upf. Universitat Pompeu Fabra Barcelona

Seminar 2 Networks, Crowds and Markets

Random Networks

Exercise 1: Betweeness centrality.

We defined the *betweenness centrality* of a vertex x in a graph G as

$$b(v) = \sum_{s,t \neq v} \frac{\sigma_{st}(v)}{\sigma_{st}},$$

where $\sigma_{st}(v)$ denotes the number of shortest s-t-paths in G that go through v, and σ_{st} denotes the number of shortest s-t-paths.

(a) Consider a path on N vertices. Calculate the betweenness centralities. At which vertex are they maximised?

(b) Consider a tree on n vertices with a vertex v of degree k, so that its removal would divide the tree into k disjoint subtrees compromising $n_1, \ldots, n_k$ vertices. Show that

$$b(v) = \tfrac{1}{2}\left((N-1)^2 - \sum_{i=1}^{k} n_i^2\right).$$

Hint: each pair of vertices in different subtrees contributes 1 to $b(v)$.

Solution to Exercise 1

(a) Path on N vertices. Label the path as $1, 2, \ldots, N$. Shortest paths are unique. A vertex i (with $1 < i < N$) lies on the unique shortest path between s and t iff

$$s < i < t \quad \text{or} \quad t < i < s.$$

Number of unordered pairs $\{s, t\}$ with $s < i < t$ is $(i-1)(N-i)$. Thus

$$b(i) = (i-1)(N-i), \qquad i = 2, \ldots, N-1,$$

and $b(1) = b(N) = 0$. The quadratic $(i-1)(N-i)$ is maximised for vertices in the middle:

- If N is odd: unique maximum at $i = \frac{N+1}{2}$.
- If N is even: two symmetric maxima at $i = \frac{N}{2}$ and $i = \frac{N}{2} + 1$.

(b) Tree split into k subtrees.
Removing v splits G into k components with sizes $n_1, \ldots, n_k$. A pair $\{s, t\}$ of vertices contributes 1 to $b(v)$ iff s and t lie in *different* components (the unique path must then pass through v).
Number of such unordered pairs is

$$\sum_{1 \le i < j \le k} n_i n_j .$$

Using

$$\Big(\sum_{i=1}^{k} n_i \Big)^2 = \sum_{i=1}^{k} n_i^2 + 2 \sum_{1 \le i < j \le k} n_i n_j$$

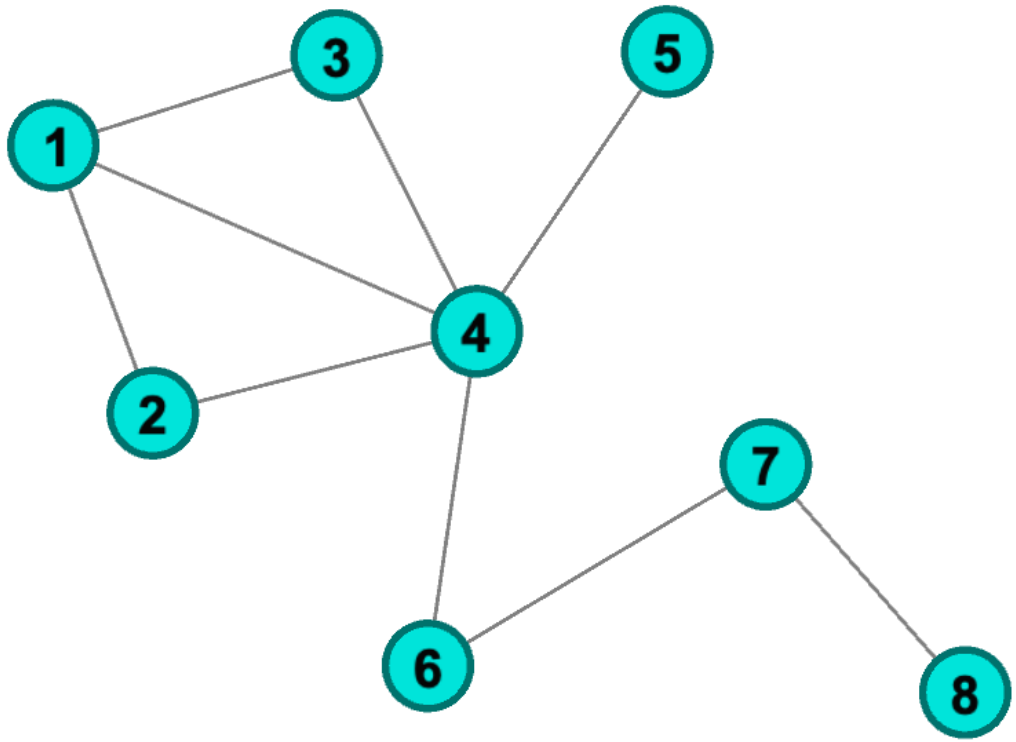
and $\sum_{i=1}^{k} n_i = N - 1$, we get

