NETWORK SCIENCE OF ONLINE INTERACTIONS

Chapter 2: Small Worlds

Joao Neto

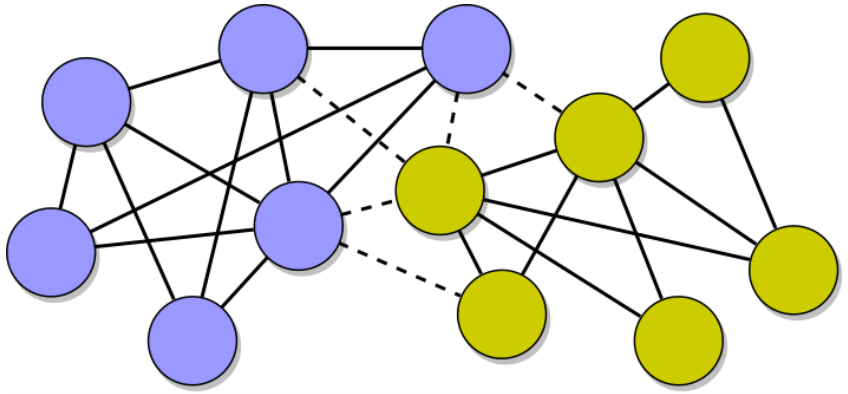
28/Apr/2023

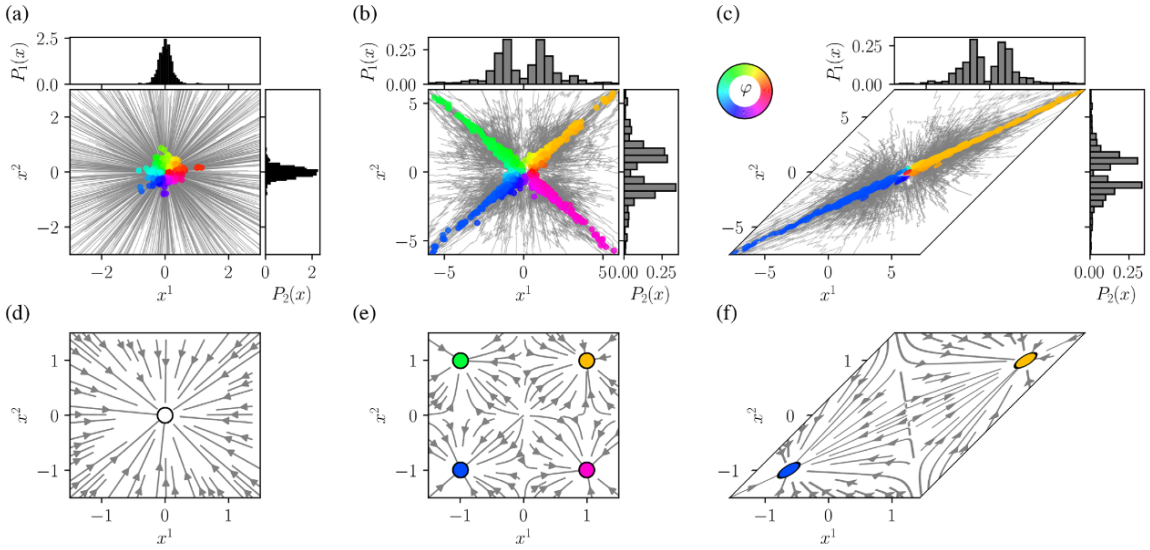
SUMMARY

- Many types of networks
  - Directed/Undirected
  - Multilayer
    - Multiplex
      - Temporal
- Basic metrics are already useful
- Network visualization is cumbersome, but there are ways
  - Copilot/chatGPT
  - Gephi

2.1 BIRDS OF A FEATHER

- Nodes can have attributes
  - Age, sex, location, personal interests, etc
- **Assortativity**: nodes tend to be more connected to similar nodes
- How?
  - **Homophilly**: similar nodes become connected
  - Social influence: connected nodes become more similar
- In more detail: Network models (Chapter 5)
- Problem: echo chambers
  - Political polarization, radicalization
  - Amplification of disinformation





(Baumann et al, 2021)

2.1 BIRDS OF A FEATHER

- Measures of assortativity
  - Node attribute
  - Degree assortativity
- Node attribute
  - Nodes have types
  - e_{ij} is the fraction of links in the network between nodes of types i and j
  - Interpreting r
    - Assortative: $0 < r \leq 1$
    - Not assortative: $r = 0$
    - Disassortative: $-1 \leq r < 0$

Assortativity

`degree_assortativity_coefficient`(G[, x, y, ...])	Compute degree assortativity of graph.
`attribute_assortativity_coefficient`(G, attribute)	Compute assortativity for node attributes.
`numeric_assortativity_coefficient`(G, attribute)	Compute assortativity for numerical node attributes.
`degree_pearson_correlation_coefficient`(G[, ...])	Compute degree assortativity of graph.

$$\sum_{ij} e_{ij} = 1, \qquad \sum_{j} e_{ij} = a_i, \qquad \sum_{i} e_{ij} = b_j,$$

$$r = \frac{\sum_i e_{ii} - \sum_i a_i b_i}{1 - \sum_i a_i b_i}$$

```
S = nx.[attribute,numeric]_assortativity_coefficient (G, attribute)
```

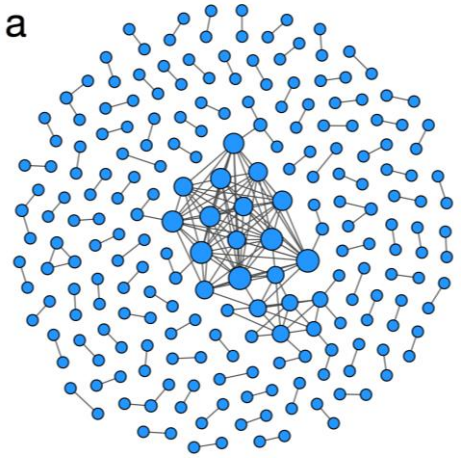
2.1 BIRDS OF A FEATHER

- Degree assortativity (degree correlation)
  - Assortative: core and periphery structure
    - Ex: social networks
  - Disassortative: hub and spoke (star) structure
    - Ex: technological networks, biological networks
- Measuring degree assortativity
  - Assortativity coefficient (Pearson correlation of degrees)

