RANDOM GRAPHS AND NETWORKS

Master Lecture FS 2019.

I. INTRODUCTION.

The aim of this lecture is to present the most important models of random graphs. These models (sometimes) come from various applications, whether concerning large real-world networks, or problems of statistical mechanics, but are really interesting mathematical models as well. Their study is often rather technical and requires combination of many different techniques, which makes it interesting.

The lecture have two main parts.

– Theory of random graphs

this field was initiated by Erdős and Rényi, in a series of papers from 1960–1s studying the following object: Let $V = [n] = \{1, \dots, n\}$ be the set of n-vertices and $p \in [0,1]$ a fixed parameter. For every pair of (distinct) vertices, draw an edge between them with probability p, independently. We denote the so-obtained random graph $ER(n,p)$.

There are many natural questions one can ask like:

- Is $ER(n,p)$ (typically) connected?
- What is the size of the largest connected component?
- How behaves the diameter of the graph?
- What is the typical distance between two vertices?
- How large is the largest complete subgraph of $ER(n,p)$?

...

The original motivation of Erdős-Rényi was mainly of combinatorial origin, and aimed to prove certain deterministic statements: "If ER(n,p) has certain property with a positive probability, then there exists a graph with n vertices having this property."

More recently, in 1990's-2000's, with growth of internet and increase of computer power, random graphs became important tool to understand various real world networks, which, while not "random" in pure sense, are so complex that their modelling via random graph is natural. As examples consider:

(a) world-wide web: vertices of the network are webpages, a (directed) edge (u,v) is present if page u contains a link to page v

(b) "low level" internet structure: nodes are servers, routers, switches, ...; an edge between u,v is present if there is a direct connection between u,v.

(c) social networks: nodes are all individuals of certain population; edge is present if u,v are "friends" (this might be real friends or in sense of e.g. Facebook)

(d) collaboration networks;

(e) biological networks: e.g. protein interaction networks, neural networks, ...

It turns out that observed properties of those networks are rather different from ER(n,p). Those include

- (a) <u>scale free phenomenon</u>: # of vertices N_k degree k decays slowly with k, like $N_k \approx C_n k^{-\alpha}$.

(b) Small world phenomenon: the largest connected component contains a significant proportion of vertices and has "small" diameter which almost does not grow with n.

We will discuss models having these properties as well.

- percolation theory:

The second part of the lecture then discusses some elements of percolation theory. The basic model here is the bond percolation on $\mathbb{Z}^d$, $d \geq 2$. Here, one considers vertex set $\mathbb{Z}^d$ and edge set

$$E_d = \{ \{x,y\} : x,y \in \mathbb{Z}^d, \ \|x-y\|_2 = 1 \}.$$

(set of nearest neighbour edges) and fixed $p \in [0,1]$. One then declares every edge open with probability p, in iid fashion. The goal is to understand properties of the random graph ($\mathbb{Z}^d$, open edges).

This model appeared for the first time in works of Broadbent & Hammersley in 1957 as a model for "permeability of inhomogeneous media".

The main feature of this model is the following phase transition: there exists $p_c = p_c(d) \in (0,1)$ s.t.

(a) for $p < p_c$

$$\theta(p) := \mathbb{P}_p [0 \leftrightarrow \infty] = 0$$

and $\mathbb{P}_p [$ there is an open con. component $] = 0$

(b) for $p > p_c$

$$\theta(p) > 0$$

and $\mathbb{P}_p [$ there is an open con. component $] = 1$.

This phase transition resembles to real phase transitions (like water-ice) and makes the model relevant in physics. Recently most of percolation studies concentrate on what happens in the vicinity of the critical point p_c. We will discuss them at the end of the lecture.

Literature:

R. v.d. Hofstad: Random Graphs & Complex Networks, CUP 201-
B. Bollobás: Random Graphs, CUP 2001
N. Alon, J. Spencer: The probabilistic method, Willey & Sons 2000
G. Grimmett: Percolation, Springer 1999.
B. Bollobás, O. Riordan: Percolation, CUP 2006
...

