Universitat
Pompeu Fabra
Barcelona

Lecture 7 · Random Networks I

Networks, Crowds and Markets

Today's Lecture

1. Random graphs, Erdős–Rényi model.
2. Probability recap: binomial and Poisson distribution.
3. Probability recap: Chebyshev and Hoeffding inequality.
4. Degree distribution in Erdős–Rényi graphs.
5. Asymptotics in networks.
6. Threshold phenomena and giant component.

Random graphs and Erdős–Rényi model

Why random graphs?

Real networks (social, economic, financial) are noisy and constantly evolving. We need a simple *baseline model* to compare against.

Definition (Erdős–Rényi (ER) model)

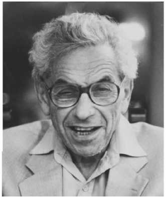
$G(N, p)$: a random graph on N nodes where each of the $\binom{N}{2}$ possible edges appears independently with probability p.

Why random graphs?

Real networks (social, economic, financial) are noisy and constantly evolving. We need a simple *baseline model* to compare against.

Definition (Erdős–Rényi (ER) model)

$G(N, p)$: a random graph on N nodes where each of the $\binom{N}{2}$ possible edges appears independently with probability p.

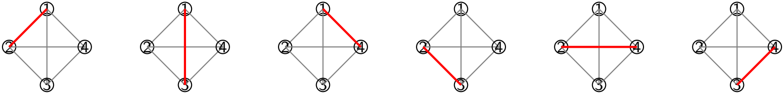


Paul Erdős (1913 - 1996)   Alfréd Rényi (1921-1970)

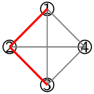
Erdős and Rényi (1959–60) launched the probabilistic study of graphs.

$G(N, p)$ Model

Take $N = 4$ then the graph can have up to six edges. Each with distribution $\mathrm{Bern}(p)$:



12 13 14 23 24 34

$\mathbb{P}($  $) = p^2(1-p)^4$

If $p = \frac{1}{2}$, each graph appears with the same probability $\frac{1}{2^6} = \frac{1}{64}$.

Probability recap: Binomial

Definition

If $X \sim \mathrm{Bin}(n, p)$ then

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \mathbb{E}[X] = np, \quad \mathrm{Var}(X) = np(1 - p).$$

Useful characterization: $X = \sum_{i=1}^{n} Z_i$ with independent $Z_i \sim \mathrm{Bern}(p)$.

In the ER graph $G(N, p)$:

- Number of edges:

$$L \;\sim\; \mathrm{Bin}\left(\binom{N}{2}, \, p\right).$$

- Degree of a fixed vertex v:

$$\deg(v) \;\sim\; \mathrm{Bin}(N - 1, p).$$

Probability recap: Poisson (as Binomial limit)

Theorem

If $X_n \sim \mathrm{Bin}(n, p_n)$ with $n \to \infty$ and $np_n \to \lambda > 0$, then

$$X_n \;\longrightarrow\; X \sim \mathrm{Pois}(\lambda), \qquad \mathbb{P}(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

The approximation $\mathrm{Bin}(n, p) \approx \mathrm{Poiss}(\lambda)$ for $\lambda = pn$ is particularly good if p is small.

Example (Quick check)

For $n = 2000$, $p = 0.003$, $\lambda = np = 6$. Compare $\mathbb{P}(X = 0)$: Binomial $\approx (1 - p)^{2000}$ vs. Poisson e^{-6} (very close).

Degree distribution in $G(N, p)$

If $p = \lambda/(N-1)$, then, for any $v \in V$,

$$\deg(v) \;\sim\; \mathrm{Bin}(N-1, p) \;\approx\; \mathrm{Pois}(\lambda).$$

- Mean degree: $\mathbb{E}[\deg(v)] = (N-1)p$.
- $\mathbb{P}\{\deg(v) = k\} \approx \dfrac{\lambda^k}{k!} e^{-\lambda}$.

Note

This gives closed forms for expectations; Poisson is a great approximation when N is large and p small.

Degree distribution: finite N concentration bounds

Concentration: Chebyshev (simple but general)

Theorem (Chebyshev inequality)

For any r.v. X with mean μ and variance σ^2,

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

For degree: $\deg(v) \sim \mathrm{Bin}(N-1, p)$, so

$$\mathbb{P}\big(|\deg(v) - (N-1)p| \geq t\big) \leq \frac{(N-1)p(1-p)}{t^2}.$$

Chebyshev already gives some concentration guarantees (e.g. take $t_0 = \sqrt{\frac{N}{\delta}p(1-p)}$ for small $\delta > 0$) but sharper results are possible.

