Statistical Network Analysis

Prof. Dr. Ingo Scholtes

Chair of Machine Learning for Complex Networks
Center for Artificial Intelligence and Data Science (CAIDAS)
Julius-Maximilians-Universität Würzburg
Würzburg, Germany

ingo.scholtes@uni-wuerzburg.de

Lecture 04
Random Graph Ensembles

November 9, 2022

Notes:

- **Lecture L04:** Random Graph Ensembles 09.11.2022
- We show how we can use simple statistical measures to characterize large networks at a macroscopic, aggregate level. We introduce a statistical approach to make probabilistic statements about networks despite lacking knowledge about topological details. For this, we define simple stochastic models that generate *random graph ensembles* that can be used to reason about the expected properties of large networks.
  - Mixing patterns and Clustering coefficient
  - Macroscopic characterization of networks
  - Statistical ensembles of networks
  - Random graph: $G(n, m)$ and $G(n, p)$ model

- **Exercise sheet 02:** Random graph models due 16.11.2022

Aggregate characterization of networks

microscopic analysis

focus: **details of a specific network topology**

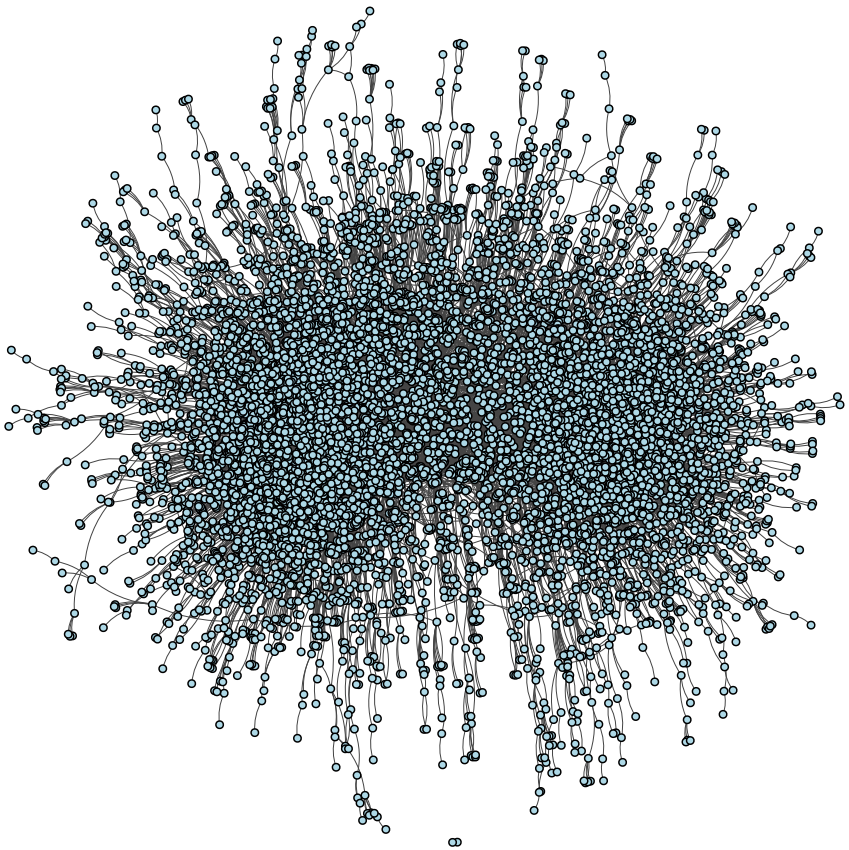
- what is the centrality of node c?
- what is the distance between nodes a and f?
- how many links does node c have?

macroscopic analysis

focus: **statistical regularities & stylized facts**

- what is the distribution of degrees?
- how long are paths on average?
- can we identify classes of similar networks?

how can we **model and characterize networks at an aggregate level**?



R & D collaborations between 8588 companies

Notes:

- In lecture L01, we distinguished between a microscopic and macroscopic perspective on complex networks. → L01 We now clarify this distinction. The **microscopic perspective** is concerned with a detailed view on the network topology, i.e. whether a link exists between a specific pair of nodes v and w exists is important. We need this detailed knowledge of the topology to make statements about specific nodes (e.g. about their centrality) or links (e.g. about their importance for shortest paths).
- But often we are not really interested in such microscopic details of a *specific* network. We are rather interested in statistical patterns that are shared by different networked systems, e.g. a certain distribution of degrees or the fact that nodes exhibit a preference to connect to other nodes that are similar to them. Addressing such settings, the **macroscopic perspective** focuses on networks "at large", i.e. we abstract from the details of a topology and aim at a large-scale characterization of patterns in complex networks.
- This allows us to identify similarities and differences between different systems, which can yield insights into universal principles that shape the evolution of complex networks. It further helps us to argue about so-called **stylized facts** (more on this later), which can be used to group networks in classes.
- But how can we characterize networks at an aggregate level? What could be suitable **macroscopic measures**?

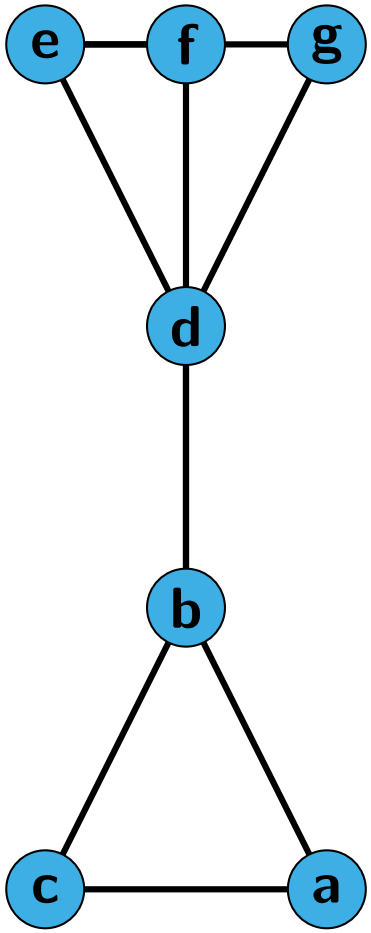
Degree sequence

- **degree sequence** of network $G = (V, E)$ is

$$S = (d_v)_{v \in V} \in \mathbb{N}^n$$

- which of the following are **degree sequences**?
  - $S = (4)$
  - $S = (1, 2, 2, 0, 2)$
  - $S = (1, 2, 2, 1, 2)$

- we call sequence $S \in \mathbb{N}^n$ **graphic** iff
  $\exists$ graph $G = (V, E) : S(G) = S$



example

$$S(G) = (2, 3, 2, 4, 2, 3, 2)$$

Notes:

- A first important macroscopic quantity, which does not depend on the details of the network topology, is the **degree sequence** of a network, i.e. the (ordered) sequence of node degrees in a network. Rather than considering the detailed topology of a network, for each node we count the number of nodes to which it is directly connected. In the example above, we assume that nodes are in alphabetical order, i.e. the ordered sequence captures degrees of nodes (a, b, c, d, e, f, g).
- Although this degree sequence contains information on the network topology, its size for a network with n nodes is only n. You can view this as a **compression** of a network into a (smaller) number sequence. This is a lossy compression because there are – in general – many different network topologies that are consistent with the same degree sequence.
- Not every sequence of n integer numbers is also a valid degree sequence of a network with n nodes. In the three examples above, only the third is a valid degree sequence. We call number sequences for which at least one graph with that degree sequence exists **graphical**. From the examples above we immediately see that graphic sequences must have an even sum and the second sequence is thus not graphic. Also, the sum of the degrees in a graphic sequence is bounded by two times the maximum number of edges (which depends on the nodes). The first sequence is thus not graphic. As an exercise, try to construct sequence of numbers that are not graphic. Which properties must a sequence of n numbers minimally fulfil to be a graphic sequence?

