Longest cycles in sparse random digraphs

Michael Krivelevich* Eyal Lubetzky† Benny Sudakov‡

Abstract

Long paths and cycles in sparse random graphs and digraphs were studied intensively in the 1980's. It was finally shown by Frieze in 1986 that the random graph $\mathcal{G}(n,p)$ with $p = c/n$ has a cycle on at all but at most $(1+\varepsilon)ce^{-c}n$ vertices with high probability, where $\varepsilon = \varepsilon(c) \to 0$ as $c \to \infty$. This estimate on the number of uncovered vertices is essentially tight due to vertices of degree 1. However, for the random digraph $\mathcal{D}(n,p)$ no tight result was known and the best estimate was a factor of $c/2$ away from the corresponding lower bound. In this work we close this gap and show that the random digraph $\mathcal{D}(n,p)$ with $p = c/n$ has a cycle containing all but $(2+\varepsilon)e^{-c}n$ vertices w.h.p., where $\varepsilon = \varepsilon(c) \to 0$ as $c \to \infty$. This is essentially tight since w.h.p. such a random digraph contains $(2e^{-c} - o(1))n$ vertices with zero in-degree or out-degree.

1 Introduction

In this paper we consider long cycles in random directed graphs, aiming to obtain estimates analogous to those derived for the undirected case. Formally, a random graph $\mathcal{G}(n,p)$ is a probability space of all graphs with vertex set $[n]$, where each pair of vertices $1 \le i < j \le n$ is an edge of $G \sim \mathcal{G}(n,p)$ independently and with probability p. The model of random directed graphs $\mathcal{D}(n,p)$ is defined as the probability space of all directed graphs with vertex set $[n]$ (without loops and without parallel edges, but possibly with anti-parallel edges), where each ordered pair (i,j), with $1 \le i \ne j \le n$, is a directed edge of $D \sim \mathcal{D}(n,p)$ independently and with probability p.

The existence of long paths and cycles in sparse random graphs was a subject of very intensive study in the eighties. Ajtai, Komlós and Szemerédi proved in [1] that with high probability[1] in the random graph $\mathcal{G}(n,c/n)$ there is a path of length $\alpha(c)n$, where $\alpha(c) > 0$ for $c > 1$ and $\lim_{c\to\infty} \alpha(c) = 1$; a similar but somewhat weaker result was proved independently by Fernandez de la Vega [7]. Then the attention has shifted to estimating the asymptotic behavior of the number of vertices uncovered by a longest path/cycle. Improving upon prior results of Bollobás [5] and Bollobás, Fenner and Frieze [6], Frieze has finally settled this problem: He showed in [9] that w.h.p. $G \sim \mathcal{G}(n,c/n)$ contains a cycle covering all but at most $(1+\varepsilon)ce^{-c}n$ vertices, where $\lim_{c\to\infty} \varepsilon(c) = 0$. This estimate is easily seen to

*School of Mathematical Sciences, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel Aviv 69978, Israel. Email address: krivelev@post.tau.ac.il. Research supported in part by USA-Israel BSF Grant 2006-322 and by grant 1063/08 from the Israel Science Foundation.

†Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399, USA. Email address: eyal@microsoft.com.

‡Department of Mathematics, UCLA, Los Angeles, CA 90095, USA. Email: bsudakov@math.ucla.edu. Research supported in part by NSF CAREER award DMS-0812005 and by a USA-Israeli BSF grant.

[1]We say that a sequence of events (A_n) in a random (di)graph model occurs with high probability, or w.h.p. for brevity, if the probability of A_n tends to 1 as the number of vertices n tends to infinity.

be asymptotically tight as $\mathcal{G}(n, c/n)$ w.h.p. contains $(1 + o(1))ce^{-c}n$ vertices of degree at most 1, all of which have to be missed by a cycle.

For random directed graphs the situation appears to be more complicated. This is to be expected as the research experience of many years has shown that problems related to long paths and cycles in directed (random) graphs are usually much more challenging than their undirected counterparts. In the aforementioned paper [9] Frieze further established that w.h.p. $\mathcal{D}(n, p)$ contains a cycle covering all but at most $(1 + \varepsilon)ce^{-c}n$ vertices, where $\lim_{c\to\infty} \varepsilon(c) = 0$. This result was derived by appealing to a general theorem of McDiarmid [11], coupling between events in $\mathcal{G}(n, p)$ and in $\mathcal{D}(n, p)$. Unlike in the undirected case, the above estimate on the number of vertices uncovered by a longest cycle is no longer asymptotically tight — the unavoidable loss in the directed case are vertices of in-degree or out-degree zero, whose number is easily seen to be asymptotic to $2e^{-c}n$.

In this paper we close the gap left by Frieze's work and obtain an asymptotically optimal result about longest cycles in sparse random digraphs.

Theorem 1. *Let $D \sim \mathcal{D}(n, p)$ be a random digraph with edge probability $p = c/n$ for fixed $c > 1$. Then w.h.p. D contains a directed cycle that covers all but at most $(2+\varepsilon)e^{-c}\,n$ vertices, where $\varepsilon = \varepsilon(c) \to 0$ as $c \to \infty$, and this is asymptotically tight as w.h.p. $(2e^{-c} - o(1))n$ vertices of D have zero in-degree or out-degree.*

The proof of the theorem is given in the next section. In certain similarity to Frieze's argument in [9] we proceed by first filtering out vertices of zero in-degree or out-degree as well as some vertices close to them. The so obtained digraph typically retains all but a negligible fraction of the vertices of positive in-degrees and out-degrees; it is then upgraded to another random digraph, containing an almost spanning cycle, by sprinkling a few more random directed edges.

Before we embark into the technicalities of the proof, we provide its outline, aiming to help the reader to parse the proof's details.

The proof has two components/stages: filtering and factoring. The filtering stage aims to filter out vertices of in- or out-degree zero and possibly some other vertices and to produce an induced subgraph D_0 of $D \sim \mathcal{D}(n, p)$, containing most of the vertices of positive degree; moreover, D_0 is constructed in a way making it rather straightforward to show that it typically contains a factor of directed cycles. In order to produce D_0, we define the following iterative process. Let $Y = \{v : d_D(v) = 0\}$, and let $Z = \{v : d_D(v) \le 3\}$, where $d_D(v) = \min\{d_D^+(v), d_D^-(v)\}$. We start with $X = \emptyset$, and then for $k \ge 1$ we obtain X_k by including vertices not in $\bigcup_{i<k} X_i$, lying on a short path connecting two vertices $x, y \in \bigcup_{i<k} X_i \cup Z$. We repeat this process till it stabilizes and set $X = \bigcup_k X_k$. Finally, the subgraph D_0 is defined by $D_0 = D[V - (X \cup Y)]$. Observe that for a given vertex v the probability that $d_D(v)$ is at most some absolute constant (independent of c) is at most $\mathrm{poly}(c)e^{-c}$. Thus, for a given v the probability of having two vertices of small degree at a constant distance from v is $\mathrm{poly}(c)e^{-2c}$. Hence, we can expect to eventually have $|X| \le \mathrm{poly}(c)e^{-2c}n$, and this is indeed what we prove. To facilitate the proof, we first get rid of short cycles (of length $O((1/c)\log n)$) in the underlying undirected graph G of D — they typically touch very few vertices. Analyzing the filtering process in a large girth graph is easier — for every $v \in X_k$ there should be an evidence for its association with X_k in the form of a tree T_v rooted at v, of prescribed order, depth and with ℓ leaves, where $k \le \ell \le 2^k$. This is proven in Lemma 2.6. Using this lemma we can bound the size of the set X_k (or rather of a set X'_k closely related to X_k and defined through an analogous filtering process) by the number of labeled rooted

trees meeting these requirements. This is done by first truncating unusually deep trees (Lemma 2.7) and then bounding from above the expected number of trees of bounded depth. The final argument invokes martingales to show the concentration of the corresponding random variable around its mean and to bound its upper tail. All this is done in Theorem 2.2.

