6.207/14.15: Networks
Lecture 3: Erdös-Renyi graphs and Branching processes

Daron Acemoglu and Asu Ozdaglar
MIT

September 16, 2009

Outline

- Erdös-Renyi random graph model
- Branching processes
- Phase transitions and threshold function
- Connectivity threshold

Reading:

- Jackson, Sections 4.1.1 and 4.2.1-4.2.3.

Erdös-Renyi Random Graph Model

- We use $G(n,p)$ to denote the undirected Erdös-Renyi graph.
- Every edge is formed with probability $p \in (0,1)$ **independently** of every other edge.
- Let $I_{ij} \in \{0,1\}$ be a Bernoulli random variable indicating the presence of edge $\{i,j\}$.
- For the Erdös-Renyi model, random variables I_{ij} are independent and

$$I_{ij} = \begin{cases} 1 & \text{with probability } p, \\ 0 & \text{with probability } 1-p. \end{cases}$$

- $\mathbb{E}[\text{number of edges}] = E[\sum I_{ij}] = \frac{n(n-1)}{2}\, p$
- Moreover, using weak law of large numbers, we have for all $\alpha > 0$

$$\mathbb{P}\left(\left| \sum I_{ij} - \frac{n(n-1)}{2}\, p \right| \geq \alpha \frac{n(n-1)}{2} \right) \to 0,$$

as $n \to \infty$. Hence, with this random graph model, the number of edges is a random variable, but it is tightly concentrated around its mean for large n.

Properties of Erdös-Renyi model

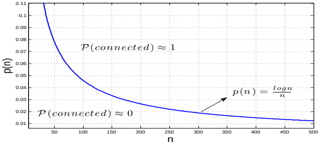
- Recall statistical properties of networks:
  - Degree distributions
  - Clustering
  - Average path length and diameter
- For Erdös-Renyi model:
  - Let D be a random variable that represents the degree of a node.
    - D is a binomial random variable with $\mathbb{E}[D] = (n-1)p$, i.e., $\mathbb{P}(D = d) = \binom{n-1}{d} p^d (1-p)^{n-1-d}$.
    - Keeping the expected degree constant as $n \to \infty$, D can be approximated with a Poisson random variable with $\lambda = (n-1)p$,

      $$\mathbb{P}(D = d) = \frac{e^{-\lambda}\lambda^d}{d!},$$

      hence the name Poisson random graph model.
    - This degree distribution falls off faster than an exponential in d, hence **it is not a power-law distribution**.
  - Individual clustering coefficient$\equiv Cl_i(p) = p$.
    - Interest in $p(n) \to 0$ as $n \to \infty$, implying $Cl_i(p) \to 0$.
  - Diameter:?

Other Properties of Random Graph Models

- Other questions of interest:
  - Does the graph have isolated nodes? cycles? Is it connected?
- For random graph models, we are interested in computing the **probabilities of these events**, which may be intractable for a fixed n.
- Therefore, most of the time, we resort to an asymptotic analysis, where we compute (or bound) these probabilities as $n \to \infty$.
- Interestingly, often properties hold with either a probability approaching 1 or a probability approaching 0 in the limit.
- Consider an Erdös-Renyi model with link formation probability $p(n)$ (again interest in $p(n) \to 0$ as $n \to \infty$).



- The graph experiences a phase transition as a function of graph parameters (also true for many other properties).

Branching Processes

- To analyze phase transitions, we will make use of branching processes.
- 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.
- This implies that

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

and therefore,

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

and $\mathbb{E}[Z_n] = \mu^n$.

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).
- We are interested in conditions and with what probabilities these events occur.
- **Two cases:**
    - Subcritical ($\mu < 1$) and supercritical ($\mu > 1$)
- Subcritical: $\mu < 1$
- Since $\mathbb{E}[Z_n] = \mu^n$, we have

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

  (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 $\mathbb{P}(\textit{extinction}) = 0$.
- Assume $p_0 > 0$.
- We have $\rho = \mathbb{P}(\textit{extinction}) \geq \mathbb{P}(Z_1 = 0) = p_0 > 0$.
- 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 $\Phi'(1) = \mu > 1$.

