Graph Mining
SD212
3. Graph structure

Thomas Bonald

2018 – 2019

Motivation

- How many clicks are needed to go from Platini to Plato on Wikipedia?

  Only 3!
  Platini → Naples → Ancient Greek → Plato

- And from Plato to Platini?

  Only 3 as well!
  Plato → Louvre → France → Platini

Motivation

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

  Using **Wikipedia for Schools** (4,591 pages):

# clicks	# nodes	proportion
1	76	1.6%
2	1505	32.8%
3	4527	98.6%
4	4584	99.8%

  In 1 click, you already get some central nodes:
  Plato → Latin, Italy, Arabic language, Iran, 19th century,...

Outline

1. Small-world property
2. Clustering coefficient
3. Watts-Strogatz graphs
4. Sparse matrices

The six degrees of separation

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

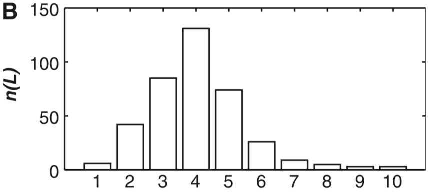
New Hampshire
Vermont
Massachusetts
Maine
North Dakota
Minnesota
Wisconsin
South Dakota
Michigan
New York
Rhode Islanc
Connecticu
Nebraska
Iowa
Pennsylvania
New Jersey
start
Illinois
Indiana
Ohio
Delaware
Kansas
Missouri
Kentucky
Virginia
Marylanc
West Virginia
North Carolina
Oklahoma
Arkansas
Tennessee
South Carolina
Alabama
Mississippi
Georgia
Texas
Louisiana
Florida

Source: Wikipedia

Emails

Dodds, Muhamad, Watts 2003

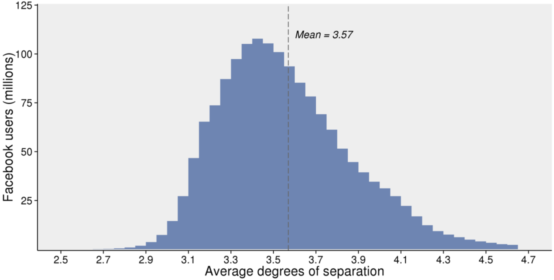
- 18 target people from all over the world
- 24,163 volunteers
- 384 successful chains
  Length of successful chains



Facebook

Bhagat, Burke, Diuk, Filiz, Edunov 2016

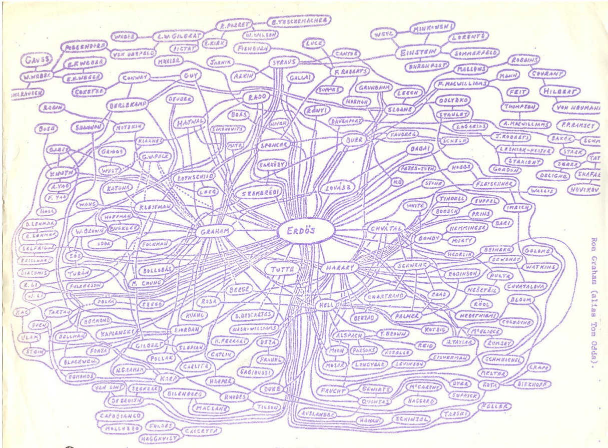
- Based on the 1.59 billion people active on Facebook
- Compute the average path length to any other people



The 3 and a half degrees of separation of Facebook

Erdös number

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

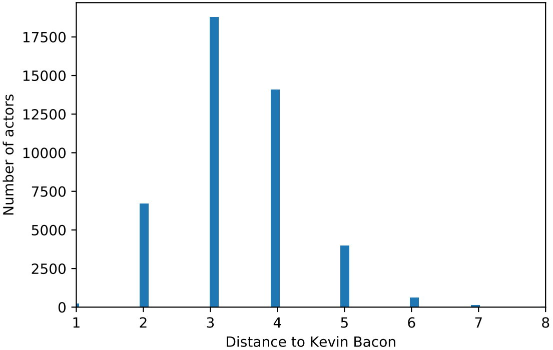


Ron Graham (alias Tom Odda).

