Network Analysis

AN INTRODUCTION FOR HUMANISTS

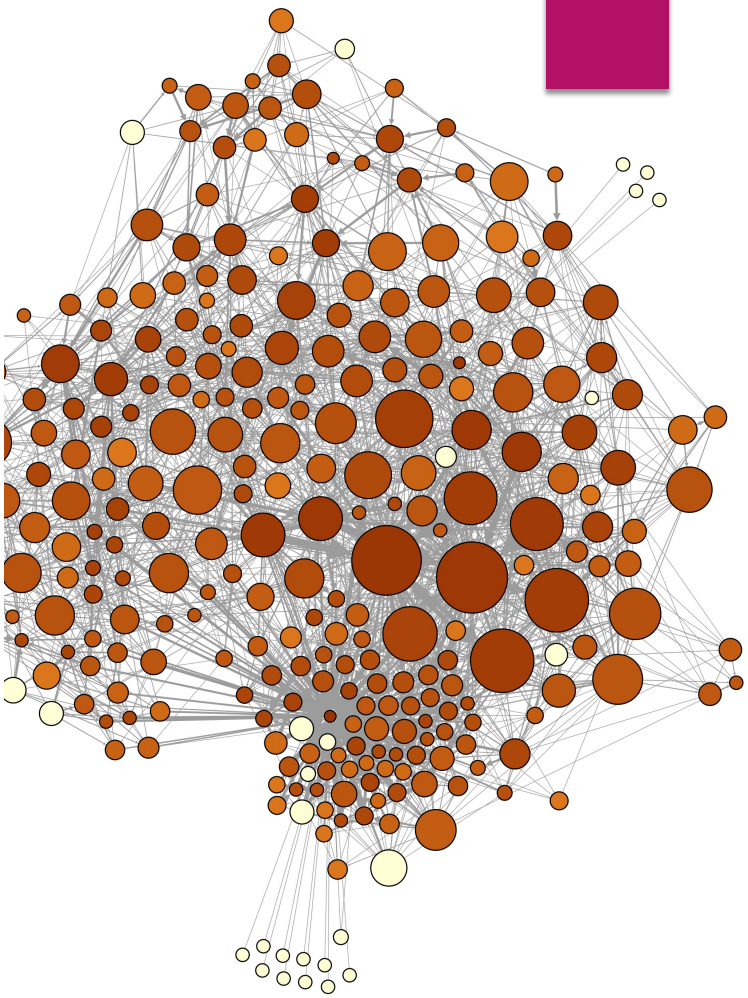
Dr Katarzyna Anna Kapitan
20 March 2025

Features of Real Networks (Recap)

- Many real networks have:
  - Short Average Path Length (Wk 3)
  - High Clustering Coefficient (Wk 3)
  - Heterogeneity & Hubs (Wk 4)

Network of Protein interaction of yeast.
Image source: Slides for Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

- Many real networks are not centred around human interactions (are not human-centric !), but have instead technical, biological, systemic, or cyber-physical in nature.
- Networks of protein interactions (for example, yeast protein interactions)
  - See: Gursoy A, Keskin O, Nussinov R. 'Topological properties of protein interaction networks from a structural perspective'. *Biochem Soc Trans*. 36 (6):1398-403. doi: 10.1042/BST0361398.
- Ecological networks (for example, everglades food-web)
  - See: Dunne JA, Williams RJ, Martinez ND, 'Food-web structure and network theory: The role of connectance and size'. *Proc. Natl. Acad. Sci. U.S.A.* 99 (20): 12917-12922. doi: 10.1073/pnas.192407699.
- Neural netwroks (for example, neural network of the roundworm c. elegans)
  - See: Ripoll-Sánchez, Lidia et al., 'The neuropeptidergic connectome of C. elegans.' *Neuron*. 111(22): 3570-3589. doi: doi: 10.1016/j.neuron.2023.09.043.



Neural Network of *c. elegans*
*I*mage source: Slides for Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

Short Average Path Length (Wk 3)

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

Average path length of a network is the average number of steps along the shortest paths between all pairs of nodes.