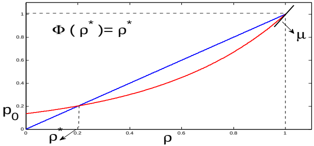


Figure: The generating function Φ has a unique fixed point $\rho^* \in [0,1)$.

Phase Transitions for Erdös-Renyi Model

- Erdös-Renyi model is completely 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.

Phase Transition Example

- Define property A as $A = \{\text{number of edges} > 0\}$.
- We are looking for a threshold for the emergence of the first edge.
- Recall $\mathbb{E}[\text{number of edges}] = \frac{n(n-1)}{2} p(n) \approx \frac{n^2}{2} p(n)$.
- Assume $\frac{p(n)}{2/n^2} \to 0$ as $n \to \infty$. Then, $\mathbb{E}[\text{number of edges}] \to 0$, which implies that $\mathbb{P}(\text{number of edges} > 0) \to 0$.
- Assume next that $\frac{p(n)}{2/n^2} \to \infty$ as $n \to \infty$. Then, $\mathbb{E}[\text{number of edges}] \to \infty$.
- This does not in general imply that $\mathbb{P}(\text{number of edges} > 0) \to 1$.
- Here it follows because the number of edges can be approximated by a Poisson distribution (just like the degree distribution), implying that

$$\mathbb{P}(\text{number of edges} = 0) = \left.\frac{e^{-\lambda}\lambda^k}{k!}\right|_{k=0} = e^{-\lambda}.$$

- Since the mean number of edges, given by λ, goes to infinity as $n \to \infty$, this implies that $\mathbb{P}(\text{number of edges} > 0) \to 1$.

Phase Transitions

- Hence, the function $t(n) = 1/n^2$ is a threshold function for the emergence of the first link, i.e.,
  - When $p(n) << 1/n^2$, the network is likely to have no edges in the limit, whereas when $p(n) >> 1/n^2$, the network has at least one edge with probability going to 1.
- How large should $p(n)$ be to start observing triples in the network?
  - We have $\mathbb{E}[\text{number of triples}] = n^3 p^2$, using a similar analysis we can show $t(n) = \frac{1}{n^{3/2}}$ is a threshold function.
- How large should $p(n)$ be to start observing a tree with k nodes (and $k-1$ arcs)?
  - We have $\mathbb{E}[\text{number of trees}] = n^k p^{k-1}$, and the function $t(n) = \frac{1}{n^{k/k-1}}$ is a threshold function.
- The threshold function for observing a cycle with k nodes is $t(n) = \frac{1}{n}$
  - Big trees easier to get than a cycle with arbitrary size!

Phase Transitions (Continued)

- Below the threshold of $1/n$, the largest component of the graph includes no more than a factor times $\log(n)$ of the nodes.
- Above the threshold of $1/n$, a giant component emerges, which is the largest component that contains a nontrivial fraction of all nodes, i.e., at least cn for some constant c.
- The giant component grows in size until the threshold of $\log(n)/n$, at which point the network becomes connected.

Phase Transitions (Continued)

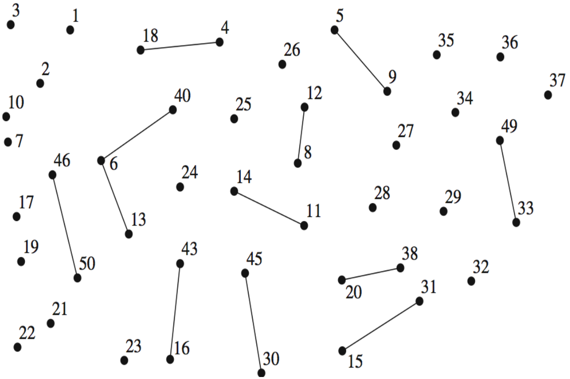


Figure: A first component with more than two nodes: a random network on 50 nodes with $p = 0.01$.