$$\sum_{1 \le i < j \le k} n_i n_j = \frac{(N-1)^2 - \sum_{i=1}^{k} n_i^2}{2},$$

which is exactly the claimed formula for $b(v)$.

Exercise 2: Centrality Measures

For the network shown below, compute all the centrality measures you know: Degree centrality, Closeness centrality, Betweenness centrality, Eigenvector centrality.



All measures except the eigenvector centrality can be computed by hand. Compare which nodes are most important under each criterion.

Solution to Exercise 2 (sketch)

Degree centrality (just degrees):

$$\deg(1) = 3,\ \deg(2) = 2,\ \deg(3) = 2,\ \deg(4) = 5,\ \deg(5) = 1,\ \deg(6) = 2,\ \deg(7) = 2,\ \deg(8) = 1.$$

So node 4 is clearly the degree hub.

Closeness centrality (rounded):

v	1	2	3	4	5	6	7	8
$C_{\text{close}}(v)$	0.50	0.47	0.47	0.70	0.44	0.58	0.44	0.32

Again node 4 is the most central; node 6 comes second.

Betweenness centrality (normalised, rounded):

v	1	2	3	4	5	6	7	8
$C_{\text{betw}}(v)$	0.02	0	0	0.74	0	0.48	0.29	0

Nodes 4 and 6 sit on many shortest paths; 4 is a strong bridge between left and right parts of the graph.

Eigenvector centrality (rounded):

v	1	2	3	4	5	6	7	8
$x(v)$	0.48	0.38	0.38	0.60	0.21	0.25	0.10	0.04

Takeaway: every reasonable centrality ranks node 4 as most important; node 6 is also central for betweenness/closeness but less so for degree alone.

Associated code

```
import networkx as nx
import numpy as np

# Define the graph
edges = [
    (1,2), (1,3), (1,4), (2,4), (3,4),
    (4,5), (4,6), (6,7), (7,8)]

G = nx.Graph()
G.add_edges_from(edges)
centrality = nx.eigenvector_centrality_numpy(G)
print("\nEigenvector centrality:")
for node, c in centrality.items():
    print(f"Node {node}: {c:.4f}")
```

Exercise 3: Bounds on $\lambda_{\max}$

Show that for any simple undirected graph with adjacency A, the largest eigenvalue $\lambda_{\max}(A)$ is:

1. at least the average degree of G,
2. at most the maximum degree.

Hint:

- $\lambda_{\max}(A) = \max_{\|\mathbf{x}\|=1} \mathbf{x}^\top A\mathbf{x}$.
- $2x_i x_j \leq x_i^2 + x_j^2$.

Solution to Exercise 3

Let d_i be degrees, $\overline{d} = \frac{1}{N}\sum_i d_i$ the average degree, and $\Delta = \max_i d_i$.

(1) Lower bound: $\lambda_{\max} \geq \overline{d}$.
Take $x = \frac{1}{\sqrt{N}}\mathbf{1}$ so that $\|x\| = 1$. Then

$$x^\top A x = \frac{1}{N}\mathbf{1}^\top A \mathbf{1} = \frac{1}{N}\sum_i d_i = \overline{d}.$$

Since $\lambda_{\max} = \max_{\|x\|=1} x^\top A x$, we obtain

$$\lambda_{\max} \;\geq\; \overline{d}.$$

(2) Upper bound: $\lambda_{\max} \leq \Delta$.
For any x,

$$x^\top A x = \sum_{(i,j)\in E} 2x_i x_j \leq \sum_{(i,j)\in E} (x_i^2 + x_j^2) = \sum_i d_i x_i^2 \leq \Delta \sum_i x_i^2 = \Delta \|x\|^2,$$

using $2x_i x_j \leq x_i^2 + x_j^2$ and $d_i \leq \Delta$.
For any unit vector x we therefore have $x^\top A x \leq \Delta$, hence

$$\lambda_{\max} = \max_{\|x\|=1} x^\top A x \;\leq\; \Delta.$$

Exercise 4: Link probability distribution.