II. BRANCHING PROCESSES / RANDOM TREES

In the theory of random graphs, branching processes often describe well the local behaviour of connected components. To prepare for this we treat them here in quite some detail. For more see classical books of Athreya-Ney (1972), Harris (1963)

Branching process is the simplest model for a population evolving in time (Galton-Watson 19th century), or for a random tree. In this model, every individual independently gives birth to a random number of children, with the same distribution $(p_i)_{i\geq 0}$

(2.1) $$p_i = \mathbb{P}(\text{individual has } i \text{ children})$$

We denote by Z_n the number of individuals in n^{th} generation, with convention $Z_0 = 1$ mostly.
Z_n satisfies recursion relation

(2.2) $$Z_n = \sum_{i=1}^{Z_{n-1}} X_{n,i}$$

where $(X_{n,i})_{n,i\geq 1}$ is a doubly infinite array of (p_i)-distributed random variables. The principal question about branching process is if it survives forever. We denote by

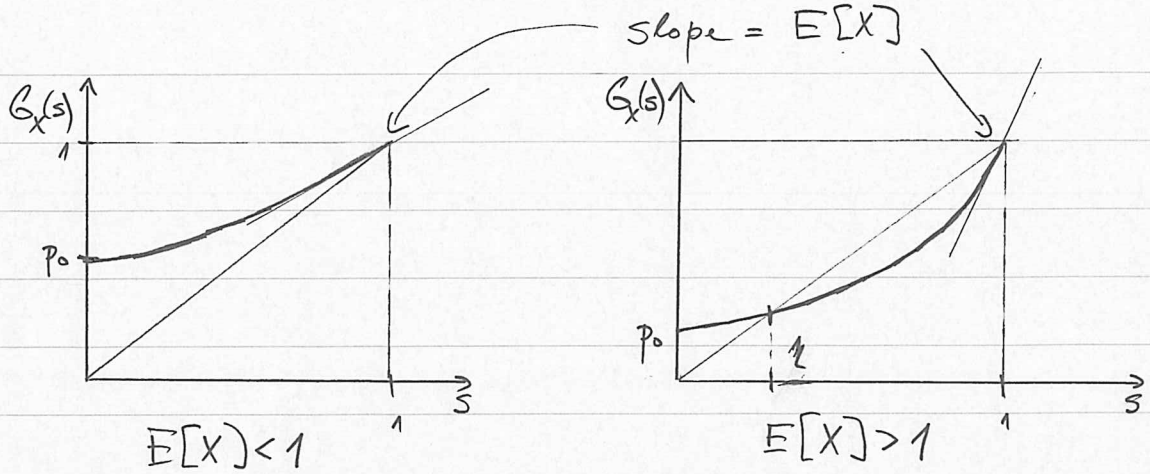
(2.3) $$\eta = \mathbb{P}(\exists n \geq 0 : Z_n = 0)$$

the extinction probability, and with X for a (p_i)-distr. r.v.

The following theorem should be known; it should be viewed as the first phase transition result.

(2.4) Theorem: Let $\mu = E(X)$ and assume $\mathbb{P}[X \leq 1] < 1$.
If $\mu \leq 1$, then $\eta = 1$
If $\mu > 1$, then $\eta < 1$ is the smallest solution in $[0,1]$ to $\eta = G_X(\eta)$
where $G_X(s)$ is the generating function of X

$$G_X(s) = E[s^X], \quad s \in [0,1].$$



(2.5) Figure: Typical slope of G_X in sat- and supercritical case.