The factoring stage takes the induced subgraph D_0, the output of the filtering stage, as an input. By Theorem 2.2 we know that with high probability D_0 contains all but a suitably small part of the vertices of D of positive degree. Moreover, one can prove (Lemma 2.3) that all vertices in D_0 have positive degree, and in addition every two vertices u, v with $d_{D_0}(u), d_{D_0}(v) \leq 2$ are at undirected distance at least 5. We then form an auxiliary bipartite graph H_0 with parts L, R, corresponding to two copies of the vertices of D_0, where an edge $(x, y) \in D_0$ becomes an edge $x_L y_R \in E(H_0)$. It is quite easy to see that the existence of a perfect matching in H_0 implies the existence of a spanning subgraph of D_0 composed of directed cycles. The probable existence of a perfect matching in H_0 is shown in Lemma 2.4 using Hall's condition and standard density/expansion arguments for random (di)graphs. The next step is to trade the factor of directed cycles in D_0 for one nearly spanning cycle using extra random edges, about $O(n/\sqrt{\log n})$ of them – a negligible quantity easily absorbed into the original random digraph. This is done using rather standard random graph arguments and extremal statements guaranteeing the existence of a long cycle in a highly connected digraph (see Lemma 2.5). The factoring stage is treated in Theorem 2.1.

The next section contains the full details of the proof of the main result, and is followed by concluding remarks in Section 3.

2 Proof of main result

2.1 Filtering and factoring

If D is a directed graph we use the notation $d_D(v)$ to denote $\min\{d_D^+(v), d_D^-(v)\}$. Similarly, we let $N_D(v) = N_D^+(v) \cup N_D^-(v)$ and in both cases may omit the subscript D when there is no danger of confusion.

For an undirected graph G and a special subset of its vertices Z we define a filtering process which produces a sequence $\{X_k\}$ of disjoint subsets of the vertices as follows:

$$
\begin{aligned}
X_0 &= \emptyset\,, \\
X_k &= \left\{ v \notin \bigcup_{j<k} X_j : \begin{array}{l} \exists\, x, y \in \left(\bigcup_{j<k} X_j\right) \cup Z \text{ such that } x \neq y \text{ and} \\ v \text{ is on a path of length } l \leq 4 \text{ between } x, y, \text{ i.e.} \\ v \in \{x = u_0, u_1, \ldots, u_l = y\} \text{ with } u_i u_{i+1} \in E(G). \end{array} \right\} \text{ for } k \geq 1\,, \\
X &= \bigcup_k X_k\,. \qquad (2.1)
\end{aligned}
$$

The first ingredient in the proof is showing that w.h.p., once we filter the set X from the graph along with the vertices with zero in/out degree, the remaining vertices may be factored into large cycles and thereafter combined into a single long cycle while losing only a negligible number of vertices in the process. This is shown in the next theorem whose proof appears in Section 2.2.

Theorem 2.1. *Let $D \sim \mathcal{D}(n,p)$ where $p = \frac{c}{n}$ for $c > 1$ fixed and let G be the undirected underlying graph of D. Let $Y = \{v : d_D(v) = 0\}$, $Z = \{v : d_D(v) \leq 3\}$, and set $X(G,Z)$ as in (2.1). Let*

D_0 be the induced subgraph of D on $V(D) \setminus (X \cup Y)$ and let D_0' be its union with a random digraph $\mathcal{D}(|D_0|, (n\sqrt{\log n})^{-1})$. If $|D_0| > n/5$ then w.h.p. D_0' contains a directed cycle on $|D_0| - o(n)$ vertices.

The following theorem, which we prove in Section 2.3, estimates the size of the filtered subset X w.r.t. vertices of low in/out degree in D.

Theorem 2.2. *Let $D \sim \mathcal{D}(n,p)$ be a random digraph with edge probability $p = c/n$ for fixed $c > 1$. Let $Z = \{v : d_D(v) \leq 3\}$ and define $X = X(G,Z)$ as in (2.1) where G is the undirected underlying graph of D. If c is sufficiently large then with high probability $|X| \leq (2c)^{10}e^{-2c}n$.*

From the above two theorems we can immediately derive our main result.

Proof of Theorem 1. Let $Y = \{v : d_D(v) = 0\}$. Note that w.h.p. $|Y| = (2e^{-c} + o(1))n$ since $d^+(v), d^-(v) \sim \mathrm{Bin}(n-1, c/n)$. For a sufficiently large c we obtain from Theorem 2.2 that w.h.p. $|X \cup Y| \leq (2e^{-c} + (2c)^{10}e^{-2c} + o(1))n$. In particular, for large c and n we have that w.h.p. D_0, the induced subgraph on $V(D) \setminus (X \cup Y)$, has at least $n/5$ vertices (with room to spare) and we deduce from Theorem 2.1 that w.h.p. $\mathcal{D}(n,p')$ has a cycle missing at most $|X \cup Y| + o(n)$ vertices, where $p' = (c/n) + (n\sqrt{\log n})^{-1} = (c + o(1))/n$. This establishes the required result for a choice of, say, $\varepsilon(c) = 2(2c)^{10}e^{-2c}$, which makes up for $|X|$ with an extra factor of 2 that readily absorbs the additive $o(n)$-term in $|X \cup Y|$ as well as the $o(1)$-term in p'. ■

2.2 Long cycles in the filtered graph

To prove Theorem 2.1 we first need to establish several properties of the graph D_0 stemming from the definition of X and the geometry of the random digraph D.

Lemma 2.3. *Let D_0 be the induced subgraph on $V(D)\setminus(X\cup Y)$ and let G_0 be its undirected underlying graph. Then*

(i) Every $u \in D_0$ has $d_{D_0}(u) \geq 1$.
(ii) Every $u, v \in D_0$ with $d_{D_0}(u), d_{D_0}(v) \leq 2$ have $\mathrm{dist}_{G_0}(u,v) \geq 5$.

Proof. To prove Part (i) assume that some $u \in V(D_0)$ has $d_{D_0}(u) = 0$ and assume without loss of generality that $d^+_{D_0}(u) = 0$.