Phase Transitions (Continued)

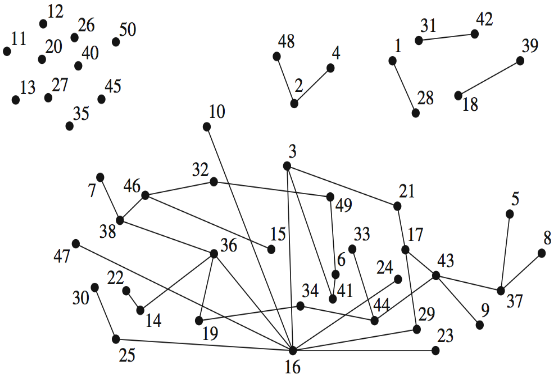


Figure: Emergence of cycles: a random network on 50 nodes with $p = 0.03$.

Phase Transitions (Continued)

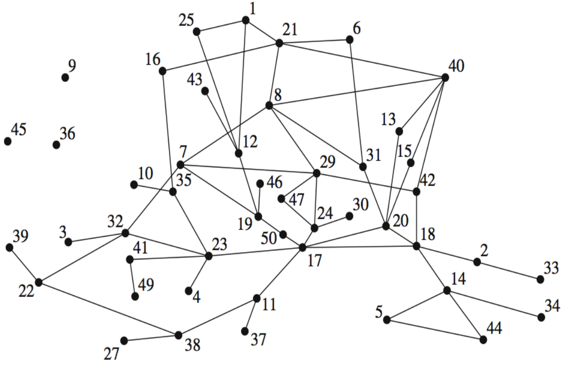


Figure: Emergence of a giant component: a random network on 50 nodes with $p = 0.05$.

Phase Transitions (Continued)

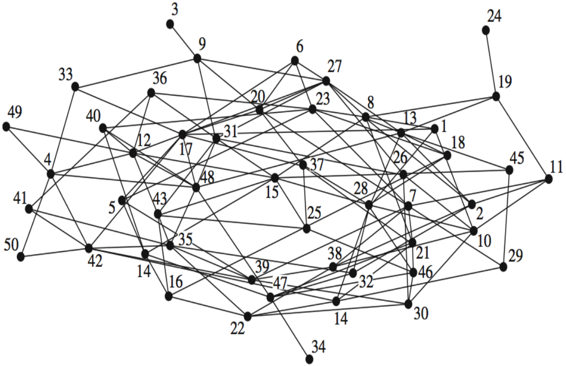


Figure: Emergence of connectedness: a random network on 50 nodes with $p = 0.10$.

Threshold Function for Connectivity

Theorem

(Erdös and Renyi 1961) A threshold function for the connectivity 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 $\mathbb{P}(\text{connectivity}) \to 0$ (and the converse).
- However, we will show a stronger result: Let $p(n) = \lambda\frac{\log(n)}{n}$.

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

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

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.
    - Can we use a Poisson approximation (as in the previous example)? No, since the random variables I_i here are dependent.
    - We show that the variance of X is of the same order as its mean.

Proof (Continued)

- We compute the variance of X, $\mathrm{var}(X)$:

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

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

- We have

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

- Combining the preceding two relations, we obtain

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

Proof (Continued)

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

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

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

- This implies that

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

  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 $\mathbb{P}(\text{connectivity}) \to 1$, or equivalently $\mathbb{P}(\text{disconnectivity}) \to 0$.
- From Eq. (4), we have $\mathbb{E}[X] = n \cdot n^{-\lambda} \to 0$ for $\lambda > 1$.
- This implies probability of 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

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

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 $k! \sim \left(\frac{k}{e}\right)^k$, which implies $\binom{n}{k} \leq \frac{n^k}{(\frac{k}{e})^k}$ in the preceding relation and some (ugly) algebra, we obtain

$$\mathbb{P}(\text{disconnected graph}) \to 0,$$

completing the proof.