(2.6) Proof: Let $\eta_n = \mathbb{P}(Z_n = 0)$. Since $\{Z_n = 0\} \subset \{Z_{n+1} = 0\}$, we have $\eta_n \uparrow \eta$. Let G_n be the generating function of Z_n.

$$G_n = E[s^{Z_n}]$$

Then $\eta_n = G_n(0)$. By conditioning on Z_1, we have

(2.7) $$G_n(s) = \sum_{i=0}^{\infty} \mathbb{P}(Z_1 = i)\, E[s^{Z_n} \mid Z_1 = i] = \sum_{i=0}^{\infty} p_i (G_{n-1}(s))^i,$$

since the individuals in 1st generation produce offsprings independently and Z_n is sum of contributions of individuals from the first generation. Denoting, $G_1(s) = G_X(s)$,

(2.8) $$G_n(s) = G_X(G_{n-1}(s)),$$

and thus by induction

(2.9) $$G_n(s) = \underbrace{(G_X \circ \dots \circ G_X)}_{n\text{-times}}(s)$$

In particular, $\eta_n = G_X(\eta_{n-1})$. Taking $n \uparrow \infty$, as G_X is continuous, we obtain

(2.10) $$\eta = G_X(\eta)$$

Recall that G_X is convex and $G_X(0) = p_0$. If $p_0 = 0$, there is nothing to show, so we assume that $G_X(0) > 0$. Also $G_X(1) = 1$, so (2.10) has exactly one solution in $[0,1]$ if $\mu \le 1$, so $\eta = 1$.

On the other hand, let $\psi \in [0,1]$ be a solution to (2.10) ($\psi = G_X(\psi)$). Then we claim $\eta \le \psi$. Indeed, obviously $\psi > 0$ and thus $\eta_0 = 0 < \psi$. Moreover, since G_X is increasing, by induction,

$$\eta_n = G_X(\eta_{n-1}) \le G_X(\psi) = \psi. \tag{2.11}$$

Hence, η is the smallest solution to (2.10) in $[0,1]$, which completes the proof. □.

We call branching process

subcritical	if	$\mu < 1$
critical	if	$\mu = 1$
supercritical	if	$\mu > 1$

Other important object is the total progeny T of the branching process,

$$T = \sum_{n=0}^{\infty} Z_n. \tag{2.12}$$

with generating function $G_T(s) = E[s^T]$.
Observe, that if $\mu > 1$, then $P[T=\infty] > 0$.
In this case $G_T(1) = \sum_{i=1}^{\infty} P(T=i) = 1 - P(T=\infty)$, as well as

$$\lim_{z \nearrow 1} G_T(z) = 1 - P(T=\infty) \tag{2.13}$$

so G_T is not a "proper" generating function of a r.v.

(2.14) Theorem: For any branching process, G_T satisfies

$$G_T(s) = s\, G_X(G_T(s)), \quad s \in [0,1)$$

Proof: Similarly as above, writing T_i for the total progeny of i^{th} individual in the 1^{st} generation, we have

$$T = 1 + \sum_{i=1}^{Z_1} T_i \tag{2.15}$$

T_i's are independent and have the same distribution

as T. Therefore

$$G_T(s) = \sum_{i=0}^{\infty} p_i \, E[s^T \mid Z_1 = i] =$$

$$= s \sum_{i=0}^{\infty} p_i \, E\left[s^{T_1 + \dots + T_i} \right] =$$

(2.16)

$$= s \sum_{i=0}^{\infty} p_i \, G_T(s)^i = s \, G_X(G_T(s)). \qquad \square.$$

We recall basic moment computations.

(2.17) Lemma: For all $n \geq 0$

$$E[Z_n] = \mu^n$$

as consequence, for $\mu < 1$,

$$P[Z_n > 0] \leq \mu^n$$

and

$$E[T] = \frac{1}{1-\mu}$$

Proof: Obviously, $E[Z_0] = 1$, and by conditioning on Z_{n-1}

$$E[Z_n] = \sum_{i=0}^{\infty} P[Z_{n-1} = i] \, E[Z_n \mid Z_{n-1} = i] =$$

$$= \sum_{i=0}^{\infty} P[Z_{n-1} = i] \underbrace{E[X_{n,1} + \dots + X_{n,i}]}_{= i \cdot \mu}$$

