Graph Mining
SD212
2. Graph structure

Thomas Bonald

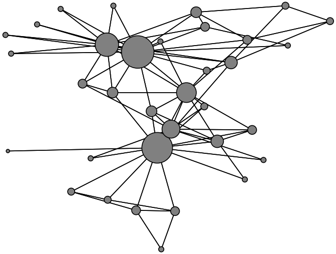
2019 – 2020

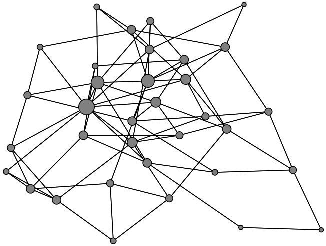
TELECOM Paris

IP PARIS

Motivation

Are real graphs **random**?



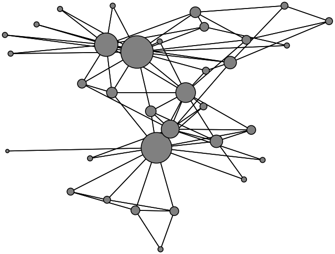


Outline

1. Degrees $\rightarrow$ power law
2. Distances $\rightarrow$ small world
3. Triangles

Power law

A few nodes have **very** high degrees (= hubs)

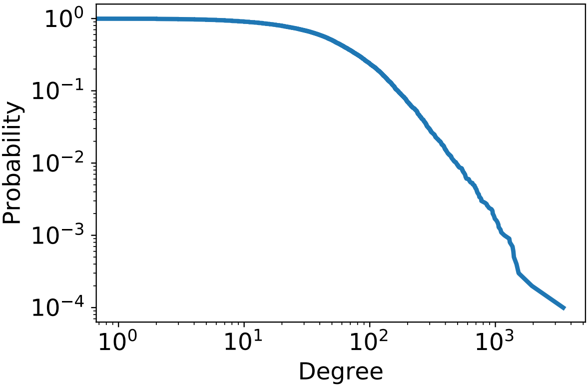


Power law:

$$\mathrm{P}(D \geq k) = \left(\frac{k_{\mathrm{m}}}{k}\right)^{\alpha} \quad \alpha > 0$$

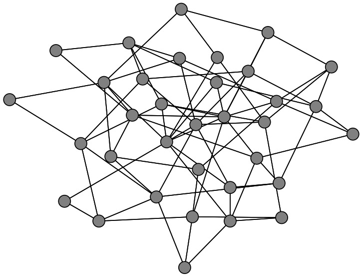
Example

In-degree distribution of Wikipedia Vitals
(10,012 nodes, average in-degree $\approx$ 80)



Erdős-Rényi model (1959)

- n nodes
- pairs connected with probability p

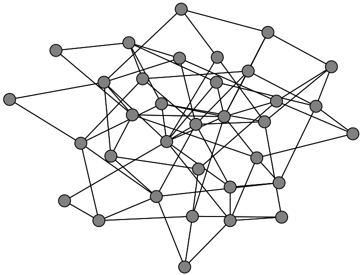


Adjacency matrix = symmetric matrix with

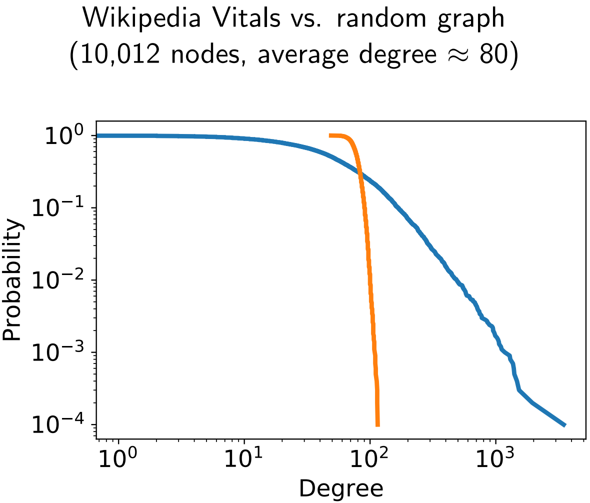
$$A_{ij} \sim \text{Bernoulli}(p) \text{ for } i < j$$