Degree distribution

- **degree distribution** $P : \mathbb{N} \to [0, 1]$ of network $G = (V, E)$ is defined as

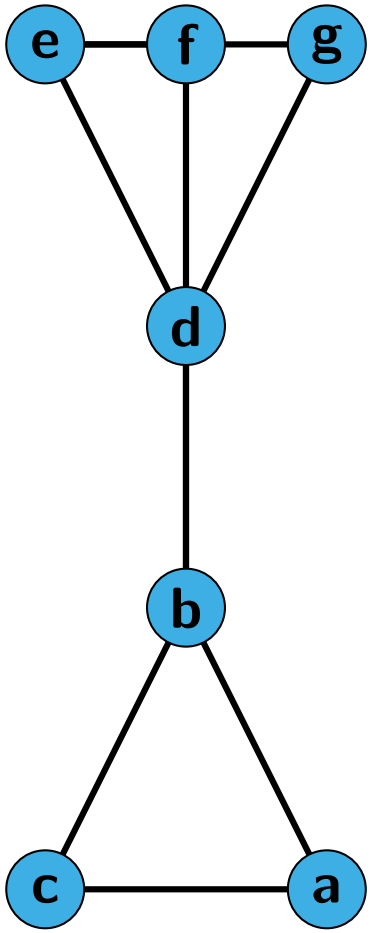
$$P(k) := \frac{N_k}{n}$$

with $N_k := |\{i \in V : d_i = k\}|$

and $n := \sum_k N_k = |V|$

- **mean degree** is defined as

$$\langle k \rangle := \sum_{k=0}^{\infty} k \cdot P(k) = \sum_{i \in V} \frac{d_i}{n} = \frac{2m}{n}$$



example

$P(2) = \frac{4}{7}, P(3) = \frac{2}{7}, P(4) = \frac{1}{7}$

$\langle k \rangle \approx 2.57$

Notes:

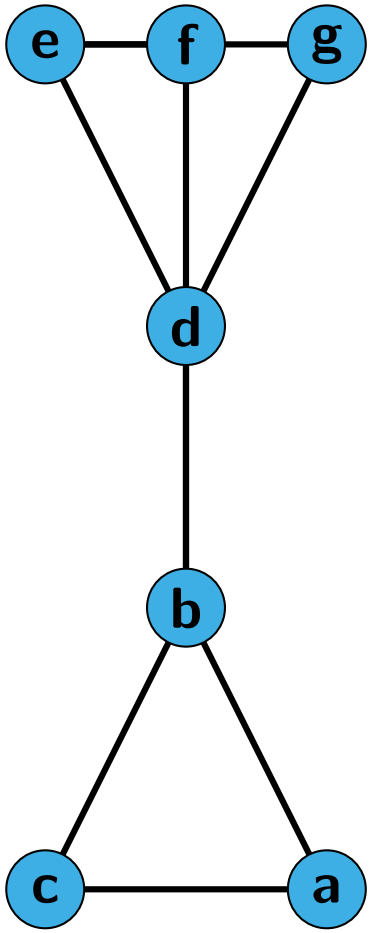
- We can go one step further and aggregate our sequence of n degrees to a **degree distribution**. I.e. rather than using the sequence of all degrees, we calculate relative frequencies $P(k)$ of unique degrees k in the sequence. For each degree k we first calculate the number of nodes N_k with degree k and divide N_k by the number of nodes n in the network. For each k we obtain the fraction of nodes with degree k. Interpreted as a probability, $P(k)$ is the probability that a node chosen uniformly at random has a given degree k. More precisely, we made a frequentist, maximum likelihood estimation of degree probabilities.
- We can use the degree distribution to calculate the **mean degree** in the network, which corresponds to the **expected degree** if we were to chose a node uniformly at random. The angle bracket notation $\langle k \rangle$ is borrowed from statistical mechanics, where angle brackets commonly refer to average quantities of "statistical ensembles".
- In the next lectures, we will see that the degree distribution of complex networks is a **simple yet powerful aggregate characterization of complex networks**. It can be used to make surprisingly strong analytical statements about expected network properties, such as diameter, average shortest path lengths, connectedness, component sizes, or robustness.

Mixing patterns

- **degree assortativity coefficient** is defined as

$$r := \frac{\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right) d_i d_j}{\sum_{ij}\left(d_i \delta_{ij} - \frac{d_i d_j}{2m}\right) d_i d_j} \in [-1, 1]$$

- **numerator**: $\mathrm{Cov}(X, Y)$ where X, Y are degrees d_i, d_j for randomly chosen edge (i, j)
- **denominator**: $\mathrm{Var}(X)$ where X is degree d_i for randomly chosen edge (i, j)
- denominator **normalizes** covariance of degrees for connected node pairs
- r is **Pearson correlation coefficient** of connected node degrees
  - $r > 0 \Rightarrow$ **assortative** mixing
  - $r < 0 \Rightarrow$ **disassortative** mixing
  - $r \approx 0 \Rightarrow$ **not assortative**



example: degree assortativity

$r \approx -0.26$

Notes:

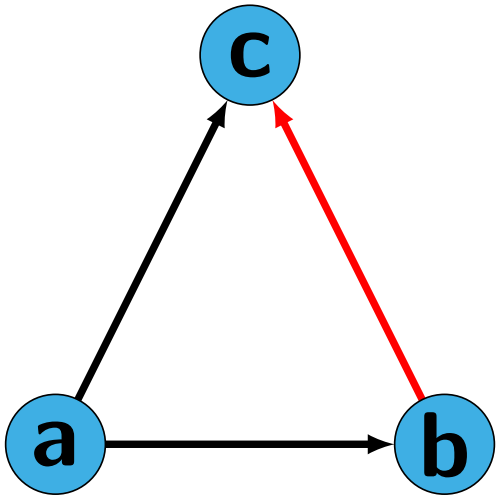
- Going beyond the degree of individual nodes, we can study how the degrees of *connected nodes* relate to each other. The **degree assortativity coefficient** is the Pearson correlation coefficient of degrees of nodes that are connected to each other. In the formula above, δ_{ij} is a delta function, i.e. $\delta_{ij} = 1$ iff $i = j$ and 0 otherwise.
- In lecture L03, we encountered a similar formula in the context of the *community assortativity coefficient*. The term *assortativity* highlights that these measures have something in common: they capture to what extent nodes "prefer" to connect to nodes that are *similar* to themselves. The community assortativity coefficient captures the preference of nodes to connect to nodes in the same community. The degree assortativity can thus be interpreted as follows:
  - positive values **(assortative mixing)** indicate that high-degree nodes are preferentially connected to other high-degree nodes and low-degree nodes are preferentially connected to other low-degree nodes
  - negative values **(disassortative mixing)** indicate that high-degree nodes are preferentially connected to low-degree nodes and vice-versa
  - values close to zero indicate that there is no preference for one over the other, in this case we say the network is **not assortative** (note that this is different from saying that it is *dis*assortative!)
- Large positive values of r tell us that high-degree nodes are likely to form a *core* of important nodes, while low degree nodes are likely in the *periphery* of the network. In social networks, assortative mixing is common, i.e. "well-connected" individuals are more likely to be connected to each other than expected at random.

Triadic closure

- consider graph $G = (V, E)$ where node a is connected to node b and node c
- does link (b, c) exist, i.e.

$$(a, b) \in E \wedge (a, c) \in E \Rightarrow (b, c) \in E?$$

- three edges (a, b), (a, c) and (b, c) are called **closed triad**
- can be defined for **directed and undirected networks**
- study of **triadic closure in social networks** was first proposed by sociologist Georg Simmel
- closed triads are an example for a **network motif**



Georg Simmel
1858 – 1918

image credit: public domain

Notes:

- Another way to study – at an aggregate, macroscopic level – how nodes are connected to each other uses the concept of triadic (or transitive) closure in networks. For this, we consider a node a that has two neighbors b and c. Given that a is connected to b and a is connected to c we can now ask whether b is also directly connected to c? If this is the case, we say that the nodes a, b, and c (or equivalently the three involved links) form a **closed triad**. A closed triad is a simple example for a so-called **network motif**. A frequent problem in graph mining is to efficiently find (and count) such motifs.
- The same pattern or motif in a network can also be studied from the perspective of **transitivity**. In this case, we study whether two links (a, b) and (b, c) imply *transitive closure*, i.e. the presence of a link (a, c) that transitively connects a to c. This notion of transitive closure in terms of *links* is different from transitive close in terms of paths. We generally assume that two links (a, b) and (b, c) imply a transitive path from a via b to c, even if no link (a, c) exists. This may sound like a trivial assumption but in a later lecture we will see that it can be violated in temporal networks.
- **Triadic closure** is common in social networks and the study of triads has first been proposed by the famous German philosopher and sociologist Georg Simmel. As an exercise, consider which mechanisms could lead to triad closure in a social network. The fact that social networks often exhibit triadic closure can actually be used as a simple recommendation scheme. How could this be used, e.g., in online social networks like Facebook, Twitter, or Instagram?

