On the connectivity of random k-th nearest neighbour graphs

Colin Cooper
School of Mathematical Sciences,
University of North London,
London N7 8DB, U.K.

Alan Frieze*
Department of Mathematics,
Carnegie-Mellon University,
Pittsburgh PA15213, U.S.A.

June 15, 1995

1 Introduction

Model 1. Consider the complete graph K_n, with vertex set $[n] = \{1, 2, ..., n\}$, in which each edge e is assigned a length X_e. Colour the k shortest edges incident with each vertex green and the remaining edges blue. The graph made up of the green edges only, will be referred to as the *k-th nearest neighbour graph*. This graph has been studied in a variety of contexts both computational and statistical.
We consider here a simple probabilistic model in which the X_e are independent uniform [0,1] random variables. We call this random model $\mathcal{O}_k$.

We remark that choosing the uniform distribution here is no restriction. The distribution of the order statistics of independent identically distributed random variables from any distribution without atoms is equivalent (by a simple transformation) to the distribution of the order statistics in the uniform case.

$\mathcal{O}_k$ is interesting to us because it arises naturally and also because it induces more severe edge dependence than the standard random graph models $G_{n,p}$ and $G_{n,m}$. Aspects of $\mathcal{O}_1$ have been studied (for example) by Holst [5] and Newman, Rinott and Tversky [7]. Our main results will be on the connectivity of these graphs, but first we will describe an equivalent but more combinatorial version of the model.

Model 2. Given the outcomes $\{X_e : e \in E(K_n)\}$ we a.s. obtain a random permutation $\left(e_i, i = 1, \ldots, N = \binom{n}{2}\right)$ of $E(K_n)$ where $X_{e_i} < X_{e_{i+1}}, 1 \leq i < N$. We now define another graph by the following process: let $E_m = \{e_1, e_2, \ldots e_m\}$ and G_m be the random

*Supported by NSF grant CCR-9225008

graph $([n], E_m)$. Let d_m denote degree in G_m. Examine the edges e_i sequentially and colour $e_{m+1} = \{u, v\}$ green if $\min\{d_m(u), d_m(v)\} < k$ and otherwise blue. The resulting green graph is also $\mathcal{O}_k$.

Thus $\mathcal{O}_k$ is the undirected counterpart of the k-out digraphs (G_{k-out}) in which a process colours the first k edges with initial vertex v, for each vertex v, using a random permutation of the edges of the complete digraph. The properties of the k-out digraphs are much better known. For example, there is a large literature on 1-out digraphs (random functional digraphs), see for instance [2] 364-373.

What can we say about $\mathcal{O}_k$? Elementary calculations show that $\mathcal{O}_1$ has approximately $3n/4$ edges, $\mathcal{O}_2$ has approximately $11n/8$ edges and in general:

Theorem 1 *For $k = o(\log n)$ the number of edges of $\mathcal{O}_k$ is* **whp**[1]

$$kn - \frac{n(n-1)}{(2n-3)} \sum_{1 \le i \le j \le k} \frac{\binom{n-2}{i-1}\binom{n-2}{j-1}}{2^{\delta(i,j)}\binom{2n-4}{i+j-2}} + O(k^2\sqrt{n}\log n)$$

where $\delta(i,j)$ is the Kronecker delta.

Thus, not surprisingly, $\mathcal{O}_1$ is almost always disconnected. But what about connectivity for $k \ge 2$? Our main theorems are that:

Theorem 2 *Let $\omega(n) \to \infty$ with n. Then* **whp**, *$\mathcal{O}_2$ is either connected, or consists of a giant component of at least $n - \omega(n)$ vertices, and one or more small cyclic components.*

Theorem 3 *There exists a positive constant γ, where $0.99081 \le \gamma \le 0.99586$ such that*

$$\lim_{n\to\infty} \mathbf{Pr}(\mathcal{O}_2 \text{ is connected }) = \gamma.$$

Theorem 4 *If $k \ge 3$ is fixed, then*

$$\lim_{n\to\infty} \mathbf{Pr}(\mathcal{O}_k \text{ is } k\text{-connected }) = 1.$$

In the following proofs of these theorems, inequalities are only claimed for sufficiently large n.

2 Proof of Theorem 1: The number of edges of $\mathcal{O}_k$.

For $1 \le i, j \le k$ let $\phi(i,j)$ be the number of green edges $e = \{u, v\}$ such that X_e is the i-th order statistic for vertex u and the j-th order statistic for vertex v. Call such an edge *shared*. Let $m = cn\log n$, for some large constant c.

[1] **whp** (*with high probability*); with probability $1 - o(1)$ as $n \to \infty$.

Then **whp** the first m edges of the process in Model 2 contain $\mathcal{O}_k$ and $\phi(i,j)$ is sharply concentrated within $O(\sqrt{n}\log n)$ of its expected value, which is

$$\frac{n(n-1)}{2^{\delta(i,j)}(2n-3)}\frac{\binom{n-2}{i-1}\binom{n-2}{j-1}}{\binom{2n-4}{i+j-2}}.$$

The subtracted terms in the statement of Theorem 1, are the correction to the degree sum (kn) due to the shared green edges.

3 Proof of Theorem 2: The giant component of $\mathcal{O}_2$.

Chernoff bounds: If $B(m,p)$ denotes a binomial random variable then the following tail estimates are well known:

$$\begin{aligned}\mathbf{Pr}(B(m,p)\le(1-\theta)mp) &\le e^{-\theta^2 mp/2},\\ \mathbf{Pr}(B(m,p)\ge(1+\theta)mp) &\le e^{-\theta^2 mp/3},\end{aligned}$$

for any $0\le\theta\le 1$. We refer to the above inequalities as *Chernoff bounds.*

We now consider the case $k=2$ and concentrate on $\mathcal{O}_2$ as defined by Model 2.
Given edges $e_i,e_j\in E_m$, we say $e_i<e_j$ if $i<j$. We extend this notation to sets of edges by saying that $S<T$ if $e\in S, f\in T$ implies $e<f$.
Similarly, from Model 1, and for any $0\le p\le 1$ we define $E_p=\{e: X_e\le p\}$ and $G_p=([n],E_p)$. Thus G_p has the same distribution as the familiar random graph $G_{n,p}$.

It is well-known (see for example [2]), that **whp** $G_{n,p}$ has minimal degree 2 for $p=(\ln n+\ln\ln n+\omega)/n$, whenever $\omega=\omega(n)\to\infty$. Therefore **whp** e_i is blue for all $i\ge n(\ln n+\ln\ln n+\omega)/2$.

Let

$$\begin{aligned}\omega &= (3\ln n)/4, & p &= \omega/n,\\ \epsilon &= \omega^{-1/4}, & \alpha &= \epsilon^{1/2},\\ n_0 &= \left\lceil\sqrt{\ln n}\right\rceil, & n_1 &= \left\lceil n^{15/16}\right\rceil.\end{aligned}$$

Let Γ_p denote the subgraph of G_p induced by the green edges of E_p.

Let $\mathcal{G}(s,\ell,t)$ denote the set of connected graphs with vertex set S, of size s, t edges and ℓ vertices of degree one, except that when $s\le 20e^{100}\ln n$ we do not include trees (see the statement of Theorem 5). We are especially interested in the case where

$$t\le(1+\epsilon)s \text{ and } \ell\le\epsilon s. \tag{1}$$

We prove the following *gap theorem* for the component structure of $\Gamma=\Gamma_p$. We use it much as in the proof by Erdős and Rényi [3] of the existence of a unique giant component in a sparse random graph.

Theorem 5 whp *no component S of Γ with s vertices satisfies*

(a) $s \le n_0$ and S contains at least $s+1$ edges, or

(b) S is not a tree, $n_0 \le s \le n_1$ and S has at most ϵs vertices of degree one, or

(c) S is a tree, $20e^{100}\ln n \le s \le n_1$ and S has at most ϵs vertices of degree one.

Proof: Part (a) follows immediately from the fact that G_p contains no subgraphs of this size and this number of edges; as the expected number of such subgraphs tends to zero as $n \to \infty$. The proof of parts (b),(c) is somewhat more substantial.

Let $I = [n_0, n_1]$ and $\Omega = \{S \subseteq [n] : |S| \in I\}$. For $|S| = s \in I$ consider the event

$$\mathcal{E}_0(S) = \{S \text{ contains at most } (1+\epsilon)s \text{ edges in } G_p\}.$$

Then (proof deferred) [2]

$$\sum_{S\in\Omega} \mathbf{Pr}(\overline{\mathcal{E}}_0(S)) = o(1). \tag{2}$$

Now let $\mathcal{D}$ denote the event that some set S satisfies (b) or (c). Then

$$\mathbf{Pr}(\mathcal{D}) \le \sum_{s\in I}\sum_{|S|=s}\left(\mathbf{Pr}\left(\overline{\mathcal{E}}_0(S)\right) + \sum_{\ell=0}^{\epsilon s}\sum_{t=s-1}^{(1+\epsilon)s}\sum_{G\in\mathcal{G}(s,\ell,t)}\mathbf{Pr}\left(\mathcal{A}(G,S)\right)\right), \tag{3}$$

where

$$\mathcal{A}(G,S) = \{\Gamma_p(S) = G \text{ and } S{:}\overline{S} \text{ is blue }\}. \tag{4}$$

Thus $\mathcal{A}(G,S)$ is the event that S induces the isolated component G in Γ_p. Here, $\Gamma_p(S)$ denotes the subgraph of Γ_p induced by S, and $S{:}\overline{S}$ is the set of G_p-edges joining S and $\overline{S}$.