First consider the case where $d^+_D(u) \geq 2$. Observe that in this case there exist distinct $x, y \notin D_0$ such that $(u,x), (u,y) \in E(D)$. Thus u is on a path of length 2 between $x, y \in X \cup Y \subset X \cup Z$, implying that $u \in X$ by definition and contradicting the fact that $u \in V(D_0)$. The case where $d^+_D(u) = 1$ is treated similarly: Here there is some vertex $v \notin V(D_0)$ such that $(u,v) \in E(D)$, $v \in X \cup Y \subset X \cup Z$, and furthermore $u \in Z$ by definition. Hence, u is on a path of length 1 between two distinct vertices $u \neq v$ in $X \cup Z$ and must thus also belong to X, in contradiction to the fact that $u \in V(D_0)$.

To prove Part (ii) let u, v be vertices satisfying $d_{D_0}(u) \leq 2$ and $d_{D_0}(v) \leq 2$. If $d_D(u) \geq 4$ then it necessarily lost at least 2 neighboring (in/out) vertices in $X \cup Y$ and hence must also belong to X. We thus conclude that $d_D(u) \leq 3$ and similarly that $d_D(v) \leq 3$.

Let G be the underlying undirected graph of D. Recalling that $u, v \in Z$ by the definition of Z, there cannot be a path of length at most 4 between u, v in G, as such a path would imply that u, v must both belong to X. In particular, the induced subgraph $G_0 \subset G$ also satisfies $\mathrm{dist}_{G_0}(u,v) \geq 5$, completing the proof. ■

Lemma 2.4. *Let H_0 be the undirected bipartite graph whose parts (L, R) correspond each to the vertices of D_0 and where $x_L y_R \in E(H_0)$ iff $(x, y) \in E(D_0)$. Then w.h.p. H_0 has a perfect matching.*

Proof. Recall that by Part (i) of Lemma 2.3 there are no isolated vertices in H_0 (neither in L nor in R). Furthermore, by Part (ii) of that lemma we know that if $x, y \in L$ have degree 1 in H_0 then $N(x) \cap N(y) = \emptyset$ (otherwise D_0 would have two vertices with out-degree 1 and an undirected distance of at most 2 between them) and similarly for $x, y \in R$ with degree 1 in H_0. In other words, if we denote by M_0 the set of edges incident to degree 1 vertices in H_0 then M_0 consists of vertex disjoint edges. Let H_1 denote the bipartite graph obtained by deleting the vertices of M_0 from H_0, i.e. $H_1 = H_0 \setminus V(M_0)$. We now claim that H_1 has minimum degree at least 2. To see this, suppose that $d_{H_1}(u) \leq 1$ and argue as follows.

First, we must have $d_{H_0}(u) > 1$ otherwise $u \in V(M_0)$ and hence does not belong to H_1. If $d_{H_0}(u) = 2$ then there must be some $w \in V(M_0)$ such that $uw \in E(H_0)$. In particular, either w has degree 1 in H_0 or it is a neighbor of such a vertex, and either way we have that there exists some degree-1 vertex $v \in H_0$ whose distance from u is at most 2. The vertices corresponding to u and v in D_0 thus satisfy $d_{D_0}(u) \leq 2$ and $d_{D_0}(v) \leq 1$ while the undirected distance between them is at most 2, contradicting Part (ii) of Lemma 2.3.

It thus remains to treat the case $d_{H_0}(u) \geq 3$. In this case u has two neighbors $w_1, w_2 \in V(M_0)$, giving rise to $v_1, v_2 \in V(M_0)$ whose distance from u is at most 2 and with $d_{H_0}(v_1) = d_{H_0}(v_2) = 1$. These correspond to two vertices v_1, v_2 in D_0 satisfying $d(v_1), d(v_2) \leq 1$ while the undirected distance between them is at most 4, again contradicting Part (ii) of Lemma 2.3.

We have thus obtained that H_1 has a minimum degree of 2, and will now derive from this fact the existence of a perfect matching on H_1. It suffices to show that w.h.p. every set $S \subset V(H_1) \cap L$ of size at most $n/2$ has $|N(S)| \geq |S|$, as the same conclusion will carry by symmetry to all sets $S \subset V(H_1) \cap R$ of size at most $n/2$, which would in turn imply Hall's condition for sets $S \subset V(H_1) \cap L$ of size larger than $n/2$.

Let S be a subset of $V(H_1) \cap L$ of size $s \leq n/5$ let $T = N(S)$ in H_1 and assume that T has size $t < s$. Identifying these vertices with those of the original digraph D we have that $e(S, T) \geq 2s$ by definition of H_1 and the fact that it has minimum degree 2.

Moreover, observe that every $u \in S$ has at most 2 neighbors in $V(D) \setminus T$. Indeed, since T includes all the neighbors of S corresponding to vertices of H_1, any other neighbor $v \in N_D^+(u) \setminus T$ must belong either to $X \cup Y$ or to the vertices corresponding to $V(M_0)$, and these satisfy:

1. The vertex u cannot have two distinct neighbors in $X \cup Y$ otherwise it would belong to X by definition and hence would be excluded from D_0.
2. The vertex u cannot have two distinct neighbors in $V(M_0)$ otherwise there would exist some x, y with $d_{D_0}(x) = d_{D_0}(y) = 1$ and $\mathrm{dist}_{G_0}(x, y) \leq 4$, contradicting Part (ii) of Lemma 2.3.

Combining these arguments we conclude that $|N_D^+(u) \setminus T| \leq 2$, and note that for a given vertex u and subset T the probability of this event is at most

$$\mathbb{P}(\mathrm{Bin}(n-t, p) \leq 2) \leq 3\binom{n-t}{2} p^2 (1-p)^{n-t-2} \leq 2c^2 e^{-\frac{4}{5}np},$$

where the last inequality used the fact $t < s \leq n/5$ and holds for any sufficiently large n as the $(1 + O(p))$-factor was absorbed into the leading constant. Further note that the event that $|N_D^+(u) \setminus$

$T| \leq 2$ depends only on the edges from u to T and therefore for distinct vertices these events are independent.

