6.207/14.15: Networks
Lecture 10: Erdös-Renyi Graphs and Phase Transitions

Outline

- Phase transitions
- Connectivity threshold
- Diameter of Erdös-Renyi graphs
- Branching processes

Phase Transitions for Erdös-Renyi Model

- Erdös-Renyi model is specified by the link formation probability $p(n)$.
- For a given property A (e.g. connectivity), we define a threshold function $t(n)$ as a function that satisfies:

$$\mathbb{P}(\text{property } A) \to 0 \qquad \text{if} \quad \frac{p(n)}{t(n)} \to 0, \text{ and}$$

$$\mathbb{P}(\text{property } A) \to 1 \qquad \text{if} \quad \frac{p(n)}{t(n)} \to \infty.$$

  - This definition makes sense for "monotone or increasing properties," i.e., properties such that if a given network satisfies it, any supernetwork (in the sense of set inclusion) satisfies it.
- When such a threshold function exists, we say that a phase transition occurs at that threshold.
- Exhibiting such phase transitions was one of the main contributions of the seminal work of Erdös and Renyi 1959.

Threshold Function for Connectivity

Theorem

(Erdös and Renyi 1961) A threshold function for the connectedness of the Erdös and Renyi model is $t(n) = \frac{\log(n)}{n}$.

- To prove this, it is sufficient to show that when $p(n) = \lambda(n)\frac{\log(n)}{n}$ with $\lambda(n) \to 0$, we have $\mathbb{P}(\text{connected}) \to 0$ (and the converse).
- However, we will show a stronger result: Let $p(n) = \lambda\frac{\log(n)}{n}$.

$$\text{If } \lambda < 1, \qquad \mathbb{P}(\text{connected}) \to 0, \tag{1}$$

$$\text{If } \lambda > 1, \qquad \mathbb{P}(\text{connected}) \to 1. \tag{2}$$

Proof:

- We first prove claim (1). To show disconnectedness, it is sufficient to show that the probability that there exists at least one isolated node goes to 1.

Proof (Continued)

- Let I_i be a Bernoulli random variable defined as

$$I_i = \begin{cases} 1 & \text{if node } i \text{ is isolated,} \\ 0 & \text{otherwise.} \end{cases}$$

- We can write the probability that an individual node is isolated as

$$q = \mathbb{P}(I_i = 1) = (1-p)^{n-1} \approx e^{-pn} = e^{-\lambda \log(n)} = n^{-\lambda}, \qquad (3)$$

where we use $\lim_{n\to\infty} \left(1 - \frac{a}{n}\right)^n = e^{-a}$ to get the approximation.

- Let $X = \sum_{i=1}^{n} I_i$ denote the total number of isolated nodes. Then, we have

$$\mathbb{E}[X] = n \cdot n^{-\lambda}. \qquad (4)$$

- For $\lambda < 1$, we have $\mathbb{E}[X] \to \infty$. We want to show that this implies $\mathbb{P}(X = 0) \to 0$.
  - In general, this is not true. But, here it holds.
  - We show that the variance of X is of the same order as its mean.

Proof (Continued)

- We compute the variance of X, var(X):

$$\text{var}(X) = \sum_i \text{var}(I_i) + \sum_i \sum_{j \neq i} \text{cov}(I_i, I_j) = n\text{var}(I_1) + n(n-1)\text{cov}(I_1, I_2)$$
$$= nq(1-q) + n(n-1)\Big(\mathbb{E}[I_1 I_2] - \mathbb{E}[I_1]\mathbb{E}[I_2]\Big),$$

where the second and third equalities follow since the I_i are identically distributed Bernoulli random variables with parameter q (dependent).

- We have

$$\mathbb{E}[I_1 I_2] = \mathbb{P}(I_1 = 1, I_2 = 1) = \mathbb{P}(\text{both 1 and 2 are isolated})$$
$$= (1-p)^{2n-3} = \frac{q^2}{(1-p)}.$$

- Combining the preceding two relations, we obtain

$$\text{var}(X) = nq(1-q) + n(n-1)\left[\frac{q^2}{(1-p)} - q^2\right]$$
$$= nq(1-q) + n(n-1)\frac{q^2 p}{1-p}.$$

Proof (Continued)

- For large n, we have $q \to 0$ [cf. Eq. (3)], or $1 - q \to 1$. Also $p \to 0$. Hence,

$$\begin{aligned} \text{var}(X) &\sim nq + n^2 q^2 \frac{p}{1-p} \sim nq + n^2 q^2 p \\ &= nn^{-\lambda} + \lambda n \log(n) n^{-2\lambda} \\ &\sim nn^{-\lambda} = \mathbb{E}[X], \end{aligned}$$

  where $a(n) \sim b(n)$ denotes $\frac{a(n)}{b(n)} \to 1$ as $n \to \infty$.