Figure 1
To appear in Topics in Graph Theory (F. Harary, ed.) New York Academy of Sciences (1979).

The Bacon number

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



Results from YaGo database (44,586 actors)

The small-world property

- Is it universal?
- Example of a ring / a grid

The small-world property

- How does it emerge?
- Example of a random graph

Outline

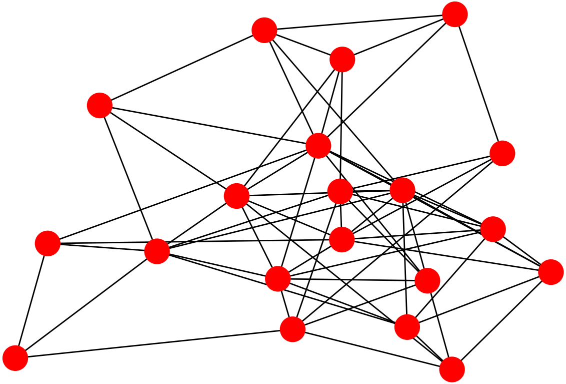
1. Small-world property
2. **Clustering coefficient**
3. Watts-Strogatz graphs
4. Sparse matrices

Structure vs. randomness

- The small-world property can be explained by randomness
- Can real graphs be considered as purely **random**?

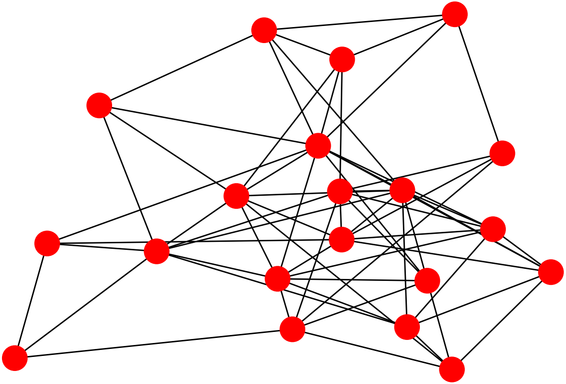
Clustering coefficient

Fraction of closed triangles



Local clustering coefficient

Proportion of my friends that are friends



Average local clustering coefficient

Some real graphs

Graph	C
Les Miserables	0.57
Openflights	0.25
Wikipedia for schools	0.28
Actor graph	0.79
Openstreet	0.001