At this point, the following straightforward first moment argument shows that w.h.p. D cannot contain sets S, T of the above sizes where S has at least $2|S|$ edges going to T and every $u \in S$ has at most 2 edges going elsewhere. Indeed, the probability that such sets exist in D for any given such s, t is at most

$$\binom{n}{s}\binom{n}{t}\binom{st}{2s}p^{2s}\left(2c^2e^{-\frac{4}{5}np}\right)^s \leq \left[\frac{en}{s}\left(\frac{en}{t}\right)^{t/s}\left(\frac{et}{2}\right)^2\frac{c^2}{n^2}2c^2e^{-4pn/5}\right]^s \leq \left[(e^4/2)c^4e^{-4c/5}(t/n)^{1-\frac{t}{s}}\right]^s$$
$$\leq \left[(e^4/2)c^4e^{-4c/5}\right]^s\frac{s-1}{n} =: \Delta(s,t)\,,$$

where we used the inequality $\binom{a}{b} \leq (ea/b)^b$ and the fact that $t < s \leq n/5$. For large enough c we have $c^4e^{-4c/5} < 2e^{-5}$ and so

$$\Delta(s,t) < e^{-s}(s-1)/n\,,$$

and summing over the possible values of s, t now gives that

$$\sum_{t<s\leq n/5}\Delta(s,t) = \sum_{t<s\leq 2\log n}\Delta(s,t) + \sum_{\substack{2\log n\leq s\leq n/5\\ t<s}}\Delta(s,t)$$
$$\leq (2\log n)^2\frac{2\log n}{n} + (n/5)^2e^{-10\log n} = o(1)\,.$$

It remains to treat sets S of size $n/5 < s \leq n/2$. Verifying Hall's condition for such sets follows immediately form that the fact that w.h.p. every two sets S, T of size $n/5$ in D have an edge from S, T, as the following calculation shows:

$$\binom{n}{n/5}^2(1-p)^{(n/5)^2} \leq \left[(5e)^2e^{-c/5}\right]^{n/5} = o(1)\,,$$

where the last inequality holds for a sufficiently large c. ■

We are now in a position to prove Theorem 2.1.

Proof of Theorem 2.1. The edges of the matching provided w.h.p. by Lemma 2.4 correspond to a spanning subgraph of D_0 comprised of disjoint directed cycles. Our first step is to delete from D_0 all cycles of length less than $\frac{1}{2}\log_c n$. Note that the number of vertices participating in such cycles in the original digraph D is w.h.p. at most

$$\sum_{l<\frac{1}{2}\log_c n} n^lp^l \leq (\log_c n)\sum_{l<\frac{1}{2}\log_c n} c^l = O(n^{1/2}\log n) = o(n)\,.$$

The remaining disjoint directed cycles, denoted by $C_1, \ldots, C_m$, thus contain $|D_0| - o(n)$ vertices. Note also that the total number of cycles m satisfies: $m \leq n/(\frac{1}{2}\log_c n) = O(n/\log n)$.

Let $\mathcal{P} = \{P_i\}_{i=1}^t$ be a maximum collection of vertex disjoint directed paths, each of length precisely $\lceil \log^{0.9} n \rceil$, formed from the edges of $\{C_i\}_{i=1}^m$. Since the number of vertices uncovered by $\mathcal{P}$ in each C_i is at most $\lfloor \log^{0.9} n \rfloor$ it follows that $\mathcal{P}$ covers all but at most $m\log^{0.9} n = O(n/\log^{0.1} n)$ vertices of D_0. Furthermore, recalling that $n/5 \leq |D_0| \leq n$, this implies that

$$(\tfrac{1}{5} - o(1))n/\log^{0.9} n \leq t \leq n/\log^{0.9} n\,. \tag{2.2}$$

For each path $P_j \in \mathcal{P}$ define its prefix A_j and suffix B_j to be its first $L = \lfloor \log^{0.8} n \rfloor$ vertices and last L vertices, respectively. Consider now the digraph $D_1 = \mathcal{D}(|D_0|, (n\sqrt{\log n})^{-1})$. We will use the edges of D_1 to weave most of the vertices covered by $\mathcal{P}$ into a long directed cycle, using the edges of the paths P_j as a backbone. Define an auxiliary digraph H where the vertex set $[t]$ corresponds to the paths $P_1, \ldots, P_t$ and $(i,j) \in E(H)$ iff D_1 contains an edge from B_i to A_j. Notice that if H contains a directed cycle $C = (i_1, \ldots, i_l)$ then $D_0 \cup D_1$ contains a directed cycle of length at least $l(\log^{0.9} n - 2\log^{0.8} n)$, obtained as follows: Start at the last vertex of A_{i_1} and proceed with the vertices/edges along P_{i_1}; use an edge from B_{i_1} to A_{i_2} to jump to P_{i_2}, then traverse the vertices/edges along P_{i_2} till an edge from B_{i_2} to A_{i_3} and so on; finally use an edge from B_{i_l} to A_{i_1} and possibly some edges of A_{i_1} to close the cycle.

The digraph H is a random digraph on $t = \Theta(n/\log^{0.9} n)$ vertices with edge probability ρ that satisfies $1 - \rho = (1 - (n\sqrt{\log n})^{-1})^{L^2}$, implying that $\rho = (1 + o(1))\frac{\log^{1.1} n}{n}$. We thus need to prove that such a random digraph contains w.h.p. an almost spanning cycle. This is an established fact, and here we derive it from the following lemma of [3], whose short proof is included for completeness.

Lemma 2.5 ([3])**.** *Let $D = (V, E)$ be a directed graph on t vertices in which for every ordered pair A, B of disjoint vertex subsets $A, B \subset V$ of size $|A| = |B| = k$ there is an edge from A to B. Then D contains a path of length at least $t - 2k$ and a cycle of length at least $t - 4k$.*

Proof. Fix an arbitrary order σ on the vertices of D and run the DFS (Depth First Search) on D, guided by σ. The DFS maintains three sets of vertices: Let S be the set of vertices which we have completed exploring, T be the set of unvisited vertices, and $U = V(G) - (S \cup T)$, where the vertices of U are kept in a stack (a last in, first out data structure). The DFS starts with $S = U = \emptyset$ and $T = V(D)$, and at each stage moves a vertex from T to U (an unvisited vertex with an incoming edge from the top of the stack U) or from U to S until eventually all vertices are in S. As such, at some point in the course of the algorithm we must have $|S| = |T|$; consider that point, and observe crucially that all the vertices in U form a directed path, and that there are no edges from S to T. We conclude that $|S| = |T| \leq k - 1$, and therefore $|U| \geq t - 2k + 2$, so there is a directed path with $t - 2k + 1$ edges in D, as required. To get a directed cycle of the desired length, take a path as above and use a directed edge from its last k vertices to its first k vertices to close a cycle. ■

In order to apply the above lemma, take $k = \lfloor n/\log n \rfloor$ while recalling that $H \sim \mathcal{D}(t, \rho)$ with t satisfying (2.2) and $\rho = (1 + o(1))\frac{\log^{1.1} n}{n}$. As $t \leq n/\log^{0.9} n$, the probability that H has two disjoint vertex sets A, B of cardinality k each with no edges from A to B is at most

$$\binom{t}{k}(1-\rho)^{k^2} \leq \left[(et/k)\, e^{-\rho k}\right]^k \leq \left[(e + o(1))\big(\log n\big)^{0.1} \cdot e^{-(1-o(1))\log^{0.1} n}\right]^k = o(1)\,,$$

thus w.h.p. H satisfies the conditions of Lemma 2.5 and in turn it contains w.h.p. a cycle of length at least $t - 4k = (1 - o(1))t$. As explained above, it follows that w.h.p. the digraph $D_0 \cup D_1$ contains a directed cycle covering all but $O(k\log^{0.9} n) + O(n/\log^{0.1} n) = o(n)$ vertices, as required. ■

2.3 Controlling the effect of the filtering process

Proof of Theorem 2.2. An important element in the proof would be to analyze the set X with respect to a subgraph of $D \sim \mathcal{D}(n, p)$ with a reasonably large undirected girth. To this end we need the following lemma.