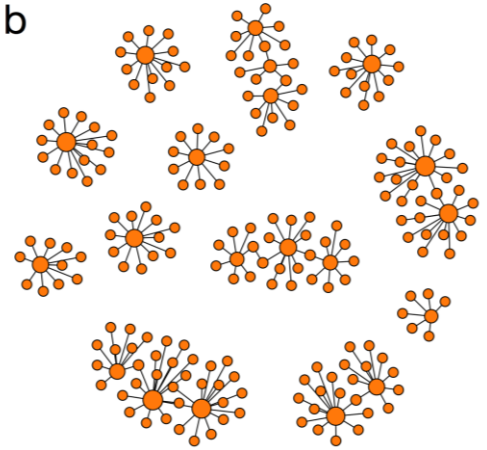
```
r = nx.degree_assortativity_coefficient(G)
```

  - Correlation between degree and average degree of neighbours of nodes with that degree

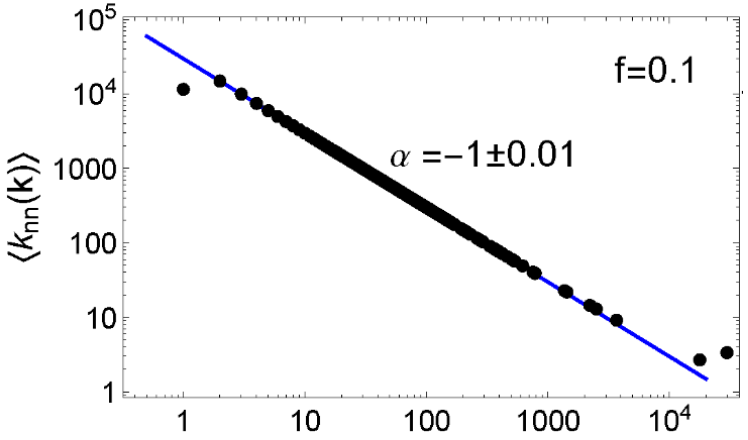
```
import  scipy.stats
knn_dict = nx.k_nearest_neighbors(G)
k, knn = list(knn_dict.keys()), list(knn_dict.values ())
r, p_value = scipy.stats.pearsonr(k, knn)
```



core periphery

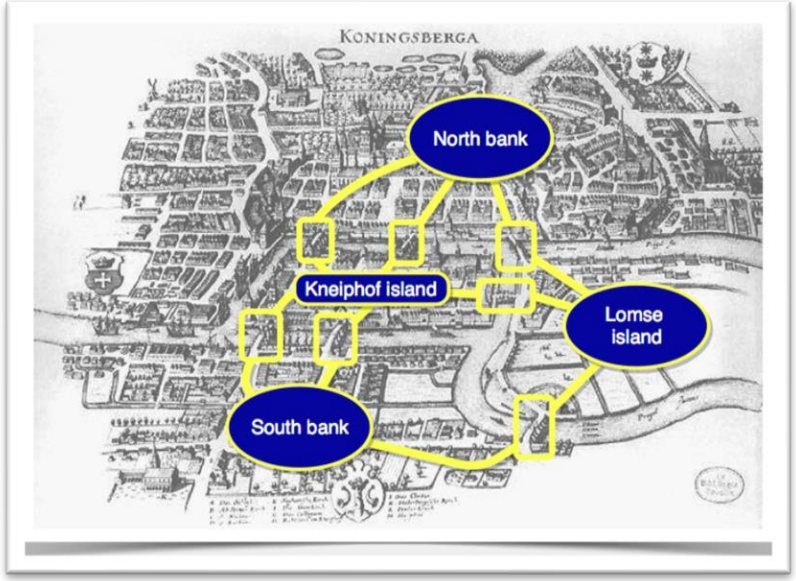


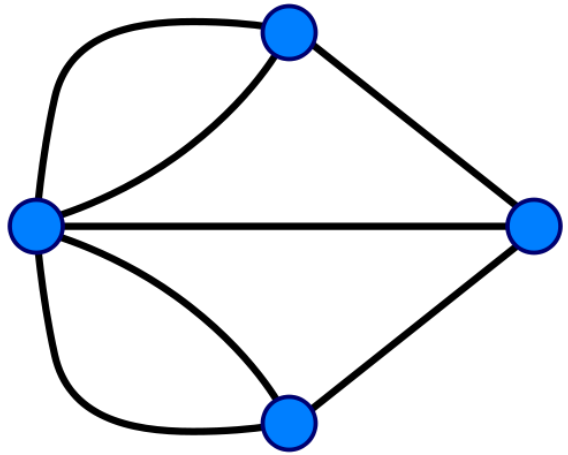
hub and spoke



2.2 PATHS AND DISTANCES

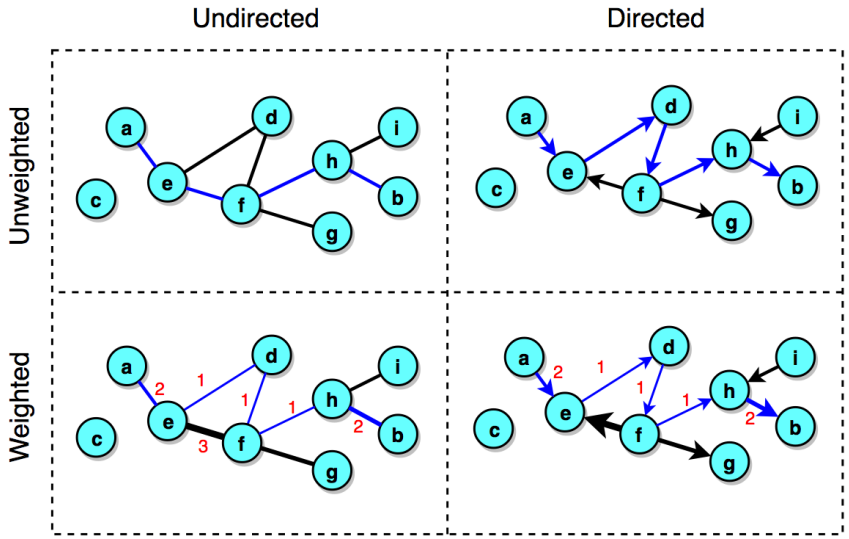
- **Path**: sequence of links traversed to go from a **source** to a **target** node
  - Directed network: streets are one-way
- **Cycle**: path where source and target node are the same
- Simple path: no traversing the same link more than once
- **Path length** ℓ_{ij} : number of links in path
- Koeningsberg bridge problem
  - Visit each place (node), cross each bridge (link) only once
  - Euler (1736): path only exists if all nodes in the middle of the path have even degree
  - Need to repeat bridges

