Lecture 2
Long paths in random graphs

1 Introduction

In this lecture we treat the appearance of long paths and cycles in sparse random graphs. We will work with the probability space $G(n,p)$ of binomial random graphs, analogous results for the sister model $G(n,m)$ can be either proven using very similar arguments, or derived using available equivalence statements between the two models.

Our goal here is two-fold: we first prove that already in the super-critical regime $p = \frac{1+\epsilon}{n}$, the random graph $G(n,p)$ contains typically a path of length linear in n; then we prove that in the regime $p = \frac{C}{n}$, the random graph $G(n,p)$ has typically a path, covering the proportion of vertices tending to 1 as the constant C increases. We will invoke the approaches and results developed in Lecture 1 (the Depth First Search Algorithm and its consequences) to achieve both of these goals, and in fact in a rather short and elegant way.

2 Linearly long paths in the supercritical regime

In their groundbreaking paper [2] from 1960, Paul Erdős and Alfréd Rényi made the following fundamental discovery: the random graph $G(n,p)$ undergoes a remarkable phase transition around the edge probability $p(n) = \frac{1}{n}$. For any constant $\epsilon > 0$, if $p = \frac{1-\epsilon}{n}$, then $G(n,p)$ has **whp** all connected components of size at most logarithmic in n, while for $p = \frac{1+\epsilon}{n}$ **whp** a connected component of linear size, usually called the giant component, emerges in $G(n,p)$ (they also showed that **whp** there is a unique linear sized component). The Erdős-Rényi paper, which launched the modern theory of random graphs, has had enormous influence on the development of the field. We will be able to derive both parts of this result very soon.

Although for the super-critical case $p = \frac{1+\epsilon}{n}$ the result of Erdős and Rényi shows a typical existence of a linear sized connected component, it does now imply that a longest path in such a random graph is **whp** linearly long. This was established some 20 years later by Ajtai, Komlós and Szemerédi [1]. In this section we present a (relatively) simple proof of their result. We will not attempt to achieve the best possible absolute constants, aiming rather for simplicity. Our treatment follows closely that of [5].

The most fundamental idea of the proof is to run the DFS algorithm on a random graph $G \sim G(n,p)$, constructing the graph "on the fly", as the algorithm progresses. We first fix the order σ on $V(G) = [n]$ to be the identity permutation. When the DFS algorithm is fed with a sequence of i.i.d. Bernoulli(p) random variables $\bar{X} = (X_i)_{i=1}^N$, so that is gets its i-th query answered positively if $X_i = 1$ and answered negatively otherwise, the so obtained graph is clearly distributed according to $G(n,p)$. Thus, studying the component structure of G can be reduced to studying the properties of the random sequence $\bar{X}$. This is a very useful trick, as it allows to "flatten" the random graph by replacing an inherently two-dimensional structure (a graph) by a one-dimensional one (a sequence of bits). In particular, observe crucially that as long as $T \neq \emptyset$, every positive answer to a query results in a vertex being moved from T to U, and thus after t queries and assuming $T \neq \emptyset$ still, we have $|S \cup U| \geq \sum_{i=1}^t X_i$. (The last inequality is strict in fact as the first vertex of each connected component is moved from T to U "for free", i.e., without need to get a positive answer to a query.) On the other hand, since the addition of every vertex, but the first one in a connected component, to U is caused by a positive answer to a query, we have at time t: $|U| \leq 1 + \sum_{i=1}^t X_i$.

The probabilistic part of our argument is provided by the following quite simple lemma.

Lemma 2.1 *Let $\epsilon > 0$ be a small enough constant. Consider the sequence $\bar{X} = (X_i)_{i=1}^N$ of i.i.d. Bernoulli random variables with parameter p.*

1. *Let $p = \frac{1-\epsilon}{n}$. Let $k = \frac{7}{\epsilon^2} \ln n$. Then* **whp** *there is no interval of length kn in $[N]$, in which at least k of the random variables X_i take value 1.*
2. *Let $p = \frac{1+\epsilon}{n}$. Let $N_0 = \frac{\epsilon n^2}{2}$. Then* **whp** $\left|\sum_{i=1}^{N_0} X_i - \frac{\epsilon(1+\epsilon)n}{2}\right| \leq n^{2/3}$.

Proof 1) For a given interval I of length kn in $[N]$, the sum $\sum_{i \in I} X_i$ is distributed binomially with parameters kn and p. Applying the Chernoff bound to the upper tail of $B(kn,p)$, and then the union bound, we see that the probability of the existence of an interval violating the assertion of the lemma is at most

$$(N-k+1)Pr[B(kn,p) \geq k] < n^2 \cdot e^{-\frac{\epsilon^2}{3}(1-\epsilon)k} < n^2 \cdot e^{-\frac{\epsilon^2(1-\epsilon)}{3}\frac{7}{\epsilon^2}\ln n} = o(1)\,,$$

for small enough $\epsilon > 0$.
2) The sum $\sum_{i=1}^{N_0} X_i$ is distributed binomially with parameters N_0 and p. Hence, its expectation is $N_0 p = \frac{\epsilon n^2 p}{2} = \frac{\epsilon(1+\epsilon)n}{2}$, and its standard deviation is of order n. Applying the Chebyshev inequality, we get the required estimate. □

Now we are ready to formulate and to prove the result of this section.

