Graph Mining
1. Graph Structure

Clement Wang

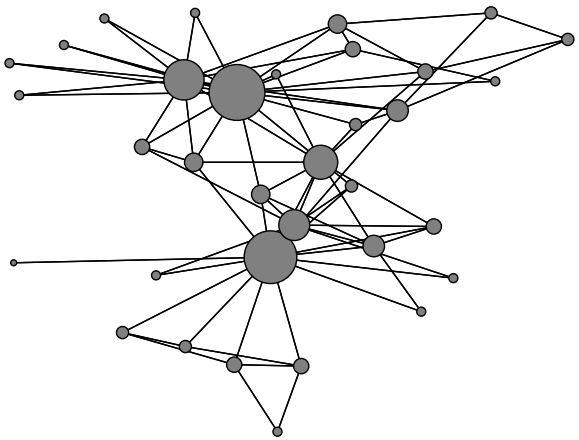
2025 – 2026

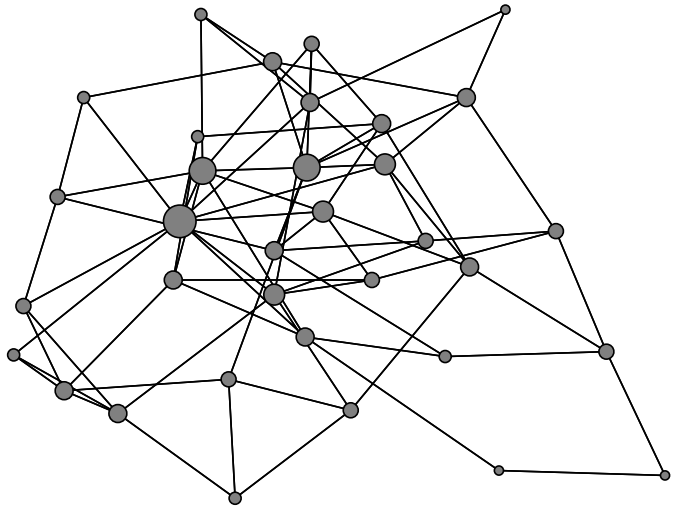


INSTITUT
POLYTECHNIQUE
DE PARIS

Motivation

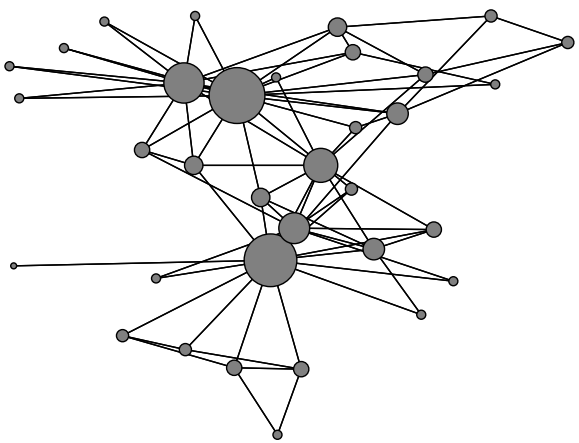
How do (large) **real graphs** look like?

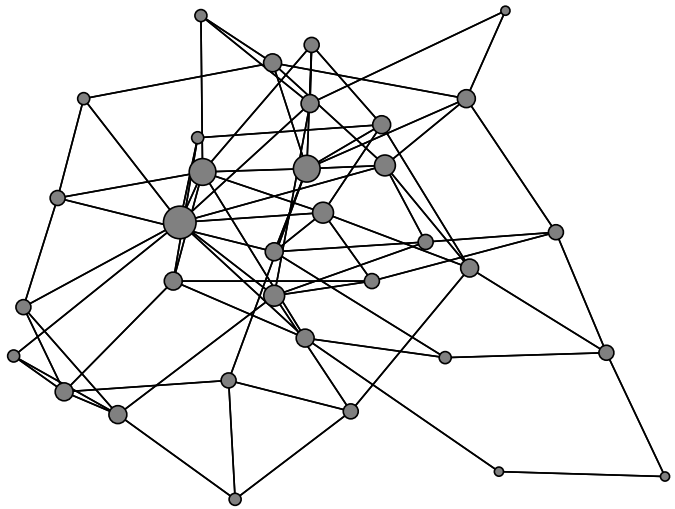




Outline

1. **The friendship paradox**
2. Scale-free property
3. Small-world property
4. Clustering property





The friendship paradox

You have less friends than your friends have.

Uniform node sampling

Sample a node X uniformly at random from V.

Expected degree

$$\mathbb{E}_U[D] = \frac{1}{n} \sum_{x \in V} d_x = \frac{2m}{n}$$

Uniform node sampling

Sample a node X uniformly at random from V.

Expected degree

$$\mathbb{E}_U[D] = \frac{1}{n} \sum_{x \in V} d_x = \frac{2m}{n}$$

Proof idea: each node has probability $\frac{1}{n}$, and $\sum_{x \in V} d_x = 2m$.

Edge sampling

Sample an edge uniformly at random, then sample one of its endpoints.

Expected degree

$$\mathbb{E}_{\text{edge}}[D] = \frac{\mathbb{E}_U[D^2]}{\mathbb{E}_U[D]} \geq \mathbb{E}_U[D]$$

Edge sampling

Sample an edge uniformly at random, then sample one of its endpoints.

Expected degree

$$\mathbb{E}_{\text{edge}}[D] = \frac{\mathbb{E}_U[D^2]}{\mathbb{E}_U[D]} \geq \mathbb{E}_U[D]$$

Proof idea: node x is sampled with probability

$$\mathbb{P}(X = x) = \frac{d_x}{2m}, \qquad \mathbb{E}_{\text{edge}}[D] = \sum_{x \in V} \frac{d_x}{2m} d_x = \frac{\sum_x d_x^2}{2m}.$$

Neighbor sampling

Sample a node Y uniformly at random, then sample one of its neighbors.

Expected degree

$$\mathbb{E}_{\mathrm{nbr}}[D] \geq \mathbb{E}_{U}[D]$$

Neighbor sampling

Sample a node Y uniformly at random, then sample one of its neighbors.

Expected degree

$$\mathbb{E}_{\mathrm{nbr}}[D] \geq \mathbb{E}_U[D]$$

Proof idea:

$$\mathbb{P}(X = x) = \frac{1}{n} \sum_{y \in V} \frac{A_{xy}}{d_y}, \qquad \mathbb{E}_{\mathrm{nbr}}[D] = \frac{1}{n} \sum_{x,y \in V} \frac{d_x}{d_y} A_{xy}.$$

$$\mathbb{E}_{\mathrm{nbr}}[D] = \frac{1}{2n} \sum_{x,y \in V} \left(\frac{d_x}{d_y} + \frac{d_y}{d_x} \right) A_{xy} \geq \frac{1}{n} \sum_{x,y \in V} A_{xy} = \frac{2m}{n} = \mathbb{E}_U[D]$$

using $z + \frac{1}{z} \geq 2$.