Lemma 2.6. *Let G be an undirected graph with girth g and let Z be a subset of its vertices. Define $X(G,Z)$ as in (2.1). For every $1 \leq k \leq g/8$ and $v \in X_k$ there is a tree $T_v \subset G$ rooted at v whose leaves are in Z and interior vertices are in $\bigcup_{j<k} X_j$. Moreover, T_v has at most $5(|T_v \cap Z| - 1)$ vertices, at most $4k$ levels (including the root) and its number of leaves ℓ satisfies $k < \ell \leq 2^k$.*

Proof. We proceed by induction on k. For the induction base recall that if $v \in X_1$ then there are 2 vertices $x, y \in Z$ such that v is on a path of length at most 4 between x, y in G. Treat this path as a tree T_v rooted at v, and notice that it has 2 leaves, at most 4 levels including the root (as $\mathrm{dist}(v,x), \mathrm{dist}(v,y) \leq 3$) and the induced subgraph on it in G is a tree by the girth assumption on G. Furthermore, T_v has at most $5 \leq 5(|T_v \cap Z| - 1)$ vertices since $|T_v \cap Z| \geq 2$, thus satisfying the statement of the lemma.

Next, let $k > 1$ and let $v \in X_k$. Let $x, y \in Z \cup \bigcup_{j<k} X_j$ be the endpoints of a shortest path P containing v (by definition (2.1) the path P has length at most 4). Suppose first that one of these vertices belongs to Z, i.e. without loss of generality $x \in Z$ whereas $y \in X_{k-1}$ (otherwise v would have belonged to some X_j with $j < k$). Define the tree T_v as a path P_y of length $\mathrm{dist}_G(v,y)$ from the root v to the sub-tree T_y, provided by the induction, together with another path P_x of length $\mathrm{dist}_G(v,x)$ from v to x. On one hand, the paths P_x, P_y are disjoint by definition, and furthermore, excluding their endpoints, their vertices do not belong to $Z \cup \bigcup_{j<k} X_j$ by the minimality of P and in particular do not belong to T_y. On the other hand, if the path P_x does intersects T_y, which was guaranteed to have at most $4(k-1)$ levels by induction, then together with P_y they complete a cycle of length at most $4(k-1)+4 < 4k \leq g/2$ in G contradicting its girth assumption. We conclude that T_v is indeed a tree, with at most $4(k-1)+4 = 4k$ levels including the root. Finally, $|T_v \cap Z| = |T_y \cap Z| + 1$, hence the induction hypothesis and the fact that T_v adds at most 4 vertices to T_y together imply that

$$|T_v| \leq |T_y| + 4 \leq 5(|T_y \cap Z| - 1) + 4 < 5(|T_v \cap Z| - 1)\,.$$

It remains to treat the case where $x \in X_j$ for some $j < k$ while $y \in X_{k-1}$. As before, if $T_x \cap T_y \neq \emptyset$ then together with the path P we obtain a cycle of length at most $8(k-1)+4 < 8k \leq g$ in G, contradicting the girth assumption. Otherwise, $|T_v \cap Z| = |T_x \cap Z| + |T_y \cap Z|$ and so our hypothesis on T_x, T_y gives that

$$|T_v| \leq |T_x| + |T_y| + 3 \leq 5(|T_x \cap Z| - 1) + 5(|T_y \cap Z| - 1) + 3 < 5(|T_v \cap Z| - 1)\,.$$

Noting that the number of leaves $\ell(T_v)$ was either $\ell(T_y)+1$ or $\ell(T_x)+\ell(T_y)$ immediately implies that $k+1 \leq \ell(T_v) \leq 2^k$ and completes the proof of the lemma. ■

Let G denote the undirected underlying graph of $D \sim \mathcal{D}(n,p)$, and define $\mathcal{C} \subset V(G)$ to be comprised of all vertices that belong to cycles of length at most

$$R = (20/c)\log n$$

in G. Since each edge appears in G with probability at most $2p$ independently of other edges, the expected number of cycles of length r in G is at most $n^r(2p)^r/r$ and thus

$$\mathbb{E}|\mathcal{C}| \leq \sum_{r<R} (2c)^r \leq \frac{(2c)^R}{c-1} < n^{1/5}\,,$$

where we used the fact that $c/\log(2c) > 100$ for sufficiently large c. In particular, $|\mathcal{C}| < n^{1/4}$ w.h.p.

Define $Z' = \mathcal{C} \cup Z$ and let D' be the graph obtained by deleting all inner edges between vertices of $\mathcal{C}$ (i.e. all edges of the induced subgraph on $\mathcal{C}$). Let G' denote the undirected underlying graph of D' and for all k let X'_k denote the set $X_k(G', Z')$ defined via (2.1). A key observation is that

$$X \subset X' \cup \mathcal{C}\,. \tag{2.3}$$

To see this, recall that $X_0 = \emptyset$ and assume by induction that

$$\textstyle\bigcup_{j<k} X_j \subset \Big(\big(\bigcup_{j<k} X'_j \big) \cup \mathcal{C} \Big) \text{ for some } k \geq 1\,.$$

Let $v \in X_k \backslash \mathcal{C}$. Let P be a shortest path containing v in G with endpoints $x, y \in \big(\bigcup_{j<k} X_j\big) \cup Z$. By the definition of X_k and the minimality of P we know that P has $1 \leq l \leq 4$ edges and none of its interior vertices belongs to $\big(\bigcup_{j<k} X_j\big) \cup Z$. Consider the two sub-paths from v to x, y (one of which is possibly empty) and let x', y' be the first vertices on these respective paths that belong to $\big(\bigcup_{j<k} X_j\big) \cup \mathcal{C} \cup Z$. Since x, y clearly belong to this set, this defines a sub-path P' of length $1 \leq l' \leq l \leq 4$ that contains v in G. Crucially, since $v \notin \mathcal{C}$ the path P' has no interior vertices in $\mathcal{C}$ and therefore all of its edges belong to G'. Finally, the induction hypothesis ensures that $x', y' \in \big(\bigcup_{j<k} X'_j\big) \cup Z'$ and we conclude that $v \in \big(\bigcup_{j\leq k} X'_j\big) \cup Z'$, completing the induction.

Our next goal is to provide an upper bound on X' which is linear in n (with a suitably small coefficient), absorbing the negligible contribution to it from vertices in $\mathcal{C}$. Observe that by definition the girth of G' is larger than $R = (20/c)\log n$ and set

$$K = \lfloor (2/c)\log n \rfloor\,. \tag{2.4}$$

Invoking Lemma 2.6 w.r.t. G', for each $k \leq K$ we can bound $|X'_k|$ from above by $|\mathcal{T}'_k|$ where

$$\mathcal{T}'_k = \left\{ T \subset G' : \begin{array}{l} \text{labeled rooted tree with } \ell \text{ leaves, } k < \ell \leq 2^k \text{, all belonging to } Z', \\ \text{a total of } t \text{ vertices for } t \leq 5(\ell-1) \text{ and at most } 4k \text{ levels.} \end{array} \right\}. \tag{2.5}$$

In particular, we will be able to assert that $X'_K = \emptyset$ by showing that $\mathcal{T}'_K$ is empty, as the next lemma establishes.