Clustering coefficient

- **local clustering coefficient** captures fraction of neighbor pairs that are connected
- for **undirected networks**

$$C_v := \frac{k(v)}{\binom{d_v}{2}} = \frac{2 \cdot k(v)}{d_v(d_v - 1)}$$

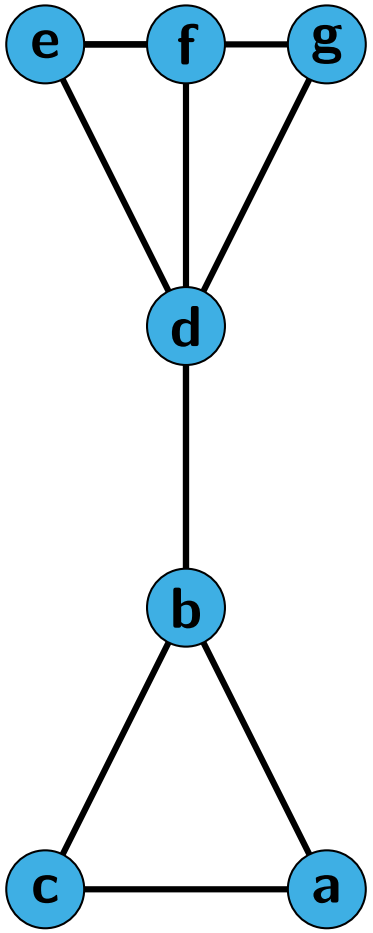
- for **directed networks**

$$C_v := \frac{k(v)}{2 \cdot \binom{d_{out}(v)}{2}} = \frac{k(v)}{d_{out}(v)(d_{out}(v) - 1)}$$

$$\text{with } k(v) := |\{(u, w) \in E : (v, u) \in E \wedge (v, w) \in E\}|$$

- **average clustering coefficient** is given as

$$C = \frac{1}{n} \sum_{v \in V} C_v$$



example

$$C_a = 1, C_b = \frac{1}{3}, C_f = \frac{2}{3}$$

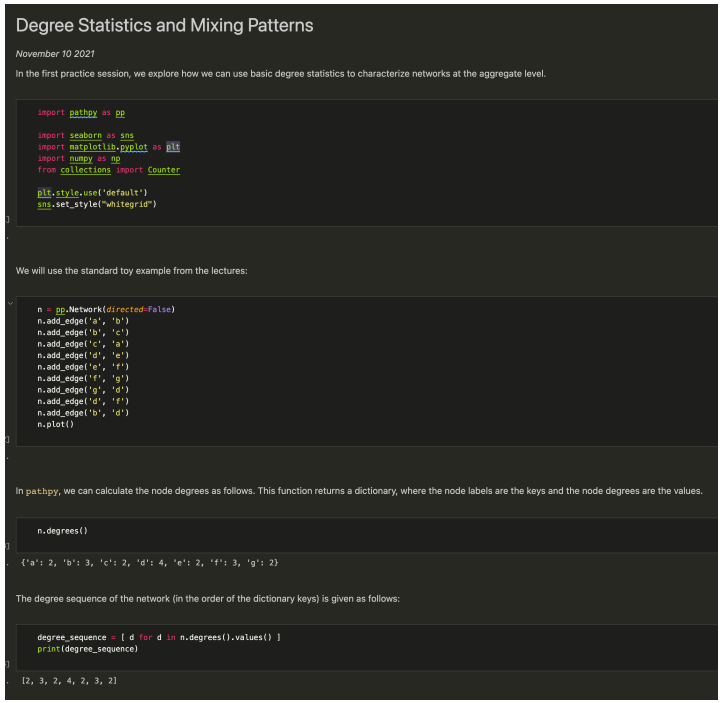
$$C \approx 0.76$$

Notes:

- We can quantify the presence of closed triads around nodes with the **local clustering coefficient**. It captures the fraction of neighbors of a node v that are connected to each other, i.e. how many of the possible triads around a given node v are closed. $k(v)$ counts the pairs of neighbors of node v that are connected. To define the local clustering coefficient we must know how many links can possibly exist between pairs of neighbors of a node i. The number of (unordered) pairs that we can draw from a set of n elements is given by the binomial coefficient $\binom{n}{2} := \frac{n(n-1)}{2!} = \frac{n(n-1)}{2}$.
- For **undirected networks** the number of pairs of neighbors of node i is thus $\binom{d_i}{2} = \frac{d_i(d_i-1)}{2}$, where d_i is the degree of node i. With $\frac{1}{\frac{1}{2}} = 2$ we get a factor of two 2 in the denominator.
- For **directed networks**, the number of neighbors is given by the out-degree of node i. Moreover, for each pair of nodes a link can exist in both directions, i.e. we have to multiply the binomial coefficient in the denominator with two (thus dividing C_v by two). This explains the definition above.
- The **local clustering coefficient** is a node-level measure. We can turn this into an aggregate, network-level measure by defining the **global clustering coefficient** as the average of local clustering coefficients. It captures whether nodes are, on average, surrounded by closed triads. In a probabilistic sense it is the expected fraction of triads that we can expect to be closed around a node chosen uniformly at random.

Practice session

- we use `pathpy` to calculate the **degree assortativity** of empirical networks
- we count closed triads and calculate the **clustering coefficient** of directed and undirected networks
- we use those measures for the **aggregate characterization** of empirical networks



Degree Statistics and Mixing Patterns

November 10 2021

In the first practice session, we explore how we can use basic degree statistics to characterize networks at the aggregate level.

```
import pathpy as pp

import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from collections import Counter

plt.style.use('default')
sns.set_style("whitegrid")
```

We will use the standard toy example from the lectures:

```
n = pp.Network(directed=False)
n.add_edge('a', 'b')
n.add_edge('b', 'c')
n.add_edge('c', 'a')
n.add_edge('d', 'e')
n.add_edge('e', 'f')
n.add_edge('f', 'g')
n.add_edge('g', 'd')
n.add_edge('d', 'f')
n.add_edge('b', 'd')
n.plot()
```

In `pathpy`, we can calculate the node degrees as follows. This function returns a dictionary, where the node labels are the keys and the node degrees are the values.

```
n.degrees()
```

```
{'a': 2, 'b': 3, 'c': 2, 'd': 4, 'e': 2, 'f': 3, 'g': 2}
```