Sampling bias

Consider a graph with n nodes and m edges. Let D denote the degree of a sampled node.

Uniform node sampling

$$\mathbb{E}_U[D] = \frac{2m}{n}$$

Edge sampling

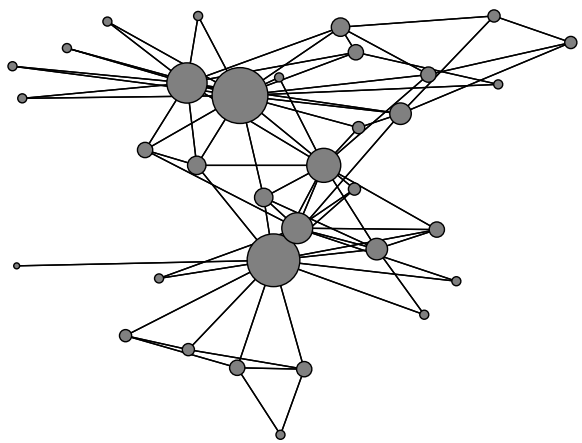
$$\mathbb{E}_{\text{edge}}[D] = \frac{\mathbb{E}_U[D^2]}{\mathbb{E}_U[D]} \geq \mathbb{E}_U[D]$$

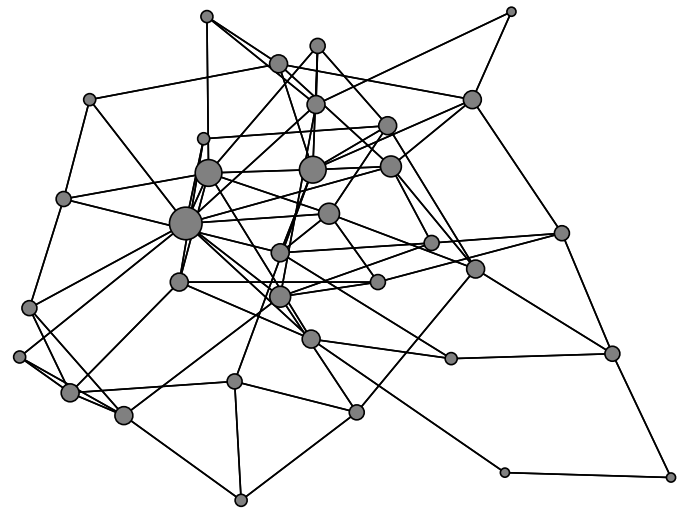
Neighbor sampling

$$\mathbb{E}_{\text{nbr}}[D] \geq \mathbb{E}_U[D]$$

Outline

1. The friendship paradox
2. **Scale-free property**
3. Small-world property
4. Clustering property





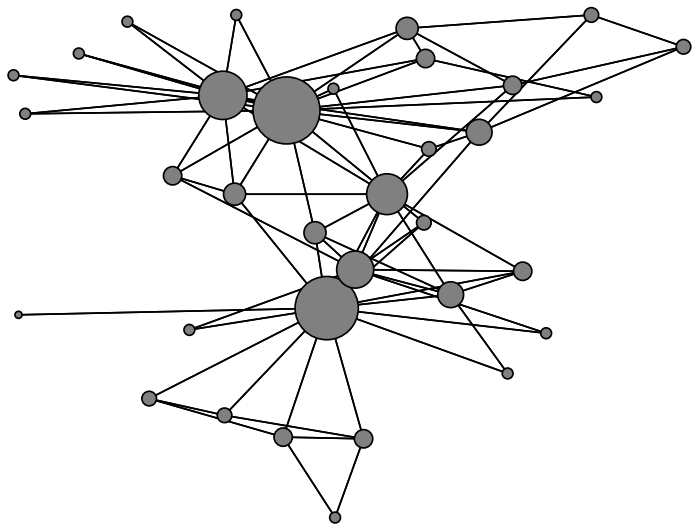
Power law

In real graphs, the degrees typically have a **power law** (= Zipf)
The degree D of a random node satisfies

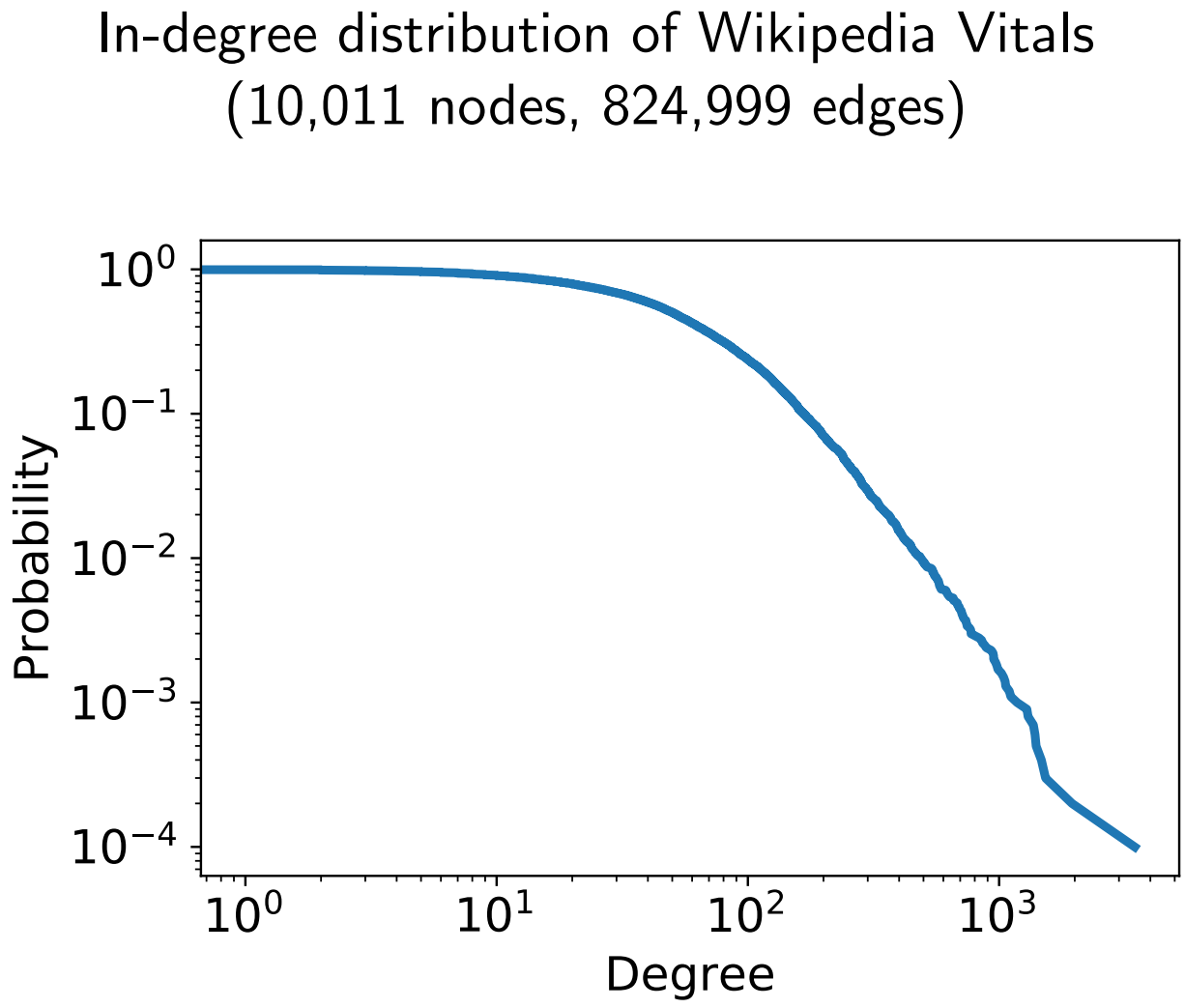
$$\mathrm{P}(D \geq k) \approx \left(\frac{k_{\min}}{k}\right)^{\alpha}$$