Lemma 2.7. *Set K as in (2.4). With high probability $X'_K = \emptyset$.*

Proof. In what follows let $L(T)$ denote the set of leaves of a tree T and recall that if $T \in \mathcal{T}'_k$ then $L(T) \subset Z' = \mathcal{C} \cup Z$ by definition.

Let $\mathcal{T}^*_k$ be the set of all trees in $\mathcal{T}'_k$ where at least $\ell - 1$ of the leaves belong to Z. We have $\binom{n}{t}$ choices for the vertices of $T \in \mathcal{T}^*_k$ on t vertices, and the well-known Cayley formula asserts that the number of labeled rooted trees on t vertices is t^{t-1}. The probability that a given labeled tree on t vertices is in G (an upper bound on the probability it belongs to $G' \subset G$) is exactly $(2p)^{t-1}$. Finally, if $u \in L(T) \cap Z$ then by definition $d_G(u) \leq 3$ and in particular $d_{G\backslash T}(u) \leq 3$. Crucially, the events $\{d_{G\backslash T}(u) \leq 3\}$ for $u \in L(T)$ are mutually independent as well as independent of all the interior edges

of T (accounted for in the probability that $T \subset G$). Altogether, for all $k \le K$,

$$\mathbb{E}|\mathcal{T}_k^*| \le \sum_{\ell=k+1}^{2^k} \sum_{t \le 5(\ell-1)} \binom{n}{t} t^{t-1} (2p)^{t-1} \ell \left(2\mathbb{P}\big(\mathrm{Bin}(n-t,p) \le 3\big)\right)^{\ell-1} .$$

$$\le \sum_{\ell=k+1}^{2^k} \sum_{t \le 5(\ell-1)} e\,(2ec)^{t-1} \left(2\mathbb{P}\big(\mathrm{Bin}(n-t,p) \le 3\big)\right)^{\ell-1} n\,, \tag{2.6}$$

where in the last inequality we used the facts that $\binom{n}{t} \le (en/t)^t$ and $t > \ell$. If c is sufficiently large then $n - t = (1 - o(1))n$ as $t < 5 \cdot 2^K = o(n)$ and in particular $\mathbb{P}\big(\mathrm{Bin}(n-t,p) \le 3\big) \le \frac{1}{2}c^3 e^{-c}$. Plugging this in (2.6) gives that for sufficiently large n,

$$\mathbb{E}|\mathcal{T}_k^*| \le \sum_{\ell=k+1}^{2^k} \sum_{t \le 5(\ell-1)} (2ec)^t \left(c^3 e^{-c}\right)^{\ell-1} n$$

$$\le \sum_{\ell=k+1}^{2^k} \left((2e)^5 c^8 e^{-c}\right)^{\ell-1} n \le n e^{-\frac{3}{4}ck}\,, \tag{2.7}$$

where the last inequality holds for large enough c and n. Substituting $k = K = \lfloor (2/c)\log n \rfloor$ now gives

$$\mathbb{E}|\mathcal{T}_K^*| \le n e^{-\frac{3}{4}cK} = O(n^{-1/2}) = o(1)\,,$$

hence w.h.p. $\mathcal{T}_K^* = \emptyset$.

Now consider $T \in \mathcal{T}'_K \setminus \mathcal{T}_K^*$. Here there exist distinct $u_i, u_j \in L(T) \cap \mathcal{C}$. As T has at most $4K$ levels and connects u_i, u_j in G' (where the inner edges between the vertices of $\mathcal{C}$ are absent) this implies the existence of a subgraph $F \subset G$ with m vertices and at least $m+1$ edges such that

$$m \le 2R + 8K + 2 \le (60/c)\log n$$

(accounting for u_i and u_j, a path of length at most $8K$ between them and up to 2 cycles in $\mathcal{C}$, with the last inequality holding for large enough c). When c is sufficiently large, the probability that such a graph F belongs to G is at most

$$\binom{n}{m}\binom{\binom{m}{2}}{m+1}(2p)^{m+1} \le \left(\frac{en}{m}\right)^m \left(\frac{em}{2}\right)^{m+1} (2c/n)^{m+1} \le \frac{m}{n}\big(e^2 c\big)^{m+1} = O(1/\sqrt{n}) = o(1) \tag{2.8}$$

implying that $\mathcal{T}'_K \setminus \mathcal{T}_K^*$, and hence also $\mathcal{T}'_K$, is w.h.p. empty. By (2.5) and the remark following that definition it now follows that $X'_K = \emptyset$ w.h.p., as required. ■

It remains to estimate $|\cup_{k<K} X'_k|$. To this end, let $B(\mathcal{C}, R/2)$ be the set of all vertices whose undirected distance from $\mathcal{C}$ in G is less than $R/2 = (10/c)\log n$. Consider some $v \in X'_k$ for some $k \le K$, let T_v be the corresponding tree provided by Lemma 2.6 and suppose first that some leaf u in T_v belongs to $\mathcal{C}$ (recall that every leaf of T_v is in $\mathcal{C} \cup Z$ by (2.5)). Since by definition T_v has at most $4k \le 4K \le (8/c)\log n$ levels it follows that $T_v \subset B(\mathcal{C}, R/2)$ and in particular $v \in B(\mathcal{C}, R/2)$. Due to

this argument, if we let $\mathcal{T}_k$ denote the set of rooted trees in $\mathcal{T}'_k$ where *all* leaves belong to Z and let Y_k denote the number of vertices serving as roots of such trees, i.e.

$$\mathcal{T}_k = \left\{ T \subset G' : \begin{array}{l} \text{labeled rooted tree with } \ell \text{ leaves, } k < \ell \le 2^k, \text{ all belonging to } Z, \\ \text{a total of } t \text{ vertices for } t \le 5(\ell-1) \text{ and at most } 4k \text{ levels.} \end{array} \right\}$$
$$Y_k = \#\big\{ v \in V(G) \,:\, v \text{ is the root of } T \text{ for some } T \in \mathcal{T}_k \big\}$$

(notice that clearly $Y_k \le |\mathcal{T}_k|$ for any k), then

$$|\cup_{k<K} X'_k| \le |B(\mathcal{C}, R/2)| + \sum_{k<K} Y_k \,.$$

To estimate the size of $B(\mathcal{C}, R/2)$ observe that each vertex v in this set corresponds to a graph on $m < 3R/2$ vertices and at least m edges. We can therefore repeat the calculation in (2.8) to get that

$$\mathbb{E}|B(\mathcal{C}, R/2)| \le \sum_{m<3R/2} \binom{n}{m} \binom{\binom{m}{2}}{m} (2p)^m \le \sum_{m<3R/2} \big(e^2 c\big)^m < \sqrt{n}\,,$$

where the last inequality is valid for large c. In particular, $|B(\mathcal{C}, R/2)| < n^{3/4}$ w.h.p. and it remains to estimate $\sum_{k<K} Y_k$.

