Universitat Pompeu Fabra
Barcelona

Lecture 6 · Centrality measures II

Networks, Crowds and Markets

Today's Lecture

1. Why random graphs? Motivation and Erdős–Rényi models.
2. Probability recap for $G(N, p)$:
   2.1 Binomial distribution (edges, degrees).
   2.2 Poisson approximation in the sparse regime.
3. Degree distribution and concentration:
   3.1 Chebyshev and Hoeffding bounds.
   3.2 Maximum degree heuristics.
4. Threshold phenomena:
   4.1 Giant component.
   4.2 Connectivity.
   4.3 Other classic thresholds.
5. Worked example + NetworkX simulation.

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)

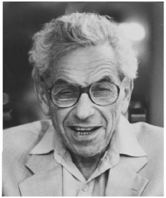
$G(N, p)$: a random graph on N nodes where each of the $\binom{N}{2}$ possible edges appears independently with prob. p.

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)

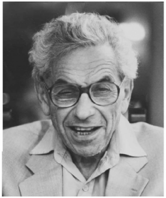
$G(N,p)$: a random graph on N nodes where each of the $\binom{N}{2}$ possible edges appears independently with prob. p.



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

Erdős and Rényi (1959–60) launched the probabilistic study of graphs. Their program connected combinatorics and probability, leading to modern random graph theory.

The main contributors

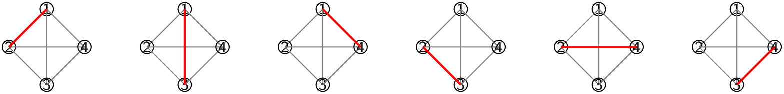


Paul Erdős (1913 - 1996)   Alfréd Rényi (1921-1970)

- Erdős and Rényi (1959–60) launched the probabilistic study of graphs.
- Their program connected combinatorics and probability, leading to modern random graph theory.
- The ER model remains the canonical baseline for testing ideas and algorithms.

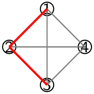
$G(N, p)$ Model

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



12 13 14 23 24 34

$$\Pr\Big(\ $$



$$\Big) = p^2(1-p)^4$$

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

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

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

In $G(N, p)$:

- 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

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

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 $\Pr(X = 0)$: Binomial $\approx (1 - p)^{2000}$ vs. Poisson e^{-6} (very close).

Degree distribution in Erdős–Rényi model

Degree distribution in $G(N, p)$

For a fixed v, if $p = \lambda/(N-1)$,

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

- Mean degree: $\mathbb{E}[\mathsf{deg}(v)] = (N-1)p$.
- Sparse regime $p = \lambda/(N-1)$: $\Pr\{\mathsf{deg}(v) = k\} \approx \dfrac{\lambda^k}{k!} e^{-\lambda}$.
- Why useful: closed forms for expectations; Poisson is a great approximation when N is large and p small.

Concentration: Chebyshev (simple but general)

Theorem (Chebyshev inequality)

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

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

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

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

Chebyshev is loose but distribution-free; good first control of deviations.

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 $t > 0$,

$$\Pr(|X - \mu| \geq t) \leq 2\exp\left(-\frac{2t^2}{n}\right).$$

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

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

Taking $t_0 = \sqrt{(N-1)\log N}$ gives

$$\Pr\left(|\deg(v) - (N-1)p| \geq t_0\right) \leq \frac{2}{N^2}.$$

A union bound over all v shows all degrees concentrate near $(N-1)p$ with high probability.

Maximum degree in $G(N, p)$

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

1. **Dense regime (p constant, not tiny):**
   - 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 (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 $p = \frac{\lambda}{N}$:

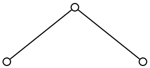
$$\begin{cases} \lambda < 1: & \text{All components have size } O(\log N) \text{ w.h.p. (no giant).} \\ \lambda > 1: & \text{There exists a unique giant component of size } \Theta(N) \text{ w.h.p.} \end{cases}$$

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

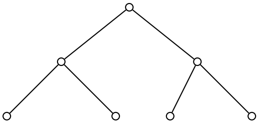
Giant component: intuition

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.
- This looks like a "chain reaction": each person you reach can connect you to more people.
- 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.