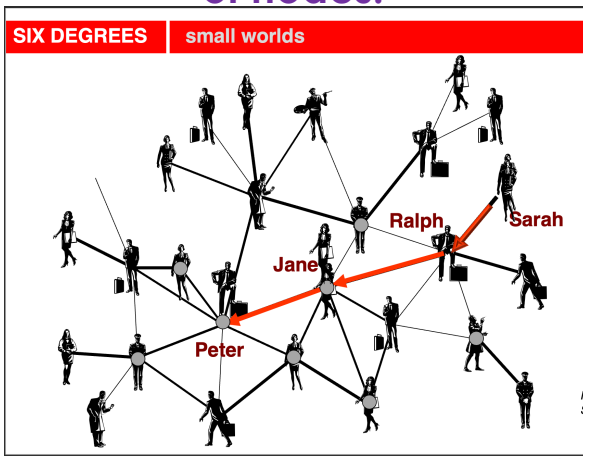


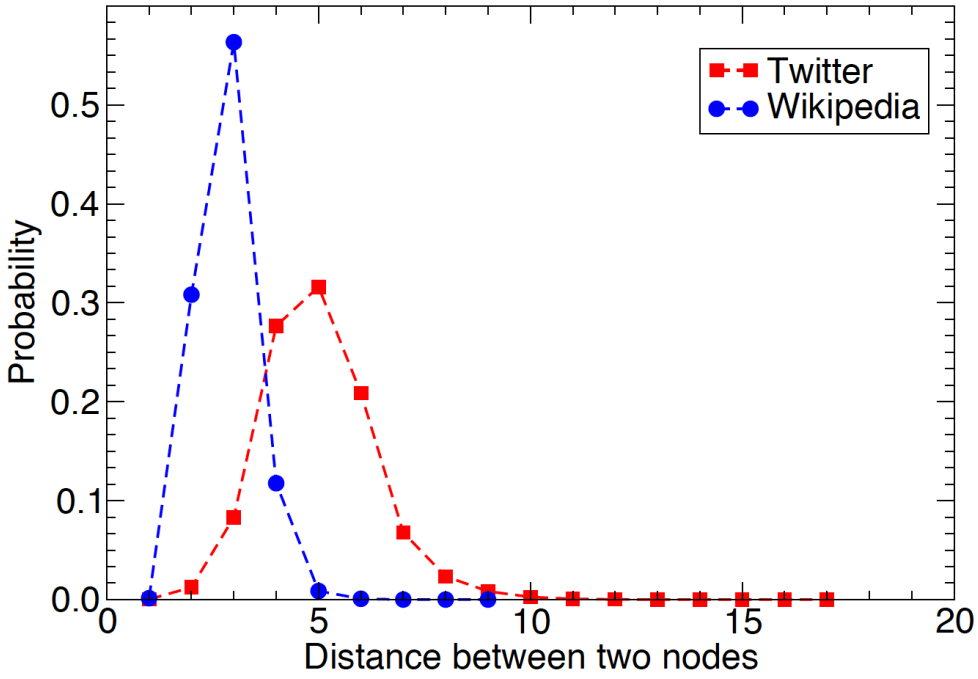
Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science* (left) & Slides for Barabási, Network Science https://networksciencebook.com (right)

Short Average Path Length (Distribution)

The small-world property is typical of most real networks of interest.

If the network has hubs, paths are ultra-short.

Many shortest paths go through hubs, for example, air transportation network



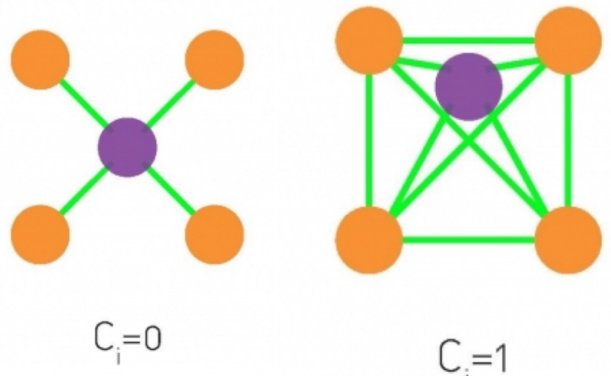
Shortest-path length distribution (Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*)

High Clustering Coefficient (Wk3)