The degree sequence of the network (in the order of the dictionary keys) is given as follows:

```
degree_sequence = [ d for d in n.degrees().values() ]
print(degree_sequence)
```

```
[2, 3, 2, 4, 2, 3, 2]
```

practice session

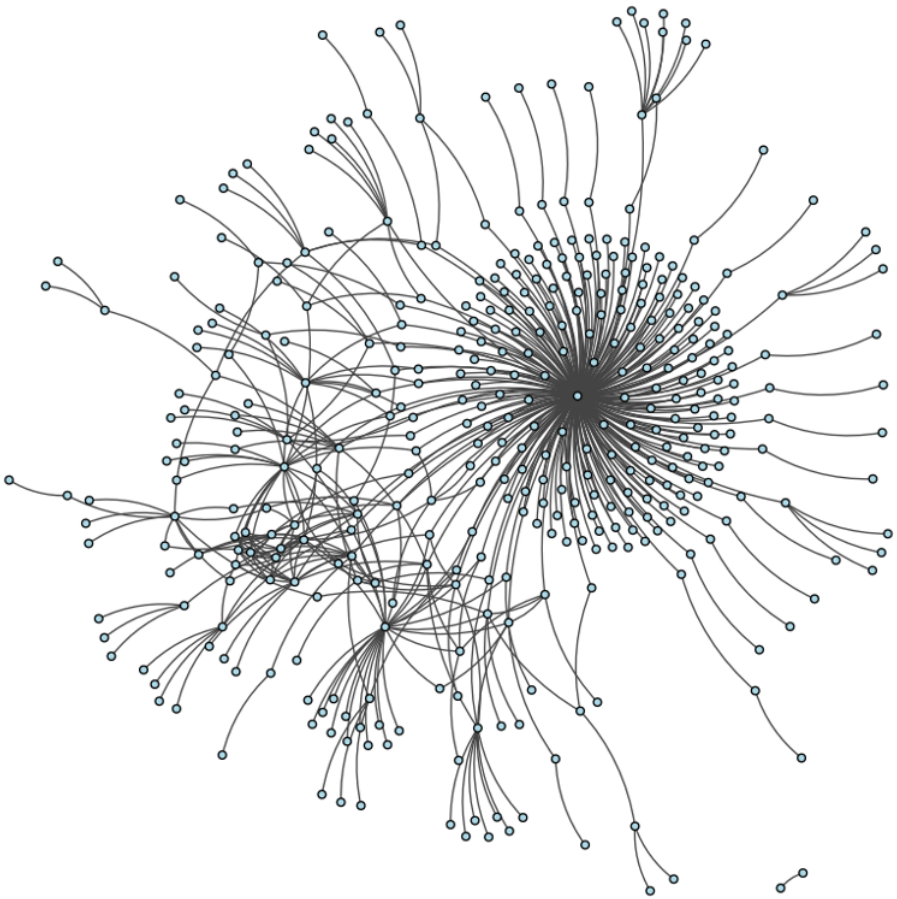
see notebooks 04-01 – 04-03 in `gitlab` repository at

→ `https://gitlab.informatik.uni-wuerzburg.de/ml4nets_notebooks/2022_wise_sna_notebooks`

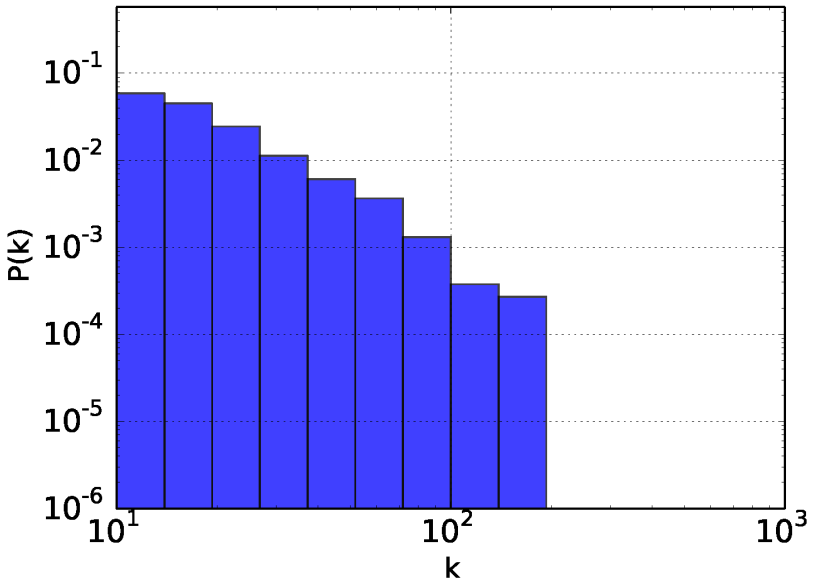
Notes:

- In the first practice session of this week, we implement functions that calculate degree statistics and the clustering coefficient for `pathpy` networks. We then apply those functions to empirical networks.

Example: gentoo collaboration network



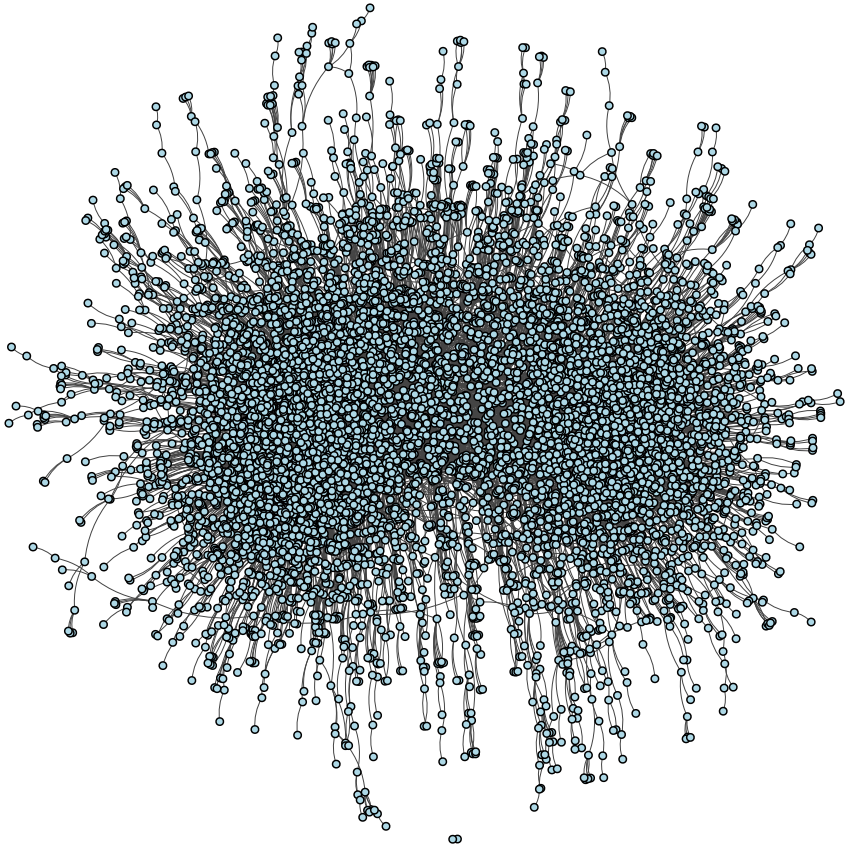
$\lvert V \rvert$	403
$\lvert E \rvert$	513
$\langle k \rangle$	2.55
diam(G)	8
$\langle l \rangle$	3.11
C	0.02
r	-0.35



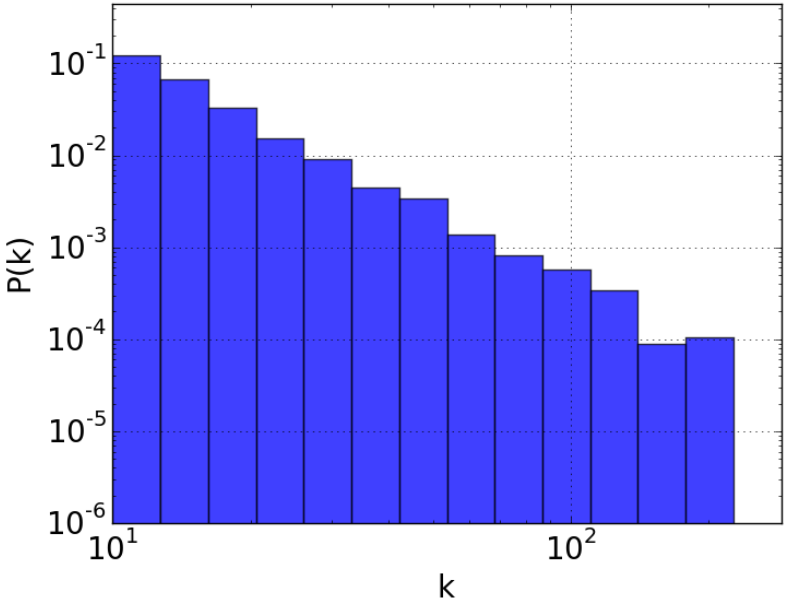
Notes:

- We apply those macroscopic characterizations to real-world examples: In the OSS collaboration network shown above, links capture the assignment of tasks between contributors to the OSS project Gentoo during a 30 day period in the year 2010. For simplicity, we have converted this to an undirected network, even though we could (and probably should) study this as a directed network.
- We observe a negative degree assortativity coefficient (i.e. this collaboration network exhibits **disassortative mixing**). Hence, high degree nodes are preferentially connected to low degree nodes. We can visually see this in the network above.
- The mean degree of this network is ≈ 2.55 and the global clustering coefficient is close to zero. Using the path-based measures introduced in lecture L02, we can further compute the diameter (diam $= 8$) and the average shortest path length $\langle l \rangle \approx 3.11$ of the network.
- The histogram above finally shows the degree distribution, where for each degree k in the network, we have the probability $P(k)$ of that degree. We see that – despite the average degree being rather small – there is still a non-zero probability to observe degrees > 100 that are much larger than the average. This broad distribution tells us that we must be careful with the interpretation of the mean degree, as we are likely to see many nodes whose degrees are smaller than the average, while for a few nodes the degree will be much larger than the average.

Example: R&D collaborations



$\lvert V \rvert$	8588
$\lvert E \rvert$	16166
$\langle k \rangle$	3.77
$\mathrm{diam}(G)$	17
$\langle l \rangle$	5.41
C	0.35
r	0



Notes:

- As a second example we study publicly declared R&D collaborations between more than 8,500 companies.
- Considering the large size of the network, we observe a relatively small diameter and average path length.
- The clustering coefficient is comparably large and the network is not assortative (with a degree assortativity coefficient r close to zero).
- The mean degree is smaller than in the previous example, and we again have a broad distribution of node degrees, with a small number of nodes having hundreds of R & D collaborations, while the majority has 10 (or less).

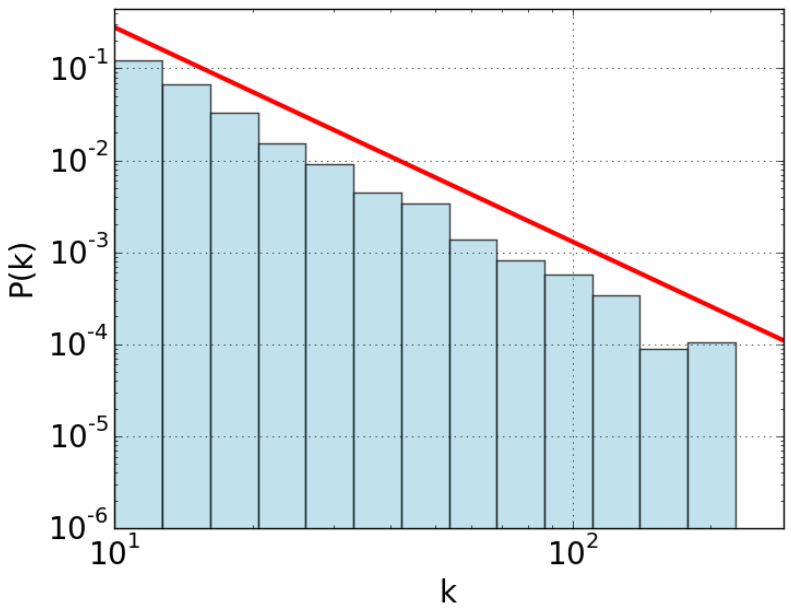
Stylised facts

- **stylised summary of empirical finding**

> "theorists should be free to start off with a stylised view of the facts - i.e. concentrate on broad tendencies, ignoring individual detail."
>
> → Nicholas Kaldor, 1961

- **example:** broad degree distribution

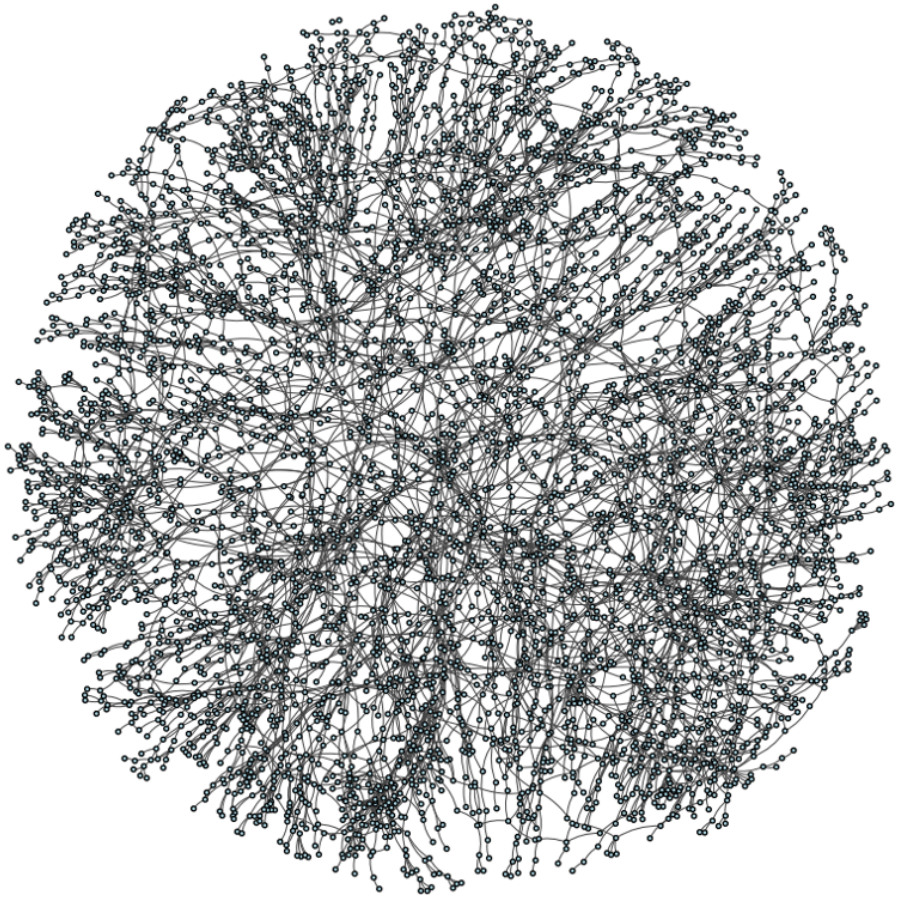
$\|V\|$	8588
$\|E\|$	16166
$\langle k \rangle$	3.77
$diam(G)$	small
$\langle l \rangle$	small
C	large
r	not assortative



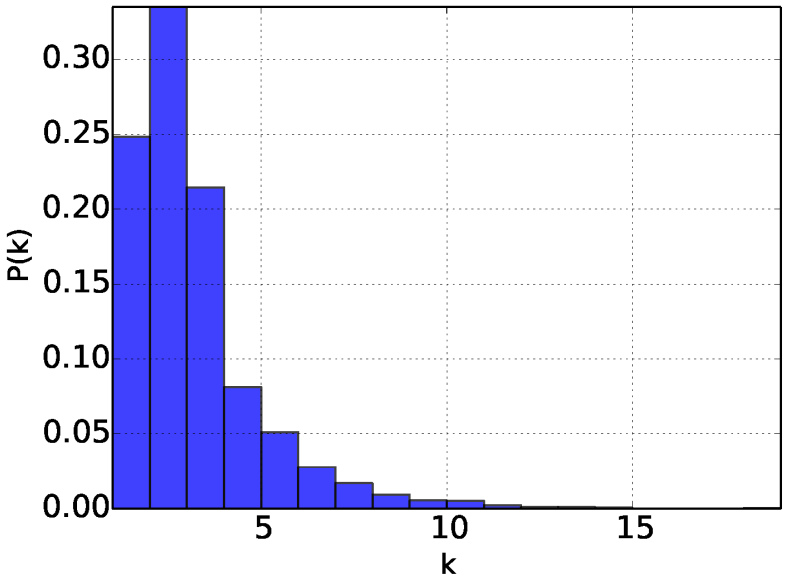
Notes:

- We often summarise such findings in terms of so-called **stylised facts**, a notion introduced by economist Nicholas Kaldor. → N Kaldor, 1961
- The idea is to focus on broad tendencies rather than specific details of a system's properties. In the example above, we could say that the clustering coefficient is small, the network is not assortative and – considering the size of the system – the diameter and average shortest path lengths are small. And just like for the OSS collaboration network, we see that the network exhibits a **broad degree distribution**, i.e. we have non-zero probabilities for large degrees, even though the mean degree is rather small.
- We can use such stylised facts to compare different systems in terms of broad tendencies and statistical patterns.