Rule of thumb in the sparse regime $p = c/N$: aim for $\mathbb{E}[\deg] = c > 1$ if you need global connectivity to start to emerge.

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

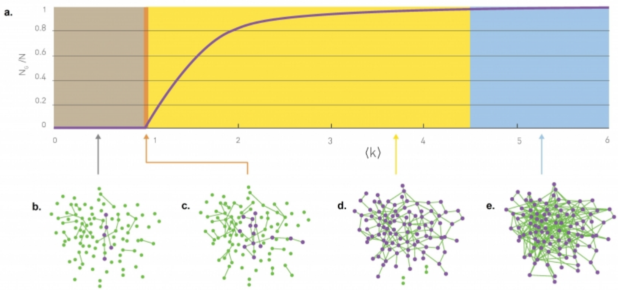
It is useful to describe random graphs in terms of the **expected degree**

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

- **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.

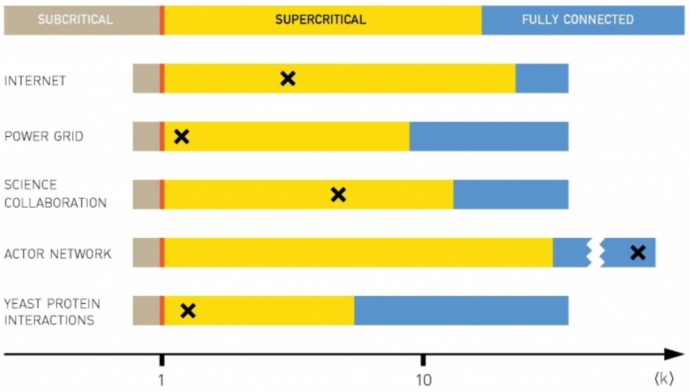
Note: For a realized graph G, the empirical mean degree $\overline{\deg}(G)$ is tightly concentrated around $\mathbb{E}[\deg]$ when N is large.

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:

$$\Pr(\deg(v) = 0) = (1-p)^{999} \approx e^{-4}, \quad \Pr(\deg(v) = 1) \approx 999p(1-p)^{998} \approx 4e$$

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 $\rightarrow$ 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.

Today's Lecture

1. Quick recap on $G(N, p)$ and degree distribution.
2. Threshold for appearance of subgraphs (example: triangles).
3. The clustering coefficient: definition, motivation, formulas.
4. Static random graph models: ER as binary vectors, ERGMs.
5. Recursive random graph models: preferential attachment.
6. Why random models matter for economics and social sciences.

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.

Clustering

Clustering coefficient: definition

For node v with degree k_v:

$$C_v = \frac{\#\text{ links among neighbors of } v}{\binom{k_v}{2}} \in [0,1].$$

- Measures "friend-of-friend closure."
- $C_v = 1$: neighbors form a clique; $C_v = 0$: none connected.
- Average clustering coefficient: $\overline{C} = \frac{1}{N}\sum_v C_v$.

Clustering in ER networks

- Pick node i and two of its neighbors u, v.
- In $G(N, p)$, edge (u, v) exists with probability p.
- Therefore $\mathbb{E}[C_i] = p$.

Implications:

- In sparse regime $p = c/N$: $\mathbb{E}[C_i] \approx c/N \to 0$.
- Prediction: clustering vanishes in large ER graphs.
- Real networks (social, financial, trade) show *much higher* clustering.
- $\Rightarrow$ Mismatch: motivates richer models.