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

Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

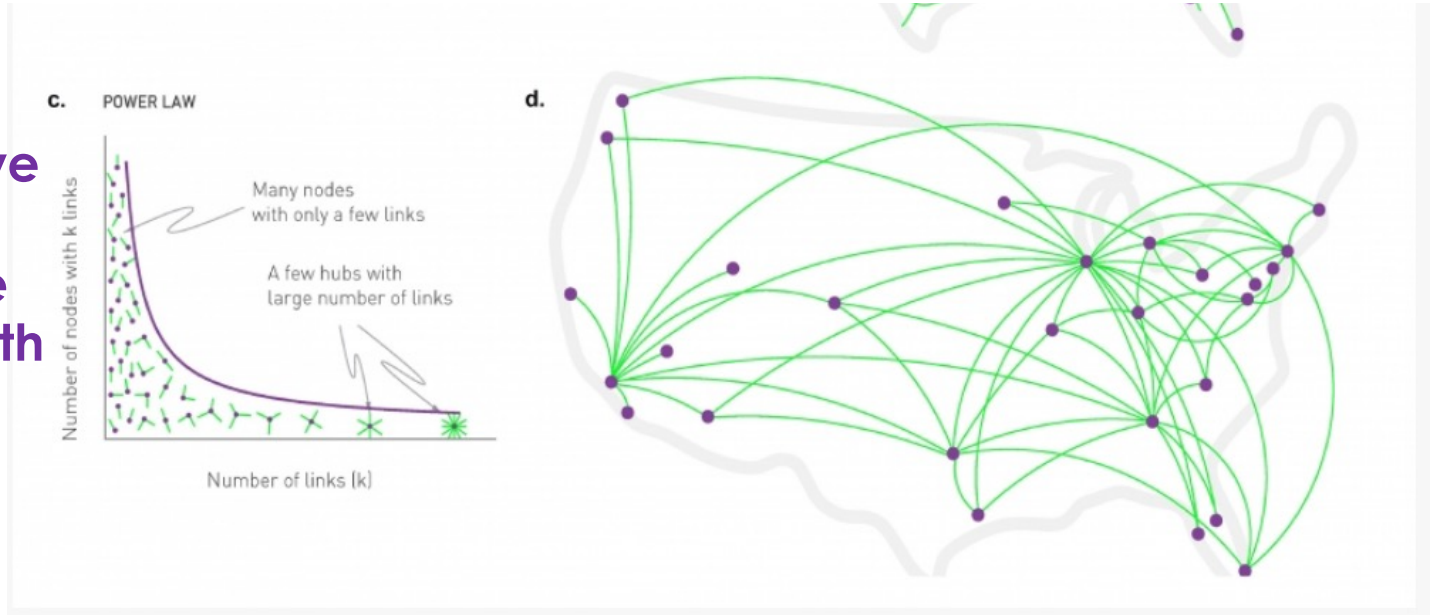
The clustering coefficient of a node is the fraction of pairs of the node's neighbours that are connected to each other



Source: Barabási, Network Science (https://networksciencebook.com)

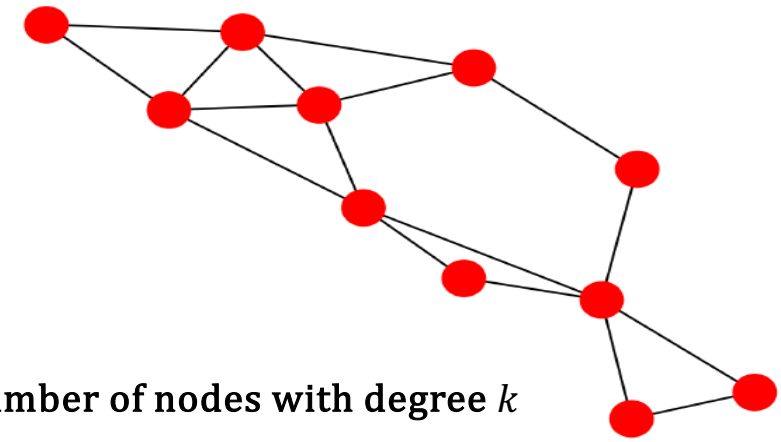
Heterogeneity & Hubs

Many real-life networks have heavy-tailed degree distribution, caused by the presence of hubs (nodes with high degree)