From (2), the term $\mathbf{Pr}(\overline{\mathcal{E}}_0(S))$ in the summation in (3) is negligible. Thus we can concentrate on the terms of the triple sum, for fixed $S \subseteq [n], s = |S|, \ell, t$ satisfying (1) and $G \in \mathcal{G}(s,\ell,t)$.

For the first part of the proof we work in G_p and only require that G is contained in $G_p(S)$, the subgraph of G_p induced by S. We consider events $\mathcal{E}_1, ..., \mathcal{E}_5$ which establish **whp** the existence of an independent set of vertices with certain regularity properties, contained in $N(S)$, the disjoint neighbour set of S.

For a graph H and $X, Y \subseteq V(H)$ we let $N(X,Y;H)$ denote the set of H-neighbours of X in Y. Also let $N_2(X,Y;H) \subseteq N(X,Y;H)$ denote the set of vertices in Y which have two or more H-neighbours in X. Let $N_2(S) = N_2(S,\overline{S};G_p)$ and for $v \in S$ let $N(v) = N(\{v\},\overline{S};G_p)$ and $N_2(v) = N(v) \cap N_2(S)$.

Let $S_0 = \{v \in S : |N(v)| \in [(1-\epsilon)\omega, (1+\epsilon)\omega]\}$ and

$$\mathcal{E}_1(S) = \left\{|S_0| \ge \left(1 - \frac{\epsilon}{6}\right)s\right\}.$$

[2] In the interests of clarity, we have deferred the derivation of the probability bounds for many of the events we consider, until Section 4. In each case we explicitly state when this has occurred.

Then (proof deferred)

$$\mathbf{Pr}(\overline{\mathcal{E}}_1(S) \mid G_p(S) \supseteq G) \le e^{-c\omega^{1/4}s}, \tag{5}$$

where, hereafter c is a *generic* positive constant, i.e. one whose value can change from formula to formula.

Now let $S_1 = \{v \in S_0 : |N(v) \setminus N_2(v)| \ge (1-\alpha)\omega\}$, and let

$$\mathcal{E}_2(S) = \{|S_1| \ge (1-3\alpha)s\}. \tag{6}$$

Then (proof deferred)

$$\mathbf{Pr}(\overline{\mathcal{E}}_2(S) \mid \mathcal{E}_1(S), G_p(S) \supseteq G) \le e^{-c\omega^{1/4}s}. \tag{7}$$

For each $v \in S_1$, choose a distinct set T_v of $\lceil (1-\alpha)\omega \rceil$ neighbours of v in $N(v) \setminus N_2(S)$. Let $T = \bigcup_{v \in S_1} T_v$. Given $\mathcal{E}_2(S)$ we have $(1-4\alpha)\omega s \le |T| \le \lceil (1-\alpha)\omega s \rceil \le \omega s$.

Next let $T_0 = \{v \in T : |N(v,\ (\overline{S} \setminus N(S)); G_p)| \le (1+\epsilon)\omega\}$ and let

$$\mathcal{E}_3(S) = \{|T_0| \ge (1-\alpha)|T|\}.$$

Then (proof deferred)

$$\mathbf{Pr}(\overline{\mathcal{E}}_3(S) \mid G_p(S) \supseteq G) \le e^{-c\omega^{1/4}s}. \tag{8}$$

Note next that given $\mathcal{E}_i(S), i = 1,2,3$, if

$$S_2 = \{v \in S_1 : |N(v) \cap T_0| \ge 9\omega/10\},$$

then $v \in S_1 \setminus S_2$ implies $|T_v \setminus T_0| \ge (1-\alpha)\omega - 9\omega/10$ and so

$$|S_1 \setminus S_2|((1-\alpha)\omega - 9\omega/10) \le |T \setminus T_0| \le \alpha|T| \le \alpha\omega s$$

which, from (6), implies

$$|S_2| \ge (1-14\alpha)s.$$

Now let

$$\mathcal{E}_4(S) = \{N(S) \text{ contains at most } s/100 \text{ edges }\}.$$

Then (proof deferred)

$$\mathbf{Pr}(\overline{\mathcal{E}}_4(S) \mid G_p(S) \supseteq G) \le e^{-c\omega s}. \tag{9}$$

Let $\mathcal{E}_5(S) = \bigcap_{i=1}^4 \mathcal{E}_i(S)$ then from (5), (7),(8), (9) we see that

$$\mathbf{Pr}(\overline{\mathcal{E}}_5(S) \mid G_p(S) \supseteq G) = e^{-c\omega^{1/4}s}. \tag{10}$$

We now consider the structure of $G \in \mathcal{G}(s,\ell,t)$ when (1) is true. Let $d(v)$ denote the degree of vertex v in G, $D_2 = \{v : d(v) = 2\}$, $n_2 = |D_2|$, $D_3 = \{v \in S : d(v) \ge 3\}$, $n_3 = |D_3|$, and $d_3 = \sum_{v \in D_3} d(v)$. Then

$$\begin{aligned} \ell + n_2 + n_3 &= s, \\ \ell + 2n_2 + d_3 &= 2t, \end{aligned}$$

where $n_3 \le d_3/3$, $\ell \le \epsilon s$ and $t \le (1+\epsilon)s$. Thus

$$\begin{aligned} d_3 &\le 2(1+\epsilon)s - \ell - 2n_2 \\ &= 2(1+\epsilon)s - \ell - 2(s - \ell - n_3) \\ &\le 3\epsilon s + \frac{2d_3}{3}. \end{aligned}$$

Thus

$$d_3 \le 9\epsilon s. \tag{11}$$

Next let $E(G) = E_2 \cup E_3$ where $E_2 = \{e : e \text{ meets a vertex of degree } \le 2\}$, and $E_3 = E(G) \setminus E_2$. Note that $|E_3| \le d_3/2$.

Starting with G, delete the vertices in D_3; unless G is a cycle, when an edge should be randomly deleted. The remaining graph is a collection of vertex disjoint paths $\mathcal{P}$. Thus, given $P \in \mathcal{P}$ we can write it as a directed path $e_1, x_1, e_2, \ldots e_r, x_r, e_{r+1}$, where $r = r(P)$. Here $e_i = x_{i-1}x_i$ for $2 \le i \le r$ and the missing endpoints of e_1, e_{r+1} are of degree 1 or 2 in G. We see that $|\mathcal{P}| \le d_3 + 1$ which is small, see (11).

Most vertices of degree two in G have both neighbours of degree two and so for each $G \in \mathcal{G}(s,\ell,t)$, we define a fixed set $B = B(G) \subseteq V(\mathcal{P})$ as the set of vertices of degree 2 in $\mathcal{P}$.

As $|\mathcal{P}| \le d_3 + 1$ we see from (11) that $|B| \ge s/2$. Let $B_1(G) = B(G) \cap S_2$. Delete from B_1 any vertex v with a neighbour $x \in N(S)$ such that x has an edge in $N(S)$, to obtain $B_2(G)$.

If we assume that $\mathcal{E}_5(S)$ occurs, then $|B_1(G)| \ge s/3$, and

$$|B_2| \ge s/3 - s/50 > s/4. \tag{12}$$

In fact replace B_2 by a subset of size exactly $\lceil s/4 \rceil$. Each vertex v in B_2 has an associated set N_v in $\overline{S}$ of size $\omega_0 = \lceil 9\omega/10 \rceil$ such that

(i) there are no edges from $\bigcup_{v \in B_2} N_v$ to $N(S)$,

(ii) each vertex in each N_v has at most $(1+\epsilon)\omega$ G_p-neighbours in $\overline{S} \setminus N(S)$.

For $v \in S$, let $A(v)$ be the set of $\{v\}$:$\overline{S}$ edges of G_p joining v to $N(v)$. We now consider events $\mathcal{A}_\mathcal{P}$, $\mathcal{B}_\mathcal{P}$, which are defined as follows.

(i) $\mathcal{A}_\mathcal{P} = \bigcap_{P \in \mathcal{P}} \mathcal{A}_P$, where $\mathcal{A}_P = \{A(x_i) \ge e_i \text{ for } 1 \le i \le r(P)\}$.