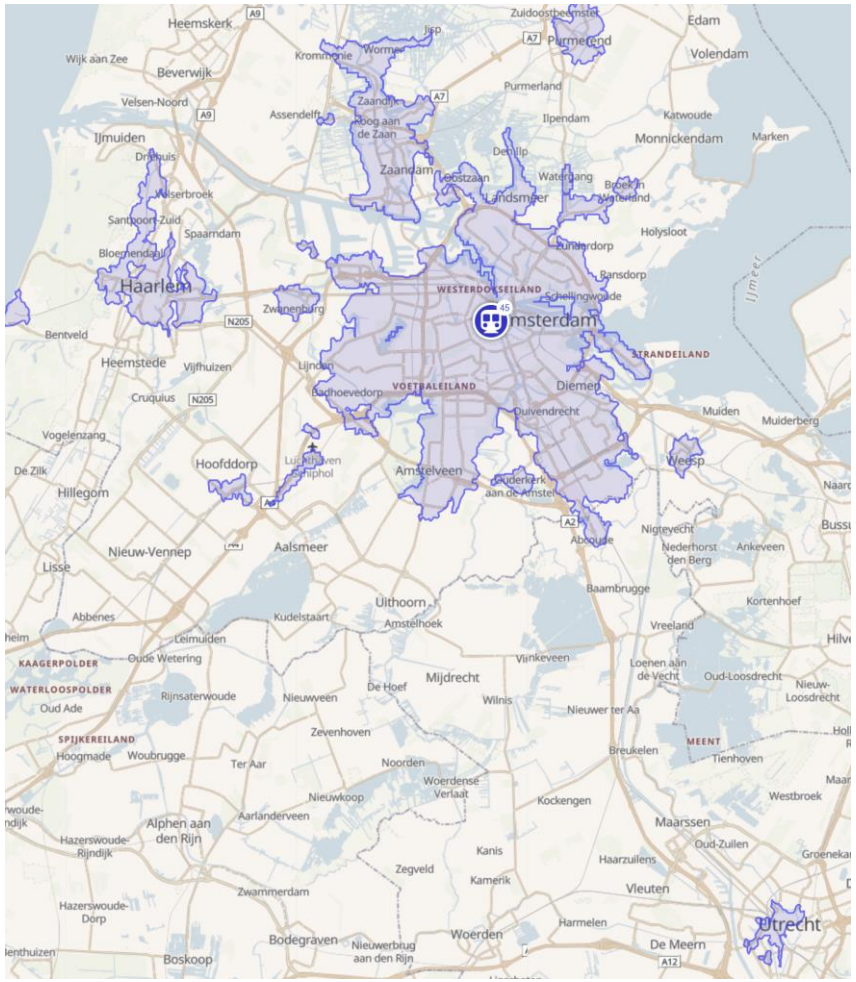




2.2 PATHS AND DISTANCES

- **Shortest path** between two nodes:
  - Path that offers the minimal length between nodes
- **Shortest path length** or **distance**: length of shortest path
  - Undefined if no path
- Weighted networks: minimize the total weight, may have more nodes





Places reachable from Amsterdam within 45min

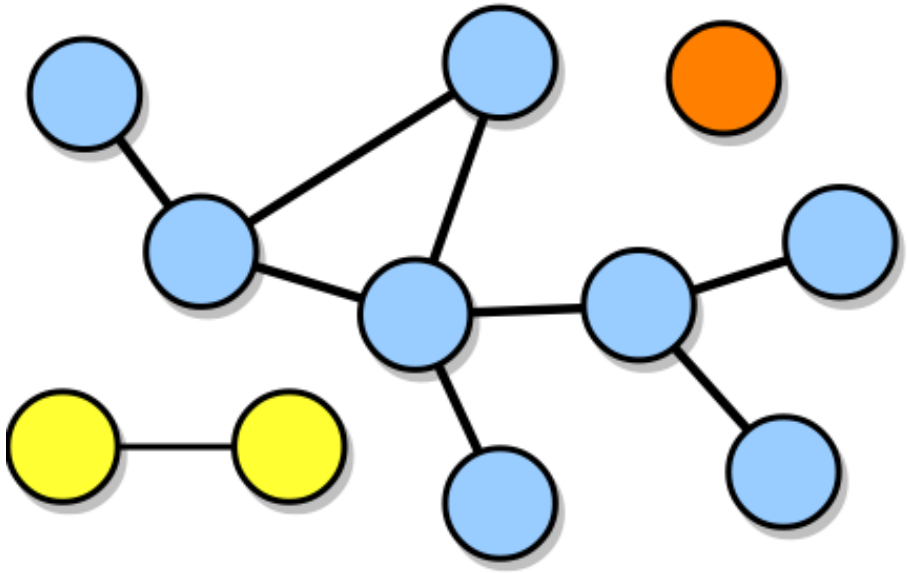
2.2 PATHS AND DISTANCES

- Characterize network with path metrics
  - **Diameter:** longest shortest path
  - **average path length:** average of all shortest paths
    - Undirected: $\langle \ell \rangle = \frac{2\sum_{i,j}\ell_{ij}}{N(N-1)}$
    - Directed: $\langle \ell \rangle = \frac{\sum_{i,j}\ell_{ij}}{N(N-1)}$
  - Undefined if network is there are unconnected nodes
- Easy to compute all of this with NetworkX

```
nx.has_path(G, 'a', 'c')

nx.shortest_path(G, 'a', 'b')

nx.shortest_path_length(G,'a','b')

nx.shortest_path(G, ‘a’)

nx.shortest_path_length(G, 'a')

nx.shortest_path(G)

nx.shortest_path_length(G)

nx.average_shortest_path_length(G)

nx.shortest_path_length(W, 'a',‘b’, ‘weight’)
```