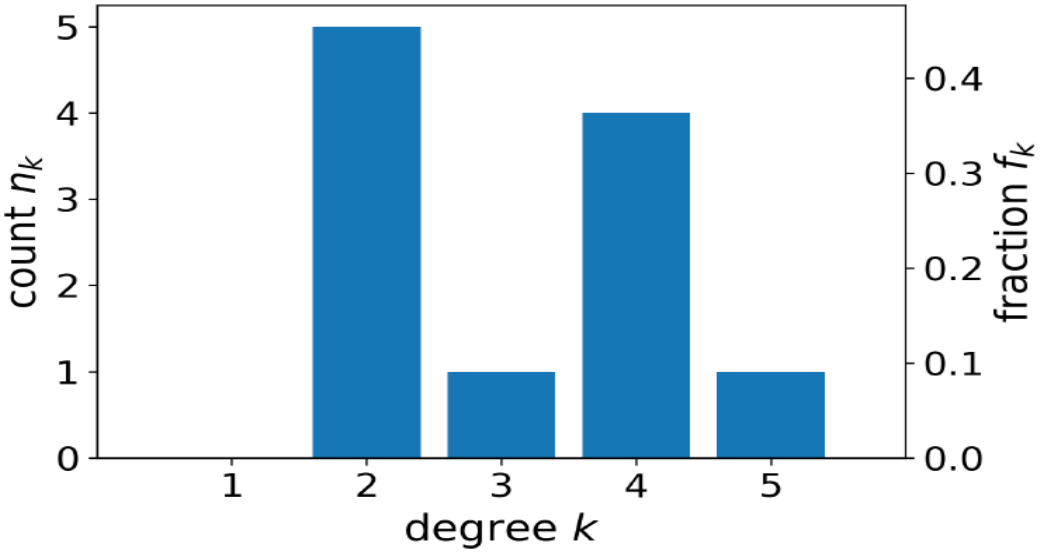
Source: Barabási, Network Science (https://networksciencebook.com)

Heterogeneity & Hubs (Degree distribution)



n_k = **number of nodes with degree** k

$f_k = \frac{n_k}{N}$ = **frequency of degree** k



Source: Menczer, Fortunato, Davis, *A First Course in Network Science*, version 3 (Cambridge University Press 2023)

Heterogeneity Parameter

Heterogeneity parameter κ is a measure of how broad the degree distribution is

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle^2}$$

⟨k⟩ is the average degree,
⟨k^2 ⟩ is the second moment of the degree distribution.
When the second moment is *much larger* than the square of the first moment, the network has *very large hubs* (high heterogeneity). So If κ is equal to 1, the network is homogeneous, and if κ is much greater than 1, the network is heterogeneous.

Network	Nodes (N)	Links (L)	Average degree ($\langle k \rangle$)	Maximum degree (k_{max})	Heterogeneity parameter (κ)
Facebook Northwestern Univ.	10,567	488,337	92.4	2,105	1.8
IMDB movies and stars	563,443	921,160	3.3	800	5.4
IMDB co-stars	252,999	1,015,187	8.0	456	4.6
Twitter US politics	18,470	48,365	2.6	204	8.3
Enron Email	36,692	367,662	10.0	1,383	14.0
Wikipedia math	15,220	194,103	12.8	5,171	38.2
Internet routers	190,914	607,610	6.4	1,071	6.0
US air transportation	546	2,781	10.2	153	5.3
World air transportation	3,179	18,617	11.7	246	5.5
Yeast protein interactions	1,870	2,277	2.4	56	2.7
C. elegans brain	297	2,345	7.9	134	2.7
Everglades ecological food web	69	916	13.3	63	2.2

Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

Modelling

WHAT AND WHY OF MODELLING

- A model is a set of rules or instructions used to generate or simulate graphs.

- The goal of network science is to develop models that generate graphs exhibiting similar characteristics to those of real-world networks.

Random Graphs (Erdős-Rényi & Gilbert models)

- Generate random graph by placing links at random between pairs of nodes

- **Algorithm**:
  - Start with *N* nodes and zero links
  - Go over all pairs of nodes;
  - for each pair of nodes *i* and *j*, generate a random number *r* between 0 and

    If $r < p \Rightarrow i$ and *j* get connected

    If $r > p \Rightarrow i$ and *j* remain disconnected

- In random graphs graphs, the average shortest-path length (L) depends on the number of nodes (N) and the probability of edge formation (p), which determines the average degree of the network.
- In **random graphs**, a **giant component** begins to form when the average degree ⟨k⟩ crosses a critical threshold determined by N.
- Around ⟨k⟩ = 1 a giant component grows very fast at the expense of other smaller components.