(ii) For $x \in B_2(G)$ let $A_x \subseteq A(x)$ be the set of $\lceil 9\omega/10 \rceil$ edges joining x to N_x. For $xy \in A_x$ let M_y be the set of at most $(1+\epsilon)\omega$ G_p-edges joining y to $\overline{S} \setminus N(S)$. Then,

$$\mathcal{B}_\mathcal{P} = \bigcap_{x \in B_2(G)} \bigcap_{y \in A_x} \{xy \not\in M_y\}.$$

If $\{S{:}\overline{S}$ is blue$\}$, then (see (i)), $x_i \in P \in \mathcal{P}$ implies that $A(x_i) \geq e_i$ and so $\{S{:}\overline{S}$ is blue$\} \subseteq \mathcal{A}_\mathcal{P}$. In addition, if $x \in B_2(G)$ and $y \in N_x$ then, since y has no neighbours in $N(S)$, $xy < M_y$ implies xy is green. Hence $\{S{:}\overline{S}$ is blue$\} \subseteq \mathcal{A}_\mathcal{P} \cap \mathcal{B}_\mathcal{P}$. It is also true that $\{\Gamma_p(S) = G\} \subseteq \{G \subseteq G_p(S)\}$.

Recall from (4) that $\mathcal{A}(G,S) = \{\Gamma_p(S) = G\} \cap \{S{:}\overline{S}$ is blue$\}$, so we may bound the last term in (3) as follows,

$$\mathbf{Pr}(\mathcal{A}(G,S)) \leq \left(\frac{\omega}{n}\right)^t \left(\mathbf{Pr}(\mathcal{A}_\mathcal{P} \cap \mathcal{B}_\mathcal{P} \cap \mathcal{E}_5 \mid G_p(S) \supseteq G) + \mathbf{Pr}(\overline{\mathcal{E}}_5 \mid G_p(S) \supseteq G)\right). \tag{13}$$

Let $\mathcal{F} = \{G_p(S) \supseteq G\}$. We claim (proof deferred) that,

$$\mathbf{Pr}(\mathcal{A}_\mathcal{P} \cap \mathcal{E}_5 \mid \mathcal{F}) \leq e^{o(s)}\omega^{-t}. \tag{14}$$

We now use Model 1 restricted to G_p, so that the edge lengths are independent uniform $[0,p]$ random variables. For $v \in B_2$, let $l(v) = X_{e(v)}$ where $e(v)$ is the $\mathcal{P}$ edge directed into v.

Let $C = \{v \in B_2 : l(v) \geq \frac{100p}{\omega}\}$ and $\mathcal{C} = \{|C| < |B_2|/2\}$.

We show (proof deferred) that

$$\mathbf{Pr}(\mathcal{B}_\mathcal{P} \mid \mathcal{C}, \mathcal{E}_5, \mathcal{A}_\mathcal{P}, \mathcal{F}) \leq e^{-\theta s}, \tag{15}$$

$$\mathbf{Pr}(\overline{\mathcal{C}} \mid \mathcal{E}_5, \mathcal{A}_\mathcal{P}, \mathcal{F}) \leq e^{-10s}, \tag{16}$$

where $\theta = \frac{1}{9}e^{-100}$.

Thus from (14), (15) and (16),

$$\begin{aligned}\mathbf{Pr}(\mathcal{A}_\mathcal{P} \cap \mathcal{B}_\mathcal{P} \cap \mathcal{E}_5 \mid \mathcal{F}) &\leq \mathbf{Pr}(\mathcal{A}_\mathcal{P} \cap \mathcal{E}_5 \mid \mathcal{F})\left(\mathbf{Pr}(\mathcal{B}_\mathcal{P} \mid \mathcal{C}, \mathcal{E}_5, \mathcal{A}_\mathcal{P}, \mathcal{F}) + \mathbf{Pr}(\overline{\mathcal{C}} \mid \mathcal{E}_5, \mathcal{A}_\mathcal{P}, \mathcal{F})\right) \\ &\leq e^{o(s)}\omega^{-t}(e^{-\theta s} + e^{-10s}). \end{aligned} \tag{17}$$

We now need an estimate of the size of $\mathcal{G}(s,\ell,t)$. We claim (proof deferred) that

$$|\mathcal{G}(s,\ell,t)| = e^{o(s)}s^t e^{-s}. \tag{18}$$

Then from (10), (13) and (17),

$$\begin{aligned}\mathbf{Pr}(\mathcal{A}(G,S)) &\leq \left(\frac{\omega}{n}\right)^t \left(e^{o(s)}\omega^{-t}(e^{-\theta s} + e^{-10s}) + e^{-c\omega^{1/4}s}\right) \\ &\leq e^{o(s)}n^{-t}e^{-\theta s}. \end{aligned} \tag{19}$$

Hence from (2), (3), (18), (19)

$$\begin{aligned}\mathbf{Pr}(\mathcal{D}) &\leq o(1) + \sum_{s\in I}\sum_{t=s-1}^{(1+\epsilon)s}(\epsilon s)\binom{n}{s}|\mathcal{G}(s,\ell,t)|\max\{\mathbf{Pr}(\mathcal{A}(G,S))\} \\ &\leq o(1) + \sum_{s\in I}\sum_{t=s-1}^{(1+\epsilon)s} e^{o(s)}\binom{n}{s}s^t e^{-s}n^{-t}e^{-\theta s} \\ &\leq o(1) + \sum_{s\in I}\sum_{t=s-1}^{(1+\epsilon)s} e^{o(s)}\left(\frac{s}{n}\right)^{t-s}e^{-\theta s} \\ &= o(1). \end{aligned}$$

provided we assume $s \geq 20e^{100} \ln n$ when $t = s - 1$, as in the statement of the theorem. This completes the proof of Theorem 5.

□

We now continue the process described in Model 2, adding edges until a subgraph of minimum degree two has been obtained. We note that **whp** G_p and hence Γ_p contains $\nu_0 \approx n^{1/4}$ isolated vertices (set V_0, say) and $\nu_1 \approx \frac{3}{4}n^{1/4} \ln n$ vertices of degree one (set V_1, say). Each of these vertices obtains at least 1 or 2 *random* green edges when the process reaches minimum degree 2.

Consider any green components of size at least $n^{15/16}$ in Γ_p. Let them be $C_1, C_2, \ldots$. Consider the graph H with vertex set $C_1, C_2, \ldots$ and an edge C_iC_j if there exists a vertex $v \in V_0$ such that v is incident with green edges vw_i, vw_j where $w_i \in C_i$ and $w_j \in C_j$. H is complete **whp** since

$$\mathbf{Pr}(C_iC_j \text{ is not an edge}) \leq \left(1 - \left(\frac{n^{15/16}}{n}\right)^2\right)^{(1-o(1))n^{1/4}} \leq e^{-(1-o(1))n^{1/8}}$$

and there are at most $n^{1/8}$ choices for i, j. Thus **whp** at the end of the process there is a green component containing $C = C_1 \cup C_2 \cup \ldots$. Assume now that Γ_p contains no component as described in Theorem 5. Then **whp** if $v \notin C$ then in Γ_p exactly one of the following is true:

(a) v is in a component with s vertices, where $n_0 \leq s \leq n^{15/16}$. This component has at least ϵs vertices of degree 1,

(b) v is in a tree of size at most $20e^{100} \ln n$,

(c) v is in a unicyclic component of size at most n_0, which is not a cycle,

(d) v is in a cyclic component of size at most n_0.

The number of vertices in (a) is at most ν_1/ϵ, in (b) at most $O(\nu_1 \ln n)$ and from (c),(d) $n^{o(1)}$ (there are few short cycles in G_p **whp**). Thus **whp** $|C| \geq n - n^{1/3}$.

Consider finally the components induced by the vertices in (a), (b) and (c) above and the vertices from V_1 that they contain. C was defined independently of the subsequent green edges that are incident with V_1. Thus the probability that there is a $v \in V_1$ whose second green edge is not incident with C is $O(|V_1|n^{-2/3}) = o(1)$ and so **whp** $\mathcal{O}_2$ only has a giant component K plus (perhaps) some cyclic components of size at most n_0. A more precise bound on the size of these components is derived in Section 5.

