 0$. There is no doubt that there is a constant c, $1/0.5139 < c < 1/0.3043$, such that if $\varepsilon > 0$ then for $n = n(k) \leq e^{(c-\varepsilon)k}$ we have $\lim_{k\to\infty} \mathbb{P}(G_{n,k} \text{ is connected}) = 1$ and for $n = n(k) \geq e^{(c+\varepsilon)k}$ we have $\lim_{k\to\infty} \mathbb{P}(G_{n,k} \text{ is connected}) = 0$. Although we cannot show the existence of this constant c, let alone determine it, in this brief section we shall show that the transition from connectedness to disconnectedness is considerably sharper than these relations indicate: the length of the window is $O(n)$ rather than $n^{1+o(1)}$. To formulate this result, for $k \geq 1$ and $0 < p < 1$, set

$$n_k(p) = \max\{\, n : \mathbb{P}(G_{n,k} \text{ is connected}) \geq p \,\}.$$

Theorem 16. *Let $0 < \varepsilon < 1$ be fixed. Then, for sufficiently large k,*

$$n_k(\varepsilon) < C(\varepsilon)(n_k(1-\varepsilon)+1)$$

where

$$C(\varepsilon) = \left\lceil \tfrac{6}{\varepsilon} \log\left(\tfrac{1}{\varepsilon}\right) + 1 \right\rceil^2.$$

Proof. Write $M = \left\lceil \frac{6}{\varepsilon} \log\left(\frac{1}{\varepsilon}\right) + 1 \right\rceil$ and $N = n_k(1-\varepsilon)+1$, so that the probability that we have at least two components in $G_{N,k}$ is at least ε. By Theorems 5 and 13, we may assume, by taking k sufficiently large, that $0.3043 \log N < k < 0.5139 \log N$. Therefore, by Lemma 8, we see that **whp** no edge in $G_{N,k}$ has length greater than $R = \sqrt{c_+(\log N)/\pi}$.

We say that a point $x \in V(G_{N,k})$ is *close* to a side s of S_N if x is less than distance $2R$ from s, and call a component G' of $G_{N,k}$ close to s if it contains points which are close to s. Further, we say that $x \in V(G_{N,k})$ is *central* if it is not close to any side s of S_N, and call a component G' of $G_{N,k}$ central if it consists entirely of central points. Finally, we call a component G' of $G_{N,k}$ *small* if it has diameter at most $c'\sqrt{\log N}$, where c' is as in Lemma 12.

By Lemma 12, with probability more than $\frac{\varepsilon}{2}$, $G_{N,k}$ contains a small component, which can be close to at most two sides of S_N. Write α for the probability that we have a small central component of $G_{N,k}$. Write β for the probability that we have a small component of $G_{N,k}$ which is close to exactly one side of S_N, and γ for the probability that we have a component of $G_{N,k}$ close to two sides of S_N (so that it lies at a corner of S_N). We have $\alpha + \beta + \gamma > \frac{\varepsilon}{2}$, and the proof of Theorem 13 shows that

$$\gamma = n^{o(1)} 3^{-k} \to 0$$

as $k \to \infty$. Therefore we may assume that at least one of α and β is greater than $\frac{\varepsilon}{6}$ (we do not know which one). If we specify one side s of S_N, the probability that we obtain a small component G' which may only be close to s is thus at least $\frac{\varepsilon}{24}$.

Now we consider the larger square S_{M^2N}, and tessellate it with copies of S_N. We only consider the small squares of the tessellation incident with the boundary of S_{M^2N}. Considering sides of these copies of S_N lying on the boundary of S_{M^2N}, we see that we have $4(M-1)$ independent opportunities to obtain a small component G' in one of the small squares S, in such a way that G' can only intersect the boundary of S on the boundary of S_{M^2N}. Such a component will also be isolated in $G_{M^2N,k}$, since **whp** no edge of $G_{M^2N,k}$ has length greater than $\sqrt{c_+(\log M^2 N)/\pi} < 2R$ for sufficiently large k (and thus N). Therefore, if p is the probability that $G_{M^2N,k}$ is connected, we have