2.3 CONNECTEDNESS AND COMPONENTS

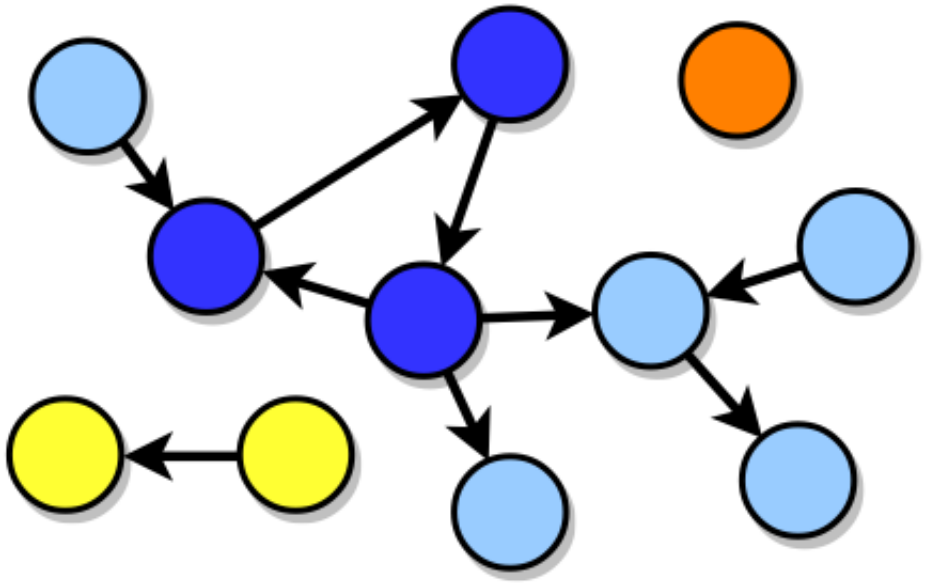
- A network is **connected** if there is a path between every pair of nodes
  - Disconnected otherwise
- A **connected component** is a connected subnetwork
- The largest one is called **giant component**; it often includes a substantial portion of the network



2.3 CONNECTEDNESS AND COMPONENTS

- Directed network
  - Strongly connected: with directions
  - Weakly connected: without directions

- Components in directed networks
  - **In-component** of a set S: can reach S, can't be reached from S
  - **Out-component** of S: the inverse



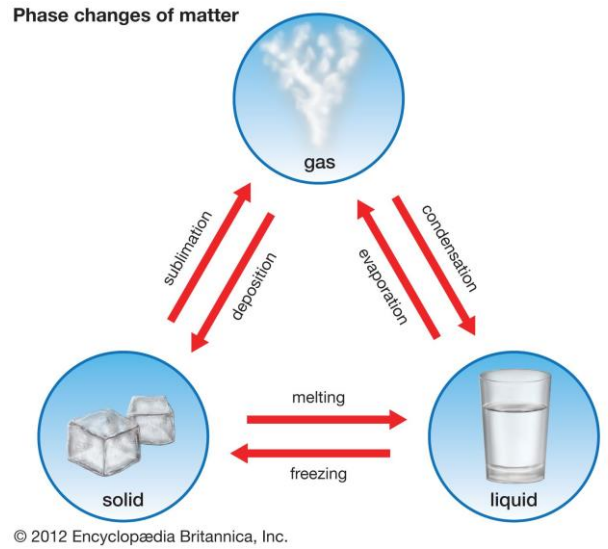
2.3 CONNECTEDNESS AND COMPONENTS

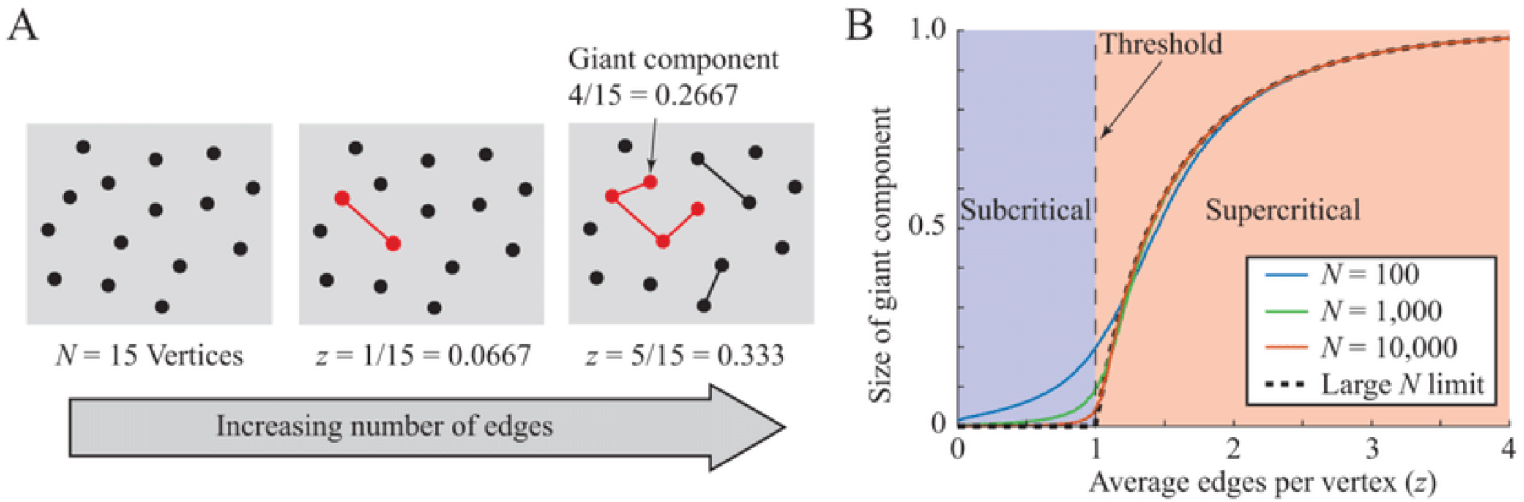
- On NetworkX

```
nx.is_connected(G)

comps = sorted(nx.connected_components(G),
               key=len, reverse=True)
nodes_in_giant_comp = comps[0]

GC = nx.subgraph(G, nodes_in_giant_comp)
nx.is_connected(GC)
nx.is_strongly_connected(D)
nx.is_weakly_connected(D) list(nx.weakly_connected_components(D))
list(nx.strongly_connected_components(D))
```