4 Deferred proofs.

Proof of (2):

$$\begin{aligned}\sum_{S\in\Omega}\mathbf{Pr}(\overline{\mathcal{E}}_0(S)) &\le \sum_{s\in I}\binom{n}{s}\binom{\binom{s}{2}}{(1+\epsilon)s}p^{(1+\epsilon)s}\\ &\le \sum_{s\in I}\left(\frac{ne}{s}\right)^s\left(\frac{se}{2}\right)^{(1+\epsilon)s}p^{(1+\epsilon)s}\\ &\le \sum_{s\in I}\left(3\left(\frac{s}{n}\right)^{\epsilon}(\ln n)^{1+\epsilon}\right)^s\\ &= o\left(\exp\left(-\sqrt{\log n}\right)\right).\end{aligned}$$

Proof of (5):

Let $\rho = \mathbf{Pr}(v \notin S_0)$, then $E(|S \setminus S_0|) = s\rho$ where $\rho \le \exp(-\epsilon^2\omega/4)$ by the Chernoff bounds on the tails of the Binomial distribution. Hence

$$\begin{aligned}\mathbf{Pr}(\overline{\mathcal{E}}_1(S) \mid G_p(S) \supseteq G) &\le \binom{s}{\epsilon s/6}(e^{-\epsilon^2\omega/4})^{\epsilon s/6}\\ &\le e^{-c\omega^{1/4}s},\end{aligned}$$

Proof of (7):

If $v \in \overline{S}$ then the probability it has exactly one neighbour in S is $p' = sp(1-p)^{s-1}$. Thus the number ν_1 of vertices in $\overline{S}$ with exactly one neighbour in S is distributed as $B(n-s,p')$. Applying the Chernoff bound and observing that $sp = o(\epsilon)$ for $s \in I$ we obtain

$$\mathbf{Pr}(\nu_1 \le (1-\epsilon)\omega s \mid G_p(S) \supseteq G) \le e^{-c\omega^{1/2}s}.$$

Assuming $\mathcal{E}_1$ occurs, $S{:}\overline{S}$ contains at most $(1+\epsilon)\omega s$ edges from S_0 to $\overline{S}$. But

$$\begin{aligned}|S_0{:}N_2(S)| &= |S_0{:}\overline{S}| - \left(\nu_1 - \left|[S\setminus S_0]{:}[N(S)\setminus N_2(S)]\right|\right)\\ &\le (1+\epsilon)\omega s - \left((1-\epsilon)\omega s - 3\omega\frac{\epsilon s}{6}\right),\end{aligned}$$

where 3ω is **whp** an upper bound on the maximum degree of any G_p. Thus with probability $1 - e^{-c\omega^{1/2}s}$ there are at most $5\epsilon\omega s/2$ edges between S_0 and $N_2(S)$.
Assume this and note that each $v \in S_0 \setminus S_1$ has at least $(\alpha-\epsilon)\omega$ neighbours in $N_2(v)$. Under these assumptions

$$|S_0 \setminus S_1|(\alpha-\epsilon)\omega \le \frac{5}{2}\epsilon\omega s$$

which implies $|S_0 \setminus S_1| \le 2.6\alpha s$, and so

$$\mathbf{Pr}(\overline{\mathcal{E}}_2(S) \mid G_p(S) \supseteq G) \le e^{-c\omega^{1/4}s}.$$

Proof of (8):

If $v \in T$ then $\mathbf{Pr}(v \notin T_0) \leq e^{-(1-o(1))\epsilon^2\omega/3}$ and the corresponding events are independent, given T. So

$$\begin{aligned}\mathbf{Pr}(\overline{\mathcal{E}}_3(S) \mid\ |T| = t) &\leq \binom{t}{\alpha t} e^{-\alpha t\epsilon^2\omega/4} \\ &\leq \left(\left(\frac{e}{\alpha}\right) e^{-\epsilon^2\omega/4}\right)^{\alpha t} \\ &\leq e^{-c\omega^{1/4}t}.\end{aligned}$$

If $\mathcal{E}_2$ occurs then $t \geq (1-o(1))\omega s$ and so

$$\mathbf{Pr}(\overline{\mathcal{E}}_3(S) \mid G_p(S) \supseteq G) \leq e^{-c\omega^{1/4}s}.$$

Proof of (9):

$$\begin{aligned}\mathbf{Pr}(\overline{\mathcal{E}}_4(S) \mid G_p(S) \supseteq G) &\leq \mathbf{Pr}(S \text{ has } \geq 3\omega s \text{ neighbours}) \\ &+ \mathbf{Pr}(\overline{\mathcal{E}}_4(S) \mid S \text{ has at most } 3\omega s \text{ neighbours}) \\ &\leq \binom{n}{3\omega s}\left(\frac{\omega s}{n}\right)^{3\omega s} + \binom{\binom{3\omega s}{2}}{s/100}\left(\frac{\omega}{n}\right)^{s/100} \\ &\leq \left(\frac{e}{3}\right)^{3\omega s} + \left(\frac{c\omega^3 s}{n}\right)^{s/100} \\ &\leq e^{-c\omega s}.\end{aligned}$$

Proof of (14):

Fix for the moment, the values of $a(v) = |A(v)| = |N(v)|$, $v \in D_2$. Then

$$\mathbf{Pr}(\bigcap_{P\in\mathcal{P}} \mathcal{A}_P) = \prod_{P\in\mathcal{P}} \mathbf{Pr}(\mathcal{A}_P),$$

where if $P = e_1, x_1, e_2, \ldots e_r, x_r, e_{r+1}$ as before, and if $A_i = A(x_i)$ then

$$\begin{aligned}\mathbf{Pr}(\mathcal{A}_P) &= \mathbf{Pr}(A_i \geq e_i \text{ for } 1 \leq i \leq r) \\ &= \prod_{i=1}^{r} \mathbf{Pr}(A_i \geq e_i) \\ &= \prod_{i=1}^{r} \frac{1}{a(x_i)+1}.\end{aligned}$$

Say $v \in S$ is *small* if $a(v) \leq (1-\epsilon)\omega$. Then

$$\mathbf{Pr}(\mathcal{A}_P) \leq ((1-\epsilon)\omega)^{-(r(P)-\sigma(P))}$$

where $\sigma(P)$ is the number of small vertices on P. Thus if σ is the total number of small vertices,

$$\begin{aligned}\mathbf{Pr}(\bigcap_{P\in\mathcal{P}}\mathcal{A}_P) &\le ((1-\epsilon)\omega)^{-\sum_{P\in\mathcal{P}}(r(P)-\sigma(P))}\\ &\le ((1-\epsilon)\omega)^{-(t-(2|E_3|+\sigma+1))}.\end{aligned}$$

Thus

$$\mathbf{Pr}(\mathcal{A}_\mathcal{P}\cap\mathcal{E}_5|\mathcal{F}) = e^{o(s)}\omega^{-t},$$

since if $\mathcal{E}_1$ occurs then $\sigma \le \epsilon s/6$, and as $G\in\mathcal{G}(s,\ell,t)$ then (11) applies.

Proof of (15):

As before, let $l(v) = X_{e(v)}$ where $e(v)$ is the $\mathcal{P}$ edge directed into v. Conditioning on $\mathcal{A}_\mathcal{P}$ means that $l(v) < \min\{X_{vw}, vw\in A(v)\}$ for all relevant $v\in V(\mathcal{P})$.

The edges vw in A_v are of length $X_{vw}=\eta_w$ say, which is at least $l(v)$ and thus $U[l(v),p]$. The edges of M_w are distributed as $U[0,p]$.

We now consider the event $\mathcal{B}_\mathcal{P}$, which requires that at least one of the (at most) $(1+\epsilon)\omega, G_p$ edges in M_w joining w to $\overline{S}\setminus N(S)$, should be green. Now suppose that $\mathcal{C}$ occurs. Let $l(B_2) = \{l(v): v\in B_2\}$ and $\omega_0 = \lceil 9\omega/10\rceil$ be the size of A_v. Recalling from (12) that $|B_2|\ge s/4$,

$$\begin{aligned}\mathbf{Pr}(\mathcal{B}_\mathcal{P}\mid l(B_2),\mathcal{A}_\mathcal{P},\mathcal{E}_5,\mathcal{F}) &\le \prod_{v\in B_2\setminus C}\prod_{j=1}^{\omega_0}\left(\int_{\eta_j=l(v)}^{p}\left(1-\left(\frac{p-\eta_j}{p}\right)^{\lceil(1+\epsilon)\omega\rceil}\right)\frac{d\eta_j}{p-l(v)}\right)\\ &\le \left(\left(1-\frac{(1-\frac{100}{\omega})^{\lceil(1+\epsilon)\omega\rceil}}{(\lceil(1+\epsilon)\omega\rceil+1)}\right)^{\omega_0}\right)^{\frac{s}{8}}\\ &\le \exp\left(-s(\tfrac{1}{9}e^{-100})\right),\end{aligned}$$

The above inequality is true for any $l(B_2)$ for which $\mathcal{C}$ occurs so that

$$\mathbf{Pr}(\mathcal{B}_\mathcal{P}\mid\mathcal{C},\mathcal{A}_\mathcal{P},\mathcal{E}_5,\mathcal{F})\le e^{-\theta s}.$$

Proof of (16):

We now consider the probability that the event $\mathcal{C}$ does not occur, so that at least half the vertices of B_2 have $l(v) > 100p/\omega$.