$$p < \left(1 - \tfrac{\varepsilon}{24}\right)^{4(M-1)} < e^{-\frac{\varepsilon}{6}(M-1)} < \varepsilon,$$

completing the proof. □

6 Coverage

Let $\mathcal{P}_n$ be a Poisson process of intensity one in the square S_n. For any $x \in \mathcal{P}_n$, let $r(x,k)$ be the distance from x to its kth nearest neighbour (infinite if this does not exist), and let $B_k(x) = D_{r(x,k)}(x) \cap S_n$. Let $\mathcal{C}_k(\mathcal{P}_n) = \bigcup_{x\in\mathcal{P}} B_k(x)$. We say that $\mathcal{P}_n$ is a *k-cover* if $\mathcal{C}_k(\mathcal{P}_n) = S_n$.

First we prove a quick lemma bounding the Poisson distribution.

Lemma 17. *Suppose that $\mathcal{P}$ is a Poisson process of intensity one in the square S_n and fix c and $\varepsilon > 0$. Then there exists $\delta > 0$ such that,* **whp**, *there does not exist a point x of the process with*

$$r(x, \lfloor c\log n\rfloor) - r(x, \lfloor (c-\varepsilon)\log n\rfloor) < \delta\sqrt{\log n}. \tag{7}$$

Proof. Let $k = \lfloor c\log n\rfloor$ and $k' = \lfloor (c-\varepsilon)\log n\rfloor$. By Lemma 8 we may assume that no edge in $G_{n,k}$ is longer than $R = c_m\sqrt{\log n}$, where $c_m = \sqrt{c_+/\pi}$ in the notation of Lemma 8. For a fixed point x, condition (7) only holds if the annulus of width $\delta\sqrt{\log n}$ and outer diameter $r(x,k)$ contains at least $\lfloor \varepsilon\log n\rfloor - 1$ points. This annulus, A, say, has area at most $2\pi R\delta\sqrt{\log n} = 2\pi\delta c_m \log n$.

The number of points in A is stochastically dominated by a Poisson distribution with mean $2\pi\delta c_m\log n$. Thus the probability p that there are more than $\lfloor\varepsilon\log n\rfloor - 1$ points in A satisfies

$$\log p \le -2\pi\delta c_m\log n - \varepsilon\log n\log\left(\frac{\varepsilon}{e2\pi\delta c_m}\right) + O(\log\log n)$$

which is less than $-\log n$ provide we choose δ small enough. Hence the probability that any point fails (7) is $o(1)$. □

Theorem 18. *Fix $c > c' > 0$.*

If **whp** $\vec{G}_{n,\lfloor c'\log n\rfloor}$ *does not have a vertex of in-degree zero. Then* **whp** $\mathcal{P}_n$ *is a $\lfloor c\log n\rfloor$-cover.*

Conversely, suppose that **whp** $\mathcal{P}_n$ *is a $\lfloor c'\log n\rfloor$-cover. Then* **whp** $\vec{G}_{n,\lfloor c\log n\rfloor}$ *does not have a vertex of in-degree zero.*

Consequently, if $c \le 0.7209$ then **whp** $\mathcal{P}_n$ *is not a $\lfloor c\log n\rfloor$-cover, while if $c \ge 0.9967$,* **whp** $\mathcal{P}_n$ *is a $\lfloor c\log n\rfloor$-cover.*

Proof. Let $k = \lfloor c\log n\rfloor$ and $k' = \lfloor c'\log n\rfloor$. Suppose that it is not true that, **whp**, $\mathcal{P}_n$ is a k-cover. Then there exists $\varepsilon > 0$, such that, for infinitely many n, the probability that $\mathcal{P}_n$ is not a k-cover is at least ε. Let $n' = n(1 + 1/\log n)$. We show that

$$\mathbb{P}(\vec{G}_{n',k'} \text{ has a vertex of in-degree zero}) > \varepsilon'$$

for some $\varepsilon' > 0$.