Example: power grid network



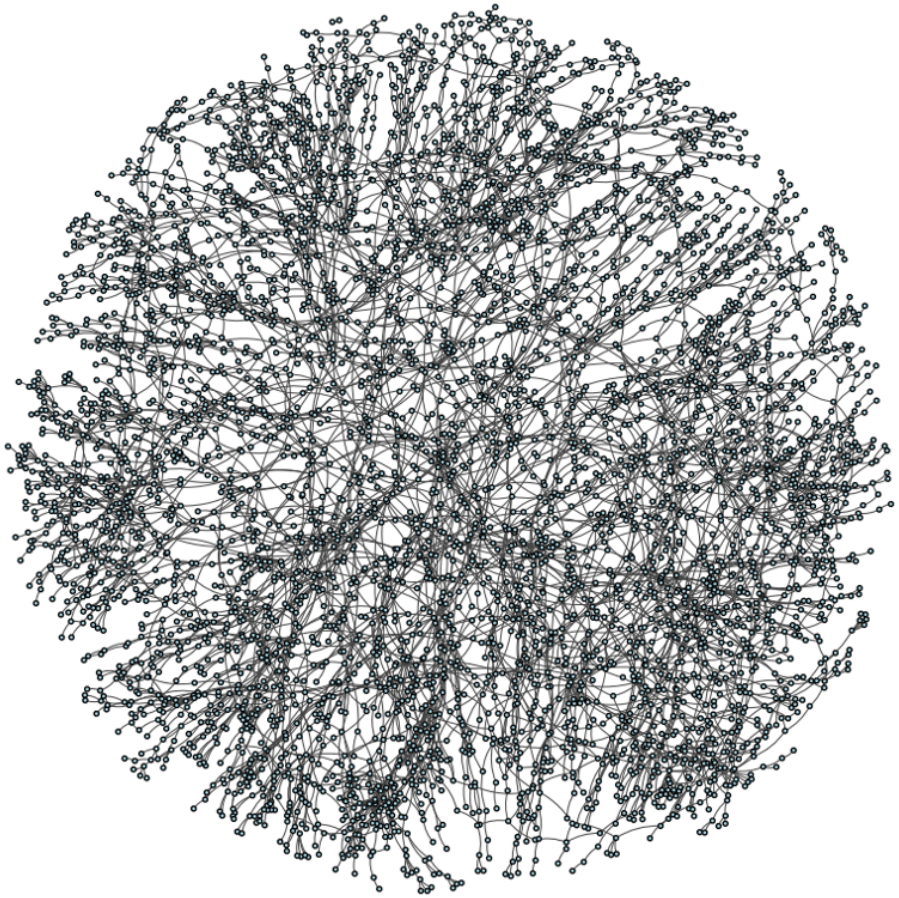
$\lvert V \rvert$	4943
$\lvert E \rvert$	6595
$\langle k \rangle$	2.67
$diam(G)$	46
$\langle l \rangle$	18.99
C	0.11
r	0



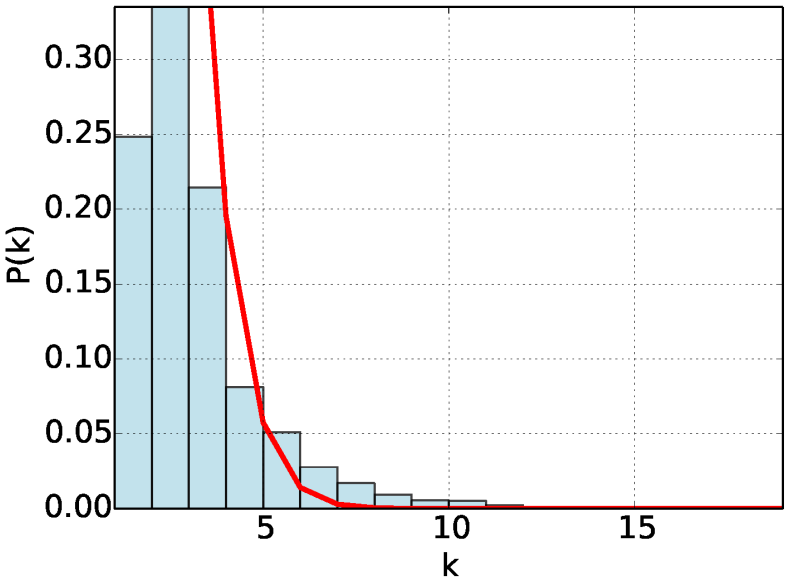
Notes:

- As a third example, we consider the power grid network of the western states in the United States. Nodes are power plants and transformer stations and undirected links are high-voltage transmissions lines
- Which stylised facts do we observe here?

Example: power grid network



$\lvert V \rvert$	4943
$\lvert E \rvert$	6595
$\langle k \rangle$	2.67
$diam(G)$	large
$\langle l \rangle$	large
C	small (?)
r	not assortative

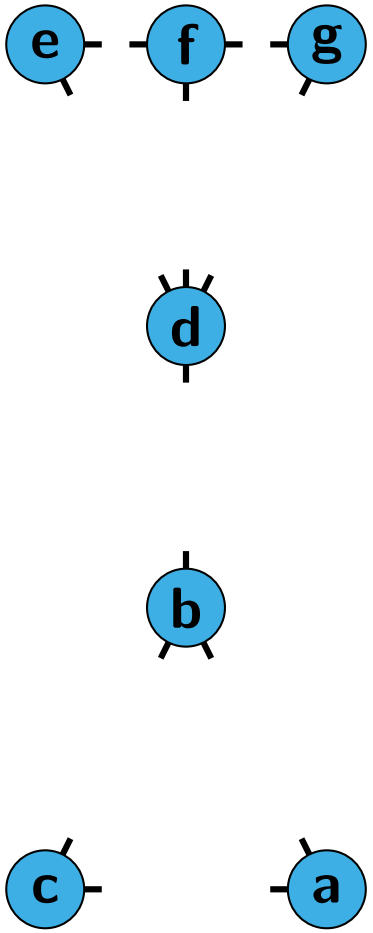


Notes:

- The diameter and the average shortest path length are – even though the size of the network is smaller than the previous example – much larger than that of the R & D network. We further observe that – different from the gentoo collaboration network – the network is not assortative. The global clustering coefficient is smaller than in the R&D collaboration network (but not zero).
- Looking at the histogram of the degrees, we observe a **narrow degree distribution**, where no node has a degree larger than ≈ 15. This is in contrast to the broad degree distributions which we observed in the two previous examples.
- In the figure above, I have fitted the probability mass function of a Poisson distribution to the observed degree distribution. In next week's lecture, we will see why this Poisson distribution is actually a good fit for the empirical distribution.
- But this example also raises an important question: We need to define what exactly we actually mean by a "stylised fact" or "statistical pattern". To do this, we must define a baseline for what we can already expect "at random". This will allow us to clarify when exactly we consider the diameter, average path length, or clustering coefficient as "small" or "large". Moreover, it helps us to say whether it is the broad or the narrow degree distribution that we should consider as an "interesting" statistical pattern.