Appendix: Proof of the Chebyshev inequality

Theorem (Markov's inequality)

If $Z \geq 0$ then $\mathbb{P}(Z \geq t) \leq \frac{1}{t}\mathbb{E}[Z]$.

Indeed,

$$\mathbb{E}[Z] \;\leq\; \mathbb{E}[Z\mathbb{1}(Z \geq t)] \;\leq\; t\mathbb{E}[\mathbb{1}(Z \geq t)] \;=\; t\mathbb{P}(Z \geq t).$$

Now, Chebyshev's inequality follows easily from Markov's. Take $Z = |X - \mu|$ then

$$\mathbb{P}(|X - \mu| \geq t) \;=\; \mathbb{P}((X-\mu)^2 \geq t^2) \;\leq\; \frac{\mathbb{E}(X-\mu)^2}{t^2} \;=\; \frac{\sigma^2}{t^2}.$$

Sharper concentration: Hoeffding for Binomial

Theorem (Hoeffding inequality)

If $X = \sum_{i=1}^{n} Y_i$ with independent $Y_i \in [0,1]$ and $\mathbb{E}X = \mu$, then for $t > 0$,

$$\mathbb{P}(|X - \mu| \geq t) \;\leq\; 2\exp\Big(-\tfrac{2t^2}{n}\Big).$$

Applied to degree: $\deg(v)$ has $N-1$ independent Bernoulli summands,

$$\mathbb{P}(|\deg(v) - (N-1)p| \geq t) \leq 2\exp\left(-\frac{2t^2}{N-1}\right).$$

Fix $v \in V$. Taking $t_0 = \sqrt{\frac{N-1}{2}\log(\frac{2}{\delta})}$ for small $\delta > 0$ gives

$$\mathbb{P}\big(|\deg(v) - (N-1)p| \geq t_0\big) \;\leq\; \delta.$$

Sharper concentration: Hoeffding for Binomial

Theorem (Hoeffding inequality)

If $X = \sum_{i=1}^{n} Y_i$ with independent $Y_i \in [0,1]$ and $\mathbb{E}X = \mu$, then for $t > 0$,

$$\mathbb{P}(|X - \mu| \geq t) \;\leq\; 2\exp\Big(-\tfrac{2t^2}{n}\Big).$$

Applied to degree: $\deg(v)$ has $N-1$ independent Bernoulli summands,

$$\mathbb{P}(|\deg(v) - (N-1)p| \geq t) \leq 2\exp\left(-\frac{2t^2}{N-1}\right).$$

Fix $v \in V$. Taking $t_0 = \sqrt{\frac{N-1}{2}\log(\frac{2}{\delta})}$ for small $\delta > 0$ gives

$$\mathbb{P}\big(|\deg(v) - (N-1)p| \geq t_0\big) \;\leq\; \delta.$$

e.g. $N = 1001$, $\delta = 0.05$, $p = 0.1$. Then with prob. ≥ 0.95

$$\deg(v) \;\in\; (100 - 42.95, 100 + 42.95) \;=\; (57.05, 142.95).$$

Uniform degree bounds

Recall: $\mathbb{P}(|\deg(v) - (N-1)p| \geq t) \leq 2\exp\left(-\frac{2t^2}{N-1}\right)$ for all $t > 0$.

Taking $t_0 = \sqrt{\frac{N-1}{2}\log(\frac{2}{\delta})}$ we get that with probability $\geq 1 - \delta$

$$\deg(v) \;\in\; \big((N-1)p - t_0, (N-1)p + t_0\big).$$

$t_0 = \sqrt{(N-1)\log N}$ A union bound over all v shows all degrees concentrate near $(N-1)p$ with high probability.

Asymptotics in networks

Asymptotic Thinking in Random Graphs

Why asymptotics?

- We study $G(N, p)$ as $N \to \infty$ to reveal general patterns.
- Precise constants matter less than the scaling behavior of p with N.

Asymptotic Thinking in Random Graphs

Why asymptotics?

- We study $G(N, p)$ as $N \to \infty$ to reveal general patterns.
- Precise constants matter less than the scaling behavior of p with N.

Note (Notation recap)

- $f(N) = o(g(N))$ means $f(N)/g(N) \to 0$.
- $f(N) = O(g(N))$ means $|f(N)| \leq C|g(N)|$ for large N (for some $C > 0$).
- $f(N) \sim g(N)$ means $f(N)/g(N) \to 1$.