By Lemma 17, there exists $\delta > 0$ such that, **whp**, $r(x,k) - r(x,k') \geq \delta\sqrt{\log n}$ for every $x \in \mathcal{P}_n$. Thus,

$$\begin{aligned}\mathbb{P}(S_n \setminus \mathcal{C}_{k'}(\mathcal{P}_n) \text{ contains a ball of radius } \delta\sqrt{\log n}) &\geq (1-o(1))\mathbb{P}(\mathcal{P}_n \text{ is not a } k\text{-cover}) \\ &\geq (1-o(1))\varepsilon.\end{aligned}$$

We identify $\mathcal{P}_{n'}$ with $\mathcal{P}_n \cup \mathcal{P}_{n/\log n}$ where all squares are scaled to be the same size as S_n. Let $R = \sqrt{c_+(\log n)/\pi} = c_m\sqrt{\log n}$ be as in Lemma 8. Fix $\mathcal{P}_n$ such that $\vec{G}_{n,k'}$ has no edge of length more than R, and that $\mathcal{C}_{k'}(\mathcal{P}_n)^c$ contains a disc of radius $\delta\sqrt{\log n}$, and let y be the centre of such a disc. The probability that the disc $D_{\delta\sqrt{\log n}}(y)$ contains exactly one point of $\mathcal{P}_{n/\log n}$ is a constant independent of n, as is the probability that the disc $D_{(c_m+\delta)\sqrt{\log n}}(y)$ contains no other point of $\mathcal{P}_{n/\log n}$. Hence there exists $\varepsilon_1 > 0$ such that

$$\mathbb{P}(\vec{G}_{n',k'} \text{ has a vertex of in-degree zero} \mid \mathcal{P}_n) \geq \varepsilon_1,$$

since this event occurs provided both the previous events occur. Combining these, we see that

$$\mathbb{P}(\vec{G}_{n',k'}) \text{ has a vertex of in-degree zero}) \geq (1-o(1))\varepsilon\varepsilon_1.$$

as claimed.

Conversely, suppose that it is not true that, **whp**, $\vec{G}_{n,k}$ does not have a vertex of in-degree zero. As before, this implies that there exists $\varepsilon > 0$ such that, for infinitely many n, the probability $\vec{G}_{n,k}$ has a vertex of in-degree zero is at least ε.

Let R be as in Lemma 8. Fix a configuration $\mathcal{P}_n$ with a point y of zero in-degree, no edge length longer than R, and no vertex with more than $c_1 \log n$ points within distance $2R$. The first condition occurs with probability at least ε, the second condition fails with probability tending to zero, as does the final condition provided that c_1 is large enough. (For the last assertion, set $c_0 = 4c_+/c_-$ and apply Lemma 8 with n replaced with n^{c_0}. Then no vertex of $S_{n^{c_0}} \cap \mathcal{P} \supset S_n \cap \mathcal{P}$ has more than $\lfloor c \log n^{c_0} \rfloor \leq cc_0 \log n$ points within a disc of area $c_- \log n^{c_0} = \pi(2R)^2$.) Fix $\delta > 0$ and let $n' = (1-\delta)n$. Similarly to before we identify $\mathcal{P}_n$ with $\mathcal{P}_{n'} \cup \mathcal{P}_{\delta n}$ (both scaled to the same size S_n) by independently assigning each vertex of $\mathcal{P}_n$ to $\mathcal{P}_{\delta n}$ with probability δ. Then

$$\mathbb{P}(\mathcal{P}_{n'} \text{ is not a } k' \text{ cover} \mid \mathcal{P}_n) \geq \varepsilon'$$

since this event occurs if the point y is in $\mathcal{P}_{\delta n}$ and no disc of radius R containing y contains more than $k - k' \geq (c - c')\log n - 1$ points of $\mathcal{P}_{\delta n}$. The number of points in $D_{2R}(y)$ is at most $c_1 \log n$, so the number of points in $D_{2R}(y) \cap \mathcal{P}_{\delta n}$ is stochastically dominated by the distribution $\mathrm{Bin}(\lfloor c_1 \log n \rfloor, \delta)$. Thus, with probability at least $1/2$, $D_{2R}(y)$ contains

at most $c_1\delta \log n$ points of $\mathcal{P}_{\delta n}$. Hence, provided that $c - c' > c_1\delta$, the latter condition is satisfied with probability at least one half for large enough n. The former condition, is independent of the latter, and occurs with probability δ. Combining these, we see that

$$\mathbb{P}(\mathcal{P}_{n'} \text{ is not a } k' \text{ cover}) \geq (1 - o(1))\delta\varepsilon'/2.$$