2.3 CONNECTEDNESS AND COMPONENTS

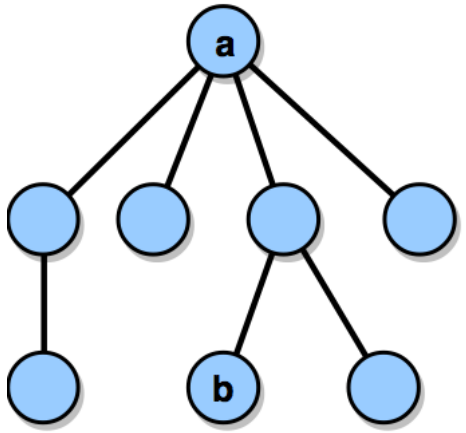
- Giant component properties
  - Generating a random network.
  - How connected is it?
  - For $\langle k \rangle > 1$, a meaningful part of the system is connected
  - For $\langle k \rangle > 3$, more than 90% of the system is connected
  - This is a **phase transition**

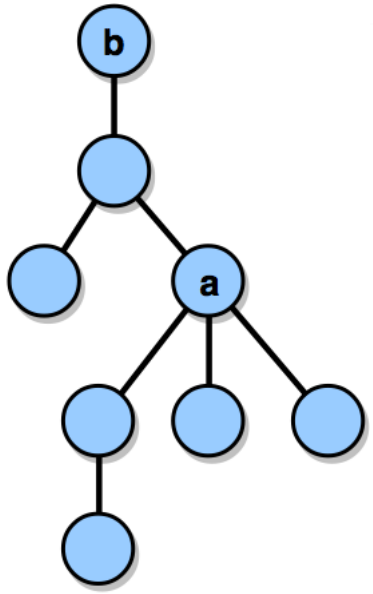


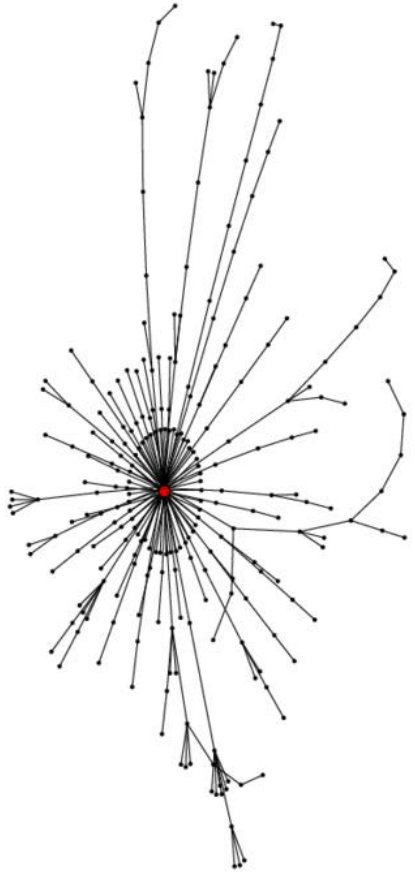


2.4 TREES

- A **tree** is a **connected** network **without cycles**
- A **tree** is a **connected** network **with *N*-1 links**
- Properties
  - Single path between nodes
  - Hierarchical:
    - Start from a **root** node
    - Each other node has a **parent** and possible **children**
    - **Roots** have no **parents**
    - **Leaves** have no **children**
- **Social media discussion threads are trees**







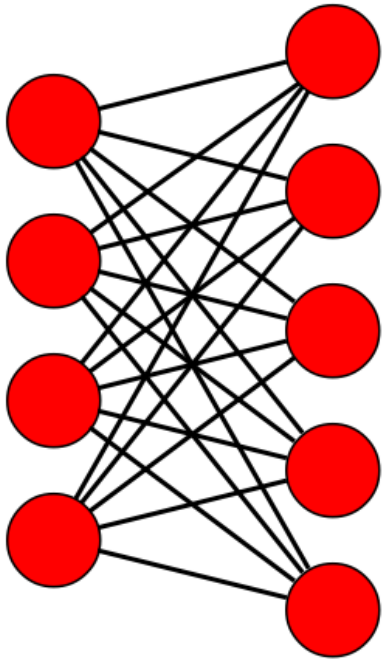
Reddit discussion tree

2.4 TREES

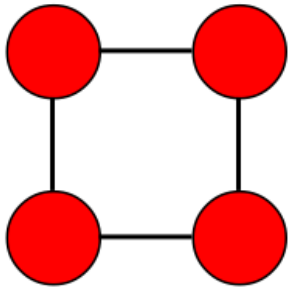
- On NetworkX

```
K4 = nx.complete_graph(4)
nx.is_tree(K4)                # False

nx.is_tree(B)                 # False
nx.is_tree(C)                 # False

nx.is_tree(S)                 # True
nx.is_tree(P)                 # True
```

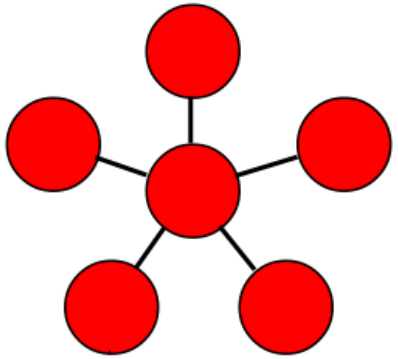
B = nx.complete_bipartite_graph(4,5)



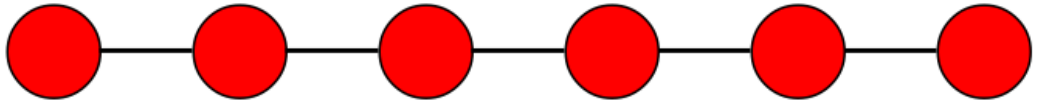
C = nx.cycle_graph(4)



S = nx.star_graph(6)