Given an ER graph with $N = 15$ and probability $p = 0.1$, determine:

a) What is the distribution of L?

b) What is the probability that $L \geq 15$?
- Computing this by hand may be hard (see next slide).
- Check that the Hoeffding bound is not very good in this case.

c) Find smallest $\ell \in \mathbb{N}$ such that $\mathbb{P}(L \geq \ell) \leq 0.05$ (use the code).
- Use the one sided Hoeffding as an alternative way to construct such interval. What do you observe?

d) Do we expect a Giant Component?

One-sided Hoeffding: $X = \sum_{i=1}^{n} Z_i$ with $Z_i \in [0, 1]$ then

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

Solution to Exercise 4

Number of possible edges: $M = \binom{15}{2} = 105$.

(a) L is the total number of present edges, so $L \sim \mathrm{Binomial}(M, p) = \mathrm{Binomial}(105, 0.1)$.

(b) $\mathbb{P}(L \geq 15) = 1 - \mathbb{P}(L \leq 14)$ can be computed numerically (see code). It is about

$$\mathbb{P}(L \geq 15) \approx 0.10.$$

Hoeffding with $n = 105$, mean $\mu = 10.5$, $t = 4.5$ gives a very loose upper bound

$$\mathbb{P}(L - \mu \geq 4.5) \leq \exp\Big(-\frac{2t^2}{n}\Big) = \exp\Big(-\frac{2 \cdot 4.5^2}{105}\Big) \approx 0.68.$$

(c) We look for the smallest ℓ with $\mathbb{P}(L \geq \ell) \leq 0.05$. Numerically one finds

$$\mathbb{P}(L \geq 16) \approx 0.058 > 0.05, \quad \mathbb{P}(L \geq 17) \approx 0.032 < 0.05,$$

so $\ell = 17$ is the first value with tail probability ≤ 0.05.
Using Hoeffding with $t = \ell - \mu$, the bound suggests a much larger threshold (around $\ell \approx 24$), again very conservative.

(d) The expected degree is

$$\mathbb{E}[\deg(v)] = (N - 1)p = 14 \cdot 0.1 = 1.4.$$

In the asymptotic $G(N, c/N)$ picture the critical value is $c = 1$. Here $c \approx 1.4 > 1$, so *for large N with the same average degree* we are in the supercritical regime and would expect a giant component. For $N = 15$ we just expect a reasonably large component, but giant is less meaningful.

Associated code

```
from scipy.stats import binom

# Parameters
n = 105     # number of trials
p = 0.1     # success probability

# Compute P(X >= 16) = 1 - P(X <= 14)
prob = 1 - binom.cdf(14, n, p)

print(prob)
```

Exercise 5: Degree and average degree in ER graphs

Let $X = \deg(v)$ for a fixed vertex in the $ER(N, p)$ graph:

- Compute the mean and the variance of $\deg(v)$.
- Compute the covariance of $\deg(u)$ and $\deg(v)$ for $u \neq v$.
- Are $\deg(u)$ and $\deg(v)$ independent?

Let $Y = \overline{\deg}(G)$ be the average degree in the $ER(N, p)$ graph.

- Show that $\mathbb{E}[Y] = (N-1)p$.
- What is the distribution of Y?
- Use the previous exercise to compute $\mathbb{P}(Y - (N-1)p \geq 2)$.
- Develop Hoeffding bound for $\mathbb{P}(Y - (N-1)p \geq t)$.

Solution to Exercise 5 (1/2)

Degrees of fixed vertices. For a fixed vertex v, the degree is $X = \deg(v) = \sum_{w \neq v} I_{\{vw\}}$, where $I_{\{vw\}}$ is the indicator of edge $\{v, w\}$. There are $N - 1$ independent Bernoulli(p) terms, so

$$X \sim \mathrm{Bin}(N-1, p), \quad \mathbb{E}[X] = (N-1)p, \quad \mathrm{Var}(X) = (N-1)p(1-p).$$

For distinct u, v,

$$\deg(u) = \sum_{w \neq u} I_{\{uw\}}, \qquad \deg(v) = \sum_{w \neq v} I_{\{vw\}}.$$

The only common indicator is $I_{\{uv\}}$. Thus

$$\mathrm{Cov}(\deg(u), \deg(v)) = \mathrm{Var}(I_{\{uv\}}) = p(1-p).$$

In particular, $\deg(u)$ and $\deg(v)$ are *not* independent.