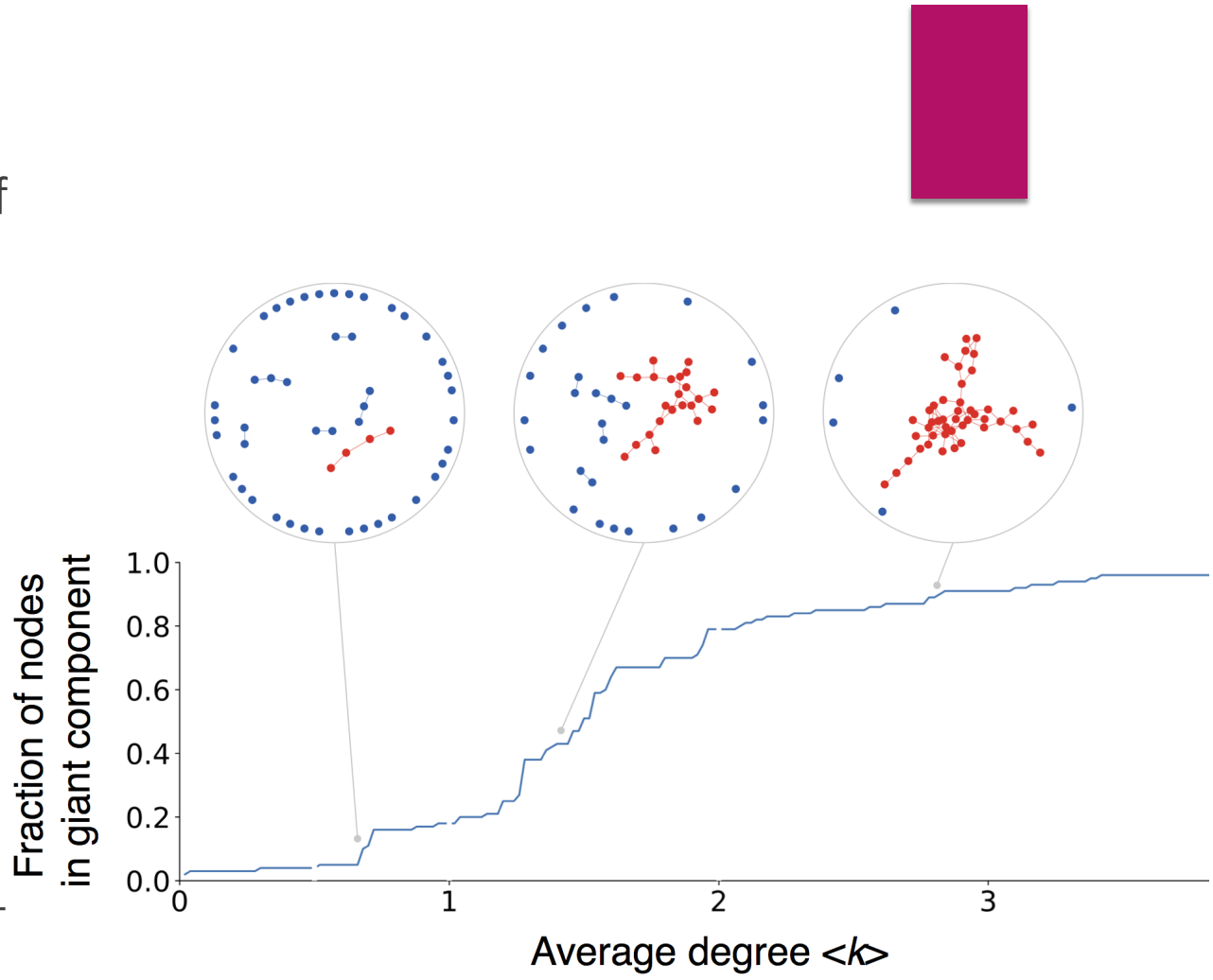


Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

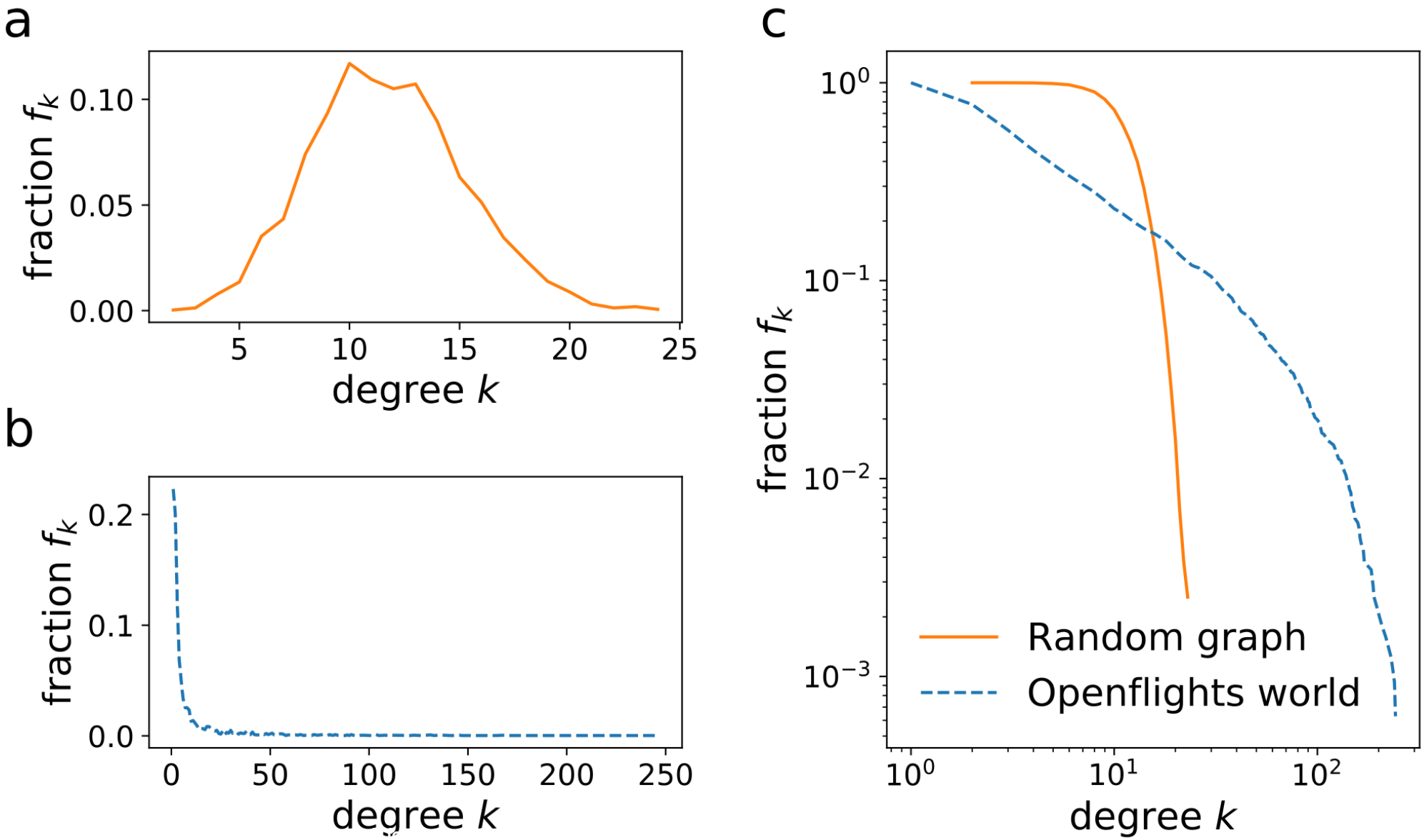


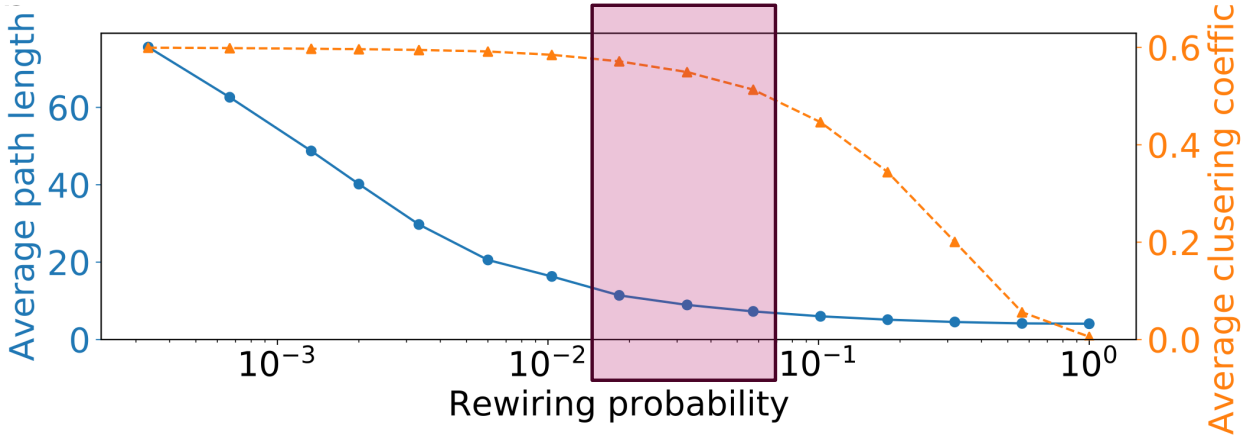
Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

Random Graphs

For a random graph to be a good model of real networks, the link probability *p* should be close to zero, as **real networks are usually sparse**, i.e. have small density.