A statistical perspective?

- knowledge of the **full network topology** allows statements about a particular system
- we often . . .
  - **do not know** the full network topology → WHY?
  - **are not interested** in particular system → WHY?
- we **lack or ignore knowledge about details**
- assuming that we only know the **degree sequence**, how can we argue about ...
  - connected components,
  - diameter and path lengths,
  - clustering coefficient?
- need **statistical methods** to model **lack of knowledge** about topological details



example

$S(G) = (2, 3, 2, 4, 2, 3, 2)$

Notes:

- This question which of the characteristics can be expected "at random" necessitates a **statistical perspective** on complex networks that is closely related to the microscopic vs. macroscopic perspective introduced before. A microscopic analysis is reasonable whenever we want to argue about a specific system for which we know the full topology. It is, however, not suitable for situations where we have **incomplete knowledge** about a system's topology, e.g. because ...
  - there is no central instance with access to the full topology (think about contact networks)
  - we can only sample portions of the system (think about the Internet topology)
  - the topology changes faster than we can possibly measure it
  - the network is too large (think about the FACEBOOK social network with more than 2 Billion nodes)
- Moreover, even if we *can* analyze the details of the topology are often not interested in the detailed properties of a particular system but we rather want to understand why *different systems* exhibit similar stylised facts. As an example consider the two collaboration networks (OSS and R&D collaborations): both have a broad degree distribution and share features like a small diameter and short average path lengths. Is this a special property of these systems or can it be explained as a result of aggregate characteristics like network size, density of links, degree distribution, etc.?
- A **statistical approach** to networks helps us to explicitly model the lack of knowledge outlined above, enabling probabilistic statements that do not depend on the details of a specific network topology.

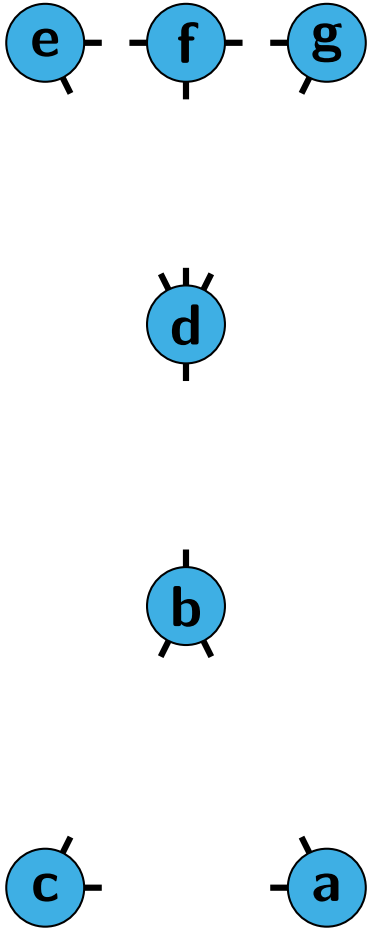
Statistical ensembles

statistical network ensemble

For a given aggregate characterization X, a **statistical network ensemble** is a probability space with sample space $\Omega(X)$ and probability measure P where

- $\Omega(X)$ is the set of all networks G consistent with X
- $P : G \to [0, 1]$ for each $G \in \Omega(X)$

X is called the **macrostate** of the ensemble. Each event $G = (V, E) \in \Omega(X)$ is called a **microstate**.



- terminology borrows from **statistical mechanics of many-particle systems**
- macrostate = **thermodynamic quantities** (particle number, volume, temperature, etc.)
- microstates = all **possible configurations** of a particle system

example 1

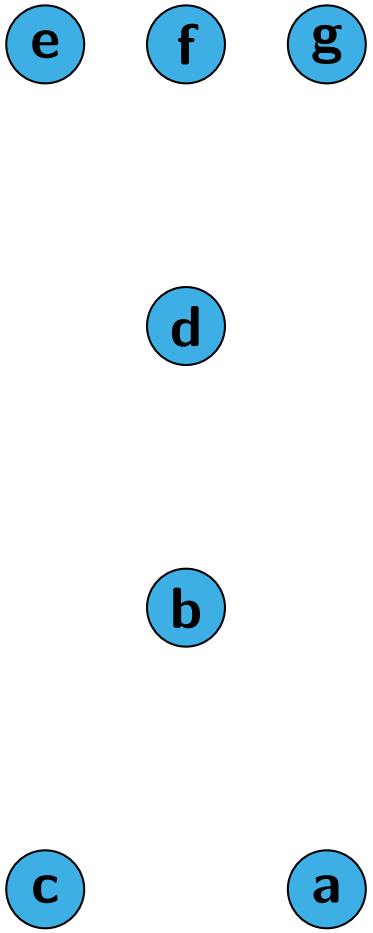
- **macrostate:** degree sequence $S = (2, 3, 2, 4, 2, 3, 2)$
- **microstates:** $\{G | S(G) = S\}$

Notes:

- The problem that we want to make statements about a network despite a lack of knowledge about its details resembles problems in statistical physics and we will actually use similar methods. In statistical mechanics,, statistical ensembles are used to study many-particle systems for which only macroscopic (thermodynamic) quantities (temperature, volume) can be measured, while the microscopic details of the underlying particles are unknown. Just like in our case, the details are not considered because (i) it is impossible to observe them, and (ii) we are eventually interested in macroscopic quantities rather than particle positions.
- Take a moment to think about the terms macro- and microstate introduced above
  - both describe the *state* of a complex network
  - at the macro-level, the state of a complex network is defined in terms of aggregate-level properties that are potentially consistent with multiple microstates
  - at the micro-level, the state of a network is defined by the detailed link topology, which we can use to calculate the macrostate
- We are usually interested in **statistical regularities** of microstates, i.e. we want to answer questions like: Given a macrostate X, what is the probability that a randomly chosen microstate has property Y? What is the expected property $\langle Y \rangle$ of a randomly chosen microstate that is consistent with macrostate X?
- A first example: most microstates consistent with the macrostate $S(G) = (2, 3, 2, 4, 2, 3, 2)$ will have a single connected component. So without further further knowledge, we can expect that a network with that degree sequence is connected (even though there are a some microstates where this is not the case).

Degree distribution and mean degree

- we can further **reduce knowledge** about microscopic details of the network
  - degree distribution rather than sequence
  - mean degree rather than distribution
- what statements can we still make about the network?
- ensemble with fixed n and $\langle k \rangle \approx 2.57$ corresponds to **ensemble with fixed n and m**
- how we can analytically study such ensembles?



example 3

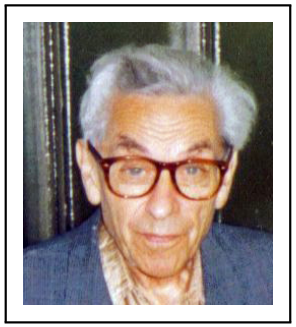
- **macrostate:** $n = 7, \langle k \rangle \approx 2.57$
- **microstates:** $\{G | \langle k \rangle = \langle k \rangle\}$

Notes:

- We can further reduce the information about the details of the network preserved in the macrostate. We can drop the constraint that we are interested in all microstates with a given degree sequence, instead considering the **statistical ensemble of networks with a given degree distribution**. If we are still able to make statements about the expected properties of microstates given this (less detailed) macrostate, these statements can be generalized to networks with the same degree distribution.
- In our example, we can define the ensemble of networks with seven nodes, where the probability that a randomly chosen node has degree k is given by the degree distribution. What statements about the expected properties of microstates can we make? Will they be connected? What is their expected diameter?
- Finally, we can go one step further and drop the constraint of s fixed degree distribution, instead studying the **statistical ensemble of networks with a given mean degree**. From $\langle k \rangle = \frac{2|E|}{|V|}$, we see that this is equivalent to fixing the number of nodes and links. This further reduces our knowledge about the details of the topology. What statements about the expected properties of a microstate can we still make? Do you still think that the network is likely to be connected? What about path lengths?
- We can answer such questions based on **random graph models**. The two simple random graph models that we will introduce today, have been the basis of statistical network analysis for more than five decades. The availability of data on real-world networks has recently led to an increasing interest in more complex models. In order to understand what these complex models add to our understanding of statistical patterns in complex networks, we start with the basic random graph models.