Definition: clustering coefficient

For node v with degree $\deg(v) = k_v$:

$$C_v = \frac{L_v}{\binom{k_v}{2}}$$

- L_v = number of actual links among i's neighbors.
- $\binom{k_v}{2}$ = maximum possible such links.
- $C_v \in [0, 1]$: fraction of “friend-of-friend” connections realized.

Clustering in ER networks

- Pick a node i and two of its neighbors u, v.
- In $G(N, p)$, the edge (u, v) exists with probability p (edges are independent).
- Therefore, each potential link among i's neighbors appears with prob. p.

$$\Rightarrow \quad \mathbb{E}[C_i] = p.$$

Implications:

- In sparse regime $p = c/N$: $\mathbb{E}[C_i] \approx c/N \to 0$ as $n \to \infty$.
- Prediction: clustering vanishes in large ER graphs.
- Real networks (social, financial, trade) show high clustering even when sparse. $\rightarrow$ Mismatch: motivates richer models.

Other random graph models

Static random graph models

- Any graph on N nodes = binary vector of length $\binom{N}{2}$.
- ER: independent Bernoulli(p) for each edge.
- **Exponential Random Graph Models (ERGMs):**

$$\Pr(G = g) \propto \exp\{\theta_1 \cdot \#\text{edges}(g) + \theta_2 \cdot \#\text{triangles}(g) + \cdots\}.$$

- θ_1 tunes density, θ_2 tunes clustering, etc.
- $\mathrm{ER}(N, p)$ = special case with $\theta_2 = \cdots = 0$.

Quick recall: exponential families

A probability distribution on $\mathcal{X}$ is an *exponential family* if

$$p_\theta(x) = h(x) \exp\left(\theta^T T(x) - \psi(\theta)\right).$$

- $T(x) =$ sufficient statistics (counts of edges, triangles, ...).
- $\theta =$ parameters controlling expected values.
- $\psi(\theta) =$ log-partition function ensures normalization.

Analogy: logistic regression, Ising models, multivariate Gaussian, and many other popular statistical models are exponential families.

Exponential Random Graph Models (ERGMs)

$$\Pr(G = g) \propto \exp\{\theta_1 \cdot \#\text{edges}(g) + \theta_2 \cdot \#\text{triangles}(g) + \cdots\}$$

- θ_1 tunes density, θ_2 tunes clustering, etc.
- $\text{ER}(N, p)$ is the special case $\theta_1 \neq 0$, $\theta_2 = \cdots = 0$.
- ERGMs allow us to encode economic/social forces: incentives for transitive closure, reciprocity, or block structures.
- But: hard to analyze, computationally challenging.

Recursive growth: preferential attachment

- Networks often grow over time.
- **Preferential attachment:** new node attaches to existing node i with probability proportional to $\deg(i)$.
- "Rich get richer" $\rightarrow$ hubs emerge.

Recursive growth: preferential attachment

- Networks often grow over time.
- **Preferential attachment:** new node attaches to existing node i with probability proportional to $\deg(i)$.
- "Rich get richer" $\rightarrow$ hubs emerge.

Result: degree distribution follows a *power law*.

- Few very large hubs.
- Many low-degree nodes.
- Matches data: web, citation networks, finance.

Why do random models matter?

- ER provides a clean *baseline* for chance fluctuations.
- Clustering & preferential attachment capture realistic features:
  - Interbank markets: dense cores, high clustering.
  - Trade: triadic closure, regional clusters.
  - Knowledge diffusion: preferential attachment in citations.
- Comparing models $\Rightarrow$ shows which properties are "non-random" in data.

Summary

- $G(N, p)$ = simplest random graph; tractable but unrealistic.
- Subgraph thresholds (triangles) show how clustering begins.
- Clustering coefficient: vanishes in ER, but high in real networks.
- Static (ERGMs) and recursive (preferential attachment) models add realism.
- Small-world phenomena + hubs: explain short distances and inequalities.

Exercise

Determine the Clustering Coefficient for nodes w and y.