If *p* is set close to zero than the **clustering co-efficient** is also close to zero, as in random graphs, the probability that a pair of neighbours of a node is connected is *p*. The link probability is the same for every pair of nodes, regardless of their having common neighbours or not.

The Watts-Strogatz model



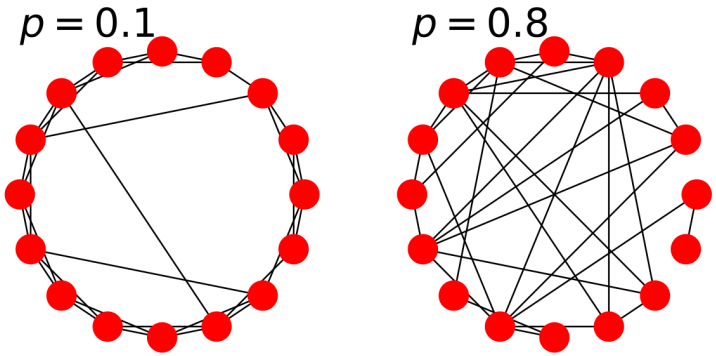


Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

N nodes form a regular ring lattice, with even degree *k*.

With probability *p*, each link is rewired randomly

The Watts-Strogatz model

- The degree distribution is peaked as most nodes have the same degree: **no hubs!**
- The Watts-Strogatz model fails to reproduce the **broad degree distributions observed in many real-world networks**