Average degree. Let L be the number of edges. Then $\overline{\deg}(G) = Y = \frac{1}{N} \sum_v \deg(v) = \frac{2L}{N}$. But $L \sim \mathrm{Bin}\left(\binom{N}{2}, p\right)$, so Y is just a scaled binomial: $Y = \frac{2}{N} L$. Hence

$$\mathbb{E}[Y] = \frac{2}{N} \mathbb{E}[L] = \frac{2}{N} \binom{N}{2} p = (N-1)p.$$

Solution to Exercise 5 (2/2)

To compute $\mathbb{P}(Y - (N-1)p \geq 2)$, note

$$Y - (N-1)p \geq 2 \quad \Longleftrightarrow \quad L - \mathbb{E}[L] \;\geq\; \frac{N}{2} \cdot 2 = N.$$

So $\mathbb{P}(Y - (N-1)p \geq 2) = \mathbb{P}\Big(L - \mathbb{E}[L] \geq N\Big)$, which can be evaluated from the binomial distribution of L.

Hoeffding for Y. Write $L = \sum_e I_e$ as sum of $m = \binom{N}{2}$ independent Bernoulli(p). Hoeffding gives

$$\mathbb{P}(L - \mathbb{E}[L] \geq s) \leq \exp\Big(-\frac{2s^2}{m}\Big).$$

Since $Y = \frac{2}{N}L$, $Y - \mathbb{E}[Y] = \frac{2}{N}(L - \mathbb{E}[L])$. Thus $Y - \mathbb{E}[Y] \geq t$ implies $L - \mathbb{E}[L] \geq \frac{tN}{2}$, and

$$\mathbb{P}(Y - \mathbb{E}[Y] \geq t) \leq \exp\Big(-\frac{2(tN/2)^2}{m}\Big) = \exp\Big(-\frac{t^2N^2}{2\binom{N}{2}}\Big).$$

For large N, $\binom{N}{2} \approx N^2/2$, so the exponent is about $-t^2$.

Exercise 6: Connectivity Threshold in $G(N, p)$

Let $\omega(N)$ be any sequence that grows to infinity (however slowly). Examples: $\log\log(N)$, $\sqrt{\log(N)}$, or even $\log\log\log(N)$.

In $G(N, p)$, the expected number of isolated vertices is

$$\mathbb{E}[N_0] = N(1-p)^{N-1} \approx Ne^{-p(N-1)}.$$

Suppose G is the $\mathrm{ER}(N, p)$

(a) Derive the formula above.

(b) Let $p = \frac{\log N - \omega(N)}{N}$ with $\omega(N) \to +\infty$. Show $\mathbb{E}[N_0] \to \infty$ and conclude G is disconnected w.h.p.

(c) Let $p = \frac{\log N + \omega(N)}{N}$ with $\omega(N) \to +\infty$. Show $\mathbb{E}[N_0] \to 0$. Can we conclude G is connected w.h.p.?
(actually $\mathbb{E}[N_k] \to 0$ for any fixed k)

Solution to Exercise 6

(a) Expected number of isolated nodes. A vertex v is isolated if none of the $N-1$ possible edges from v is present: $\mathbb{P}(v \text{ isolated}) = (1-p)^{N-1}$. Let I_v be the indicator of v isolated. Then

$$N_0 = \sum_{v=1}^{N} I_v, \qquad \mathbb{E}[N_0] = \sum_{v} \mathbb{E}[I_v] = N(1-p)^{N-1}.$$

Using $(1-p)^{N-1} \approx e^{-p(N-1)}$ gives the approximation.

(b) $p = \frac{\log N - \omega(N)}{N}$, $\omega(N) \to \infty$. Then

$$p(N-1) = (\log N - \omega(N))\Big(1 - \frac{1}{N}\Big) = \log N - \omega(N) + o(1).$$

Hence

$$\mathbb{E}[N_0] = N(1-p)^{N-1} \approx Ne^{-p(N-1)} \approx Ne^{-(\log N - \omega(N))} = e^{\omega(N)} \to \infty.$$

So the expected number of isolates diverges. Hence with high probability there are isolated vertices, so G is disconnected w.h.p.