where

- $k_{\min}$ is the **minimum degree**
- α is the **exponent**, typically between 1 and 2



Example


In-degree distribution of Wikipedia Vitals
(10,011 nodes, 824,999 edges)
Probability
10^{0}
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{0}
10^{1}
10^{2}
10^{3}
Degree

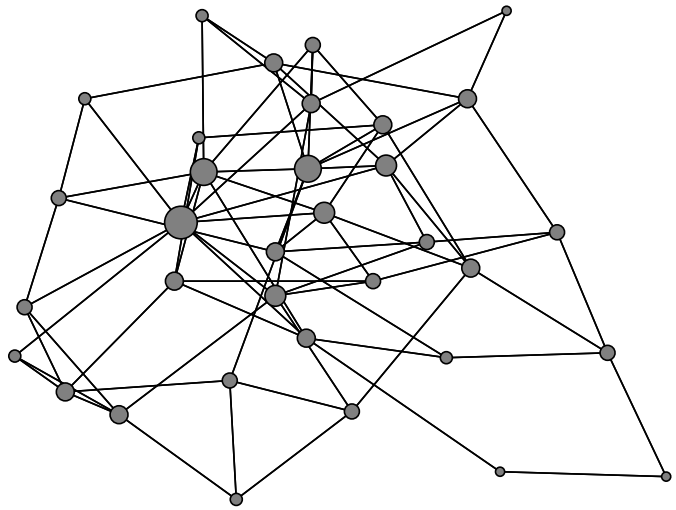
A random graph

Erdős-Rényi (1959)

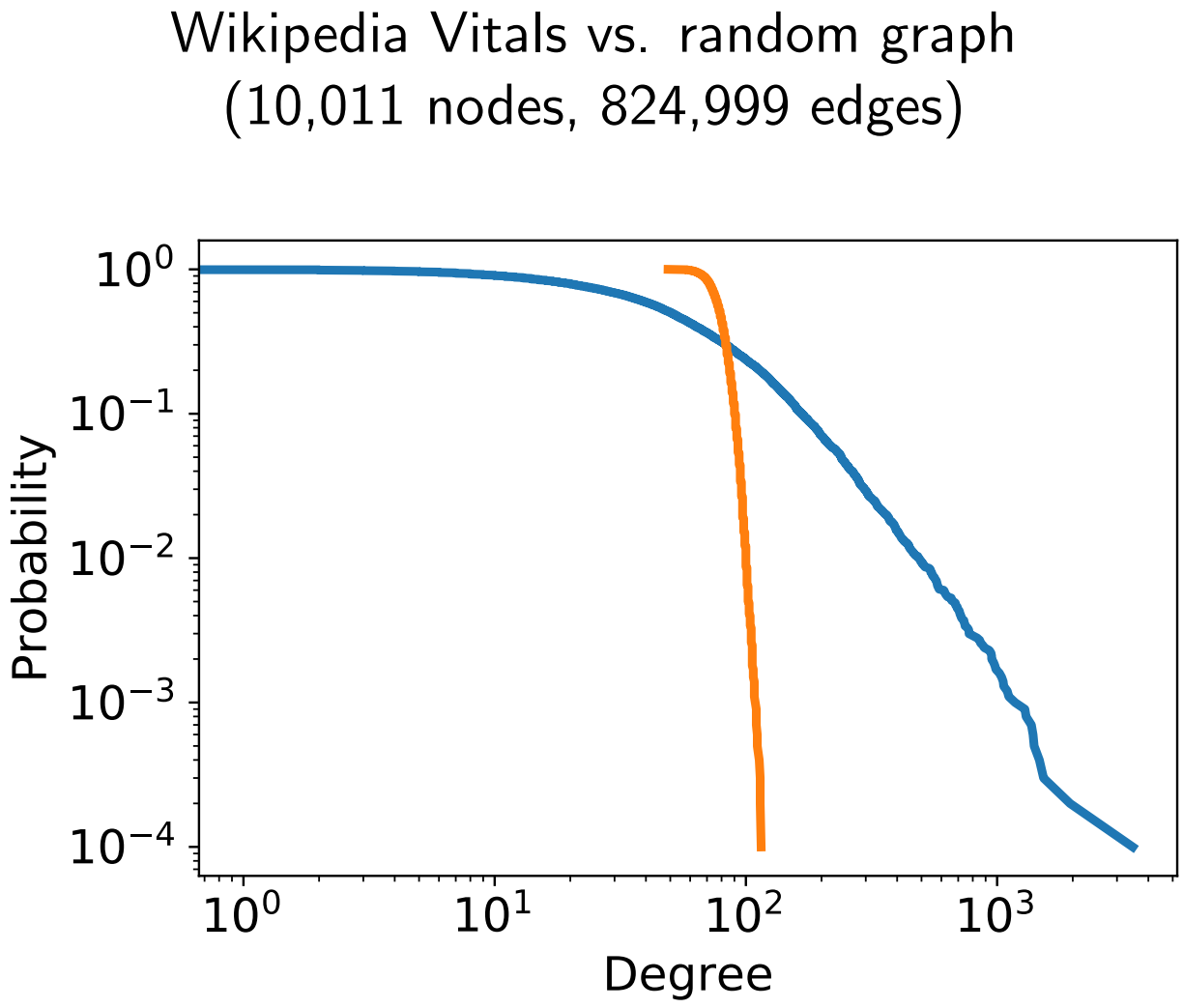
Consider n nodes, with pairs connected with probability p