Degree distribution

- Each node pair is connected with probability p
- Degree $\sim$ **Binomial** with parameters $n-1, p$
- For large graphs, $n \to +\infty$ with $np \to \lambda$, this tends to a **Poisson** distribution with parameter λ

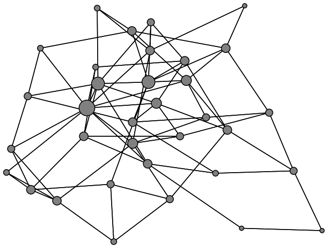


Example



Wikipedia Vitals vs. random graph
(10,012 nodes, average degree ≈ 80)
Probability
Degree
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^0
10^1
10^2
10^3

Edge sampling in random graphs



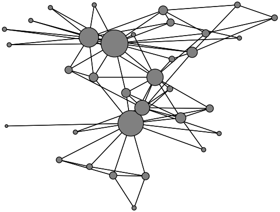
Biased Poisson distribution:

$$\mathrm{P}_\infty(D = k) \propto k\mathrm{P}_0(D = k) \propto \mathrm{P}_0(D = k | D \geq 1)$$

Expected degree:

$$\mathrm{E}_\infty(D) = \mathrm{E}_0(D) + 1$$

Edge sampling in power-law graphs



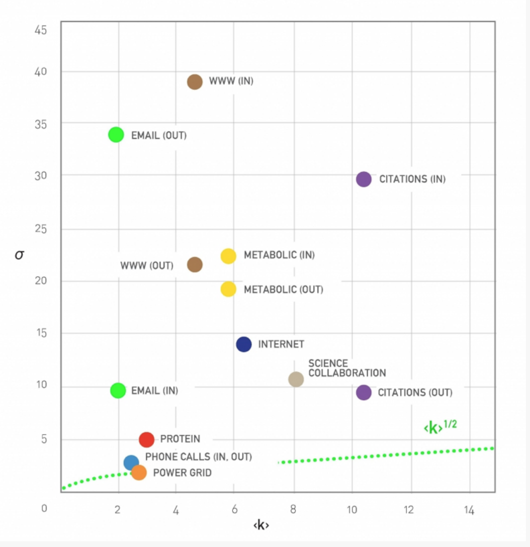
Expected degree:

$$\mathrm{E}_\infty(D) = \frac{\mathrm{E}_0(D^2)}{\mathrm{E}_0(D)} = \mathrm{E}_0(D)(1 + c_v^2)$$

where c_v is the coefficient of variation:

$$c_v = \frac{\sigma_0(D)}{\mathrm{E}_0(D)} = \frac{1}{\alpha(\alpha - 2)} \quad \alpha > 2$$

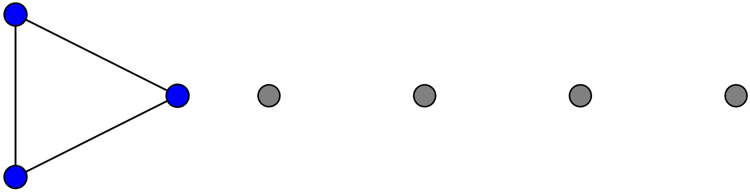
Scale-free graphs



Source: Barabasi, Network Science, 2016

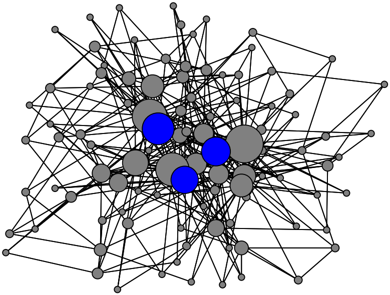
Barabasi-Albert model (1999)

- 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 ($n = 100, d = 3$)



Outline

1. Degrees $\rightarrow$ power law
2. Distances $\rightarrow$ small world
3. Triangles

Small world

How many pages are accessible in k clicks from Plato on Wikipedia?