- This implies that

$$\mathbb{E}[X] \sim \text{var}(X) \geq (0 - \mathbb{E}[X])^2 \mathbb{P}(X = 0),$$

  and therefore,

$$\mathbb{P}(X = 0) \leq \frac{\mathbb{E}[X]}{\mathbb{E}[X]^2} = \frac{1}{\mathbb{E}[X]} \to 0.$$

- It follows that $\mathbb{P}(\text{at least one isolated node}) \to 1$ and therefore, $\mathbb{P}(\text{disconnected}) \to 1$ as $n \to \infty$, completing the proof.

Converse

- We next show claim (2), i.e., if $p(n) = \lambda \frac{\log(n)}{n}$ with $\lambda > 1$, then $\mathbb{P}(\text{connected}) \to 1$, or equivalently $\mathbb{P}(\text{disconnected}) \to 0$.
- From Eq. (4), we have $\mathbb{E}[X] = n \cdot n^{-\lambda} \to 0$ for $\lambda > 1$.
- This implies probability of having isolated nodes goes to 0. However, we need more to establish connectivity.
- The event "graph is disconnected" is equivalent to the existence of k nodes without an edge to the remaining nodes, for some $k \leq n/2$.
- We have

$$\mathbb{P}(\{1, \dots, k\} \text{ not connected to the rest}) = (1-p)^{k(n-k)},$$

and therefore,

$$\mathbb{P}(\exists \text{ k nodes not connected to the rest}) = \binom{n}{k}(1-p)^{k(n-k)}.$$

Converse (Continued)

- Using the union bound [i.e. $\mathbb{P}(\cup_i A_i) \leq \sum_i \mathbb{P}(A_i)$], we obtain

$$\mathbb{P}(\text{disconnected graph}) \leq \sum_{k=1}^{n/2} \binom{n}{k} (1-p)^{k(n-k)}.$$

- Using Stirling's formula $k! \sim \left(\frac{k}{e}\right)^k$

$$\binom{n}{k} \approx \exp\left(n \log n - k \log k - (n-k)\log(n-k)\right) \quad = \exp\left(nH(k/n)\right),$$

where $H(x) = -x \log x - (1-x)\log(1-x)$ is the entropy function

- For $p = \lambda \log n/n$, using $(1-p) \approx \exp(-p)$

$$(1-p)^{k(n-k)} \approx \exp\left(-n \log n \, \lambda \, \frac{k}{n}\left(1 - \frac{k}{n}\right)\right)$$

Converse (Continued)

- Using these approximations, we obtain

$$\mathbb{P}(\text{disconnected graph}) \leq \sum_{k=1}^{n/2} \exp\left(nH\left(\frac{k}{n}\right) - n\log n\ \lambda\ \frac{k}{n}\left(1-\frac{k}{n}\right)\right)$$
$$\approx \int_{1/n}^{n/2} \exp\left(nf_n(r)\right)$$

where $f_n(r) = H(r) - \log n\lambda r(1-r)$.

- Invoking *Laplace's principle*, we approximate

$$\int_{1/n}^{n/2} \exp\left(nf_n(r)\right) \approx C \exp\left(\max_{\frac{1}{n}\leq r\leq \frac{n}{2}} f_n(r)\right).$$

for some constant C, independent of n

Converse (Continued)

- It can be checked that maximum over $[1/n, n/2]$ is achieved at $r = 1/n$.

$$f_n(1/n) \approx -(\lambda - 1)\frac{\log n}{n}.$$

- Therefore, we obtain

$$\begin{aligned}\mathbb{P}(\text{disconnected graph}) &\leq C \exp\Big(-(\lambda-1)\log n\Big) \\ &= Cn^{-1+\lambda} \\ &\stackrel{\lambda>1}{\rightarrow} 0.\end{aligned}$$