Adjacency matrix = symmetric matrix with

$$A_{ij} \sim \text{Bernoulli}(p) \text{ for } i < j$$



Example

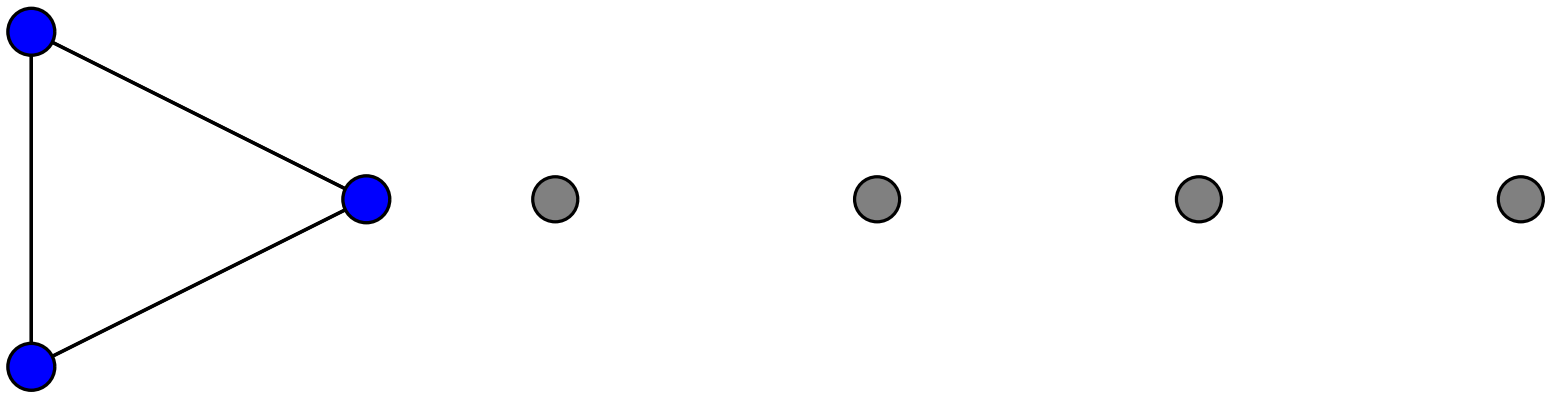

Wikipedia Vitals vs. random graph
(10,011 nodes, 824,999 edges)
Probability
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^0
10^1
10^2
10^3
Degree

Why the power law?

Barabasi-Albert model (1999)

Start from a clique of d nodes (with $d \geq 1$)

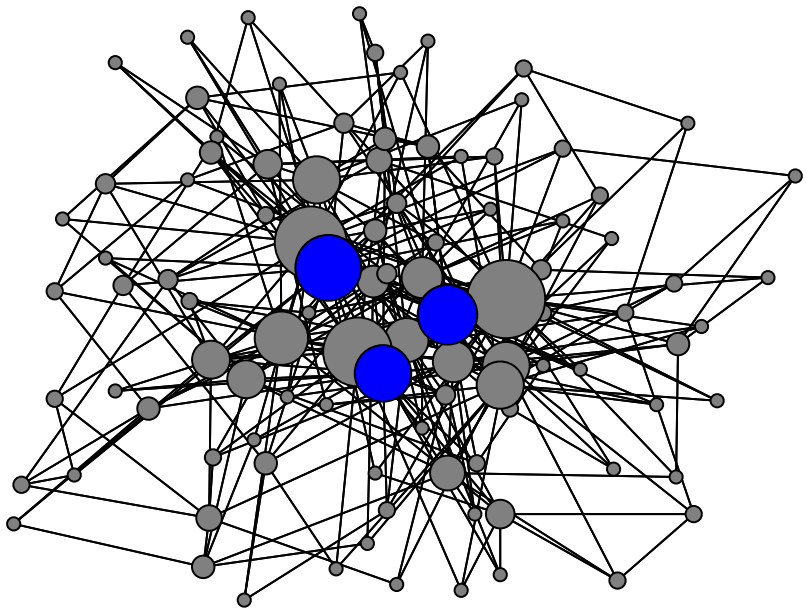
Add new nodes one at a time, each of degree d and with **preferential attachment**



"rich get richer"

Example

Graph generated from the **Barabasi-Albert model** ($n = 100, d = 3$)



The scale-free property

Let D be the degree distribution:

$$\mathrm{P}(D \geq k) = \left(\frac{k_{\min}}{k} \right)^{\alpha}$$

The scale-free property

Let D be the degree distribution:

$$\mathrm{P}(D \geq k) = \left(\frac{k_{\min}}{k} \right)^{\alpha}$$

For a typical power exponent $\alpha \in (1, 2]$, we have:

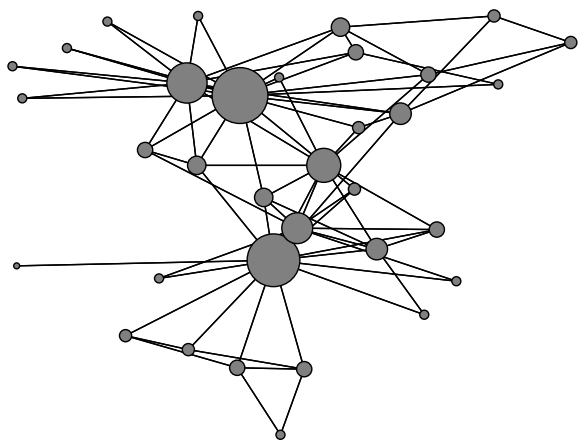
$$\mathrm{E}(D) < \infty$$

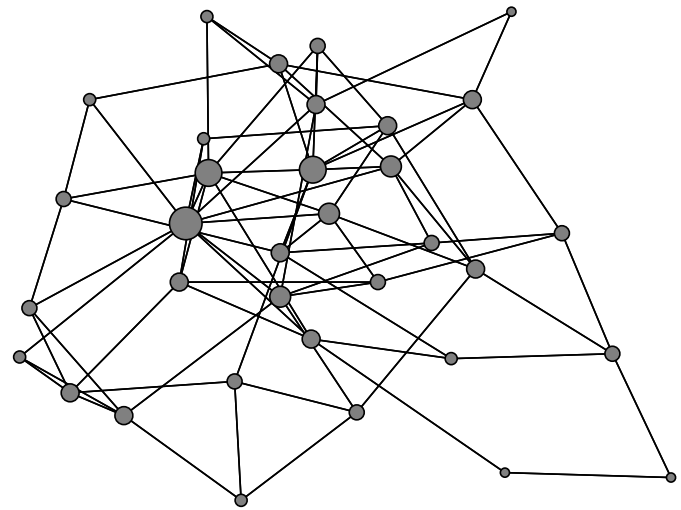
$$\mathrm{var}(D) = \infty$$

$\rightarrow$ The average degree is **not** informative!