Consider $|\mathcal{T}_1|$, counting rooted labeled trees in G with 2 leaves (i.e. paths with a distinguished vertex) and at most 5 vertices and where both leaves are in Z. Conditioned on the existence of a given labeled path P in G, the probability that its endpoints are in Z is less than the probability that each endpoint has an at most 3 in-neighbors or at most 3 out-neighbors in $D \setminus P$. Altogether,

$$\begin{aligned} \mathbb{E}Y_1 \le \mathbb{E}|\mathcal{T}_1| &\le \sum_{2\le t\le 5} t n^t (2p)^{t-1} \big(2\mathbb{P}\big(\mathrm{Bin}(n-t,p) \le 3\big)\big)^2 \\ &\le 4 \cdot 5(2c)^4 \big(2\mathbb{P}\big(\mathrm{Bin}(n-5,p) \le 3\big)\big)^2 n \le 20(2c)^4 \left(c^3 e^{-c}\right)^2 n < 600\, c^{10} e^{-2c} n\,, \end{aligned}$$

where the last inequality holds for sufficiently large n.

Next examine $|\mathcal{T}_k|$ for $2 \le k < K$, which counts trees with at most $4k$ levels, $\ell \in \{k+1, \dots, 2^k\}$ leaves and a total of $t \le 5(\ell-1)$ vertices, where all leaves are in Z. The calculation in (2.6),(2.7), with the single change that now all leaves (rather than $\ell - 1$) belong to Z, yields

$$\mathbb{E}Y_k \le \mathbb{E}|\mathcal{T}_k| \le \sum_{\ell=k+1}^{2^k} \left((2e)^5 c^8 e^{-c}\right)^\ell n \le e^{-\frac{3}{4}c(k+1)} n\,, \tag{2.9}$$

and combining the above inequalities we deduce that for large enough c and n we have

$$\sum_{k<K} \mathbb{E}Y_k \le 1000\, c^{10} e^{-2c} n\,. \tag{2.10}$$

To assess the deviation of the Y_k's from their mean, set $K_0 = \lfloor \log\log n \rfloor$ and observe that

$$\sum_{K_0 \le k < K} \mathbb{E}Y_k \le 2e^{-\frac{3}{4}cK_0} n < n/\log^2 n\,,$$

with the last inequality easily holding for c large. Applying Markov's inequality we deduce that

$$\mathbb{P}\Big(\sum_{K_0 \le k < K} Y_k \ge n/\log n\Big) \le 1/\log n = o(1)\,. \tag{2.11}$$

It remains to estimate the Y_k's for $k < K_0$. To this end, define Y'_k to be the number of roots of trees $T \in \mathcal{T}_k$ such that every vertex in T has degree less than $\log^2 n$ in G:

$$Y'_k = \#\big\{v \in V(G) \,:\, v \text{ is the root of } T \text{ for some } T \in \mathcal{T}_k \text{ and } d(u) < \log^2 n \text{ for all } u \in T\big\}\,.$$

Recall that the underlying graph G is obtained from D by erasing its edge directions. Therefore, G itself is a random undirected graph $\mathcal{G}(n,p')$ with edge probability $p' = 1-(1-p)^2 = (1+o(1))2p$. As we will formally state later, $G \sim \mathcal{G}(n,p')$ has maximum degree less than $\log^2 n$ except with extremely low (super-polynomial) probability, and so $Y_k = Y'_k$ w.h.p. We will show that Y'_k is concentrated about its mean and then use it to derive concentration for Y_k.

Let (M_t) be the edge-exposure Doob's martingale for D; that is, let $e_1, \dots, e_{\binom{n}{2}}$ be an arbitrary ordering of the edges of the complete graph on n vertices and set $M_t = \mathbb{E}\left[Y_k \mid \mathcal{F}_t\right]$ where $\mathcal{F}_t$ is the σ-algebra corresponding to revealing the indicators $\{\mathbb{1}_{\{e_i \in E(D)\}} : i \le t\}$. We are interested in bounds on the increments of the martingale (M_t) in L^∞ and L^2.

Consider the effect of modifying one of the indicators $\mathbb{1}_{\{e \in E\}}$; clearly this can create or destroy a tree $T \in \mathcal{T}_k$ only if that tree includes an endpoint of e as one of its vertices. Since Y'_k counts roots of such trees where every vertex has degree less than $\log^2 n$ and by the definition of $\mathcal{T}_k$ each such tree has at most $4k$ levels (including the root), it follows that modifying e can alter Y'_k by at most $(\log^2 n)^{4k}$. In other words, Y'_k is B-Lipschitz as a function of the edges of D, where

$$B = (\log^2 n)^{4k} \le (\log^2 n)^{4K_0} \le \exp\big(8(\log\log n)^2\big) = n^{o(1)}\,.$$

It is a well-known (and easy to show) corollary that in this case $|M_{t+1} - M_t| \le B$ for all t (see, e.g. [2] for the standard coupling argument deriving this for Doob's martingale of Lipschitz functions).

Now assume that we have exposed $\mathbb{1}_{\{e_1 \in E\}}, \dots, \mathbb{1}_{\{e_t \in E\}}$ and are about to reveal whether or not $e_{t+1} \in E$. We wish to bound $\mathrm{Var}(M_{t+1} \mid \mathcal{F}_t)$. If we let $\theta = \mathbb{E}[Y'_k \mid \mathcal{F}_t,\, e_{t+1} \notin E]$ then the shifted variable $Q = M_{t+1} - \theta$ satisfies $\mathbb{P}(Q \ne 0 \mid \mathcal{F}_t) = \mathbb{P}(e_{t+1} \in E \mid \mathcal{F}_t) = p' \le 2p$ whereas $|Q| \le B$ by the assumption that $\mathbb{P}(|M_{t+1} - M_t| \le B) = 1$ (in fact, even more precisely, one has $|Q| \le B$ due to the B-Lipschitz property of Y'_k). Thus,

$$\mathrm{Var}\left(M_{t+1} \mid \mathcal{F}_t\right) = \mathrm{Var}(Q \mid \mathcal{F}_t) \le 2p(B)^2 = n^{-1+o(1)}\,.$$

and we conclude that for some $L = n^{1+o(1)}$ we have $\sum_t \mathrm{Var}\left(M_t \mid \mathcal{F}_{t-1}\right) \le L$ with probability 1.