Phase Transitions — Connectivity Threshold

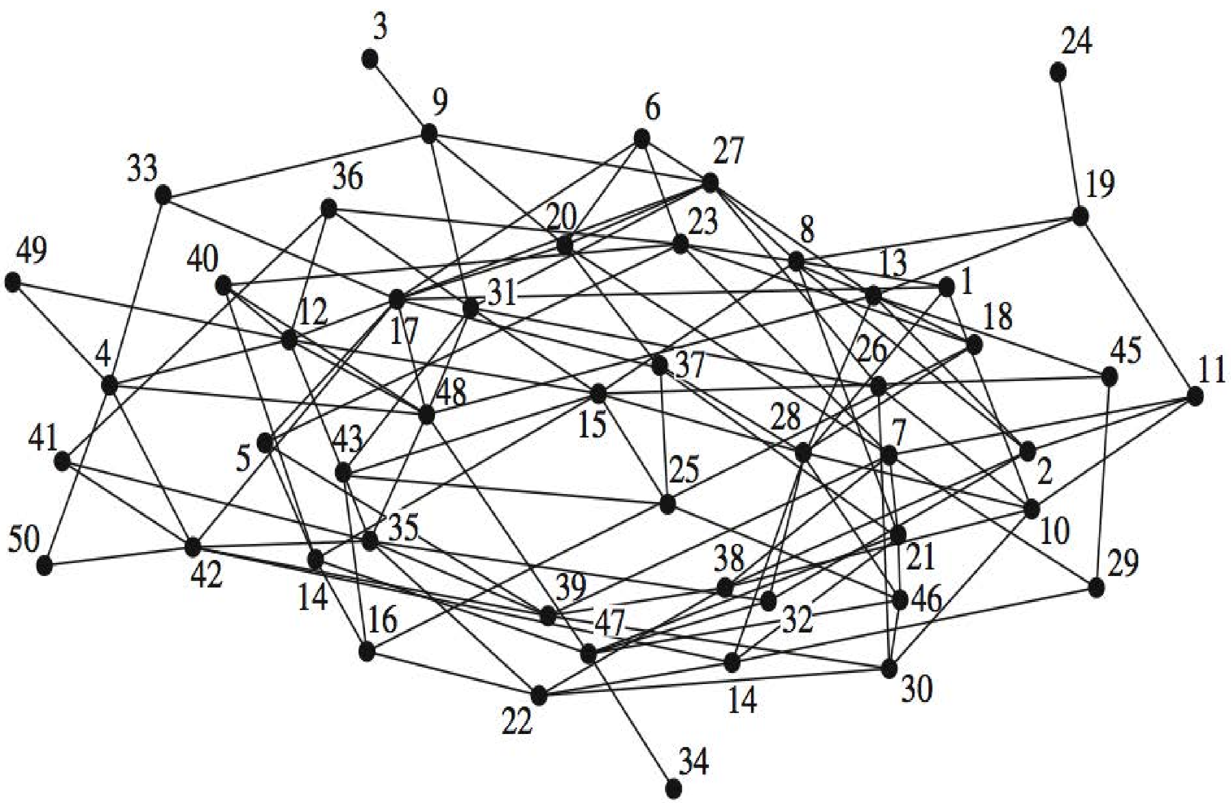

Figure: Emergence of connectedness: a random network on 50 nodes with $p = 0.10$.

Diameter

- Recall the diameter of a graph: let d_{ij} be the distance between nodes i and j (i.e., length of the shortest path between i and j).

$$\text{diameter} = \max_{i,j} d_{ij}.$$

- We will show that the diameter of the ER graph varies as $\ln n$.
- Heuristic Argument:
  - Let c denote the average degree of a node, $c = (n-1)p$.
  - The average number of nodes s steps away from a randomly chosen node is c^s.
  - The number of nodes reached is equal to the total number of nodes when $c^s \approx n$, or $s \approx \frac{\ln n}{\ln c}$
  - Every node is within s steps of the starting point, implying that the diameter is approximately $\frac{\ln n}{\ln c}$.
  - This argument works when s is small (breaks down when c^s become comparable with n since number of nodes within distance s cannot exceed number of nodes in the whole graph).

Diameter

- Consider two different starting nodes i and j. The average number of nodes s and t steps away from them will be equal to c^s and c^t (assume both remain smaller than order n).

- We have $d_{ij} > s + t + 1$ if and only if there is no edge between the surfaces. Since there are on average $c^s \times c^t$ pairs of nodes between surfaces, this implies $P(d_{ij} > s + t + 1) = (1 - p)^{c^{s+t}}$. Denoting $l = s + t + 1$, we have
$$P(d_{ij} > l) = (1 - p)^{c^{l-1}} \approx \left(1 - \frac{c}{n}\right)^{c^{l-1}}.$$

Diameter

- Taking logs of both sides, we find

$$\ln P(d_{ij} > l) = c^{l-1} \ln\left(1 - \frac{c}{n}\right) \approx -\frac{c^l}{n},$$

where we used $\ln(1 + x) \approx x$ (which holds for large n). Therefore,

$$P(d_{ij} > l) = exp\left(-\frac{c^l}{n}\right).$$

- The diameter is the smallest l such that $P(d_{ij} > l)$ is zero. The preceding will tend to zero only if c^l grows faster than n, i.e., $c^l = an^{1+\epsilon}$ for some constant a and $\epsilon \to 0$ (note that this can be achieved while keeping both c^s and c^t smaller than n).
- Rearranging for l, we obtain the diameter as

$$l = \frac{\ln a}{\ln c} + \lim_{\epsilon \to 0} \frac{(1+\epsilon)\ln n}{\ln c} = A + \frac{\ln n}{\ln c},$$

- Example: Let $n = 7 \times 10^9$ and $c = 1000$. Then, $l = \frac{\ln n}{\ln c} = 3.3$.