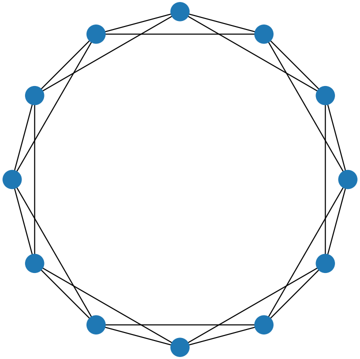
Case of Erdös-Rényi graphs

Outline

1. Small-world property
2. Clustering coefficient
3. **Watts-Strogatz graphs**
4. Sparse matrices

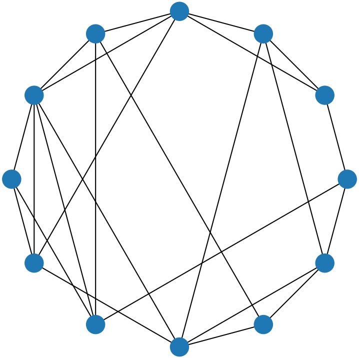
Watts-Strogatz graphs

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



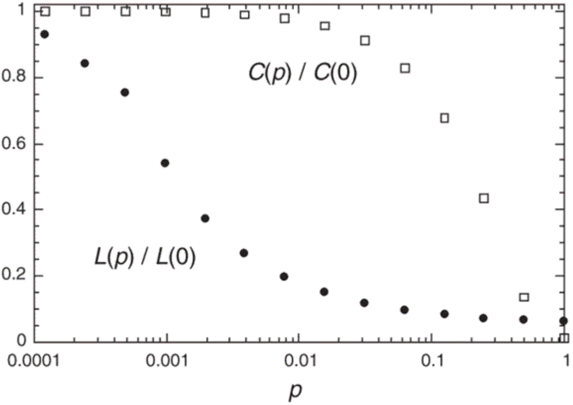
$n = 12, d = 4$

Example



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

Small-world vs clustering structure



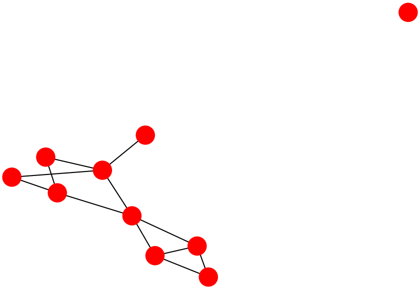
$n = 1000, d = 10$

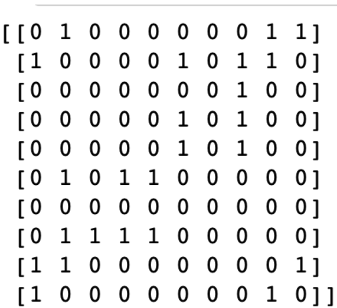
Source: Watts & Strogatz 1998

Outline

1. Small-world property
2. Clustering coefficient
3. Watts-Strogatz graphs
4. **Sparse matrices**

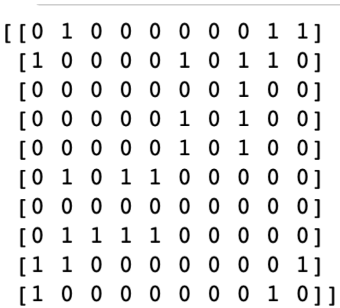
Adjacency matrix



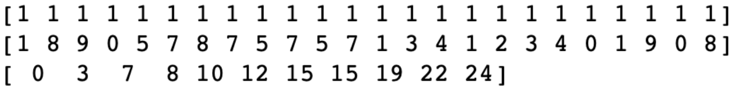


```
[[0 1 0 0 0 0 0 0 1 1]
 [1 0 0 0 0 1 0 1 1 0]
 [0 0 0 0 0 0 0 1 0 0]
 [0 0 0 0 0 1 0 1 0 0]
 [0 0 0 0 0 1 0 1 0 0]
 [0 1 0 1 1 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0]
 [0 1 1 1 1 0 0 0 0 0]
 [1 1 0 0 0 0 0 0 0 1]
 [1 0 0 0 0 0 0 0 1 0]]
```

Compressed Sparse Row (CSR) format



```
[[0 1 0 0 0 0 0 0 1 1]
 [1 0 0 0 0 1 0 1 1 0]
 [0 0 0 0 0 0 0 1 0 0]
 [0 0 0 0 0 1 0 1 0 0]
 [0 0 0 0 0 1 0 1 0 0]
 [0 1 0 1 1 0 0 0 0 0]
 [0 0 0 0 0 0 0 0 0 0]
 [0 1 1 1 1 0 0 0 0 0]
 [1 1 0 0 0 0 0 0 0 1]
 [1 0 0 0 0 0 0 0 1 0]]
```



```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]
[1 8 9 0 5 7 8 7 5 7 5 7 1 3 4 1 2 3 4 0 1 9 0 8]
[ 0  3  7  8 10 12 15 15 19 22 24]
```

Pros and cons

Pros

- Efficient storage
- Fast row slicing
- Fast matrix-vector product

Cons

- Slow column slicing
- Slow modification (e.g., add an entry)

Some applications

- Neighbors / degrees
- Path lengths
- Breadth-first search (BFS)
- Clustering coefficient

Summary

- Most real graphs have both the **small-world property** and a **high clustering** coefficient
- **Watts-Strogatz graphs** have both properties (for properly chosen parameters)
- **Sparse matrices** are key to efficient computation of path lengths and clustering coefficients