Outline

1. The friendship paradox
2. Scale-free property
3. **Small-world property**
4. Clustering property





Small world

Which fraction of the articles are **accessible in** k **clicks** from Plato on Wikipedia?

Small world

Which fraction of the articles are **accessible in** k **clicks** from Plato on Wikipedia?

Using Wikipedia Vitals (10,011 articles):

# clicks	# nodes	proportion
1	392	4%
2	5866	59%
3	9939	99%
4	9990	99.8%

The six degrees of separation

- First mention in *Chains*, a short story by Karinthy in 1929

The six degrees of separation

- First mention in *Chains*, a short story by Karinthy in 1929
- Verified experimentally by Milgram in 1967



Source: Wikipedia

The six degrees of separation

- First mention in *Chains*, a short story by Karinthy in 1929
- Verified **experimentally** by Milgram in 1967

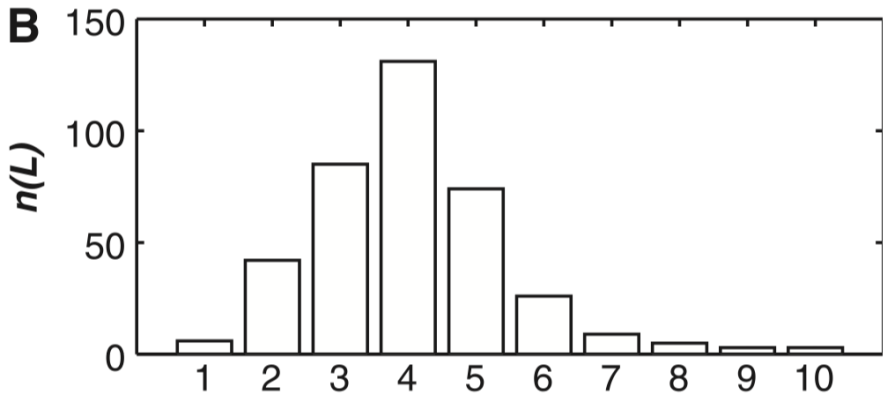


Source: Wikipedia

Emails

Dodds, Muhamad, Watts 2003

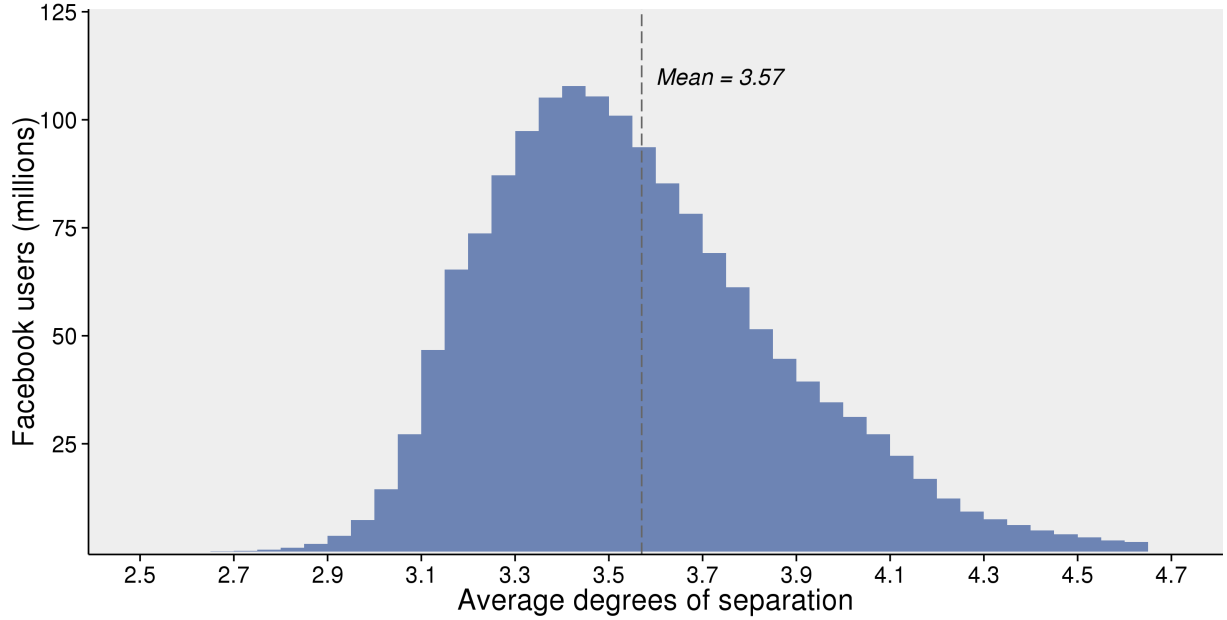
- 18 target people from all over the world
- 24,163 volunteers
- 384 successful chains
  Length of successful chains



Facebook

Bhagat, Burke, Diuk, Filiz, Edunov 2016

- Based on the 1.6 billion people active on Facebook
- Compute the average path length to any other people



The 3.5 degrees of separation of Facebook

Erdős number

- Graph of co-authors of scientific papers
- Distance to Erdős (1913-1996)