Using **Wikipedia Vitals** (10,012 pages):

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

The six degrees of separation

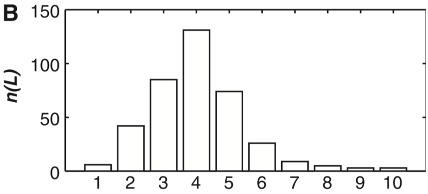
- Stated by Karinthy in 1929!
- Verified experimentally by Milgram in 1967

Source: Wikipedia

Emails

Dodds, Muhamad, Watts 2003

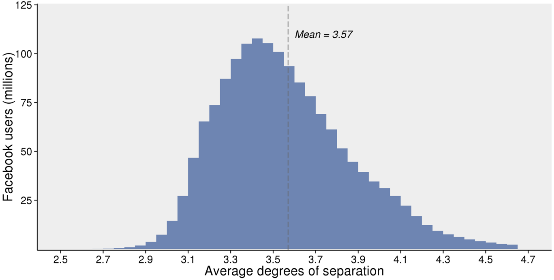
- 18 target people from all over the world
- 24,163 volunteers
- 384 successful chains
  Length of successful chains



Facebook

Bhagat, Burke, Diuk, Filiz, Edunov 2016

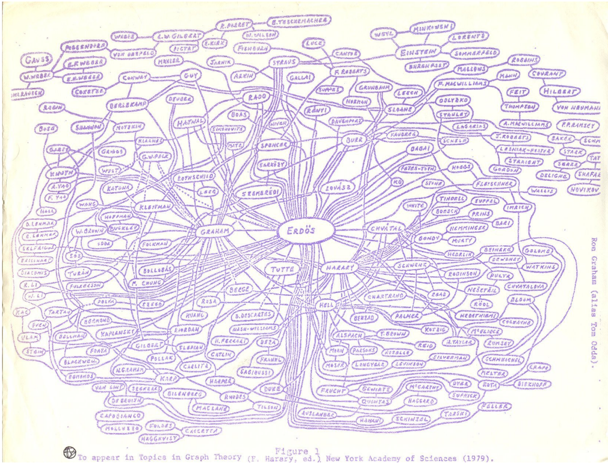
- Based on the 1.6 billion people active on Facebook
- Compute the average path length to any other people



The 3.5 degrees of separation of Facebook

Erdős number

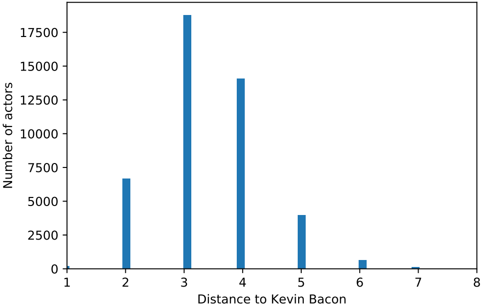
- Graph of co-authors of scientific papers
- Distance to Erdős (1913-1996)



The Bacon number

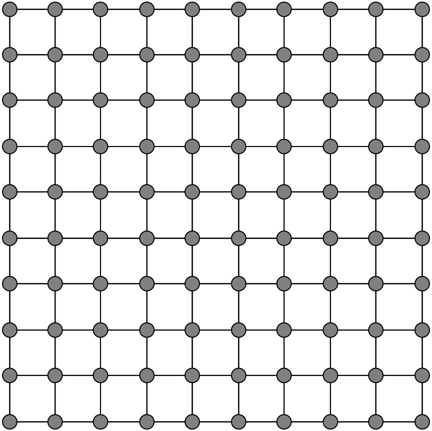
See The Oracle of Bacon

- 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)