(c) $p = \frac{\log N + \omega(N)}{N}$, $\omega(N) \to \infty$. Now

$$p(N-1) = (\log N + \omega(N))(1 - o(1)) = \log N + \omega(N) + o(1).$$

Thus $\mathbb{E}[N_0] \approx Ne^{-(\log N + \omega(N))} = e^{-\omega(N)} \to 0$. So isolated vertices disappear in expectation. In fact one can show $\mathbb{P}(N_0 > 0) \to 0$, and more strongly $\mathbb{E}[N_k] \to 0$ for each fixed k, which implies that w.h.p. there are no small components at all, hence G is connected w.h.p.
But note: $\mathbb{E}[N_0] \to 0$ *alone* does not automatically imply connectivity; one needs to rule out small components of size ≥ 2 as well.

Exercise 7: Triangles in ER models

Let T be the number of triangles in the $\mathrm{ER}(N, p)$ graph.

What is the expected number of T?

Bonus: Compute $\mathrm{Var}(T)$.

Solution to Exercise 7

Let $I_{\{i,j,k\}}$ be the indicator that vertices i, j, k form a triangle.

Expectation.
There are $\binom{N}{3}$ triples, each forms a triangle with probability p^3 (all three edges must be present). So

$$T = \sum_{i<j<k} I_{\{i,j,k\}}, \qquad \mathbb{E}[T] = \binom{N}{3} p^3.$$

Variance (sketch).

$$\mathrm{Var}(T) = \sum_{i<j<k} \mathrm{Var}(I_{\{i,j,k\}}) + 2 \sum_{\{i,j,k\} \neq \{i',j',k'\}} \mathrm{Cov}(I_{\{i,j,k\}}, I_{\{i',j',k'\}}).$$

Each I is Bernoulli(p^3), so

$$\mathrm{Var}(I_{\{i,j,k\}}) = p^3(1 - p^3).$$

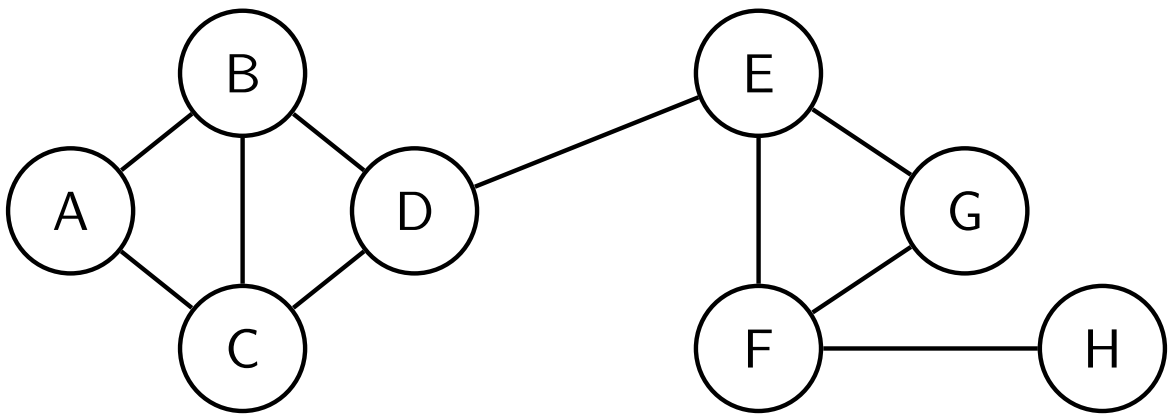
Covariances are 0 unless the two triangles share at least one edge. One then enumerates the overlapping cases (share exactly one edge, or share a path of length 2) to obtain an explicit formula. The key point: expectations and variances can be computed from independence of edges and combinatorics of overlapping triples.

Additional exercises

Exercise: Degree vs. Eigenvector Centrality

Compute the **eigenvector centrality** of all nodes in the undirected graph below (you may use Python/NetworkX). Then compare with **degree centrality**.

- Which nodes are important under each measure?
- Why can these rankings differ?



Solution: Degree vs Eigenvector Centrality

Degree centrality.
Left cluster: A, B, C, D all have degree 2 or 3; right cluster: E, F, G also relatively high degree; H is a leaf. Many nodes have similar degrees.

Eigenvector centrality.
Eigenvector centrality rewards nodes connected to already central nodes. The bridge edge $D-E$ links two dense clusters, so:

- D and E get very high eigenvector centrality (they connect two well-connected groups).
- Nodes like A, B, C and F, G are also central, but a bit less than D, E.
- H (a leaf hanging off F) gets low eigenvector centrality.

Takeaway: degree centrality just counts neighbours, while eigenvector centrality prefers nodes that connect to *well-connected* neighbours. Bridge nodes between dense parts can become very important under eigenvector centrality even if their degree is not the largest.