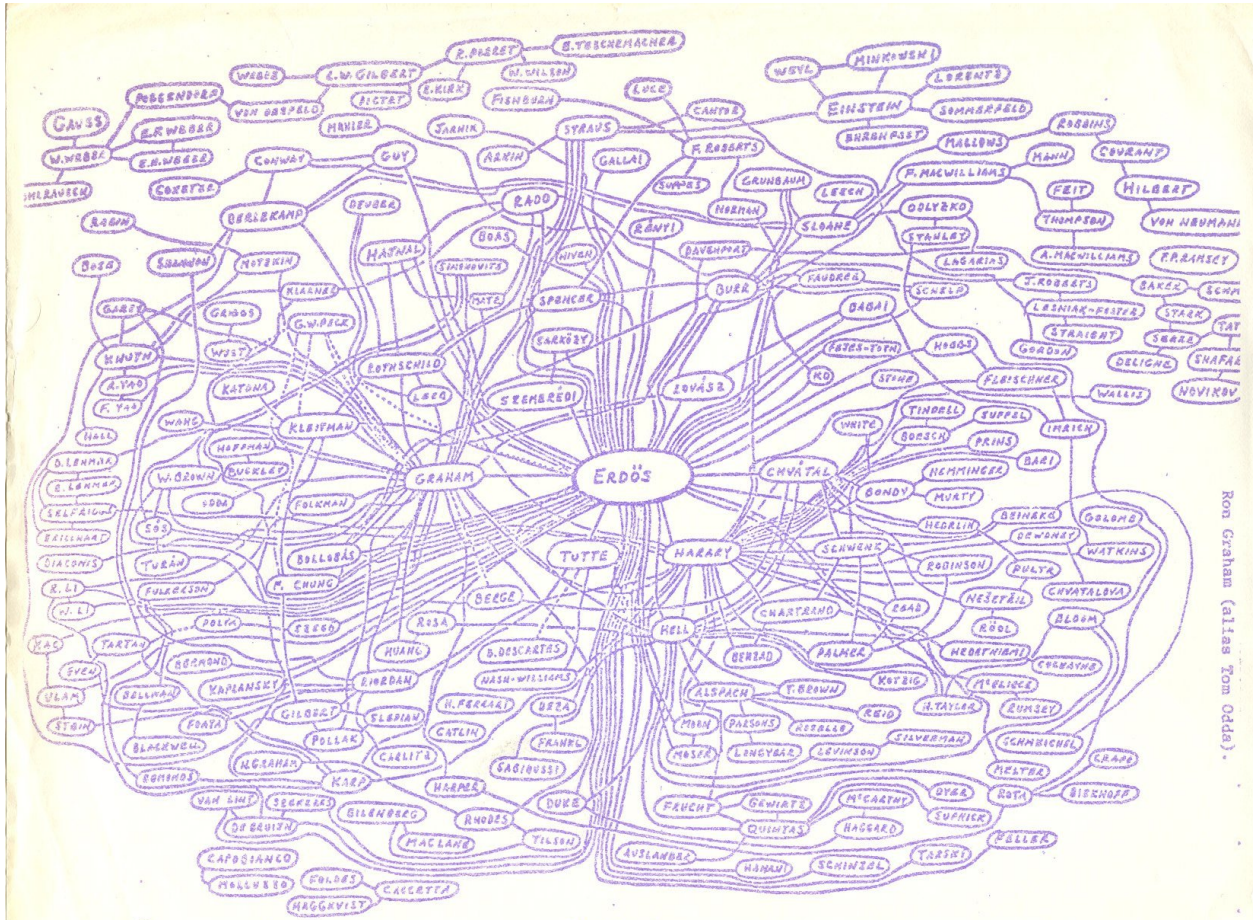


Figure 1

To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

The Bacon number

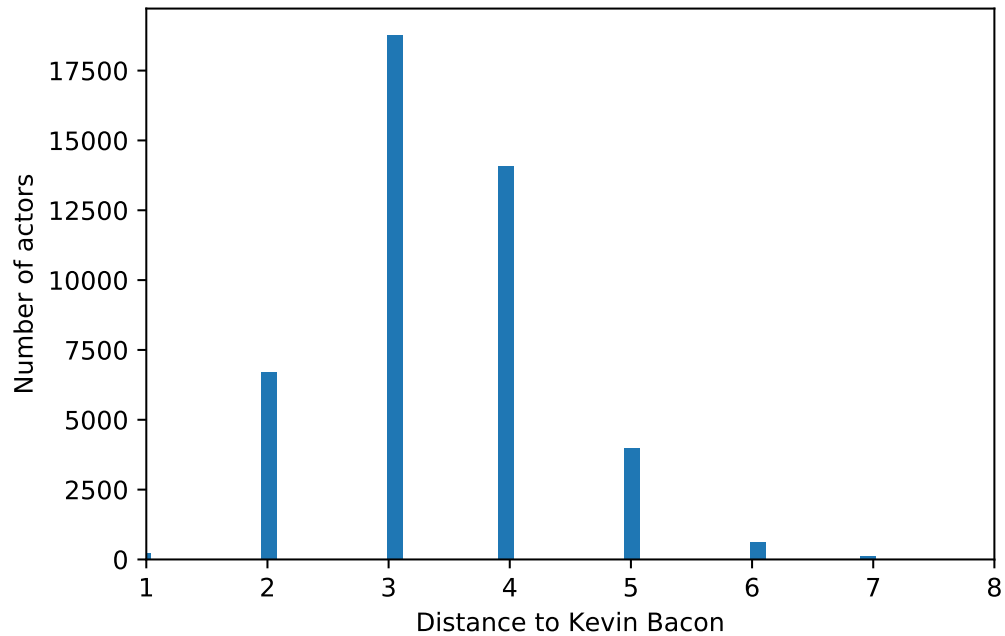
See The Oracle of Bacon

- Originated from an interview of Kevin Bacon by Premiere Magazine in 1994

The Bacon number

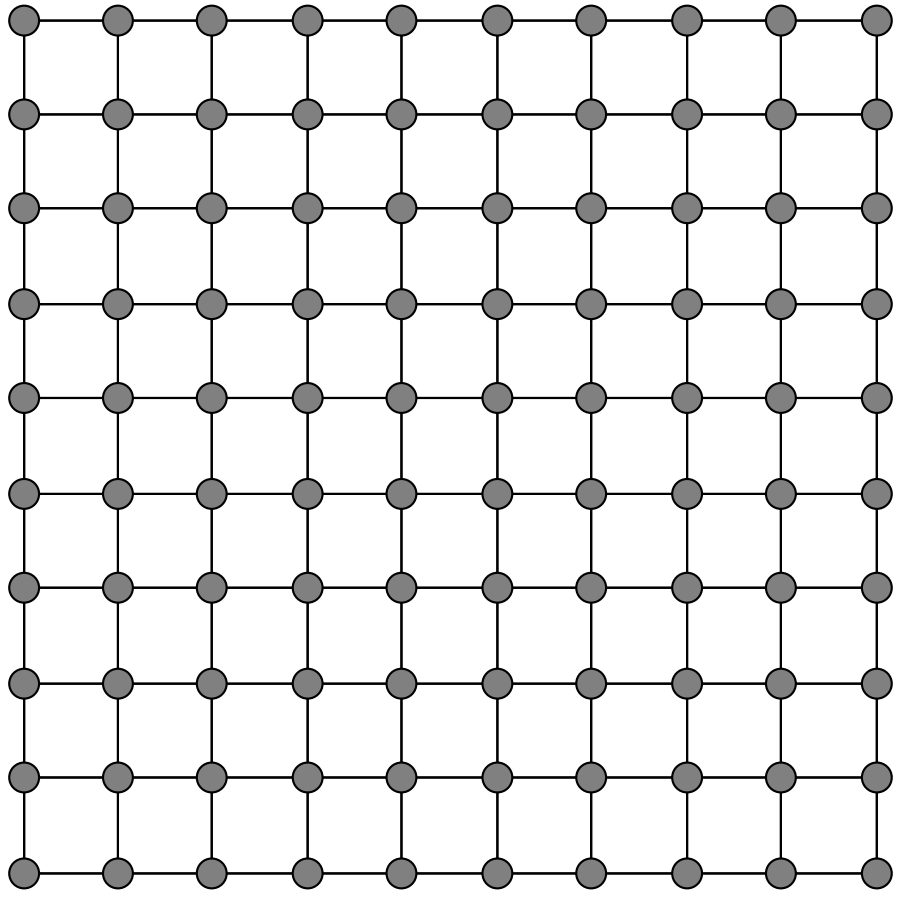
See The Oracle of Bacon

- Originated from an interview of Kevin Bacon by Premiere Magazine in 1994
- Graph of co-starring in movies