□

7 Numerical results

Computer simulations suggest that for $k \geq 3$ there exists a giant component in $G_{n,k}$ which contains almost all of the vertices (over 98.5% for $k = 3$) with a few isolated small components. On the other hand, for $k \leq 2$ all components are small. As we are interested mainly in large k we have confined our numerical results to $k \geq 3$, since these are more likely to reflect the situation when k is large.

For $k \geq 3$ the small components are relatively few and far between (more so for larger k). As a result one would expect that for a large rectangular region A, the small components would be roughly Poisson distributed with constant density throughout the area A, with perhaps a somewhat different density near the sides and corners of A. Hence we would expect the average number of small components in A to be approximately Poisson distributed with mean $\alpha_k|A| + \beta_k|\partial A| + 4\gamma_k$, where α_k represents the density of components far from the boundary of A, β_k gives a correction for "edge effects", and γ_k gives a correction for "corner effects". By considering rectangles with various sizes and aspect ratios, one can investigate numerically the constants α_k, β_k and γ_k. Computer simulations were performed on large rectangular regions for $3 \leq k \leq 8$ and the number and sizes of the small components were recorded. The numbers of components found were fitted by the linear formula $\alpha_k|A| + \beta_k|\partial A| + 4\gamma_k$ and for all k considered this did indeed fit the data extremely well. In total an area of over 10^{12} was simulated for each k from 3 to 8. Estimates of α_k, β_k and γ_k are given in Table 1.

The values of β_k and γ_k were positive, indicating that small components are more common near the boundary and corners of A. Figure 6 plots the probability that $G_{n,k}$ is connected and the average number of components against n for $3 \leq k \leq 8$. The predictions based on the number of components being distributed as $1 + \mathrm{Po}(\alpha_k n + 4\beta_k\sqrt{n} + 4\gamma_k)$ are also given and are in excellent agreement for large n. We know from Theorem 13 that $\gamma_k \to 0$, however it also appears that $\beta_k << \sqrt{\alpha_k}$. Hence, if A is the square S_n, when n is large enough so that the k nearest neighbour model has a reasonable chance of being disconnected, the expected number of components is dominated by the term $\alpha_k n$. One would therefore expect that the probability that the model is connected to be approximated

k	$-\log \alpha_k$	$-\log \beta_k$	$-\log \gamma_k$	$\mathbb{E}\lvert C\rvert$
3	6.2259 [1]	4.9876 [3]	2.8685 [13]	7.1031 [2]
4	9.1828 [1]	7.1871 [6]	4.6905 [22]	6.7519 [3]
5	12.0917 [4]	9.3145 [13]	6.2918 [33]	7.3551 [9]
6	15.0052 [17]	11.4542 [31]	7.8476 [53]	8.1728 [30]
7	17.9340 [71]	13.6015 [79]	9.4211 [93]	9.0659 [116]
8	20.8979 [310]	15.7770 [221]	11.0057 [179]	10.0022 [425]

Table 1: Best fit data for α_k, β_k, γ_k, and the average size of small components. Numbers in [ ] indicate 1 standard deviation error in last digit.