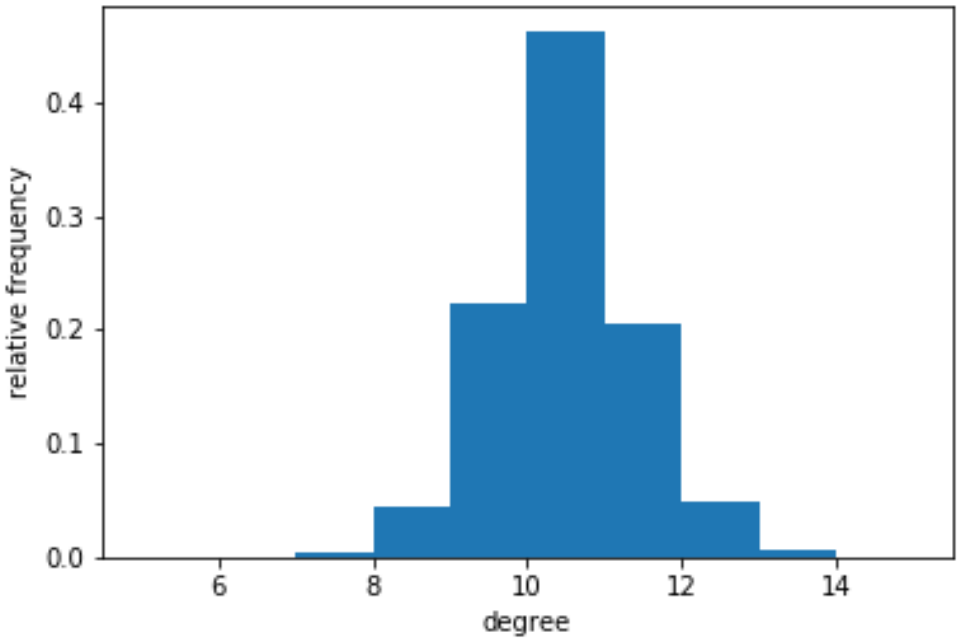


Image Source: Slides for Menczer, Fortunato, Davis, *A First Course in Network Science*

The Barabási-Albert Model

Algorithm:

- Start with a group of m_0 nodes, usually fully connected (clique)
- At each step a new node *i* is added to the system, and sets *m* links with some of the older nodes ($m \leq m_0$) **(Growth)**
- Each new link is wired to an old node j with probability which is proportional to the degree k_j of *j*. **(Preferential Attachment)**

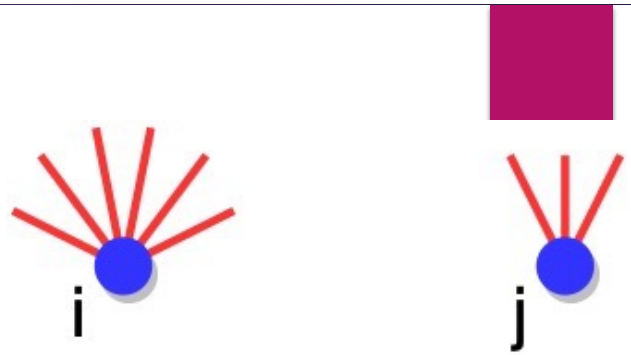
The Barabási-Albert Model

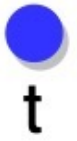
Generates graphs where a small number of nodes accumulate the majority of the links (hubs) => skewed degree distribution that fits the power-law observed in many real-world networks.

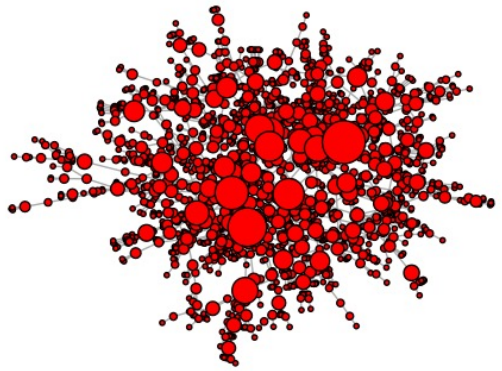
Perfect model?

Not really.

What's the issue?







Katarzyna Anna Kapitan, Network Analysis for Humanists,
Paris 2025

Lab

Lab5 feedback => Final Project

- The rationale behind your decisions regarding the construction of your datatset and network (nodes, links, weights, directed or not) need to be described and explained in in prose.
  - Tell me what is a node, why is it a node and not an edge? Could it be an edge instead? Why not? Why your network it weighted/directed?
- If you use additional packages and scripts to pre-process your data, they need to be referenced (and included in the *requirements* of your code)
- Your code needs to be commented!
  - Your comments explain what each of your functions is doing at each step.
  - It is a way of demonstrating that you understand what is going on.

Lab5 feedback => Final Project

- The goal of your project and your research question(s) need to be clearly specified.
- You don't need to create and analyse a huge network or try to come up with very ambitious research questions; something analogous to the comparison of wikipedia pages (Lab5) is fine.
- If you are working with someone else's data,
  - you still need to describe the datatset **AND** reference its source.
  - you need to describe how the question you are trying to answer with your project is **different** from the one for which this dataset was originally created.
  - You need to design your own code to process the data, so simple re-use and reproduction of someone else's results **won't be accepted.**
- Keep these points in mind while preparing your final project, in addition to the points outlined in the Assesment Guidelines (Kapitan_2025_NA_Final_Assignement_Guidelines.md).