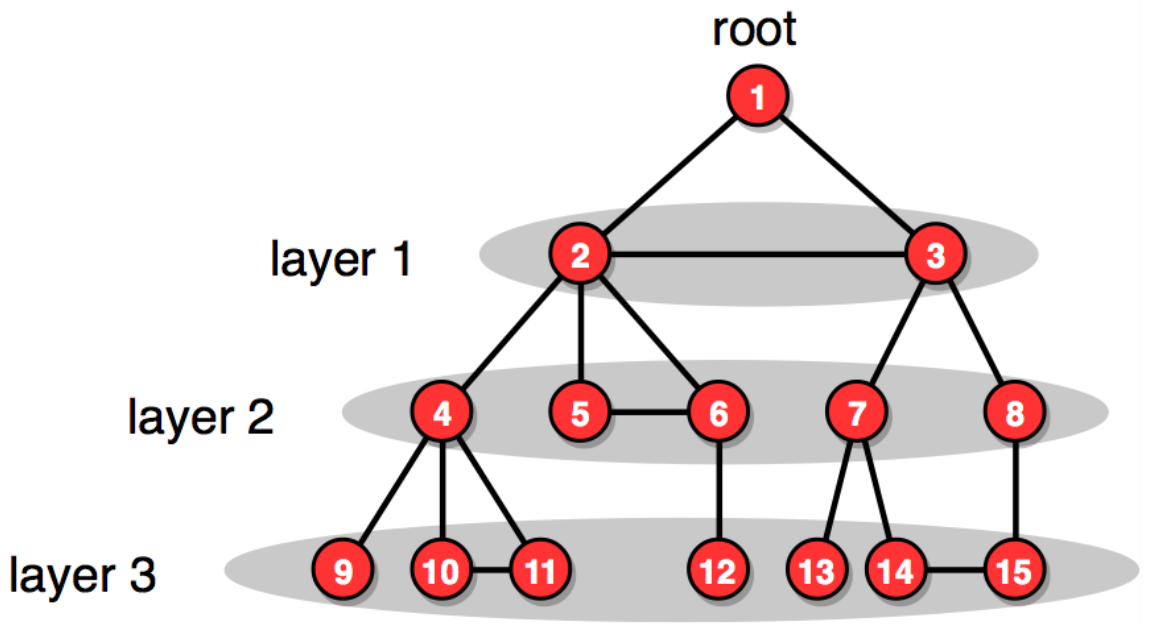


P = nx.path_graph(5)



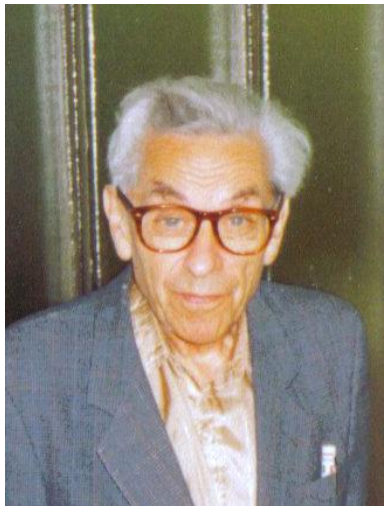
2.5 FINDING SHORTEST PATHS

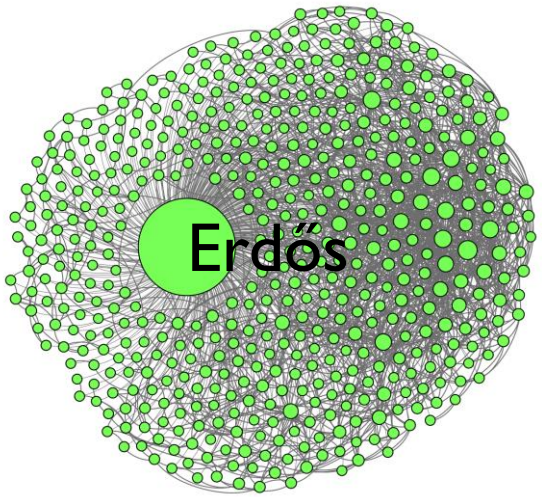
- Finding paths is computationally expensive
- Lots of algorithms specialized by domain
- Standard: **breadth-first search**
  - Calculate distance to all nodes at a certain depth, moves on to deeper depth
  - Doing that for all nodes: $O(N + L)$
    - But L may scale up to N^2
- Weighted: Dijkstra's algorithm
  - More expensive: $O(N + L \log L)$
- NetworkX has **a lot** of algorithms (check docs)
  - General ones
  - Faster ones for undirected, unweighted graphs



2.6 SOCIAL DISTANCE

- How topologically distant are social beings?
- One of key questions in social networks
- Paul Erdős (1913-1996)
  - father of graph theory (with Alfréd Rényi)
  - 511 coauthors
- Erdős number: how far from Paul Erdős is someone?
  - Erdős 0: 1
  - Erdős 1: 511
  - Erdos 2: 11,000+
  - Median value around 5