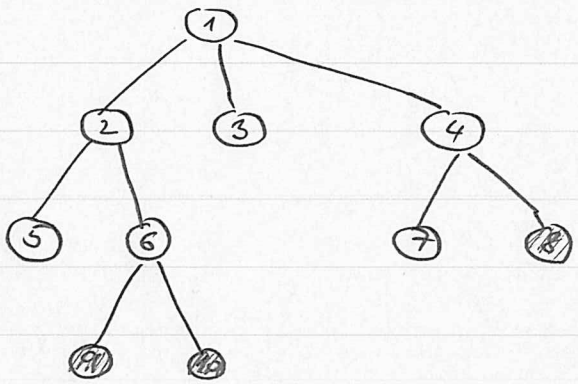
$$= \mu \cdot E[Z_{n-1}]$$

and the first claim follows by induction. The second one is then follows from Markov inequality and the last one by summing geometrical series. $\square$.

Branching Processes & Random Walks

Previously, we explored the branching process generation by generation. It is useful to explore it individual by individual, as it leads to a connection to random walk.

We start with a collection $(X_i)_{i\geq 1}$ of iid (p_i)-distributed r.v's and one "active" individuum (which will become the root of the tree. At time i we select one active individuum and give it X_i children. The children (if any) become active and the individual itself becomes inactive.



(2.18) Figure: Exploration of the tree in "breadth-first" order: node labeled i is explored on i^{th} step. This tree corresponds to state of the algorithm after 7 steps with $X_1=3, X_2=2, X_3=0, X_4=2, X_5=0, X_6=2, X_7=0$. Vertices 8, 9, 10 are active and will be explored later.

We now set

$$S_0 = 1, \quad S_i = S_{i-1} + X_i - 1 = \sum_{j=1}^{i} X_j - (i-1), \quad i \geq 1 \tag{2.19}$$

Observe that S_i is the number of active vertices at the end of the i-th step.

(2.20)

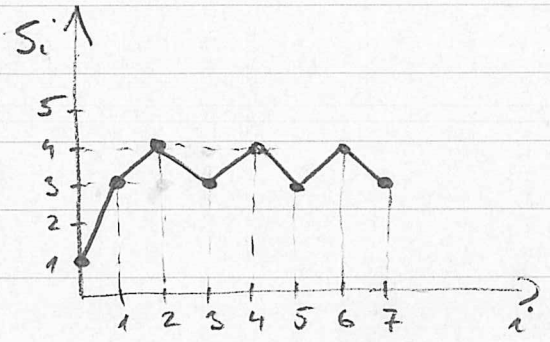


Figure Random walk (S_i) corresponding to Figure (2.18)

Denoting by $T = \inf\{i \geq 0 : S_i = 0\}$ the hitting time of 0, we see that T has the same law as the total progeny of the branching process, since at every step we explored exactly one vertex and at time T there is nobody new to explore, that is the tree is constructed.

While it is difficult to see e.g. the sizes Z_i of various generation from the representation (2.19), it is useful otherwise, e.g. to understand the total progeny.

We write

(2.21) $$H = (X_1, \dots, X_T)$$

for the history of the process up to T. A sequence $(x_1, \dots, x_t)$ is a possible history iff $x_i \geq 0$, $i \leq t$, and

(2.22) $$\sum_{i=1}^{t} x_i - (t-1) = 0, \quad \text{while} \quad \sum_{i=1}^{k} x_i - (k-1) > 0 \quad \forall k < t.$$

For any such $(x_i)_{i=1}^{t}$ with $t < \infty$ one has

(2.23) $$P(H = (x_1, \dots, x_t)) = \prod_{i=1}^{t} p_{x_i}$$

We now explore the distribution of the branching process conditioned on extinction. For $(p_i)_{i \geq 0}$ we define <u>conjugate distribution</u>

(2.24) $$p_i' = \eta^{i-1} p_i, \qquad i \geq 0$$

where η is as in (2.3). Observe that $\sum_{i=0}^{\infty} p_i' = \sum_{i=0}^{\infty} \eta^{i-1} p_i = \eta^{-1} G_X(\eta) = 1$ by (2.10), so (p_i') is a probability distribution.