We are now in a position to apply the following large-deviation inequality which is a special case of a result of Freedman [8, Theorem 1.6] (see also [12, Theorem 3.15]):

Theorem 2.8. *Let $(S_0, S_1, \dots, S_N)$ be a martingale with respect to the filter $(\mathcal{F}_i)$. Assume that $S_{i+1} - S_i \le B$ for all i and that $\sum_{i=1}^N \mathrm{Var}(S_i \mid \mathcal{F}_{i-1}) \le L$ with probability 1 for some $L > 0$. Then for any $s > 0$ we have $\mathbb{P}\left(\bigcup_{i=1}^N \{S_i \ge S_0 + s\}\right) \le \exp\left[-\frac{1}{2}s^2/(L + Bs)\right]$.*

Plugging in our estimate for $|M_{t+1} - M_t|$ and $\sum_t \mathrm{Var}(M_t \mid \mathcal{F}_{t-1})$ while recalling that by definition of the Doob martingale $M_0 = \mathbb{E}Y'_k$ while $M_{\binom{n}{2}} = Y'_k$ it now follows that

$$\mathbb{P}(|Y'_k - \mathbb{E}Y'_k| > s) \leq 2\exp\left[-\tfrac{1}{2}s^2/\left(n^{1+o(1)} + n^{o(1)}s\right)\right],$$

and in particular

$$\mathbb{P}(|Y'_k - \mathbb{E}Y'_k| > n^{3/4}) \leq \exp\left(n^{-1/2+o(1)}\right). \tag{2.12}$$

To complete the proof, recall that the probability that any vertex in $G \sim \mathcal{G}(n,p')$ would have degree at least $\log^2 n$ is at most

$$n\mathbb{P}\left(\mathrm{Bin}(n,2c/n) \geq \log^2 n\right) \leq n\exp\left(-c'\log^2 n\right) < n^{-10},$$

where the last inequality holds for large enough n. In particular, $Y'_k = Y_k$ except with probability n^{-10} and since by definition $0 \leq Y'_k \leq Y_k \leq n$ we further have $\mathbb{E}[Y_k] = \mathbb{E}[Y'_k] + O(n^{-9})$. Combining these inequalities with (2.12) now gives

$$\mathbb{P}(|Y_k - \mathbb{E}Y_k| > 2n^{3/4}) \leq 2n^{-10},$$

where the extra factors of 2 absorbed the $O(n^{-9})$ and $\exp(n^{-1/2+o(1)})$ error terms. In particular, taking a union bound over the $K_0 \leq \log\log n$ values of k we deduce that w.h.p.

$$\sum_{k<K_0} Y_k - \sum_{k<K_0} \mathbb{E}Y_k \leq 2n^{3/4}\log\log n.$$

Finally, combining this inequality with (2.10) and (2.11) we conclude that w.h.p.

$$\sum_{k<K} Y_k \leq 1000c^{10}e^{-2c}n + 2n^{3/4}\log\log n + n/\log n = (1000+o(1))c^{10}e^{-2c}n,$$

where the last inequality holds for large enough n. Together with the aforementioned bounds on X in terms of X' and in turn of X' in terms of $\sum Y_k$ we conclude that w.h.p.

$$|X| \leq |\mathcal{C}| + |B(\mathcal{C},R/2)| + \sum_{k<K} Y_k < n^{1/4} + n^{3/4} + (1000+o(1))c^{10}e^{-2c}n \leq (2c)^{10}e^{-2c}n,$$

where the last inequality is valid for any sufficiently large n, as required. ■

3 Concluding remarks

We have proved that a random directed graph $\mathcal{D}(n,c/n)$ contains with high probability a directed cycle including all but at most $(2+\varepsilon)e^{-c}n$ vertices, where $\varepsilon = \varepsilon(c) \to 0$ as $c \to \infty$. In fact, our proof shows that the relative error term $\varepsilon(c)$ is exponentially small in c, namely $\varepsilon(c) \leq \mathrm{poly}(c)e^{-c}$. The main term in the result is asymptotically optimal as such a random digraph typically contains $(2e^{-c} - o(1))n$ vertices with zero in-degree or out-degree.

It would be very interesting to derive accurate estimates for the length of a longest cycle in $\mathcal{D}(n,c/n)$ for small(er) values of the constant c, starting perhaps as low as the threshold for the

appearance of a linear length cycle in such a random digraph. See the related work [10] where Łuczak studied the length of the longest cycle in the undirected random graph near its critical window, showing lower and upper bounds that are tight up to a factor of $1 + \log(3/2) \approx 1.41$.

Compared to the situation in undirected graphs, the toolkit available for the case of directed graphs is rather poor at present, thus making the progress in a variety of questions about directed random and pseudo-random graphs much harder to achieve. In particular, the absence of any form of a direct analogue of the famed Pósa's rotation-extension technique, widely applied for undirected graphs, is felt throughout. It would be very useful to derive some directed version of it.

In general, the field of random and pseudo-random directed graphs is largely an uncharted territory, compared to the situation for the undirected case. Although this is certainly partly due to its relative difficulty, we believe enough knowledge and technology have been accumulated now to start exploring it in a systematic way. One recent such result is the paper [4], where global resilience type results with respect to long cycles have been derived for sparse random and pseudo-random directed graphs. It would be interesting to explore further resilience type questions in directed graphs.

Acknowledgment. A major part of this work was carried out when the first and the third authors were visiting Microsoft Research at Redmond, WA. They would like to thank the Theory Group at Microsoft Research for hospitality and for creating a stimulating research environment.

References

[1] M. Ajtai, J. Komlós, and E. Szemerédi, *The longest path in a random graph*, Combinatorica **1** (1981), no. 1, 1–12.

[2] N. Alon and J. H. Spencer, *The probabilistic method*, 3rd ed., John Wiley & Sons Inc., 2008.

[3] I. Ben-Eliezer, M. Krivelevich, and B. Sudakov, *The size Ramsey number of a directed path*, preprint. Available at `arXiv:1005.5171` (2010).

[4] I. Ben-Eliezer, M. Krivelevich, and B. Sudakov, *Long cycles in subgraphs of (pseudo)random directed graphs*, preprint. Available at `arXiv:1009.3721` (2010).

[5] B. Bollobás, *Long paths in sparse random graphs*, Combinatorica **2** (1982), no. 3, 223–228.

[6] B. Bollobás, T. I. Fenner, and A. M. Frieze, *An algorithm for finding Hamilton paths and cycles in random graphs*, Combinatorica **7** (1987), no. 4, 327–341.

[7] W. Fernandez de la Vega, *Long paths in random graphs*, Studia Sci. Math. Hungar. **14** (1979), 335–340.

[8] D. A. Freedman, *On tail probabilities for martingales*, Ann. Probability **3** (1975), 100–118.

[9] A. M. Frieze, *On large matchings and cycles in sparse random graphs*, Discrete Math. **59** (1986), no. 3, 243–256.

[10] T. Łuczak, *Cycles in a random graph near the critical point*, Random Structures Algorithms **2** (1991), no. 4, 421–439.

[11] C. McDiarmid, *Clutter percolation and random graphs*, Math. Programming Study **3** (1980), 17–25.

[12] C. McDiarmid, *Concentration*, Probabilistic methods for algorithmic discrete mathematics, Algorithms Combin., vol. 16, Springer, Berlin, 1998, pp. 195–248.