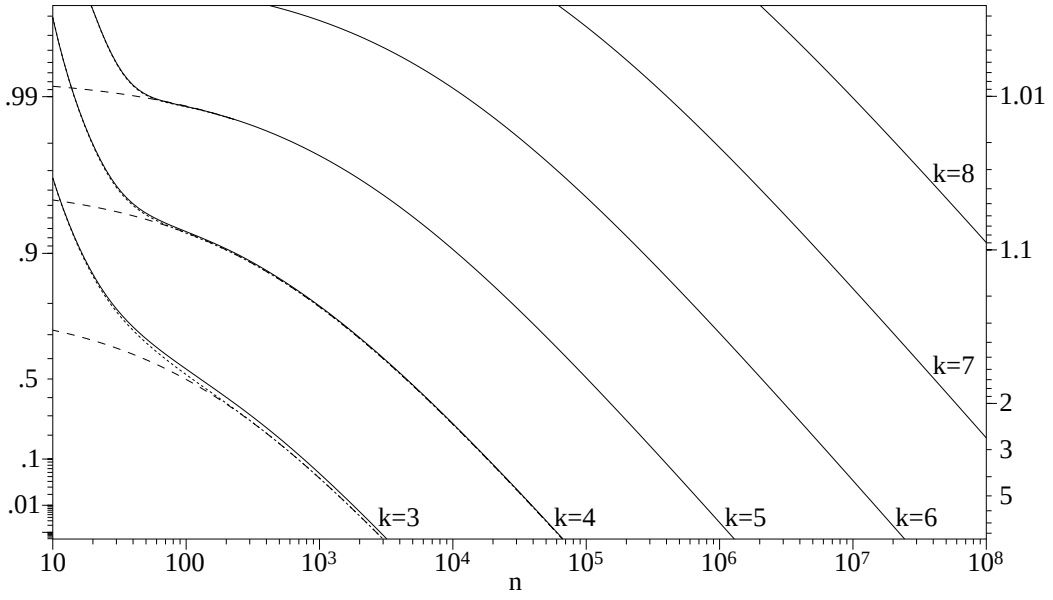

Figure 6: Probability that $G_{n,k}$ is connected (solid line, left scale), average number of components (dotted line, right scale), and theoretical predictions based on number of components being given by $1 + \mathrm{Po}(\alpha_k n + 4\beta_k\sqrt{n} + 4\gamma_k)$ (dashed line, either scale). Note that lines are indistinguishable for $k > 5$. The left hand scale is exponentially related to the right hand scale.

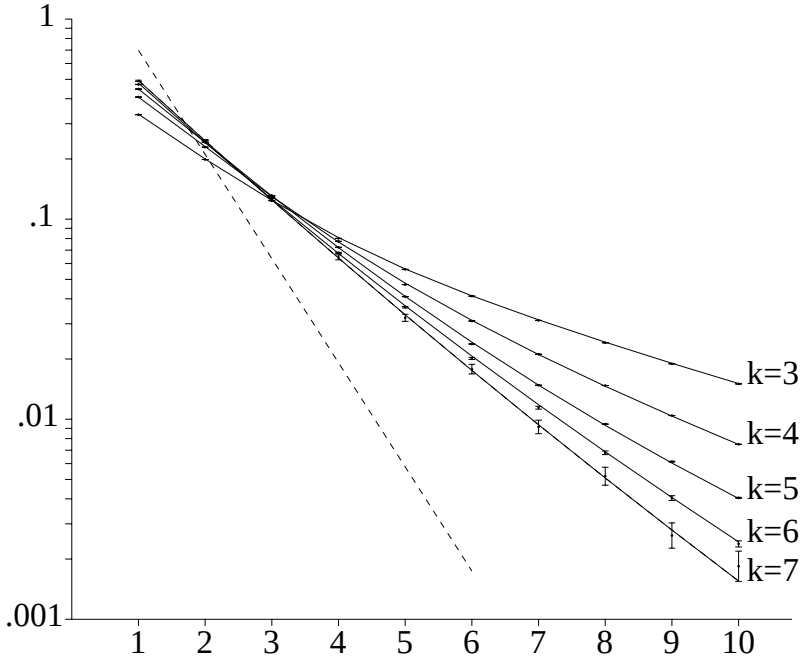

Figure 7: Proportion of small components that are of size $k + x$. The dotted line is the theoretical prediction for large k based on the lower bound argument. Error bars represent 1 standard deviation.

very well by $\exp\{-\alpha_k n\}$, and to be fairly insensitive to the shape of the region S_n, provided the boundary is reasonably smooth and not excessively long. One would also expect that for fixed n the critical value of k occurs when $\alpha_k \sim 1/n$. The data suggests that this critical k is between about $0.3 \log n$ and $0.4 \log n$, consistent with the theoretical bounds, and closer to the lower bound.