Theorem 2.2 *Let $\epsilon > 0$ be a small enough constant. Let $G \sim G(n,p)$.*

1. *Let $p = \frac{1-\epsilon}{n}$. Then* **whp** *all connected components of G are of size at most $\frac{7}{\epsilon^2} \ln n$.*
2. *Let $p = \frac{1+\epsilon}{n}$. Then* **whp** *G contains a path of length at least $\frac{\epsilon^2 n}{5}$.*

In both cases, we run the DFS algorithm on $G \sim G(n,p)$, and assume that the sequence $\bar{X} = (X_i)_{i=1}^N$ of random variables, defining the random graph $G \sim G(n,p)$ and guiding the DFS algorithm, satisfies the corresponding part of Lemma 2.1.

Proof 1) Assume to the contrary that G contains a connected component C with more than $k = \frac{7}{\epsilon^2}\ln n$ vertices. Let us look at the epoch of the DFS when C was created. Consider the moment inside this epoch when the algorithm has found the $(k+1)$-st vertex of C and is about to move it to U. Denote $\Delta S = S \cap C$ at that moment. Then $|\Delta S \cup U| = k$, and thus the algorithm got exactly k positive answers to its queries to random variables X_i during the epoch, with each positive answer being responsible for revealing a new vertex of C, after the first vertex of C was put into U in the beginning of the epoch. At that moment during the epoch only pairs of edges touching $\Delta S \cup U$ have been queried, and the number of such pairs is therefore at most $\binom{k}{2} + k(n-k) < kn$. It thus follows that the sequence $\bar{X}$ contains an interval of length at most kn with at least k 1's inside – a contradiction to Property 1 of Lemma 2.1.

2) Assume that the sequence $\bar{X}$ satisfies Property 2 of Lemma 2.1. We claim that after the first $N_0 = \frac{\epsilon n^2}{2}$ queries of the DFS algorithm, the set U contains at least $\frac{\epsilon^2 n}{5}$ vertices (with the contents of U forming a path of desired length at that moment). Observe first that $|S| < \frac{n}{3}$ at time N_0. Indeed, if $|S| \geq \frac{n}{3}$, then let us look at a moment t where $|S| = \frac{n}{3}$ (such a moment surely exists as vertices flow to S one by one). At that moment $|U| \leq 1 + \sum_{i=1}^t X_i < \frac{n}{3}$ by Property 2 of Lemma 2.1. Then $|T| = n - |S| - |U| \geq \frac{n}{3}$, and the algorithm has examined all $|S| \cdot |T| \geq \frac{n^2}{9} > N_0$ pairs between S and T (and found them to be non-edges) – a contradiction. Let us return to time N_0. If $|S| < \frac{n}{3}$ and $|U| < \frac{\epsilon^2 n}{5}$ then, we have $T \neq \emptyset$. This means in particular that the algorithm is still revealing the connected components of G, and each positive answer it got resulted in moving a vertex from T to U (some of these vertices may have already moved further from U to S). By Property 2 of Lemma 2.1 the number of positive answers at that point is at least $\frac{\epsilon(1+\epsilon)n}{2} - n^{2/3}$. Hence we have $|S \cup U| \geq \frac{\epsilon(1+\epsilon)n}{2} - n^{2/3}$. If $|U| \leq \frac{\epsilon^2 n}{5}$, then $|S| \geq \frac{\epsilon n}{2} + \frac{3\epsilon^2 n}{10} - n^{2/3}$. All $|S||T| \geq |S|\left(n - |S| - \frac{\epsilon^2 n}{5}\right)$ pairs between S and T have been probed by the algorithm (and answered in the negative). We thus get:

$$\begin{aligned} \frac{\epsilon n^2}{2} &= N_0 \geq |S|\left(n - |S| - \frac{\epsilon^2 n}{5}\right) \geq \left(\frac{\epsilon n}{2} + \frac{3\epsilon^2 n}{10} - n^{2/3}\right)\left(n - \frac{\epsilon n}{2} - \frac{\epsilon^2 n}{2} + n^{2/3}\right) \\ &= \frac{\epsilon n^2}{2} + \frac{\epsilon^2 n^2}{20} - O(\epsilon^3)n^2 > \frac{\epsilon n^2}{2} \end{aligned}$$

(we used the assumption $|S| < \frac{n}{3}$), and this is obviously a contradiction, completing the proof. □