The associated code

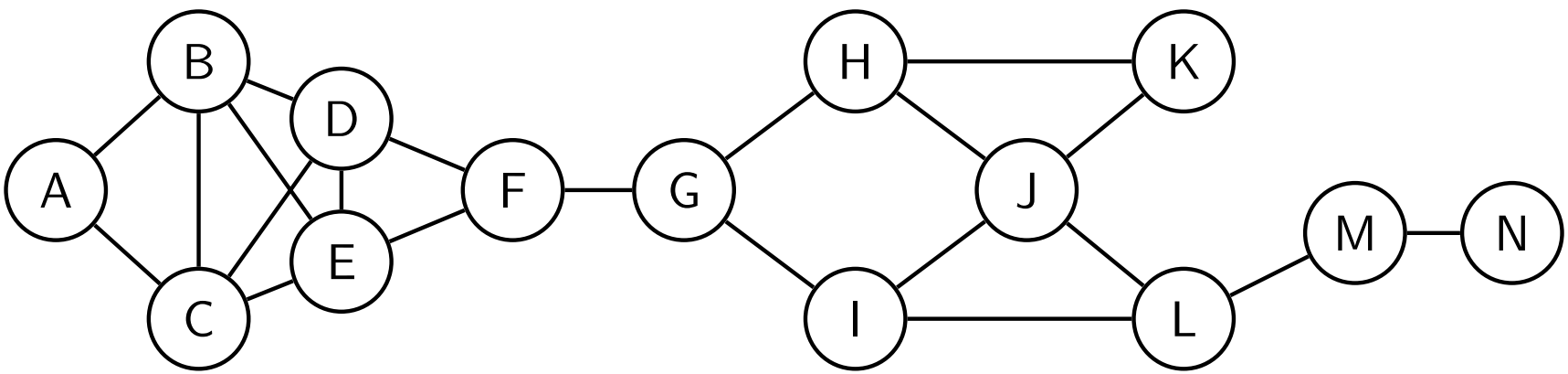
```
import networkx as nx

G = nx.Graph()
edges = [
 ('A','B'),('A','C'),('B','C'),('B','D'),('C','D'),
 ('D','E'),('E','F'),('E','G'),('F','G'),('F','H')
]
G.add_edges_from(edges)
x = nx.eigenvector_centrality(G, max_iter=1000, tol=1e-6)
x_rounded = {node: round(val, 3) for node, val in x.items()}
print(x_rounded)
```

Exercise: Closeness and Betweenness centrality

For the graph shown below, compute for every node:

- **Closeness centrality** $C_{\text{close}}(v)$
- **Betweenness centrality** $C_{\text{betw}}(v)$

Which measure better identifies bridge nodes in this network?



Solution: Closeness vs Betweenness

Closeness. Nodes roughly in the middle of the whole graph (around F, G, J) tend to have the largest closeness: they have small average distance to all others.
Betweenness. Nodes that lie on many shortest paths between left and right parts have high betweenness:

- F and G are clear bridges between the left and right clusters.
- J also lies on many shortest paths inside the right cluster and towards the tail MN.

So betweenness centrality highlights the bridge nodes F, G, J more sharply than closeness, which also rewards nodes that are central inside dense parts of the graph.

The associated code

```
G = nx.Graph()
edges = [
    ('A','B'),('B','C'),('C','A'),('B','D'),('D','E'),
    ('E','C'),('B','E'),('C','D'),('D','F'),('E','F'),
    ('F','G'),('G','H'),('G','I'),('H','J'),('I','J'),
    ('J','K'),('J','L'),('H','K'),('I','L'),('L','M'),
    ('M','N')]
G.add_edges_from(edges)

# --- Centralities ---
close = nx.closeness_centrality(G)                    # closeness
betw  = nx.betweenness_centrality(G, normalized=True)  # between

print("Node  Closeness   Betweenness")
for v in sorted(G.nodes(), key=lambda x: betw[x], reverse=True):
    print(f"{v:>4}  {close[v]:8.3f}   {betw[v]:10.3f}")
```

Exercise: PageRank (small directed web)

Consider the directed network with adjacency matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

(a) Write down the transition matrix P.

(b) Find the stationary vector satisfying $P^\top \pi = \pi$, $\mathbf{1}^\top \pi = 1$.