Branching Processes

- Brief history of branching processes
  - Genesis in work by Thomas Malthus (1798)
  - *An Essay on the Principle of Population*
  - Led to *Malthusianism*: one of the key premises
    - *Unchecked population grows exponentially; resources (e.g. food) don't*

  which is justified through the study of branching processes
- John Keynes, *Economic consequences of the Peace* (1919)
  - Argues that European political economy of that time is unstable
  - Due to premise of Malthusianism
- Study of extinction or growth of species in Ecology
  - Branching processes play crucial role
  - The Galton-Watson (1875) was one of the first such approach
- General branching process theory
  - T. E. Harris, *The Theory of Branching Processes* (1963)
  - K. B. Athreya and P. E. Ney, *Branching Processes* (1972)

Branching Processes

- We'll use branching process
  - To analyze the *emergence of giant component* in ER graph
- The Galton-Watson Branching process is defined as follows:
- Start with a single individual at generation 0, $Z_0 = 1$.
- Let Z_k denote the number of individuals in generation k.
- Let ξ be a nonnegative discrete random variable with distribution p_k, i.e.,

$$P(\xi = k) = p_k, \qquad \mathbb{E}[\xi] = \mu, \qquad var(\xi) \neq 0.$$

- Each individual has a random number of children in the next generation, which are independent copies of the random variable ξ. That is,

$$Z_1 = \xi, \qquad Z_2 = \sum_{i=1}^{Z_1} \xi^{(i)} (\text{sum of random number of rvs}).$$

$$\mathbb{E}[Z_1] = \mu, \quad \mathbb{E}[Z_2] = \mathbb{E}[\mathbb{E}[Z_2 \mid Z_1]] = \mathbb{E}[\mu Z_1] = \mu^2,$$
$$\mathbb{E}[Z_n] = \mu^n.$$

Branching Processes (Continued)

- Let Z denote the total number of individuals in all generations, $Z = \sum_{n=1}^{\infty} Z_n$.
- We consider the events $Z < \infty$ (extinction) and $Z = \infty$ (survive forever).
- Our interest: when and with what probability do these events occur.
- Two cases:
  - Subcritical ($\mu < 1$) and supercritical ($\mu > 1$)
- Subcritical: $\mu < 1$
- Since $\mathbb{E}[Z_n] = \mu^n$, we have

$$\mathbb{E}[Z] = \mathbb{E}\left[\sum_{n=1}^{\infty} Z_n\right] = \sum_{n=1}^{\infty} \mathbb{E}\left[Z_n\right] = \frac{1}{1-\mu} < \infty,$$

  (some care is needed in the second equality).
- This implies that $Z < \infty$ with probability 1 and $\mathbb{P}(\textit{extinction}) = 1$.

Branching Processes (Continued)

- Supercritical: $\mu > 1$
- Recall $p_0 = \mathbb{P}(\xi = 0)$. If $p_0 = 0$, then $\mathbb{P}(extinction) = 0$.
- Let $p_0 > 0$. We have $\rho = \mathbb{P}(extinction) \geq \mathbb{P}(Z_1 = 0) = p_0 > 0$.
- We can write the following fixed-point equation for ρ:

$$\rho = \sum_{k=0}^{\infty} p_k \rho^k = \mathbb{E}[\rho^\xi] \equiv \Phi(\rho).$$

- We have $\Phi(0) = p_0$ (using convention $0^0 = 1$) and $\Phi(1) = 1$
- Φ is a convex function ($\Phi''(\rho) \geq 0$ for all $\rho \in [0,1]$), and $\Phi'(1) = \mu > 1$.

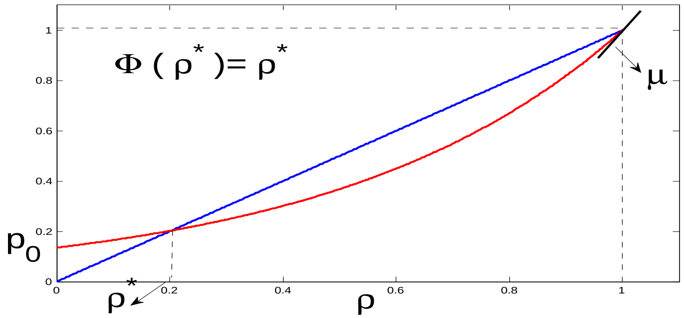

Figure: The generating function Φ has a unique fixed point $\rho^* \in [0,1)$.