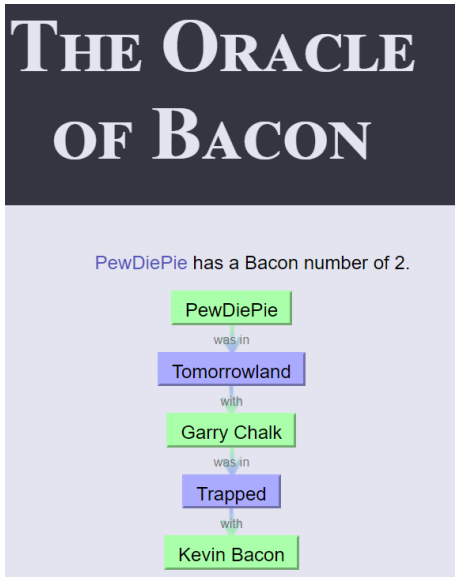


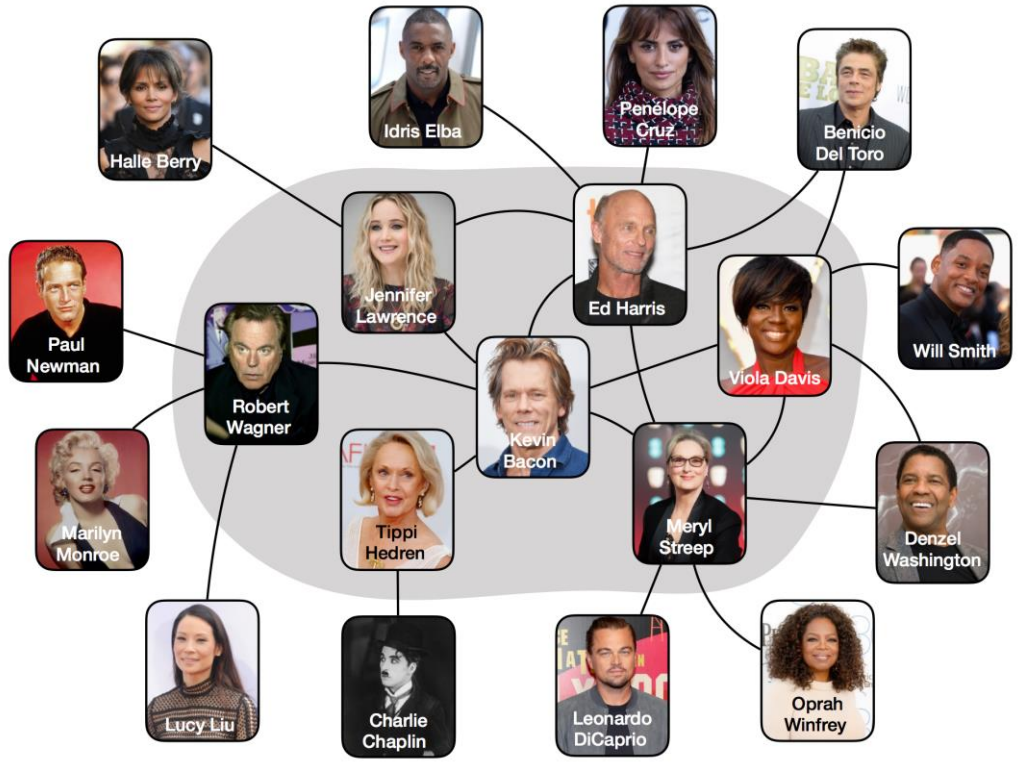
MR Erdos Number = 5

David García Becerra	coauthored with	Frank Schweitzer	MR3420956
Frank Schweitzer	coauthored with	Fernando Vega-Redondo	MR2548303
Fernando Vega-Redondo	coauthored with	Matteo Marsili	MR2029176
Matteo Marsili	coauthored with	László A. Székely	MR3076123
László A. Székely	coauthored with	Paul[1] Erdős	MR1209184

2.7 SIX DEGREES OF SEPARATION

- Another social hub: Kevin Bacon
- Bacon number:
  - 1: ~2800
  - 2: ~300,000
  - 3: ~1,000,000
  - Mean around 3



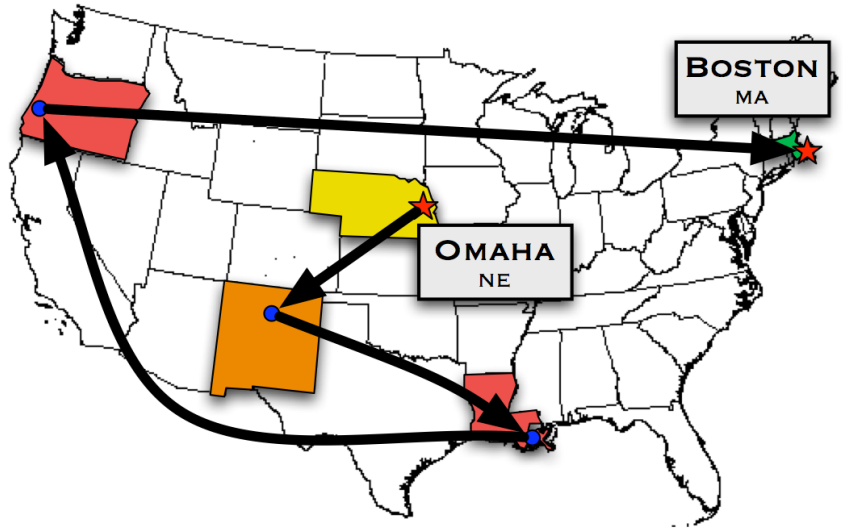


2.7 SIX DEGREES OF SEPARATION

- Social networks tend to have very **short paths**

- **Six degrees of separation**
  - The idea that any two people are at most six steps away from each other in the social network

- The Milgram experiment (1967)
  - Instructions: send to personal acquaintance who is more likely to know target
  - 160 letters to people in Omaha, NE and Wichita, KS
  - 2 targets in Massachusetts
  - 42 letters delivered (26%)
  - Average: **6.5** steps (range: 3-12 steps)

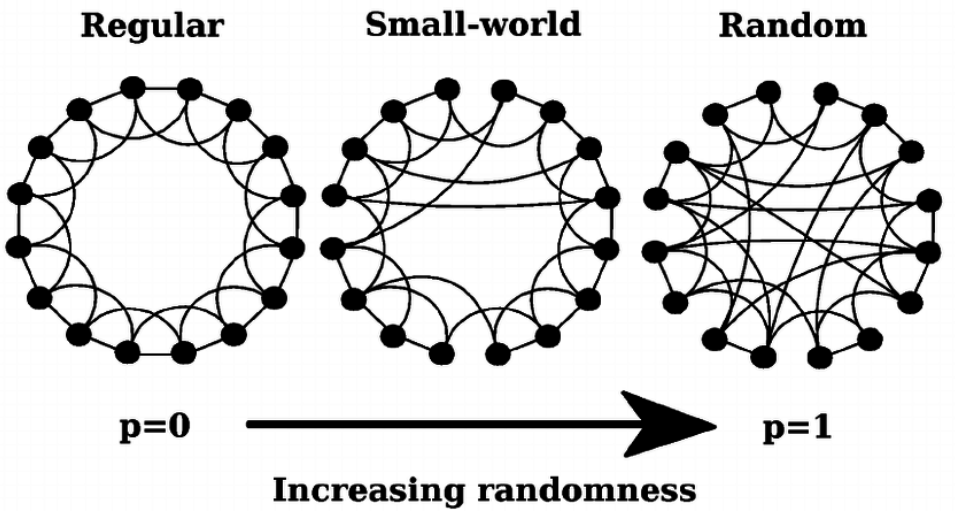


2.7 SIX DEGREES OF SEPARATION

- Replications
  - Yahoo 2003
    - 18 targets emailed in 13 countries
    - 384 completed chains out of more than 24 thousand started
    - Average path length estimated around **5-7**
  - Facebook 2011
    - 721M users
    - 69B links
    - Average path length of **4.74**