(c) With teleportation $\alpha = 0.85$,

$$P_\alpha = \alpha P + (1-\alpha)\frac{1}{4}\mathbf{1}\mathbf{1}^\top,$$

compute the PageRank vector π (solve $P_\alpha^\top \pi = \pi$, $\mathbf{1}^\top \pi = 1$).

Solution: PageRank (small directed web)

(a) Transition matrix P. The random walk on this directed graph has:

$$P = \begin{pmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

(b) Stationary vector of P. Solve $P^\top \pi = \pi$ with $\mathbf{1}^\top \pi = 1$. Since 4 is an absorbing state and every node can reach 4, we get

$$\pi = (0, 0, 0, 1)^\top.$$

All probability mass eventually accumulates at node 4.

(c) PageRank with teleportation $\alpha = 0.85$. With teleportation,

$$P_\alpha = \alpha P + (1 - \alpha)\frac{1}{4}\mathbf{1}\mathbf{1}^\top,$$

and we solve $P_\alpha^\top \pi = \pi, \qquad \sum_i \pi_i = 1$. Numerically,

$$\pi \approx (0.079,\ 0.071,\ 0.098,\ 0.752)^\top.$$

Node 4 still has the largest PageRank, but nodes 1 and 3 get non-zero scores because teleportation rescues probability mass from the absorbing state.

Follow-up: Understanding PageRank qualitatively

The network is

$$1 \to \{2, 4\}, \quad 2 \to 3, \quad 3 \to \{1, 4\}, \quad 4 \to \varnothing.$$

Questions:

(a) Which node acts as a sink in the random walk defined by P? What happens to probability mass over time if there is no teleportation?

(b) After adding teleportation ($\alpha = 0.85$), which nodes PageRank values *increase* the most? Why does this happen?

(c) What is the qualitative effect of changing α?

- As $\alpha \to 1$, what happens to π?
- As $\alpha \to 0$, what does π converge to?

(d) Suppose we add one new edge $4 \to 1$. How would that affect the PageRank scores? (Hint: which node now becomes a stronger hub?)

Solution: PageRank follow-up (qualitative)

(a) Node 4 is a sink (no outgoing links). Without teleportation, probability mass drifts towards 4 and stays there; the stationary distribution is $\pi = (0, 0, 0, 1)$.

(b) With teleportation, some mass is periodically redistributed to all nodes. Nodes that receive flow from the sink (via teleportation) and are well-connected (like 1 and 3) gain PageRank relative to the no-teleportation case (where they had zero in the limit).

(c) As $\alpha \to 1$, P_α approaches P and PageRank approaches the stationary distribution of the raw random walk (concentrated on sinks / dead ends). As $\alpha \to 0$, P_α tends to the uniform matrix $\frac{1}{4}\mathbf{1}\mathbf{1}^\top$, and PageRank converges to the uniform distribution $(1/4, 1/4, 1/4, 1/4)$.

(d) Adding an edge $4 \to 1$ removes the sink. Node 1 becomes a stronger hub (receiving links from 3 and 4), so its PageRank increases, and the scores become more balanced across $\{1, 4\}$.

Exercise: Random Walk Stationary Distribution

Let G be a connected, undirected graph with adjacency A and degree matrix D. The random walk transition is $P = D^{-1}A$.

(a) Show that $\pi_i = \dfrac{\deg(i)}{\sum_j \deg(j)}$ is a stationary distribution: $P^\top \pi = \pi$.

(b) Why is this stationary distribution unique when G is connected?

Solution: Random walk stationary distribution (1/2)

(a) Stationarity. Let $\deg(i)$ be the degree of node i and m the number of edges. Then

$$\pi_i = \frac{\deg(i)}{\sum_j \deg(j)} = \frac{\deg(i)}{2m}.$$

Recall $P_{ij} = \mathbb{P}(i \to j) = \frac{A_{ij}}{\deg(i)}$. Compute $(P^\top \pi)_j$:

$$(P^\top \pi)_j = \sum_i P_{ij}\, \pi_i = \sum_i \frac{A_{ij}}{\deg(i)} \cdot \frac{\deg(i)}{2m} = \frac{1}{2m} \sum_i A_{ij} = \frac{\deg(j)}{2m} = \pi_j.$$

So $P^\top \pi = \pi$.