If one believes that the lower bound construction of Theorem 5 is in fact asymptotically correct, then the sizes of the components in the interior should be geometrically distributed with minimum value $k + 1$ and ratio about $e^{-\mu} \approx 0.3016$, where μ is the constant found in the proof of Theorem 5. Of course, this assumes that k is very large. For more modest values of k, the lower bound construction suggests that the density of components of size $t \geq k+1$ should be about $\exp\{-\eta_k \sqrt{t}\}$ for some constant η_k. To see this, consider a disc of area t with t points in it and insist that a vertex-free annulus of constant width surrounds it. If this width is large enough, the t points inside the disc should form a component, and the vertex-free region is of area $O(\sqrt{t})$, so this configuration has probability about

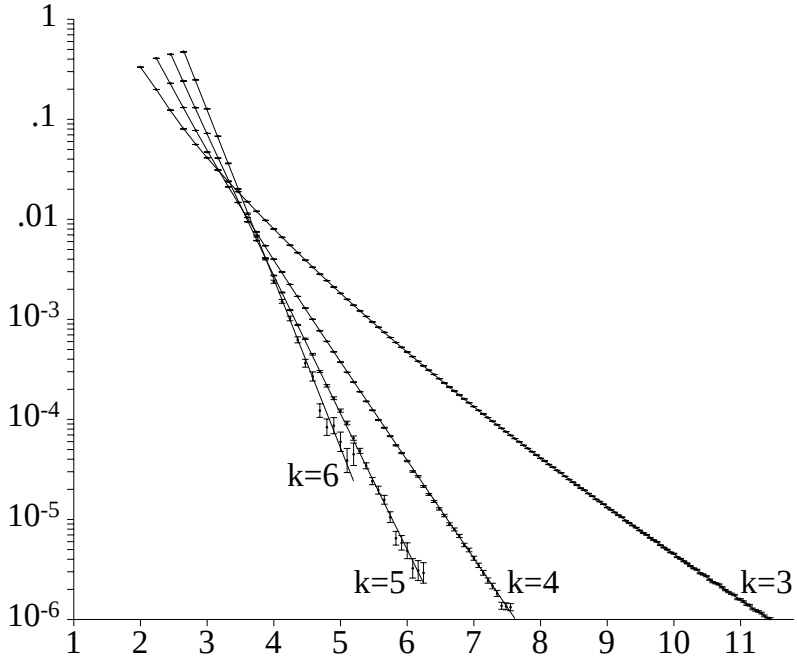

Figure 8: Proportion of small components that are of size t versus $\sqrt{t}$ for $3 \leq k \leq 6$. Error bars represent 1 standard deviation.

	$k=3$	$k=4$	$k=5$	$k=6$	$k=7$	$k=8$
n_C	2,174,360,691	113,019,084	6,163,109	334,633	17,923	924
$\max\|C\|$	547	106	65	37	27	20

Table 2: Number and maximum size of small components in simulation results in area of size $2^{40} \approx 10^{12}$.

$\exp\{-\eta_k\sqrt{t}\}$. The component size distribution for components near the edge of A is different than for components near the centre of A, so we only considered components far from the boundary of A. (Numerical evidence suggests that the components near the boundary are on average slightly larger than components far from the boundary.) Table 2 gives the total number of components found in our simulations and the maximum size of a small component. Figures 6 and 7 plot the proportion of small components found against their size, first using a linear scale in component size and second versus $\sqrt{t}$. For $k \geq 4$, the plot against $\sqrt{t}$ does indeed appear to be close to linear, however for $k = 3$ there does seem to be some deviation from linearity. The average small component sizes for components far from the boundary are given in Table 1.

8 Conjectures

We end with three extremely natural conjectures we would very much like to see solved. The first was mentioned briefly in the introduction.