$$\begin{aligned}\mathbf{Pr}(\overline{\mathcal{C}}\mid\mathcal{A}_\mathcal{P},\mathcal{E}_5,\mathcal{F}) &\le \sum_{\substack{C\subseteq B_2\\ |C|\ge|B_2|/2}}\prod_{v\in C}\mathbf{Pr}(l(v)\ge 100p/\omega\mid l(v)\le X_{vw},\, vw\in A_v)\\ &\le \sum_{\substack{C\subseteq B_2\\ |C|\ge|B_2|/2}}\prod_{v\in C}\left(1-\frac{100}{\omega}\right)^{\omega_0+1}\\ &\le \sum_{\substack{C\subseteq B_2\\ |C|\ge|B_2|/2}} e^{-90|C|}.\end{aligned}$$

Finally as $|B_2| = s/4$, we have,

$$\begin{aligned}\mathbf{Pr}(\overline{\mathcal{C}} \mid \mathcal{A}_{\mathcal{P}}, \mathcal{E}_5, \mathcal{F}) &\leq \sum_{u=s/8}^{s/4} \binom{s/4}{u} e^{-90u} \\ &\leq 2^{\frac{s}{4}} e^{-11s} \\ &\leq e^{-10s}.\end{aligned}$$

Proof of (18):

We claim first that

$$|\mathcal{G}(s,\ell,t)| \leq \binom{s}{\ell}\binom{t}{\ell}\frac{\ell!\,2^\ell}{t!\,2^t}\, h_{2t-\ell,s-\ell}, \tag{20}$$

where $h_{m,n}$ is the number of ways of putting m labelled balls into n labelled boxes with at least two balls per box.

We consider a natural mapping f from $[s]^{2t} \to \mathcal{MG}(s,t)$, where $\mathcal{MG}(s,t)$ is the set of multigraphs with vertex set $[s]$ and t edges. If $\mathbf{x} = (x_1, x_2, \ldots, x_{2t}) \in [s]^{2t}$ then we let $f(\mathbf{x})$ be the multigraph with edge-set $\{\{x_{2i-1}, x_{2i}\} : 1 \leq i \leq t\}$. Observe now that each $G \in \mathcal{G}(s,\ell,t)$ is the image of precisely $t!\ 2^t$ members of $[s]^{2t}$. So we need only prove that the remaining factors on the RH S of (20) are an upper bound on $|f^{-1}(\mathcal{G}(s,\ell,t))|$. For $\mathbf{x} \in [s]^{2t}$ let $d_{\mathbf{x}}(j) = |\{i : x_i = j\}|, 1 \leq j \leq s$. To construct $\mathbf{x} \in f^{-1}(\mathcal{G}(s,\ell,t))$ we may

(i) choose $J_1 = \{j : d_{\mathbf{x}}(j) = 1\}$ in $\binom{s}{\ell}$ ways,

(ii) we then choose, in $\binom{t}{\ell}\ell!2^\ell$ ways, the set of indices i such that $x_i \in J_1$, noting that, by connectivity any selected edge has only a single vertex of degree 1, and

(iii) fill in the remaining $2t-\ell$ positions, using $s-\ell$ values so that $d_{\mathbf{x}}(j) \geq 2$ for $j \notin J_1$.

This yields an upper bound only, as we have not ensured that $f(\mathbf{x})$ will be connected or simple, but verifies (20).

Now $h_{m,n}$ can be expressed as $e^{o(n)}n^m\phi(\alpha)^n$ where $\alpha = 2m/n \geq 2$. The exact form of the function ϕ is not important to us, only that it is continuous and that $\phi(2) = 2e^{-2}$ as $h_{2n,n} = 2n!/2^n$ and so $\phi(2+o(1)) = 2e^{-2}(1+o(1))$.

For $\ell \leq \epsilon s$ and $t \leq (1+\epsilon)s$, we have

$$\binom{s}{\ell}, \binom{t}{\ell} = e^{o(s)}; \quad (s-\ell)^{2t-\ell}\ell!2^\ell \leq s^{2t} \text{ and } \phi\left(\frac{2t-\ell}{s-\ell}\right) = 2e^{-2}(1+o(1)).$$

Hence

$$|\mathcal{G}(s,\ell,t)| = e^{o(s)}s^t e^{-s}.$$

5 Proof of Theorem 3: The limiting probability that $\mathcal{O}_2$ is connected.

We have shown in Section 3 that **whp** either $\mathcal{O}_2$ is connected or consists of a giant component and some cyclic components of size at most $n_0 = \sqrt{\log n}$. We now show that the expected number of cyclic components of size at least s tends to zero as s tends to infinity, and that the number of cyclic components of constant size is asymptotically Poisson with constant parameter. Thus **whp** the size of the giant component is at least $n - \omega(n)$ for any $\omega(n) \to \infty$.

Working with Model 1, we first establish some sharp inequalities for the number of second order statistic edges which fall in the interval $[0, y]$. Let $\Omega_k = \Omega_k(y)$ be the event that there are exactly k out of $\binom{n}{2}$ edges whose length is in $[0, y]$, and let Ω be the union of these events. Thus Ω is the entire sample space for Model 1. In either case, any edge with length at most y, has length distributed as $U[0, y]$.

Let X be some random variable on Ω with expectation $E(X)$, and let $E_k(X)$ be the expectation of X conditional on Ω_k.

Lemma 1 *For* $0 < \epsilon < 1$ *let*

$$\begin{aligned} K &= \{k : (1-\epsilon)\binom{n}{2}y \le k \le (1+\epsilon)\binom{n}{2}y\}, \\ \rho &= \mathbf{Pr}(\bigcup_{k \in V \setminus K} \Omega_k), \\ \delta &= \max_{k,l \in K} |E_l(X) - E_k(X)| + \rho \max_{l \in V \setminus K} E_l(X), \end{aligned}$$

then

$$\mathbf{Pr}(|X - E(X)| > \theta) \le \max_{k \in K} \mathbf{Pr}(|X - E_k(X)| > \theta - \delta \;\Big|\; \Omega_k) + \rho. \tag{21}$$

Proof: Using

$$E(X) = \sum_{l \in K} E_l(X)\mathbf{Pr}(\Omega_l) + \sum_{l \in V \setminus K} E_l(X)\mathbf{Pr}(\Omega_l),$$

we see that

$$|E(X) - E_k(X)| \le \max_{k,l \in K} |E_l(X) - E_k(X)| + \rho \max_{l \in V \setminus K} E_l(X) = \delta.$$

Now for fixed k,

$$\begin{aligned} |X - E(X)| &\le |X - E_k(X)| + |E_k(X) - E(X)| \\ &\le |X - E_k(X)| + \delta, \end{aligned}$$

thus

$$\mathbf{Pr}(|X - E(X)| > \theta \;\Big|\; \Omega_k) \le \mathbf{Pr}(|X - E_k(X)| + \delta > \theta \;\Big|\; \Omega_k).$$

However

$$\mathbf{Pr}(|X - E(X)| > \theta) \leq \sum_{k \in K} \mathbf{Pr}(|X - E(X)| > \theta \quad \Big| \, \Omega_k)\mathbf{Pr}(\Omega_k) + \rho,$$

and the result follows. □

Now given $\omega \in \Omega$, let $V_y(\omega)$ be the set of vertices v whose second order statistic edge length $x_{(2)v}$, is at most y. Let $N(y) = |V_y|$ and $S(y) = \sum_{v \in V_y} x_{(2)v}$; thus $N(y)$ and $S(y)$ are the number of such vertices and the total length of their second order statistic edges respectively. We adopt the convention that the parameter y is omitted from the random variable if no confusion arises from this omission. We find that

$$E(N(y)) = n\left(1 - (1-y)^{n-2}\left(1 + (n-2)y\right)\right), \tag{22}$$

$$E(S(y)) = 2\left(1 - (1-y)^{n-2}\left(1 + (n-2)y + \binom{n-1}{2}y^2\right)\right). \tag{23}$$

The probability density function for the length, z, of the second order statistic edge at a vertex is given by

$$g(z) = (n-1)(n-2)\, z(1-z)^{n-3},$$

thus, (22) is $n\int_0^y g(z)dz$ and (23) is $n\int_0^y zg(z)dz$.

We now derive asymptotic approximations for N, S for $y \leq 3\log n/n$ in terms of their expected values at interpolated points.