2.7 SIX DEGREES OF SEPARATION

- The average path length is short or long depending on **scaling**
- It is **short** if $\langle \ell \rangle \sim \log N$
- **The small-world phenomena**: many networks have short average paths
  - Hubs act as shortcuts in the network
  - Usefulness depends on what the network models
    - Very useful for travel by intelligent agents
  - Social networks are pretty much always small-world



2.7 SIX DEGREES OF SEPARATION

- Path length of many empirical networks is small
- Counter-example: grid or lattice-like networks

Table 2.1 Average path length and clustering coefficient of various network examples. The networks are the same as in Table 1.1, their numbers of nodes and links are listed as well. Link weights are ignored. The average path length is measured only on the giant component; for directed networks we consider directed paths in the giant strongly connected component. To measure the clustering coefficient in directed networks, we ignore link directions.

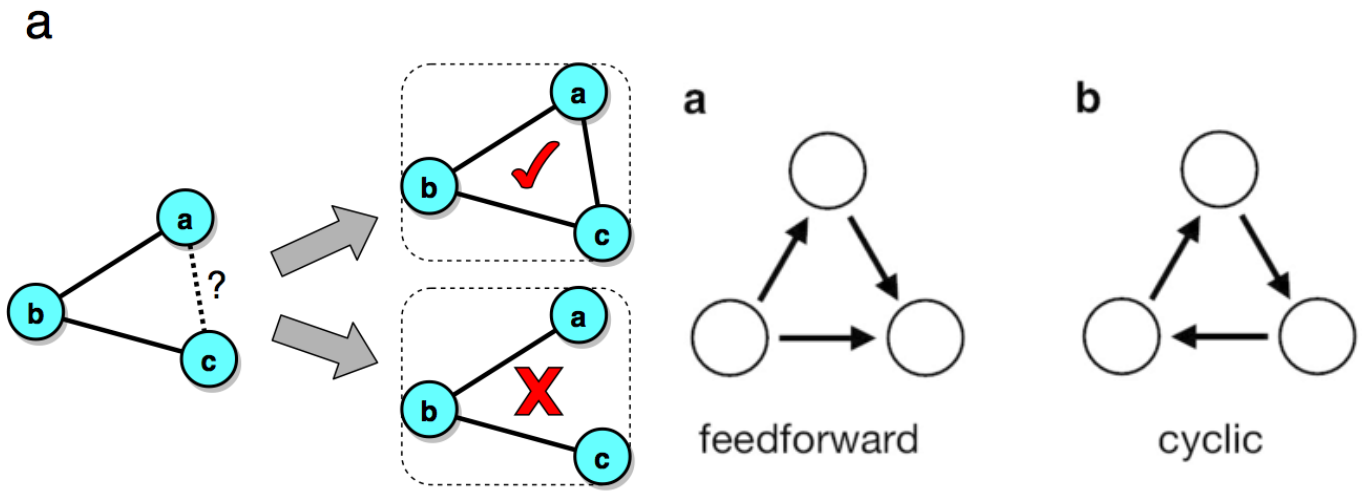
Network	Nodes (N)	Links (L)	Average path length ($\langle \ell \rangle$)	Clustering coefficient (C)
Facebook Northwestern Univ.	10,567	488,337	2.7	0.24
IMDB movies and stars	563,443	921,160	12.1	0
IMDB co-stars	252,999	1,015,187	6.8	0.67
Twitter US politics	18,470	48,365	5.6	0.03
Enron Email	87,273	321,918	3.6	0.12
Wikipedia math	15,220	194,103	3.9	0.31
Internet routers	190,914	607,610	7.0	0.16
US air transportation	546	2,781	3.2	0.49
World air transportation	3,179	18,617	4.0	0.49
Yeast protein interactions	1,870	2,277	6.8	0.07
C. elegans brain	297	2,345	4.0	0.29
Everglades ecological food web	69	916	2.2	0.55

2.8 FRIEND OF A FRIEND

- Triangles are common in networks
  - "friend of a friend is also a friend"
- Measurement: **clustering coefficient** C
  - Fraction of all possible triangles that exists
  - For an undirected graph:

$$C_i = \frac{\sum_{jk} a_{ij} a_{jk} a_{ki}}{k_i(k_i - 1)}$$

  - C: mean of C_i for nodes with $k_i > 1$
- Directed networks:
  - Two types of triangles: cyclic and feedforward
  - No NetworkX built-in function




```
nx.triangles(G)
nx.clustering(G, node)
nx.clustering(G)
nx.average_clustering(G)
```

2.8 FRIEND OF A FRIEND

- In social networks, high C can come from **triadic closure**
  - You meet friends through friends
- Bipartite, tree-like networks have low C

Table 2.1 Average path length and clustering coefficient of various network examples. The networks are the same as in Table 1.1, their numbers of nodes and links are listed as well. Link weights are ignored. The average path length is measured only on the giant component; for directed networks we consider directed paths in the giant strongly connected component. To measure the clustering coefficient in directed networks, we ignore link directions.

Network	Nodes (N)	Links (L)	Average path length ($\langle \ell \rangle$)	Clustering coefficient (C)
Facebook Northwestern Univ.	10,567	488,337	2.7	0.24
IMDB movies and stars	563,443	921,160	12.1	0
IMDB co-stars	252,999	1,015,187	6.8	0.67
Twitter US politics	18,470	48,365	5.6	0.03
Enron Email	87,273	321,918	3.6	0.12
Wikipedia math	15,220	194,103	3.9	0.31
Internet routers	190,914	607,610	7.0	0.16
US air transportation	546	2,781	3.2	0.49
World air transportation	3,179	18,617	4.0	0.49
Yeast protein interactions	1,870	2,277	6.8	0.07
C. elegans brain	297	2,345	4.0	0.29
Everglades ecological food web	69	916	2.2	0.55

SUMMARY

- Networks can be **assortative**
- **Path length** is an important characteristic of a network
- **Connected component** and **giant component** are two others
- Most networks have low **average path length** and are **small-world**