Asymptotic Thinking in Random Graphs

Why asymptotics?

- We study $G(N, p)$ as $N \to \infty$ to reveal general patterns.
- Precise constants matter less than the scaling behavior of p with N.

Note (Notation recap)

- $f(N) = o(g(N))$ means $f(N)/g(N) \to 0$.
- $f(N) = O(g(N))$ means $|f(N)| \leq C|g(N)|$ for large N (for some $C > 0$).
- $f(N) \sim g(N)$ means $f(N)/g(N) \to 1$.

Probabilistic language:

- "With high probability" (w.h.p.) means $\mathbb{P}(\text{event}) \to 1$ as $N \to \infty$.
- Example: in $G(N, p)$ with $p = \frac{\log N}{N}$, the graph is connected *w.h.p.*

Asymptotic regimes: dense vs sparse graphs

When N grows, the connection probability $p = p_N$ can scale differently.

Dense regime: p_N tends to a constant $c > 0$; $p_N = O(1)$.

- $\mathbb{E}[\deg(v)] \approx cN$ grows linearly with N.
- The number of edges $L \approx cN^2/2$ — a positive fraction of all pairs are connected.
- Not a realistic large network, but a useful contrast.

Sparse regime: $p_N = \lambda/N$ (or smaller); $p_N = O(N^{-1})$.

- $\mathbb{E}[\deg(v)] \approx \lambda$ stays constant as $N \to \infty$.
- The total number of edges $L \approx \lambda N/2$ grows linearly with N.
- This captures the idea that most nodes have only a few links even in huge networks.

Language note:

- Saying "real networks are sparse" means that as they grow, the *average degree* stays bounded, not that p is small for a fixed N.
- The scaling of p_N determines which asymptotic regime we are in.

Maximum degree in $G(N, p)$

Let $\Delta = \max_v \deg(v)$ be the **maximum degree**.

1. **Dense regime:**
   - Each $\deg(v) \sim \mathrm{Bin}(N-1, p)$ with mean $\mathbb{E}\deg(v) \approx Np$.
   - With high probability:

$$\Delta \;=\; Np \,+\, O\big(\sqrt{N \log N}\big).$$

2. **Sparse regime ($p = \lambda/N$):**
   - Each $\deg(v) \approx \mathrm{Pois}(\lambda)$ — mean λ.
   - By extreme–value theory for Poisson tails:

$$\Delta \;\approx\; \frac{\log N}{\log\log N}.$$

Takeaway: Even in purely random graphs, a few nodes will look like *"hubs"* simply due to chance.

Notation: average degree vs expected degree

For a graph G with N vertices and L edges:

- The **empirical average degree** is (a random variable)

$$\overline{\deg}(G) \;=\; \frac{1}{N}\sum_{v\in V}\deg(v) \;=\; \frac{2L}{N}.$$

- The **expected degree** under a random graph model is

$$\mathbb{E}[\deg] \;:=\; \mathbb{E}[\overline{\deg}(G)].$$

Example (Erdős–Rényi $G(N,p)$):

$$\overline{\deg}(G) \approx (N-1)p, \qquad \mathbb{E}[\deg] = (N-1)p.$$

We saw that for large N, $\overline{\deg}(G)$ is tightly concentrated around $\mathbb{E}[\deg]$.

Threshold phenomena and giant component

Threshold phenomena in ER (concept)

Definition

A **threshold** for a graph property $\mathcal{P}$ is a function $p^*(N)$ such that:

$$p \ll p^*(N) \Rightarrow G(N, p) \text{ has } \neg\mathcal{P} \text{ w.h.p.},$$

$$p \gg p^*(N) \Rightarrow G(N, p) \text{ has } \mathcal{P} \text{ w.h.p.}$$

ER graphs display many sharp thresholds:

- Emergence of a giant component.
- Connectivity (no isolated vertices).
- Appearance of fixed subgraphs (e.g., triangles).

Giant component: where it appears

Theorem (Giant component threshold)

In $G(N, p)$ with $p = \frac{\lambda}{N}$:

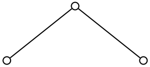
$$\begin{cases} \lambda < 1 : & \text{All components have size } O(\log N) \text{ w.h.p. (no giant).} \\ \lambda > 1 : & \text{There exists a unique giant component of size } \Theta(N) \text{ w.h.p.} \end{cases}$$

Interpretation: $\lambda = 1$ is the phase transition. Above it, a macroscopic fraction of nodes are mutually reachable.