Lemma 2 *Let* $y_0 = \Delta y = n^{-3/2}$, $y_1 = 3\log n/n$, $\epsilon^2 = 3\log^4 n/(n(n-1)y)$,
$\theta(N(y)) = 4n(\log^2 n)\sqrt{y}$, *and* $\theta(S(y)) = 6n(\log^2 n)y\sqrt{y}$.
Let the random variable X *denote either* N *or* S.
(i) For $y_0 \leq y \leq y_1$,

$$\mathbf{Pr}(|X(y) - E(X(y))| \geq \theta(X(y))) \leq 4e^{-\frac{1}{2}\log^4 n}. \tag{24}$$

(ii) Let $y = i\Delta y$, $i = 1, 2, ..., \lceil 3\sqrt{n}(\log n)\rceil$, *and* $z \in [y, y + \Delta y]$, *then* **whp**,

$$E(X(z)) - 2\theta(X(y)) \leq X(z) \leq E(X(z)) + 2\theta(X(y)). \tag{25}$$

Proof: (i) The expected number of edges falling in $[0, y]$ is $\binom{n}{2}y$. Thus if k is the actual number of edges,

$$\begin{aligned}\rho = \mathbf{Pr}(\bigcup_{k \in V \setminus K} \Omega_k) &\leq 2\exp\left\{-\frac{\epsilon^2}{3}\binom{n}{2}y\right\} \\ &\leq 2\exp\left\{-\log^4 n\right\}.\end{aligned}$$

We use the bounded martingale difference method (see, for example Alon and Spencer [1] or McDiarmid [6]) to show that N and S are sharply concentrated for fixed k, y. In particular we

see that $|N_i - N_{i+1}| \leq 2$ and $|S_i - S_{i+1}| \leq 2y$, where N_i is the evaluation of $E(N)$ conditional on the first i edges of the process defined by Model 2. Hence,

$$\mathbf{Pr}\left(|N - E_k(N)| > \sqrt{2k\log^4 n}\right) \quad \leq \quad 2e^{-\log^4 n}$$

Now, $\max_{l\in V\setminus K} E_l(N) \leq n$. Also $\max_{l\in K} |E_l(N) - E_k(N)| \leq |K|$ as we are adding at most $|K|$ edges going from Ω_k to Ω_l, $(k < l)$. Thus $\delta(N) \leq |K| + \rho n \leq n\log^2 n\sqrt{3y}(1 + o(1))$, and the result follows from (21). We note that similar approximations can be made for S.

(ii) If $X = N, S$ then both $X(z)$ and $E(X(z))$ are monotone nondecreasing for $z \in [0, 1]$. Thus for any $z \in [y, y + \Delta y]$, we have from part (i) **whp** that

$$E(X(y)) - \theta(X(y)) \leq X(z) \leq E(X(y + \Delta y)) + \theta(X(y + \Delta y)).$$

Setting $X = N$ and using a Taylor series expansion of $E(N(y + \Delta y))$ we see that

$$E(N(y)) \leq E(N(z)) \leq E(N(y)) + (n)_3\Delta y[\xi(1 - \xi)^{n-3}], \qquad \xi \in (y, y + \Delta y).$$

The final term is less than $\theta(N(y))$ for all y in the range.
A similar proof holds for S. □

Let $M = \{1, 2, ..., m\}$, and $C = 12...m1$ be the cycle with edges

$$e_1 = \{1, 2\}, e_2 = \{2, 3\}, ..., e_m = \{m, 1\}$$

of lengths $x_1, x_2, ..., x_m$ respectively, where $x_1 \leq x_2 \leq \cdots \leq x_m$. We shall call this the *natural ordering* $\alpha = (e_1, e_2, ..., e_m)$ of the edges of C induced by $x_1 \leq x_2 \leq \cdots \leq x_m$.

As before, let $n_0 = \sqrt{\log n}$, and let $m \leq n_0$ be a fixed natural number. Let $\mathcal{E}_\alpha$ be the event

(i) C is a cyclic component of $\mathcal{O}_2$ with the natural ordering α, where the edge lengths satisfy $x_1 \leq x_2 \leq \cdots \leq x_m \leq y_1 = (3\log n)/n$,

(ii) $|X(y) - E(X(y))| \leq \theta(X(y))$ for $y \leq y_1, X = N, S$ as in Lemma 2.

The conditional probability density function $g(x_1, ..., x_m)$ for the edge lengths $(x_1, ..., x_m)$ of C given the event $\mathcal{E}_\alpha$ is well defined and corresponds to a function

$$f(x_1, ..., x_m) = g(x_1, ..., x_m)\mathbf{Pr}(\mathcal{E}_\alpha).$$

We will call $f(x_1, ..., x_m)$ the incomplete probability density function (ipdf) of $\mathcal{E}_\alpha$.

For brevity, denote $[n]$ by V, and let $x_{(2)j}$ denote the length of the second shortest edge incident with vertex j. Let $i \in M$, and $j \in V - M$, then the edges $\{i, j\}$ in $M \times (V - M)$ have length $d_{j,i}$ where $d_{j,i} \sim U[0, 1]$, independently of each other.

Suppose we are given $\mathcal{O}_2[M]$ and $\mathcal{O}_2[V - M]$, then for the edges of $\mathcal{O}_2$ to be the union of those of $\mathcal{O}_2[M]$ and $\mathcal{O}_2[V - M]$ we require for all $i \in M$ and for all $j \in V - M$ that $d_{i,j} > \max\{x_{(2)i}, x_{(2)j}\}$.

Lemma 3 *Let $f(x_1, x_2, ..., x_m)$ be the ipdf for the event $\mathcal{E}_\alpha$. Then if $\phi(y) = e^{-ny-(2+ny)e^{-ny}}$, we have*

$$f(x_1, x_2, ..., x_m) = \phi(x_2)\phi(x_3)\cdots\phi(x_{m-1})\phi^2(x_m)\left(1 + o\left(\frac{\log^5 n}{\sqrt{n}}\right)\right).$$

Proof: If we consider the vertex set M in isolation, then the edges in M are $U[0,1]$ and the ipdf for the event $A(C)$ that $\mathcal{O}_2[M]$ is the cycle C with the natural ordering is

$$\psi(x_1, x_2, ..., x_m)dx_1...dx_m = (1 - x_m)^{m-3} \times \prod_{l=4}^{m}(1 - x_l)^{l-3}dx_1...dx_m, \tag{26}$$

as we require that $x_1 \le x_2 \le ... \le x_m$ and thus the probability that an edge $\{j, i\}$, $k < i - 1$ is longer than x_i is $1 - x_i$. For the values given by $\mathcal{E}_\alpha$ we have that $\psi = 1 - O(m^2 \log n/n)$.

We now estimate the probability that $\mathcal{O}_2 = \mathcal{O}_2[M] \cup \mathcal{O}_2[V - M]$, when $A(C)$ has occurred and the structure and edge lengths of $\mathcal{O}_2[M]$ and $\mathcal{O}_2[V - M]$ are known.

Let $i \in M$, and let $z = x_{(2)i}$. Let

$$\mathcal{H}(i, z) = \{d_{j,i} > z \text{ and } d_{j,i} > x_{(2)j},\ \forall j \in V - M \mid A(C),\ \mathcal{O}_2[M],\ \mathcal{O}_2[V - M]\},$$

and let $h(i, z) = \mathbf{Pr}(\mathcal{H}(i, z))$.
By the independence of the edges $\{i, j\}$, the ipdf $f(x_1, ..., x_m \mid \mathcal{O}_2[V - M])$ of $\mathcal{E}_\alpha$ given $\mathcal{O}_2[V - M]$ is

$$\begin{aligned} f(x_1, ..., x_m \mid \mathcal{O}_2[V - M])dx_1...dx_m &= h(2, x_2)...h(m - 1, x_{m-1})h(m, x_m)h(1, x_m) \\ &\quad \times \psi(x_1, ..., x_m)dx_1...dx_m. \end{aligned} \tag{27}$$

The sharp estimates for N and S given in Lemma 2, allow us to estimate $h(i, z)$ (**whp**) as follows.

Let $\eta = |V - M|$, and let N, S in Lemma 2 refer to $V - M$. Let $y_0 \le y \le y_1$, and $z \in [y, y + \Delta y]$, as defined in Lemma 2. Then,

$$\begin{aligned} h(i, z) &= \prod_{\substack{j \in V - M \\ x_{(2)j} \le z}} (1 - z) \prod_{\substack{j \in V - M \\ x_{(2)j} > z}} (1 - x_{(2)j}) && (28) \\ &= (1 - z)^{N(z)} \prod_{x_{(2)j} > z} (1 - x_{(2)j}) \\ &= \exp\left(-(N(z)z + O(N(z)z^2)) - \sum_{x_{(2)j} > z} (x_{(2)j} + O(x_{(2)j}^2))\right). && (29) \end{aligned}$$

Now, **whp** $E(S(y_1)) = E(S(1))$ which is $2 + O(\frac{1}{\eta})$ and

$$\mathbf{Pr}\left(|S(y_1) - E(S(y_1))| > \log^3 \eta/\sqrt{\eta}\right) = O(\eta^{-\log^2 \eta}).$$

Thus, as $S(1) - S(z) = \sum_{x_{(2)j} > z} x_{(2)j}$, we have (**whp**) that

$$\sum_{x_{(2)j} > z} x_{(2)j} = 2 - S(z) + O(\frac{\log^3 \eta}{\sqrt{\eta}}).$$