(2.25) **Theorem:** (Duality of branching processes) The branching process with offspring distribution (p_i) conditioned on extinction (i.e. on $T<\infty$) has the same distribution as branching process with offspring distribution (p_i').

Proof: By previous reasoning, it is sufficient that the two measures agree on finite histories. For $t<\infty$ and a $(x_1,\dots,x_t)$ as in (2.22),

$$P^{(p_i)}\big(H=(x_1,\dots,x_t)\,\big|\,T<\infty\big) = \frac{P^{(p_i)}\big(\overbrace{H=(x_1,\dots,x_t)}^{=\{H=(x_1,\dots,x_t)\}}\cap T<\infty\big)}{P(T<\infty)} =$$

$$= \zeta^{-1}\cdot P^{(p_i)}\big(H=(x_1,\dots,x_t)\big) = \zeta^{-1}\prod_{i=1}^{t} p_{x_i}$$

$$\overset{(2.24)}{=} \zeta^{-1}\prod_{i=1}^{t} p'_{x_i}\cdot\zeta^{1-x_i} = \zeta^{(t-1)-\sum_{i=1}^{t}x_i}\prod_{i=1}^{t} p_i' \overset{(2.22)}{=} \prod_{i=1}^{t} p_0'$$

$$= P^{(p_i')}\big[H=(x_1,\dots,x_t)\big]. \qquad \square$$

One also easily obtains the following estimate

(2.26) **Proposition:** For a supercritical branching process

(2.27) $$P(k\le T<\infty) \le \frac{e^{-Ik}}{1-e^{-I}} \quad \text{with}$$
$$I = \sup_{t\le 0}\big(t-\log E[e^{tX}]\big) > 0.$$

(it is unlikely, that T is large but the process goes extinct)

Proof: $T=s$ implies $S_s=0$ and thus $X_1+\dots+X_s = s-1 \le s$.

Hence,

(2.28) $$P(k\le T<\infty) = \sum_{s=k}^{\infty} P(T=s) = \sum_{s=k}^{\infty} P(X_1+\dots+X_s\le s)$$

Using exponential Markov, with $t\le 0$

$$P(X_1+\dots+X_s\le s) \le P\big[e^{t(X_1+\dots+X_s)}\ge e^{ts}\big] \le e^{-ts}\,E\Big[\prod_{i=1}^{s} e^{tX_i}\Big]$$

(2.29) $$= \exp\big\{-ts + s\cdot\log E(e^{tX})\big\} \le e^{-Is}$$

where the last inequality is optimizing over $t\le 0$. (2.28)+(2.29) imply (2.27). The fact $I>0$ follows from continuity of $(-\infty,0)\ni t\overset{g}{\mapsto} t-\log E[e^{tX}]$ and $g'(0) = 1-E[X] < 0$. $\square$

Supercritical branching processes

(2.30) **Exercise**: Prove that $M_n = \mu^{-n} Z_n$ is a martingale wrt. filtration $\mathcal{F}_n = \sigma(Z_1, \dots, Z_n)$. Deduce that

(2.31) $$\mu^{-n} Z_n \xrightarrow{a.s.} W_\infty$$

for some a.s. finite random variable W_∞. Show that $W_\infty = 0$ if $\mu \le 1$ and that $P(W_\infty = 0) \ge \eta$.

This exercise does not answer the question how fast (Z_n) grow when the process survives, it might be that $T = \infty$ but $W_\infty = 0$ in which case (2.31) is rather uninformative. The answer is provided by

(2.32) **Theorem**: (Kesten–Stigum 1966) If $\mu > 1$, then TFAE

(i) $E[X \log X] < \infty$

(ii) $P(W_\infty = 0) = \eta$, i.e. $W > 0$ on $\{T = \infty\}$ a.s.

(iii) $E[W_\infty] = 1$

When $E[X \log X] = \infty$, then $P(W_\infty = 0) = 1$.