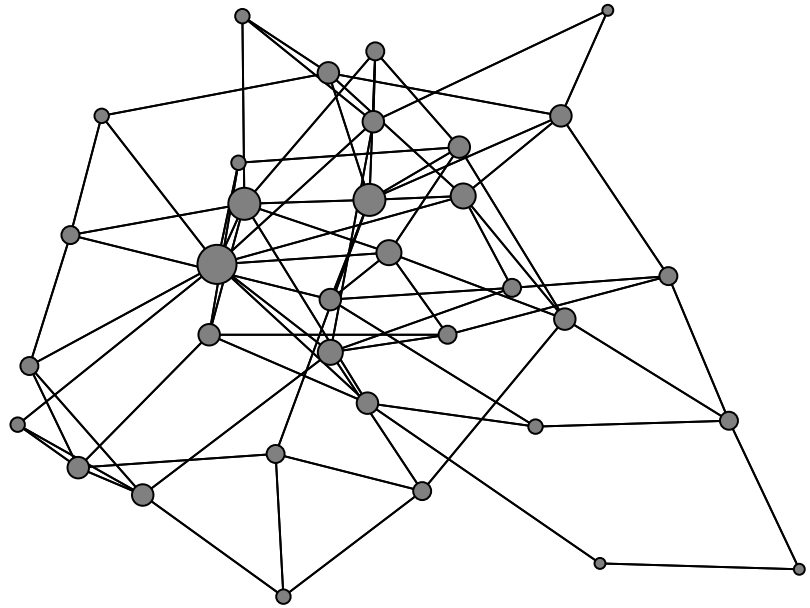


Results from YAGO database (44,586 actors)

