Universitat Pompeu Fabra Barcelona

# Lecture 8 · Random Networks II

## Networks, Crowds and Markets

# Today's Lecture

1. Probability recap: Chebyshev and Hoeffding inequality.
2. Degree distribution in Erdős–Rényi graphs.
3. Threshold phenomena and giant component.

# Degree distribution: finite $N$ concentration bounds

# Concentration: Chebyshev (simple but general)

**Theorem ( Chebyshev inequality )**

For any r.v. $X$ with mean $\mu$ and variance $\sigma^2$,

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

**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

Markov's inequality: If $Z \geq 0$ then $\mathbb{P}(Z \geq t) \leq \frac{1}{t}\mathbb{E}[Z]$.

Markov's inequality follows immediately from the following calculation,

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

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 $X = \sum_{i=1}^{n} Z_i$ with independent $Z_i \in [0,1]$ 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

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

note much better behavior of $t_0$ as a function of $\delta$

# Sharper concentration: Hoeffding for Binomial

**Theorem (Hoeffding inequality)**

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

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

**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

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

note much better behavior of $t_0$ as a function of $\delta$

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

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

# 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 $t > 0$.

Suppose we now want to provide a bound for the degrees all $v \in V$.

# 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 $t > 0$.

Suppose we now want to provide a bound for the degrees all $v \in V$.

Take $t_0 = \sqrt{\frac{N-1}{2}\log(\frac{2N}{\delta})}$ we get that, for any fixed $v \in V$,

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

# 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 $t > 0$.

Suppose we now want to provide a bound for the degrees all $v \in V$.

Take $t_0 = \sqrt{\frac{N-1}{2}\log(\frac{2N}{\delta})}$ we get that, for any fixed $v \in V$,

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

**Union bound**: For any two events $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

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

# 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 $t > 0$.

Suppose we now want to provide a bound for the degrees all $v \in V$.

Take $t_0 = \sqrt{\frac{N-1}{2}\log(\frac{2N}{\delta})}$ we get that, for any fixed $v \in V$,

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

**Union bound**: For any two events $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

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

e.g. $N = 1001$, $\delta = 0.05$, $p = 0.1$. Then with prob. $\geq 0.95$ all degrees lie in $(100 - 72.8, 100 + 72.8) \;=\; (27.2, 172.8)$.

# 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 $N$.

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

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

# Average degree: dense vs sparse graphs

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

**Dense regime:** $(p_N)$ tends to a constant $c > 0$.

- $\mathbb{E}[\deg(v)] \approx cN$ grows linearly with $N$.
- The number of edges $L \approx c\binom{N}{2}$.
- Not a realistic large network, but a useful contrast.

**Sparse regime:** $p_N = \lambda/N$ (or smaller).

- $\mathbb{E}[\deg(v)] \approx \lambda$ stays constant as $N \to \infty$.
- The total number of edges $L \approx \lambda N/2$ grows linearly with $N$.

# Average degree: dense vs sparse graphs

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

**Dense regime:** $(p_N)$ tends to a constant $c > 0$.

- $\mathbb{E}[\deg(v)] \approx cN$ grows linearly with $N$.
- The number of edges $L \approx c\binom{N}{2}$.
- Not a realistic large network, but a useful contrast.

**Sparse regime:** $p_N = \lambda/N$ (or smaller).

- $\mathbb{E}[\deg(v)] \approx \lambda$ stays constant as $N \to \infty$.
- The total number of edges $L \approx \lambda N/2$ grows linearly with $N$.