Thus (**whp**)

$$h(i,z)=\exp\Big(-N(z)z-(2-S(z))\Big)\exp\left(O\left(\frac{\log^3\eta}{\sqrt{\eta}}\right)+O(N(z)z^2)+O(\sum_{x_{(2)j}>z}x_{(2)j}^2)\right),$$

and the terms in the second exponent are $O(\log^3\eta/\sqrt{\eta})$ as $z,x_{(2)j}<y_1$ **whp**. We now use the results of Lemma 2 (25) to substitute the expected values of $N(z)$, $S(z)$ into the first exponent to give

$$h(i,z)=\exp\left\{-E(N(z))z-(2-E(S(z)))\right\}(1+O[z\theta(N(y))+\theta(S(y))])\left(1+O(\tfrac{\log^3\eta}{\sqrt{\eta}})\right).$$

From (22), (23) we see that, on replacing η by $n-m$,

$$E(N(z))z+(2-E(S(z)))=nz+(1-z)^{n-2}(2+nz)+O(mz+nz^2+\frac{m}{n}).$$

Finally, recalling that $m\le\sqrt{\log n}$ and $z\le 3\log n/n$, we see that

$$h(i,z)=\exp\left(-nz-(2+nz)e^{-nz}\right)\left(1+o\left(\frac{\log^4 n}{\sqrt{n}}\right)\right). \tag{30}$$

The result follows from (26), (27) and (30). □

Consider the cycle $C=12...m1$ with edges $e_1,e_2,...,e_m$ on M. In the previous lemma we used the *natural ordering* $\alpha=(e_1,e_2,...,e_m)$ of the edges induced by $x_1\le x_2\le\cdots x_m$. In general let $\Phi=\{\omega:\omega=(e_{(1)},e_{(2)},...,e_{(m)})\}$ be the set of all m tuples, where $e_{(i)}$ is the edge of length $x_{(i)}$, and $x_{(1)}\le x_{(2)},\cdots\le x_{(m)}$.
If $\mathcal{E}_\omega$ is the event that C with edge ordering ω is an isolated cycle of $\mathcal{O}_2$; then for $m\ge 3$, the expected number of such cyclic components ν_m is

$$\nu_m=\binom{n}{m}\frac{m-1!}{2}\sum_{\omega\in\Phi}\mathbf{Pr}(\mathcal{E}_\omega). \tag{31}$$

We note that as $m!=o(\exp(\frac{1}{2}\log^4 n))$ the bounds for N,S derived in (24) of Lemma 2 hold simultaneously for all entries in this expression. The limiting ipdf $f(x_1,x_2,...,x_m;\omega)$ is derived in the same way as for the natural ordering α, except now in $h(i,z)$, the length of the second order statistic edge of vertex i in $\mathcal{O}_2[M]$ is given by $z=\max\{d_{i-1,i},d_{i,i+1}\}$ whereas previously $z=d_{i,i+1}$.
The explicit form of this ipdf and the details of the bounds for ν_m are given in Lemmas 5,6 below. Before we prove the bounds in Theorem 3, we need to indicate why the number of small cyclic components tends to a Poisson distribution.

Lemma 4 *The number of cyclic components of $\mathcal{O}_2$ tends to a Poisson distribution with parameter $\nu=\sum_{m\ge 3}\nu_m$.*

Proof: Let W be a random variable counting the number of isolated cycles, where $E(W) = \nu = \sum_{m\geq 3} \nu_m$. What can we say about the rth factorial moment of W? Consider the case for $r = 2$. We wish to count the number of ordered pairs of isolated cycles.
Let $C_i, i = 1, 2$ have edge orders ω_i and vertex sets M_i, where $M = M_1 \cup M_2$ and $M_1 \cap M_2 = \emptyset$. In order to compare $\mathbf{Pr}(\mathcal{E}_{\omega_1}\mathcal{E}_{\omega_2})$ with $\mathbf{Pr}(\mathcal{E}_{\omega_1})\mathbf{Pr}(\mathcal{E}_{\omega_2})$ we examine $\mathbf{Pr}(\mathcal{E}_{\omega_2} \mid \mathcal{E}_{\omega_1})$. Specifically, we consider (27) but condition on the existence of C_1.

If $k \notin M$ and $j \in M_2$, then the term in the product for k in $h_2(j, z|C_1)$ is the same as in $h_2(j, z)$. If $k \in M_1$ then the product term (either $(1 - z)$ or $(1 - x_{(2)k})$) is missing from $h_2(j, z|C_1)$ see (28). Thus,

$$\mathbf{Pr}(\mathcal{E}_{\omega_1}\mathcal{E}_{\omega_2}) = \mathbf{Pr}(\mathcal{E}_{\omega_1})\mathbf{Pr}(\mathcal{E}_{\omega_2})(1 + O(m_1 m_2 y_1))$$

as we have already noted that $\psi(.)$ given in (26) is $1 - O(m^2 \log n/n)$.
Thus $E(W)_2 = E^2(W)(1 + o(\log^3 n/n))$.
This generalizes to $E(W)_r = E^r(W)(1 + o(\log^{r+1} n/n))$ which gives the required convergence. □

As before, let $\phi(y)$ denote the asymptotic value of $h(i, y)$, thus from (30),

$$\phi(y) = \exp(-ny - (2 + ny)e^{-ny})$$

Lemma 5 *Let $\omega = (e_{(1)}, e_{(2)}, ..., e_{(m)})$ be an ordering of the edges of C by increasing length, and let $\alpha = (e_1, e_2, ..., e_m)$ be the natural ordering. Then*

$$I_L = \int_0^1 \frac{z^{m-1}}{m-1!}\phi^m(z)\, dz \leq \mathbf{Pr}(\mathcal{E}_\omega) \leq \mathbf{Pr}(\mathcal{E}_\alpha) \leq \int \prod_{i=1}^{m} \phi(x_i)\, dx_i = I_U. \tag{32}$$

Proof: We have already shown that, asymptotically

$$\mathbf{Pr}(\mathcal{E}_\alpha) = \int dx_1 \left(\prod_{i=2}^{m-1} \phi(x_i)\, dx_i\right) \phi^2(x_m)\, dx_m. \tag{33}$$

Consider adding the edges $e_{(i)}$ to the cycle C in order of increasing length. The variable j_i counts the number of vertices whose degree becomes 2 on addition of the edge $e_{(i)}$.
We now claim that, asymptotically

$$\mathbf{Pr}(\mathcal{E}_\omega) = \int dx_{(1)} \left(\prod_{i=2}^{m-1} \phi^{j_i}(x_{(i)})\, dx_{(i)}\right) \phi^2(x_{(m)})\, dx_{(m)},$$

where $0 \leq j_i \leq 2$, $j_2 + \cdots + j_i \leq i - 1$, $j_2 + \cdots + j_{m-1} = m - 2$.
In particular let $e_{(i)} = \{k, k + 1\}$. There will be a ϕ entry for vertex k corresponding to $h(k, x_{(i)})$ if and only if $e_{(i)}$ is the second edge in the sequence ω incident with vertex k, if and only if $x_{(i)} = \max\{d_{k-1,k}, d_{k,k+1}\}$. Hence $0 \leq j_i \leq 2$.
For $i \leq m - 2$, $j_2 + \cdots + j_i \leq i - 1$ as this sum gives the number of vertices of degree 2 in the paths formed by the edges $e_{(1)}, e_{(2)}, ..., e_{(i)}$. Moreover $j_2 + \cdots + j_{m-1} = m - 2$ as this counts vertices of degree 2 on the path formed by deleting the longest edge from the cycle.

Let i be the lowest index such that $j_i = 0$ in $\mathbf{Pr}(\mathcal{E}_\omega)$, and k the first index such that $j_k = 2$, where $i < k$ by the above discussion. The function ϕ is non-increasing on $[0,1]$, and $x_{(i)} \le x_{(k)}$ so

$$\int \cdots \phi^2(x_{(k)})\, dx_{(i)}\, dx_{(k)} \le \int \cdots \phi(x_{(i)})\phi(x_{(k)})\, dx_{(i)}\, dx_{(k)}.$$

An induction based on successive rearrangements of this form gives the required inequalities. □

We now give some bounds for ν, the parameter of the Poisson distribution for the asymptotic number of cyclic components.

Lemma 6

$$0.004152 \le \nu \le 0.009228$$

Proof: *Lower bound.*

We remark that for cycles of very short length, the different possible orderings of the edges can be represented explicitly and the Poisson parameters obtained by direct integration. In particular, for cycles of length 3, all cycles can be relabelled to have the natural ordering.

Thus, from (31) and (33) we see that

$$\begin{aligned} \nu_3 &= (n)_3 \int_{z=0}^{1} \int_{y=0}^{z} y e^{-(ny+(2+ny)e^{-ny})} e^{-2(nz+(2+nz)e^{-nz})}\, dy\, dz \\ &= 0.00415239, \end{aligned}$$

by numerical integration. Thus the number of triangles is asymptotically Poisson with parameter $\nu_3 = 0.00415239$. We note that the lower bound ν_L on ν must be at least ν_3.