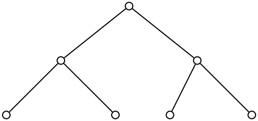
Giant component: intuition

How does a "large" connected component emerge in $G(N, p)$ with $p = c/N$?

- Pick one node and start exploring its neighbors. Each neighbor brings along its own neighbors, and so on.
- If on average each node connects to **less than one new person** ($c < 1$), the process fizzles out quickly $\Rightarrow$ only small groups.
- If on average each node connects to **more than one new person** ($c > 1$), the process can keep expanding $\Rightarrow$ one very large group forms (the "giant component").



$c < 1$ (dies out)



$c > 1$ (keeps growing $\Rightarrow$ giant)

Why the giant component matters (econ/social)

Consider the world's friendship network:

- Clearly disconnected (think small remote communities)
- But our component is large, spans most of the world.
- There should be no two big components.

Giant components are important:

- **Contagion & diffusion:** A giant component enables large cascades (diseases, information, bank runs).
- **Market connectivity:** Sufficient density is needed for trade/payment networks to connect most participants.
- **Infrastructure design:** Tuning p (or expected degree c) above 1 ensures large-scale reachability.

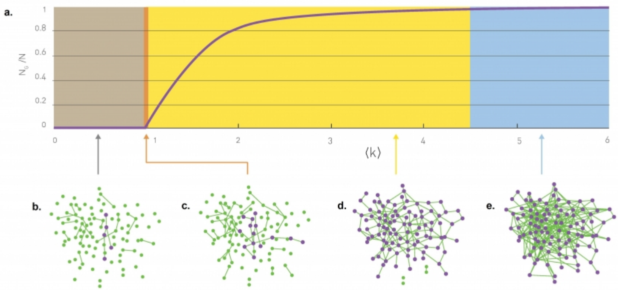
Regimes of $G(N, p)$ (sparse case $p = c/N$)

It is useful to describe random graphs in terms of the **expected degree**

$$\mathbb{E}[\deg] \approx c.$$

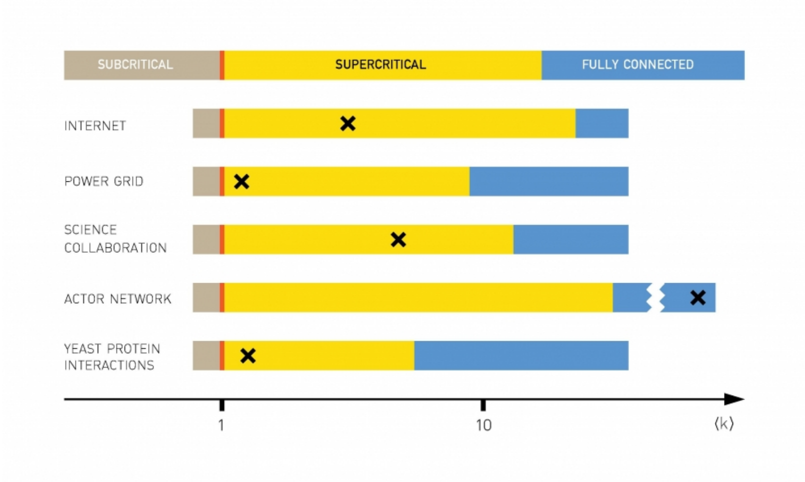
- **Subcritical regime ($c < 1$):** only small tree-like components; largest size $\sim \log N$.
- **Critical point ($c = 1$):** largest component has size $\sim N^{2/3}$; no giant yet.
- **Supercritical regime ($c > 1$):** a unique **giant component** emerges, containing a positive fraction of nodes.
- **Connected regime ($c \gtrsim \log N$):** almost surely the whole graph becomes connected.

Illustration of regimes



Interpretation: As c increases, the largest connected component grows from negligible size, through a sudden phase transition ($c = 1$), and eventually absorbs almost all nodes.

Where are real networks?



- Most real-world social, economic, and technological networks live **well above the critical point**.
- They are highly connected (often even "superconnected"), yet they

Connectivity threshold

Theorem

In $G(N, p)$ the threshold for connectivity is

$$p^*(N) = \frac{\log N}{N}.$$

More precisely:

$$\begin{cases} p = \frac{\log N + \omega(N)}{N}, & G(N, p) \text{ connected w.h.p.,} \\ p = \frac{\log N - \omega(N)}{N}, & G(N, p) \text{ disconnected w.h.p..} \end{cases}$$

Intuition: At this density, isolated vertices disappear. Since isolated vertices are the last obstacle to connectivity, once they vanish, the whole graph connects.