**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**.

**Dense regime:** $(p_N)$ tends to a constant $c > 0$.

- With high probability (remember we ignore constants here):

$$\Delta \;=\; (N-1)p \,+\, O\big(\sqrt{N \log N}\big).$$

(use Slide 6 to argue for this asymptotic formula)

# Maximum degree in $G(N, p)$

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

**Dense regime:** $(p_N)$ tends to a constant $c > 0$.

- With high probability (remember we ignore constants here):

$$\Delta \;=\; (N-1)p \;+\; O\big(\sqrt{N \log N}\big).$$

(use Slide 6 to argue for this asymptotic formula)

**Sparse regime:** $p_N = \lambda/N$ (or smaller).

- Each $\deg(v) \approx \mathrm{Pois}(\lambda)$ — mean $\lambda$.
- By extreme–value theory for Poisson tails:

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

This is very thin tailed: $N = 10^3, 10^6, 10^{12}$ gives $\frac{\log N}{\log \log N} = 4.3, 6.3, 9.2.$
In real networks we observe "hubs".

# 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).

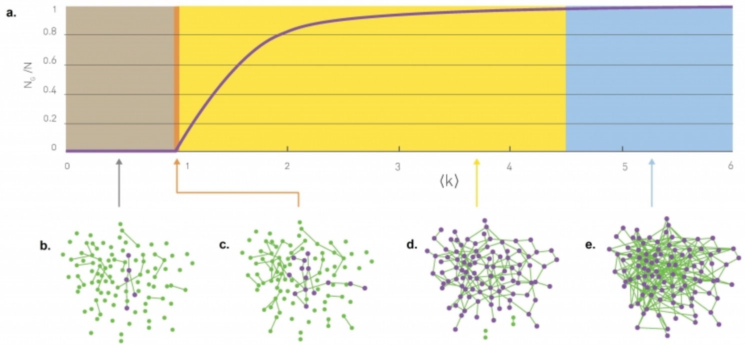
# Regimes of $G(N,p)$ (sparse case $p = c/N$)

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

$$\mathbb{E}[\deg(v)] = c.$$

- **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.

# Why the giant component matters (econ/social)

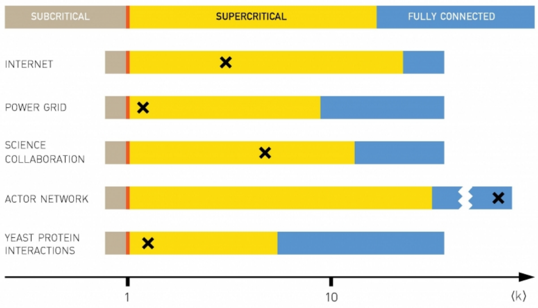
Consider the world's friendship network:

- 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.

# Where are real networks?




Most real-world networks live **well above the critical point**.

They are highly connected (often even "superconnected"), yet they also exhibit additional structure (clustering, hubs, communities).

The ER model a *baseline*: it shows that above $c = 1$, large-scale connectivity is the default, but real networks have richer features.

# Connectivity threshold in $G(N, p)$

**Theorem**

The threshold for connectivity in $G(N, p)$ is

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

More precisely:

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

Here, $\omega(N)$ means any function that grows to infinity (however slowly).
Examples: $\log \log N$, $\sqrt{\log N}$, or even $\log \log \log N$.

# Idea of proof (intuition)

- A vertex is isolated with probability

$$\mathbb{P}(v \text{ isolated}) \; = \; (1-p)^{N-1} \; \approx \; e^{-pN}.$$

- Expected number of isolated vertices:

$$\mathbb{E}[N_0] \approx N e^{-pN}.$$

- If $p = c\frac{\log N}{N}$, then

$$\mathbb{E}[N_0] \approx N^{1-c}.$$

- For $c < 1$, $\mathbb{E}[N_0] \to \infty$; many isolated vertices $\to$ disconnected.

  For $c > 1$, $\mathbb{E}[N_0] \to 0$; isolated vertices disappear.

**Careful:** No isolated vertices do not automatically imply connectivity. However, one can show that once all isolated vertices disappear, all other components merge into one giant component w.h.p.

# 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 $\overline{\mathrm{deg}}(G)$ close to theoretical $\mathbb{E}[\mathrm{deg}]$.