Let us discuss the obtained result and its proof. First, given the probable existence of a long path in $G(n,p)$, that of a long cycle is just one short step further. Indeed, we can use sprinkling as follows. Let $p = \frac{1+\epsilon}{n}$ for small $\epsilon > 0$. Write $1 - p = (1 - p_1)(1 - p_2)$ with $p_2 = \frac{\epsilon}{2n}$; thus, most of the probability p goes into $p_1 \geq \frac{1+\epsilon/2}{n}$. Let now $G \sim G(n,p)$, $G_1 \sim G(n,p_1)$, $G_2 \sim G(n,p_2)$, we can represent $G = G_1 \cup G_2$. By Theorem 2.2, G_1 **whp** contains a linearly long path P. Now, the edges

of G_2 can be used to close most of P into a cycle – there is **whp** an edge of G_2 between the first and the last $n^{2/3}$ (say) vertices of P.

The dependencies on ϵ in both parts of Theorem 2.2 are of the correct order of magnitude – for $p = \frac{1-\epsilon}{n}$ a largest connected component of $G(n,p)$ is known to be **whp** of size $\Theta(\epsilon^{-2}) \log n$ while for $p = \frac{1+\epsilon}{n}$ a longest cycle of $G(n,p)$ is **whp** of length $\Theta(\epsilon^2)n$.

Observe that using a Chernoff-type bound for the tales of the binomial random variable instead of the Chebyshev inequality would allow to claim in the second part of Lemma 2.1 that the sum $\sum_{i=1}^{N_0} X_i$ is close to $\frac{\epsilon(1+\epsilon)n}{2}$ with probability exponentially close to 1. This would show in turn, employing the argument of Theorem 2.2, that $G(n,p)$ with $p = \frac{1+\epsilon}{n}$ contains a path of length linear in n with exponentially high probability, namely, with probability $1 - \exp\{-c(\epsilon)n\}$.

As we have mentioned in Lecture 1, the DFS algorithm is applicable equally well to directed graphs. Hence essentially the same argument as above, with obvious minor changes, can be applied to the model $D(n,p)$ of random digraphs. It then yields the following theorem:

Theorem 2.3 *Let $p = \frac{1+\epsilon}{n}$, for $\epsilon > 0$ constant. Then the random digraph $D(n,p)$ has* **whp** *a directed path and a directed cycle of length $\Theta(\epsilon^2)n$.*

This recovers the result of Karp [4].

3 Nearly spanning paths

Consider now the regime $p = \frac{C}{n}$, where C is a (large) constant. Our goal is to prove that **whp** in $G(n,p)$, the length of a longest path approaches n as C tends to infinity. This too is a classical result due to Ajtai, Komlós and Szemerédi [1], and independently due to Fernandez de la Vega [3]. It is fairly amusing to see how easily it can be derived using the DFS-based tools we developed in Lecture 1.

Theorem 3.1 *For every $\epsilon > 0$ be a small enough constant there exists $C = C(\epsilon) > 0$ such that the following is true. Let $G \sim G(n,p)$, where $p = \frac{C}{n}$. Then* **whp** *G contains a path of length at least $(1-\epsilon)n$.*

Proof Let $k = \lfloor \frac{\epsilon n}{2} \rfloor$. By Proposition 2.2 from Lecture 1 it suffices to show that $G \sim G(n,p)$ contains **whp** an edge between every pair of disjoint subsets of size k of $V(G)$. For a given pair of disjoint sets S, T of size $|S| = |T| = k$, the probability that G contains no edges between S and T is exactly $(1-p)^{k^2}$ (all k^2 pairs between S and T come out non-edges in G). Using the union bound, we obtain that the probability of the existence of a pair violating this requirement is at most

$$\binom{n}{k}\binom{n-k}{k}(1-p)^{k^2} < \binom{n}{k}^2 (1-p)^{k^2} < \left(\frac{en}{k}\right)^{2k} e^{-\frac{p(k-1)k}{2n}} < \left[\left(\frac{en}{k}\right)^2 \cdot e^{-\frac{C(k-1)}{2n}}\right]^k .$$

Recalling the value of k and taking $C = \frac{3\ln(5/\epsilon^2)}{\epsilon}$ guarantees that the above estimate vanishes (in fact exponentially fast in n), thus establishing the claim. □

A similar statement holds for the probability space $D(n,p)$ of random directed graphs, with an essentially identical proof.

References

[1] M. Ajtai, J. Komlós and E. Szemerédi, *The longest path in a random graph*, Combinatorica 1 (1981), 1–12.

[2] P. Erdős and A. Rényi, *On the evolution of random graphs*, Publ. Math. Inst. Hungar. Acad. Sci. 5 (1960), 17–61.

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

[4] R. Karp, *The transitive closure of a random digraph*, Random Structures and Algorithms 1 (1990), 73–93.

[5] M. Krivelevich and B. Sudakov, *The phase transition in random graphs – a simple proof*, Random Structures and Algorithms, to appear.