Upper bound. Let

$$\psi(y) = \begin{cases} e^{-2}, & 0 \le y \le \frac{2}{n} \\ e^{-ny}, & \frac{2}{n} < y \le 1, \end{cases}$$

then $\psi(y) \ge \phi(y)$ for $y \in [0,1]$. Replacing ϕ by ψ in the integral I_U of the previous lemma, we see that $I_U(\psi) = \sum_{i=0}^{m} L(i)$ where

$$L(i) = \int_{K_{2,i}} \left(\int_{K_{1,i}} e^{-2i} dx_1 \cdots dx_i \right) e^{-n(x_{i+1}+\cdots+x_m)} dx_{i+1} \cdots dx_m$$

and the range of integration of $L(i)$ is over $K_{1,i} = \{0 \le x_1 \le \cdots \le x_i \le 2/n\}$ and $K_{2,i} = \{2/n \le x_{i+1} \le \cdots \le x_m \le 1\}$. Thus each $L(i)$ is the product of two integrals. The first $A(1,i)$ has range of integration $K_{1,i}$ and each of the i entries for ψ is e^{-2}. The second $B(i+1,m)$ has range of integration $K_{2,i}$ and thus each of the $m-i$ entries for ψ is of the form e^{-ny}.
It is immediate that $A(1,i) = \frac{e^{-2i}}{i!}(\frac{2}{n})^i$ and a tedious induction shows that $B(i+1,i+k) = \frac{e^{-2k}}{n^k k!}$ asymptotically. Hence $I_U(\psi) = \frac{3^m e^{-2m}}{n^m m!}$ and it follows from (31), (32) that, asymptotically

$$\nu_U = \frac{1}{2}\left(\log \frac{1}{1-3e^{-2}} - 3e^{-2} - \frac{(3e^{-2})^2}{2} - \frac{(3e^{-2})^3}{3} \right) + \nu_3.$$

□

It is an immediate consequence of Lemma 4 and 6 that the expected number of cyclic components of length m is at most

$$\binom{n}{m}\frac{(m-1)!}{2}I_U(\psi) \le \frac{(3e^{-2})^m}{2m},$$

and thus **whp** there are no cycles of length $\omega(n)$ for any $\omega(n) \to \infty$, and at most $\omega(n)$ vertices on cyclic components of constant length. This completes the proof of Theorems 2 and 3.

6 Proof of Theorem 4: k-Connectivity of $\mathcal{O}_k$.

Assume now that $k \ge 3$ is fixed. For $T \subseteq [n]$ let

$$\mathcal{D}(T) = \{\mathcal{O}_k \setminus T \quad \text{is not connected}\}.$$

We need to prove that

$$\mathbf{Pr}\left(\bigcup_{|T|=k-1} \mathcal{D}(T)\right) = o(1). \tag{34}$$

Putting $T_0 = [k-1]$ and $\mathcal{D}_0 = \mathcal{D}(T_0)$, we prove (34) by showing

$$\mathbf{Pr}(\mathcal{D}_0) = o(n^{-(k-1)}). \tag{35}$$

It is important to note that proving connectivity in $\mathcal{O}_k[[n] - T_0]$ is not sufficient, as green edges in this graph may be recoloured blue on addition of T_0, and hence deleted.

Suppose now that $\Gamma_{p,k}$ denotes the subgraph of $\mathcal{O}_k \setminus T_0$ induced by the edges of G_p. Reworking the calculations of Sections 3 and 4 we see that there exists a constant c_k such that with probability $1 - o(n^{-k})$, $\Gamma_{p,k}$ has the following property:

There is no component of size s between $c_k \ln n$ and n_1 which has fewer than ϵs vertices of degree 1.

One hardly notices the effect of edges incident with T_0 in the calculations relating to the event $\mathcal{E}_5$ of Theorem 5, where now $S{:}\overline{S}$ denotes the set of edges from S to $[n] \setminus (S \cup T_0)$. We now extend our definition of B_2, (refer to the paragraph of equation (12)), to require not only that these neighbour vertices are independent, but also have no edge to T_0 in G_p.

Continuing the analysis following the proof of Theorem 5 in Section 3, we obtain a subgraph of $\mathcal{O}_k$ of minimum degree 2, induced by G_q, where $q = (\log n + \log\log n + \omega)/n$, which we will denote by $\mathcal{O}_k[G_q]$. With probability $1 - o(n^{-k})$, $\mathcal{O}_k[G_q] - T_0$ is (a) connected, or (b) contains one giant component K of size at least $n - n^{1/3}$ plus small components of size at most $c_k \ln n$.

We must consider Case (b). These small components are either

(i) isolated components C of $\mathcal{O}_k[G_q]$, not necessarily cyclic, or

(ii) span a set of vertices W, such that there are no green edges from W to K, but $\mathcal{O}_k[G_q[W \cup T_0]]$ is connected.

Case (i) is dealt with below by showing that either C does not exist **whp** in G_q as it contains too many edges, or that at least k additional random green edges must be incident with C in $\mathcal{O}_k$, as the degree sum of C in $\mathcal{O}_k[G_q]$ is too small. The probability that none of these random green edges is incident with K is $O(n^{-k})$. Specifically,

If $|C| = 3$ then there are at least $3(k-2) \geq k$ additional green edges incident with C.

If $|C| = 4$ and $k \geq 4$ then there are at least $4k - 12 \geq k$ additional edges incident with C.

If $|C| = 4$ and $k = 3$ then either (i) there are at least 4 additional edges incident with C, or (ii) there are at least 5 edges contained in C. The probability that G_q contains a set of 4 vertices spanning 5 or 6 edges is $o(1)$.

If $|C| = s \geq 5$ then there are at least

$$sk - 2(s + a_s) + 1 = s(k - 2 - o(1)) - 2k + 1 \geq k$$

additional edges incident with C by Lemma 7, stated below.

Case (ii) can now be dealt with in G_r, $(r = (\log n + (k-1)\log\log n + \omega)/n)$. Let $|W| = w$. We will show **whp** that the set $W \cup T_0$ contains too many edges. Suppose that in $\mathcal{O}_k$, W has x internal edges and there are y edges between W and T_0. The probability in G_r of this event is at most

$$\binom{n-(k-1)}{w}\binom{(k-1)w}{y}\binom{\binom{w}{2}}{x}r^{x+y} \quad = \quad n^{w-(x+y)}\left(\frac{e}{w}\right)^w\left(\frac{kwe\ln n}{y}\right)^y\left(\frac{w^2 e\ln n}{2x}\right)^x$$

Now $2x + y \geq wk$, where $y \geq 1$ and $x \leq \min\{\binom{w}{2},\ w + a_w\}$. By arguments similar to those used above, we see that the probability of the event is bounded by $O(n^{-k})$ as required.

Thus (35) and the rest of Theorem 4 follows.

Lemma 7 *Suppose $s \leq c_k \ln n$. Let $\rho \leq 2\log n/n$*

$$\begin{aligned} a_s &= \frac{k\ln n + s\ln\ln n + s\ln 8}{\ln n - \ln s - \ln 8 - \ln\ln n} \\ &= k + s\frac{\ln\ln n}{\ln n} + O(s/\ln n) + o(1). \end{aligned}$$

Then the probability that G_ρ contains a set of s vertices spanning $s + a_s$ edges is $o(n^{-k})$.

Proof: The expected number of sets of size s with at least $s + a$ edges is

$$\binom{n}{s}\binom{\binom{s}{2}}{s+a}\left(\frac{2\ln n}{n}\right)^{s+a} \leq \left(\frac{s}{n}\right)^a (8\ln n)^{s+a}.$$

(See (2) for a similar calculation.) Substituting a_s for a and simplifying yields the result. □

7 Acknowledgement

We are indebted to the anonymous referees for many useful suggestions.

References

[1] Alon, N. & Spencer, J.H. *The Probabilistic Method.* John Wiley, 1991.

[2] Bollobás, B. *Random Graphs.* Academic Press, 1985.

[3] Erdős, P. & Rényi, A. *On the strength of connectedness of a random graph.* Acta Math. Acad. Sci. Hungar. 12, (1961) 261-267.

[4] Hardy, G.H., Littlewood, J.E. & Pólya, G. *Inequalities.* Cambridge University Press (reprinted 1989).

[5] Holst, L., *On the number of vertices in the complete graph with a given vertex as nearest neighbour.* Random Graphs, Volume 2 (Edited by A.M.Frieze and T.Łuczak), 91-99, John Wiley, 1992.

[6] McDiarmid C. *On the method of bounded differences.* LMS Surveys in Combinatorics, 1989. 148-184 Cambridge University Press.

[7] Newman, C.M., Rinott, Y. & Tversky, A. *Nearest neighbours and Voronoi regions in certain point processes.* Advances in Applied Probability 15 (1983) 726-751.