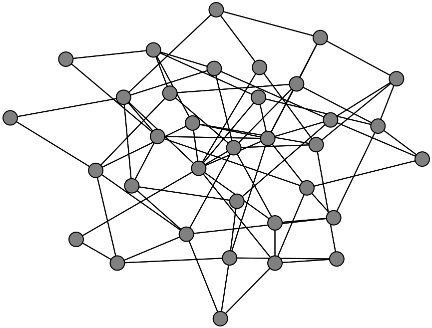
Planar graphs



Distance $= O(\sqrt{n})$

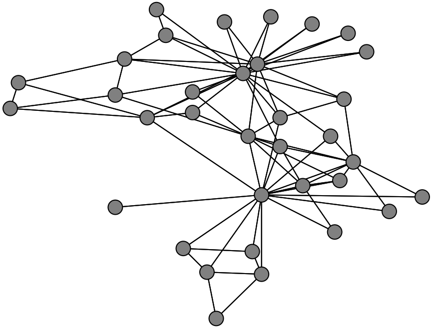
Random graphs

$$n \to +\infty \quad np \to \lambda > 1$$



Distance $= O(\ln n)$

Power-law graphs

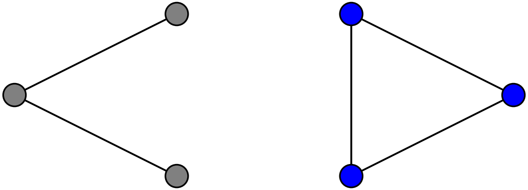


Distance $= O(1)$ (for $\alpha < 3$)

Outline

1. Degrees $\rightarrow$ power law
2. Distances $\rightarrow$ small world
3. Triangles

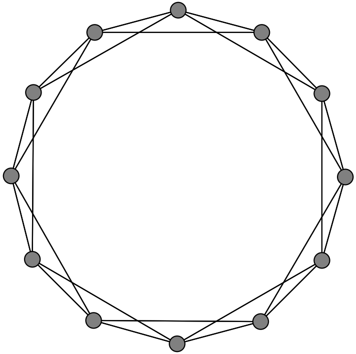
Clustering coefficient



Graph	C
Karate Club	0.26
Les Miserables	0.50
Openflights	0.25
WikiVitals	0.19

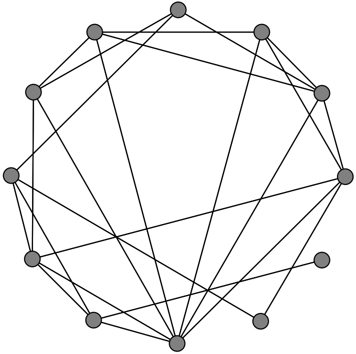
Watts-Strogatz model (1998)

1. Start from a ring of n nodes where each node is connected to its d nearest neighbors (d even)
2. Modify each edge at random with probability p



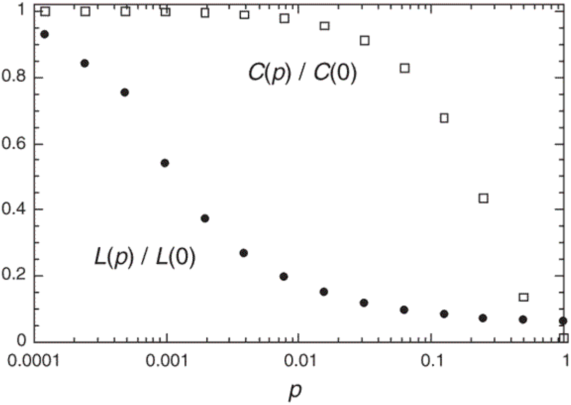
$n = 12, d = 4$

Example



$n = 12, d = 4, p = 0.4$

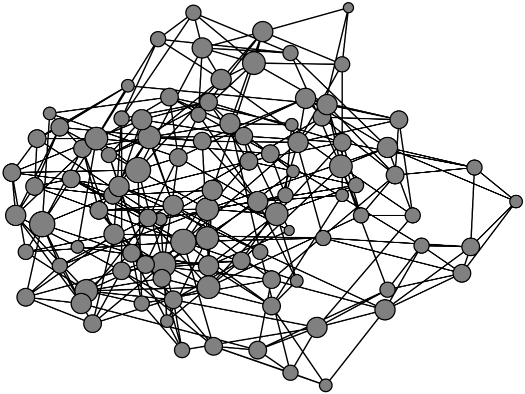
Small-world vs clusters



$n = 1000, d = 10$

Source: Watts & Strogatz 1998

Small-world with clusters



$n = 100, d = 6, p = 0.2$