Other classic thresholds (very brief)

Let $p = N^{-\alpha}$:

- **Fixed subgraph** H**:** appearance when $p \gg N^{-1/m(H)}$ (where $m(H) = \max_{H' \subseteq H} e(H')/v(H')$).
- **Triangles:** threshold $p \sim N^{-1}$ (expected count $\sim \binom{N}{3} p^3$).
- **Hamiltonian cycle:** appears around $p \approx (\log N)/N$ (up to constant factors).

These give a menu of “phase transitions” that help calibrate model realism for given N, p.

Worked example: Poisson approximation in $G(N, p)$

Example (Binomial vs Poisson)

Let $N = 1000$, $p = 0.004$ so $Np = 4$. For a fixed v:

$$\mathbb{P}(\deg(v) = 0) \;=\; (1-p)^{999} \;\approx\; e^{-4},$$

$$\mathbb{P}(\deg(v) = 1) \;\approx\; 999p(1-p)^{998} \;\approx\; 4e^{-4}.$$

The Poisson(4) values e^{-4}, $4e^{-4}$ match closely.

Simulation in NetworkX (Colab) — generate and inspect

Python (run in Google Colab)

```
import networkx as nx
import matplotlib.pyplot as plt

n, p = 200, 0.015  # try also p = 0.005, 0.02, 0.05
G = nx.erdos_renyi_graph(n, p)

print("Nodes:", G.number_of_nodes())
print("Edges:", G.number_of_edges())

# Empirical vs expected average degree
deg = [d for _, d in G.degree()]
print("Empirical mean degree:", sum(deg)/n)
print("Theoretical mean degree:", (N-1)*p)

# Largest component size
components = list(nx.connected_components(G))
largest = max(components, key=len)
print("Largest component size:", len(largest))

# Draw (small n looks better)
plt.figure(figsize=(5,5))
pos = nx.spring_layout(G, seed=7)
nx.draw(G, pos, node_size=30, edge_color="#cccccc")
plt.show()
```

Simulation in NetworkX — degree histogram

Python (run in Google Colab)

```
import numpy as np
import matplotlib.pyplot as plt

deg = np.array([d for _, d in G.degree()])
print("Empirical mean degree:", deg.mean())
print("Theoretical mean degree:", (N-1)*p)

plt.figure(figsize=(5,4))
bins = np.arange(deg.max()+2) - 0.5
plt.hist(deg, bins=bins)
plt.xlabel("Degree k"); plt.ylabel("Count")
plt.title("Degree distribution in G(N,p)")
plt.show()
```

Observation. For $p = c/N$ the histogram should resemble a Poisson(c), with empirical mean degree $\overline{\text{deg}}(G)$ close to theoretical $\mathbb{E}[\text{deg}]$.

Summary

- ER $G(N, p)$ is the baseline random network: tractable degrees and component structure.
- Degrees: Binomial $\to$ Poisson in sparse regime; strong concentration via Hoeffding.
- Phase transitions: giant component at $p \sim 1/N$; connectivity at $p \sim (\log N)/N$.
- Why we care: gives parameter ranges where large-scale behavior becomes plausible.

Recap: degree distribution in $G(N, p)$

- For fixed vertex v, $\deg(v) \sim \mathrm{Bin}(N-1, p)$.
- In sparse regime $p = c/N$: $\deg(v) \approx \mathrm{Pois}(c)$.
- ER networks give **tractable formulas** for degrees.
- Baseline question: how much variability in data is due to pure chance?

Threshold for subgraphs

Definition

Threshold:

probability p at which a fixed subgraph H typically appears in $G(N, p)$.

$$\mathbb{E}[X_H] = \binom{N}{h} p^m \approx N^h p^m,$$

where H has h vertices and m edges.

Threshold for subgraphs

Definition

Threshold:

probability p at which a fixed subgraph H typically appears in $G(N, p)$.

$$\mathbb{E}[X_H] = \binom{N}{h} p^m \approx N^h p^m,$$

where H has h vertices and m edges.

Example: triangles

$$\mathbb{E}[\#\triangle] = \binom{N}{3} p^3 \approx N^3 p^3.$$

- If $p \ll 1/N$: almost surely no triangles.
- If $p \gg 1/N$: many triangles appear.

Interpretation: $p \sim 1/N$ is the threshold for local clustering to begin.