More on total progeny:

We now give a nicer formula for the law of T in terms of random walk. We define $Y_i = X_i - 1$ and write

(2.33) P_k for the law of Markov chain $S_0 = k$, $S_i = S_{i-1} + Y_i$, $i \ge 1$.

If $H_0 = \inf\{n \ge 0 : S_n = 0\}$, then as on page 9-10,

(2.34) $$P[T = k] = P_1[H_0 = k].$$

(2.35) **Lemma** Let S_n be random walk with incr. integer valued steps Y_i, with $Y_i \ge -1$. Then for any $k \ge 1$, $n \ge 1$

$$P_k[H_0 = n] = \frac{k}{n} P[S_n = 0].$$

Proof: For $0 \le l < n$ consider random variables $(Y_i^l)_{i=1}^n$

(2.36)
$$Y_i^l = \begin{cases} Y_{i+l} & \text{if } i+l \le n \\ Y_{i+l-n} & \text{if } i+l > n \end{cases} \qquad i = 1, \dots, n.$$

i.e. Y^l is a cyclic permutation of Y's. Hence $(Y_i^l)_{i \le n}$ has the same distribution as $(Y_i)_{i \le n}$. Setting

$$S_i^l = S_{i-1}^l + Y_i^l, \quad H_0^l = \inf\{j \ge 0 : S_j^l = 0\}$$

$$P_k[H_0 = n] = \frac{1}{n} \sum_{l=0}^{n-1} P_k[H_0^l = n].$$

Now observe that for any realisation $Y_1, \dots, Y_n$ of Y's such that $k + \sum_{i=1}^{n} Y_i = 0$ there are exactly k values of l such that $H_0^l = n$.

(2.37)

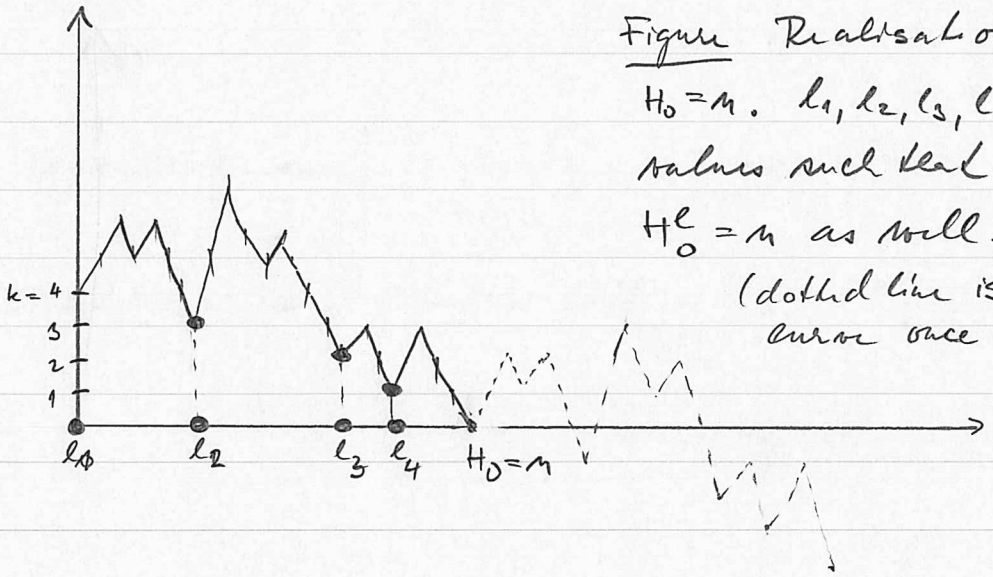


Figure Realisation of $H_0 = n$. l_1, l_2, l_3, l_4 are values such that $H_0^l = n$ as well. (dotted line is the same curve once more)

Hence $$\sum_{k=0}^{n-1} P[H_0^l = 0] = k \cdot P[S_n = 0]$$ □.

(2.38) **Remark:** Another proof of (2.35), by induction, can be found in van der Hofstad's book (Thm 3.14).