Conjecture 1. *Is there a critical value of c such that, for $c' < c$, $G_{n,\lfloor c' \log n \rfloor}$ is disconnected* **whp***, and, for $c'' > c$, $G_{n,\lfloor c'' \log n \rfloor}$ is connected* **whp***? In the terminology introduced in the introduction, is it true that $c_l = c_u$? Is it true for the directed graphs $\vec{G}_{n,k}$?*

Conjecture 2. *For the directed graphs $\vec{G}_{n,k}$, write*

$$\vec{c}_l = \sup\{c : \mathbb{P}(\vec{G}_{n,\lfloor c \log n \rfloor} \text{ is connected}) \to 0\}, \text{ and}$$
$$\vec{c}_{iso} = \sup\{c : \mathbb{P}(\vec{G}_{n,\lfloor c \log n \rfloor} \text{ contains a vertex with zero in-degree}) \to 1\}.$$

Trivially, we have $\vec{c}_l \geq \vec{c}_{iso}$. Is it in fact true that $\vec{c}_l = \vec{c}_{iso}$?

Conjecture 3. *Is the threshold for connectivity of $G_{n,k}$ sharp in k? In other words, setting*

$$k_n(p) = \min\{\, k : \mathbb{P}(G_{n,k} \text{ is connected}) \geq p \,\},$$

is it true that, for any $0 < \varepsilon < 1$*, there exists* $C(\varepsilon)$ *such that, for all sufficiently large* n*,*

$$k_n(1-\varepsilon) < C(\varepsilon) + k_n(\varepsilon)?$$

"Sharpness in n" was proved in Section 5, but perhaps this is more natural.

9 Acknowledgements

We would like to thank Michael Lemmon and Martin Haenggi for drawing this problem to our attention.

References

[1] P. Balister, B. Bollobás and M. Walters, *Continuum percolation with steps in the square or the disc*, to appear in Random Structures and Algorithms.

[2] B. Bollobás, Random Graphs, second edition, Cambridge University Press, 2001.

[3] B. Bollobás and I. Leader, *Edge-isoperimetric inequalities in the grid*, Combinatorica **11** (1991), 299–314.

[4] E.N. Gilbert, *Random plane networks*, Journal of the Society for Industrial Applied Mathematics **9** (1961), 533–543.

[5] J.M. Gonzáles-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), 23–34.

[6] B. Hajek, *Adaptive transmission strategies and routing in mobile radio networks*, Proceedings of the Conference on Information Sciences and Systems (1983), 373–378.

[7] T. Hou and V. Li, *Transmission range control in multihop packet radio networks*, IEEE Transactions on Communications **COM-34** (1986), 38–44.

[8] L. Kleinrock and J.A. Silvester, *Optimum transmission radii for packet radio networks or why six is a magic number*, IEEE Nat. Telecommun. Conf., December 1978, 4.3.1–4.3.5.

[9] R. Mathar and J. Mattfeldt, *Analyzing routing strategy NFP in multihop packet radio network on a line*, IEEE Transactions on Communications **43** (1995), 977–988.

[10] R.D. Maudlin (Ed.), The Scottish Book, Birkhäuser Verlag, Boston, Basel, Stuttgart, 1979.

[11] R. Meester and R. Roy, Continuum Percolation, Cambridge University Press, 1996.

[12] J. Ni and S. Chandler, *Connectivity properties of a random radio network*, Proceedings of the IEE – Communications **141** (1994), 289–296.

[13] M.D. Penrose, *The longest edge of the random minimal spanning tree*, Annals of Applied Probability **7** (1997), 340–361.

[14] M.D. Penrose, Random Geometric Graphs, Oxford University Press, 2003.

[15] J. Quintanilla, S. Torquato and R.M. Ziff, *Efficient measurement of the percolation threshold for fully penetrable discs*, J. Phys. A **33** (42): L399–L407 (2000).

[16] J.A. Silvester, *On the spatial capacity of packet radio networks*, Department of Computer Science, UCLA, Engineering Report UCLA-ENG-8021, May 7 1980.

[17] H. Takagi and L. Kleinrock, *Optimal transmission ranges for randomly distributed packet radio terminals*, IEEE Transactions on Communications **COM-32** (1984), 246–257.

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