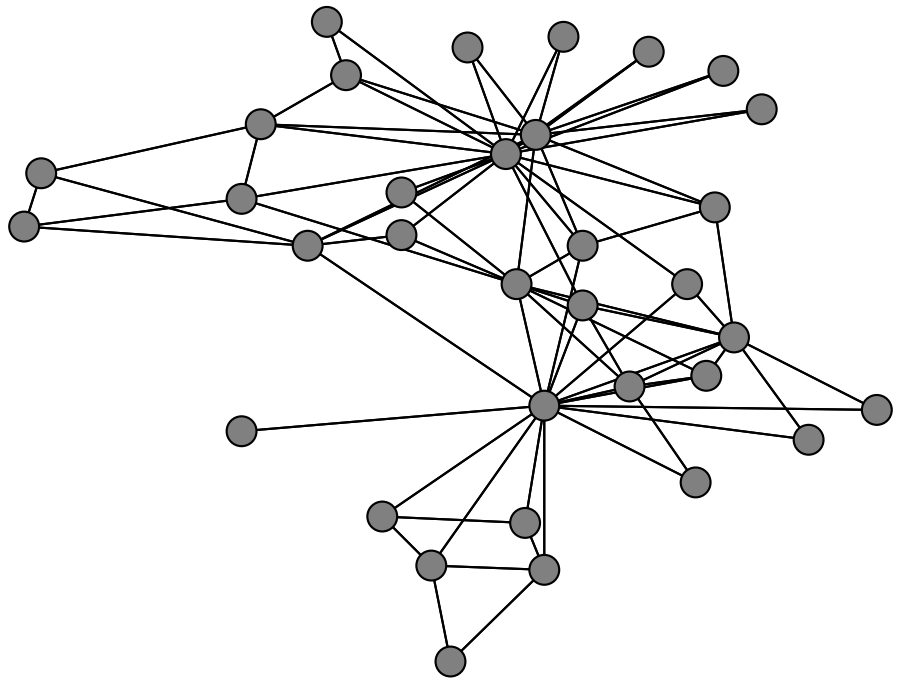
Planar graphs



Random graphs

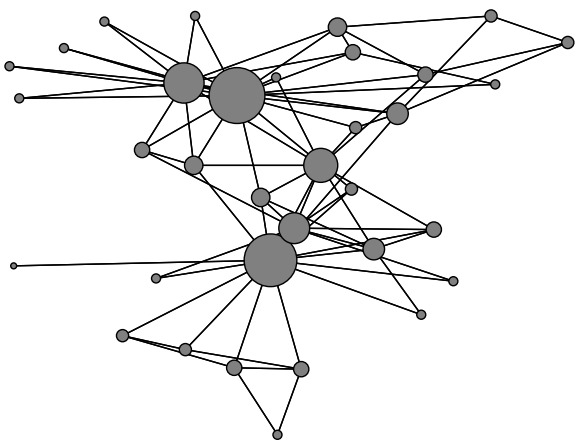


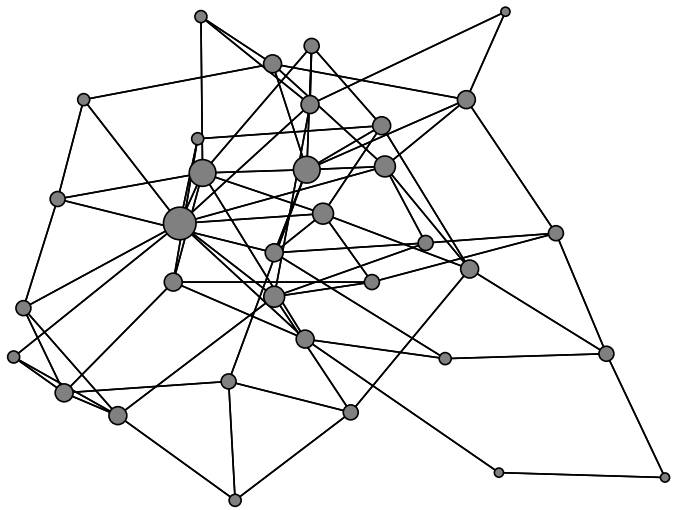
Power-law graphs



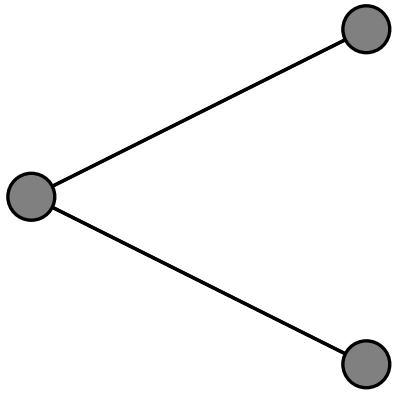
Outline

1. The friendship paradox
2. Scale-free property
3. Small-world property
4. **Clustering property**

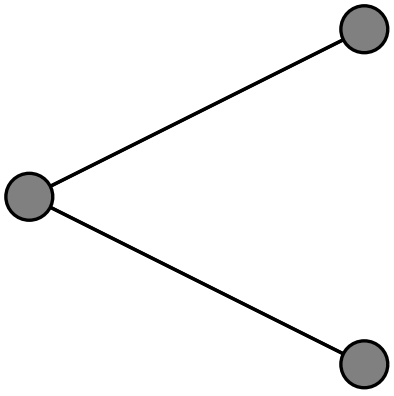


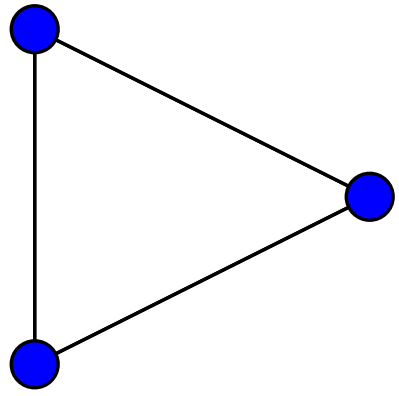


Clustering coefficient

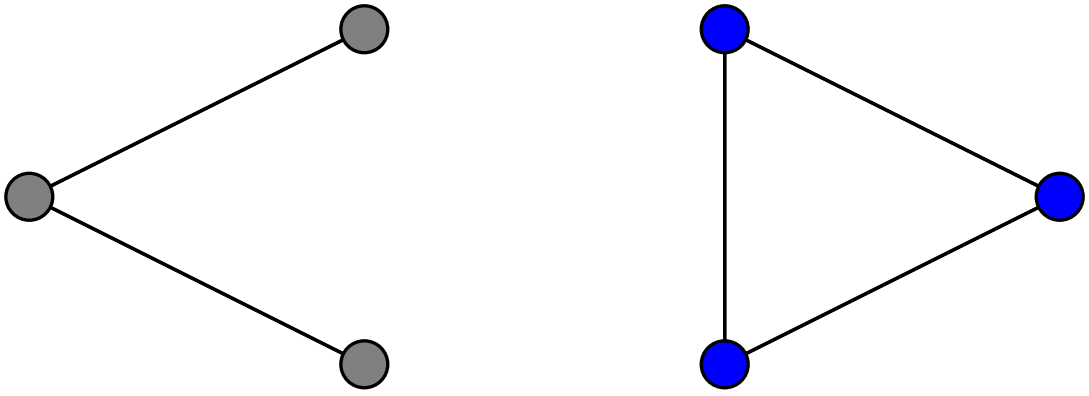


Clustering coefficient





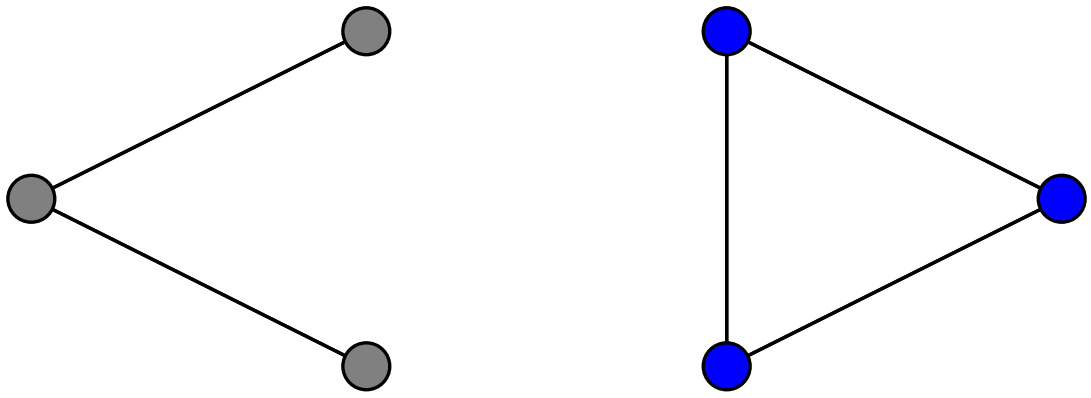
Clustering coefficient



Fraction of **closed** triplets:

$$C = \frac{3 \ \#\text{triangles}}{\sum_i \binom{d_i}{2}}$$

Clustering coefficient



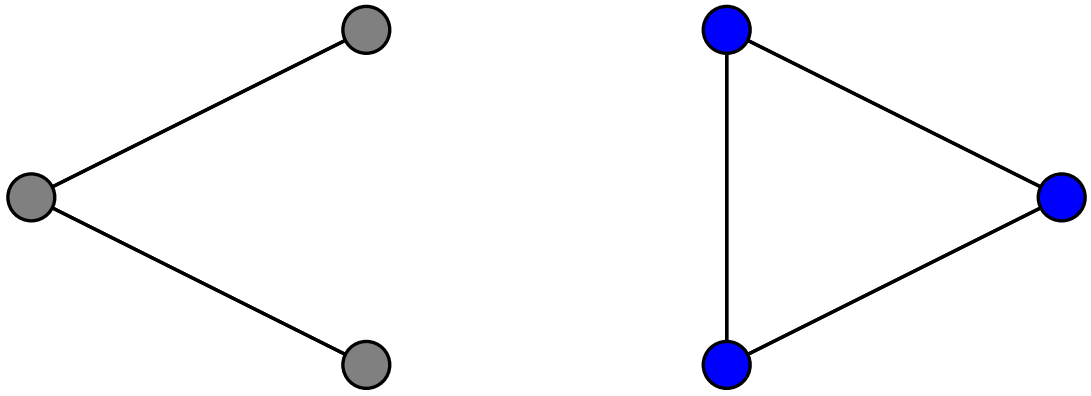
Fraction of **closed** triplets:

$$C = \frac{3 \ \#\text{triangles}}{\sum_i \binom{d_i}{2}}$$

For one node i:

$$C_i = \frac{\#\text{edges between neighbors of } i}{\binom{d_i}{2}}$$

Clustering coefficient



Graph	C
Karate Club	0.26
Les Miserables	0.50
Openflights	0.25
WikiVitals	0.21

Summary

Graph structure

1. **Friendship paradox** (sampling)
2. **Power law** (degrees)
3. **Small world** (distances)
4. **Clustering** (triangles)