Solution: Random walk stationary distribution (2/2)

(b) Uniqueness when G is connected. Let

$$S \;=\; D^{1/2} P D^{-1/2} \;=\; D^{-1/2} A D^{-1/2}.$$

Then P and S are similar, so $P^\top x = x$ iff $Sy = y$ with $y = D^{-1/2}x$. Set $L_{\text{norm}} := I - S$. For any vector y,

$$y^\top L_{\text{norm}} y = y^\top (I - S) y = \frac{1}{2} \sum_{i,j} A_{ij} \Big(\frac{y_i}{\sqrt{\deg(i)}} - \frac{y_j}{\sqrt{\deg(j)}} \Big)^2 \;\geq 0.$$

Now suppose x is a stationary vector: $P^\top x = x$. Then $y = D^{-1/2}x$ satisfies $Sy = y$, i.e. $L_{\text{norm}} y = 0$, so

$$y^\top L_{\text{norm}} y = 0.$$

By the formula above, every term in the sum must be zero, hence for every edge (i, j),

$$\frac{y_i}{\sqrt{\deg(i)}} = \frac{y_j}{\sqrt{\deg(j)}}.$$

Because G is connected, this propagates along paths, so

$$\frac{y_i}{\sqrt{\deg(i)}} = c \quad \text{for all } i$$

for some constant c. Thus

$$y_i = c \sqrt{\deg(i)} \quad \Rightarrow \quad x_i = \sqrt{\deg(i)}\, y_i = c \, \deg(i).$$

So any stationary vector x is a scalar multiple of the degree vector $(\deg(i))_i$. Imposing the normalisation $\sum_i x_i = 1$ fixes c uniquely, and hence the stationary distribution is unique:

$$\pi_i = \frac{\deg(i)}{\sum_j \deg(j)}.$$

Exercise: Spectrum of $P = D^{-1}A$

Let G be a connected, undirected graph. Show / verify that the eigenvalues of $P = D^{-1}A$ lie in $[-1, 1]$.

- Hint: Relate P to the symmetric matrix $S = D^{1/2}PD^{-1/2} = D^{-1/2}AD^{-1/2}$.
- Why does $\lambda = 1$ correspond to the stationary distribution? Why simple (multiplicity 1) if G is connected?

(Optional: verify numerically on a medium graph.)

Solution: Spectrum of $P = D^{-1}A$

Consider

$$S = D^{1/2} P D^{-1/2} = D^{-1/2} A D^{-1/2}.$$

Then S is real symmetric, and P is similar to S, so they have the same eigenvalues.

Eigenvalues in $[-1, 1]$.
For any vector x with $\|x\| = 1$,

$$x^\top S x = \sum_{(i,j)\in E} \frac{2x_i x_j}{\sqrt{\deg(i)\deg(j)}} \le \sum_{(i,j)\in E} \left(\frac{x_i^2}{\deg(i)} + \frac{x_j^2}{\deg(j)} \right) = \sum_i x_i^2 = 1,$$

using $2ab \le a^2 + b^2$ and counting degrees. Thus the Rayleigh quotient of S is at most 1, so $\lambda_{\max}(S) \le 1$. Applying the same argument to $-S$ gives $\lambda_{\min}(S) \ge -1$. Hence all eigenvalues of S (and therefore of P) lie in $[-1, 1]$.

$\lambda = 1$ and stationarity.
Take u with components $u_i = \sqrt{\deg(i)}$. Then

$$(Su)_i = \sum_j \frac{A_{ij}}{\sqrt{\deg(i)\deg(j)}} u_j = \sum_{j\sim i} \frac{1}{\sqrt{\deg(i)\deg(j)}} \sqrt{\deg(j)} = \frac{1}{\sqrt{\deg(i)}} \sum_{j\sim i} 1 = \sqrt{\deg(i)} = u_i.$$

So u is an eigenvector of S with eigenvalue 1. Transforming back, $D^{-1/2}u$ is an eigenvector of P with eigenvalue 1, and its entries are proportional to $\deg(i)$, i.e. to the stationary distribution from the previous exercise.
If G is connected, the random walk is irreducible, so the eigenvalue 1 is simple (eigenspace is one-dimensional). This corresponds to the uniqueness of the stationary distribution.

Exercise: Simulating the Giant Component (ER)

Simulate $G(N, p)$ with $N = 500$ and $p = c/N$ for $c \in \{0.5, 1, 1.5, 2, 3, 4\}$.

- For each c, estimate the fraction $\frac{|C_{\max}|}{N}$ of nodes in the largest component.
- Plot $\frac{|C_{\max}|}{N}$ vs. c and mark roughly where the phase transition occurs.

(You may do this as a short *optional homework* using NetworkX.)