Random graphs: $G(n, m)$ model

- consider a simple **generative stochastic model** for random graphs → P Erdős and A Rényi, 1959
- **model parameters**
  - number of nodes n
  - number of links m
- **generative procedure**
  - generate empty network with n nodes
  - add m links to pairs of nodes chosen uniformly at random
- **microstates**: all networks $G = (V, E)$ with $|V| = n$ and $|E| = m$
- for all microstates we have $\langle k \rangle = \frac{2m}{n}$
- probability $P(G)$ of a given microstate G?



Pál Erdős
1913 – 1996

Alfred Rényi
1921 – 1970

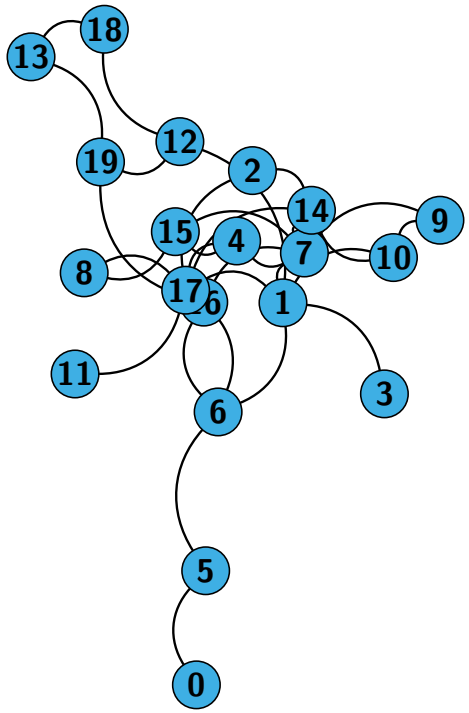


image credit: alchetron.com / fair use, user Kmhkmh, Wikimedia Commons, CC-BY-SA 2.0

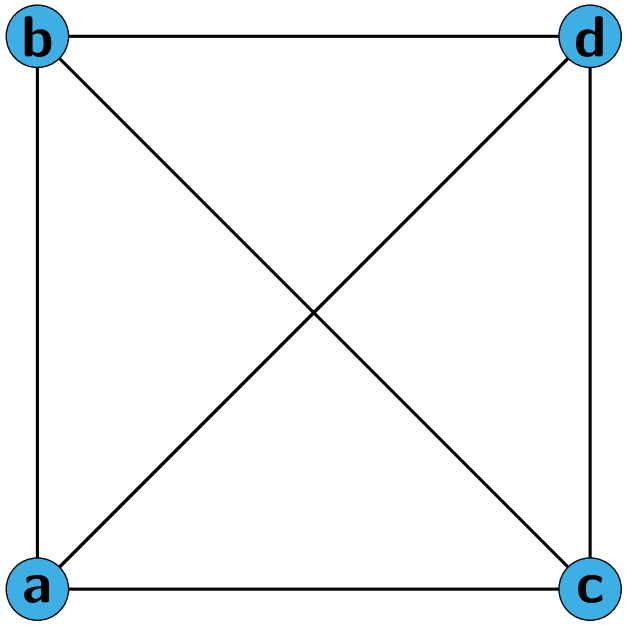
Notes:

- The $G(n, m)$ **model** allows us to directly address the question which properties we can expect in a network, given that we only know the number of nodes and links. It was introduced in 1959 by Pál Erdős and Alfréd Rényi, even though other researchers have studied random graph models before. The $G(n, m)$ assumes that the number of nodes and links of a network are fixed, while the topology of the network is randomly generated as follows:
- We start with an empty network with n nodes and distribute m links to pairs of nodes chosen uniformly at random.
- This is a standard urn problem in combinatorics, where the "urn" contains all possible links in a network with n nodes. From this urn, we draw m links uniformly at random (typically without replacement).
- How many links can a network possibly have, i.e. what is the size of the urn? This depends on whether we study directed or undirected networks with or without self-loops. Hence, to make further statements about the microstate probabilities in this model we need a primer in (network) combinatorics.

Combinatorics primer: undirected networks

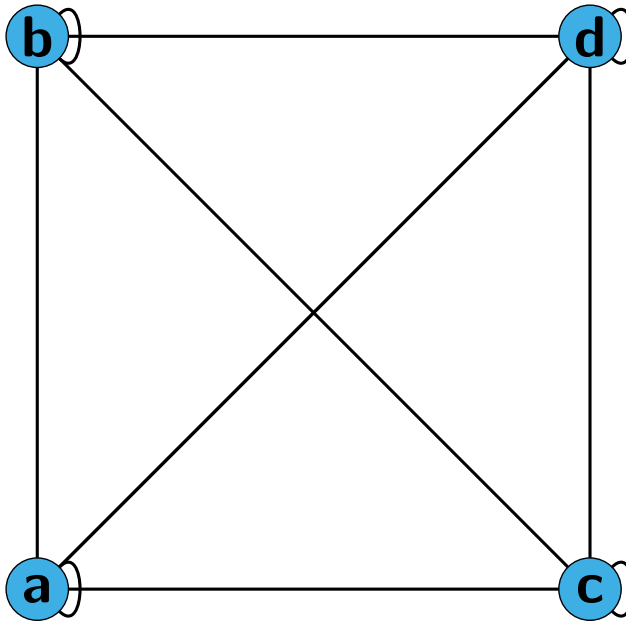
Maximum number of links in **undirected** network with n nodes?

without self-loops



$$\binom{n}{2} = \frac{n \cdot (n-1)}{2}$$

with self-loops



$$\binom{n+1}{2} = \frac{n \cdot (n+1)}{2}$$

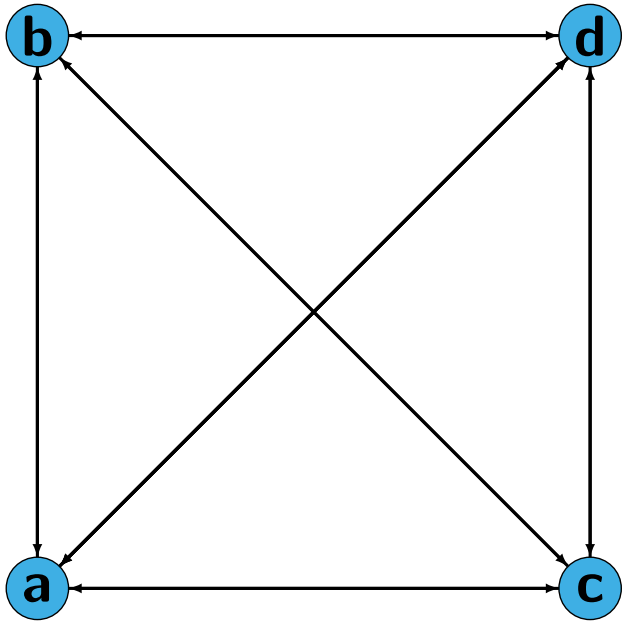
Notes:

- We first have to distinguish whether or not we allow self-loops, i.e. a link of a node to itself. For undirected networks drawing random links then corresponds to the following sampling processes:
- Without self-loops:
  - corresponds to a sampling of node pairs without ordering and without replacement
  - There are $\binom{n}{2}$ different combinations in which we can draw 2-element subsets from an urn with n elements
  - For an undirected network without self-loops and $n = 4$ we have: $\binom{n}{2} = \frac{12}{2} = 6$ possible links
- With self-loops:
  - corresponds to a sampling of node pairs without ordering but with replacement
  - For this process, the general formula is $\binom{n+k-1}{k}$, so for $k = 2$ we have the formula above
  - For an undirected network with self-loops and $n = 4$ we have: $\binom{n+1}{2} = \frac{20}{2} = 10$ possible links

Combinatorics primer: directed networks

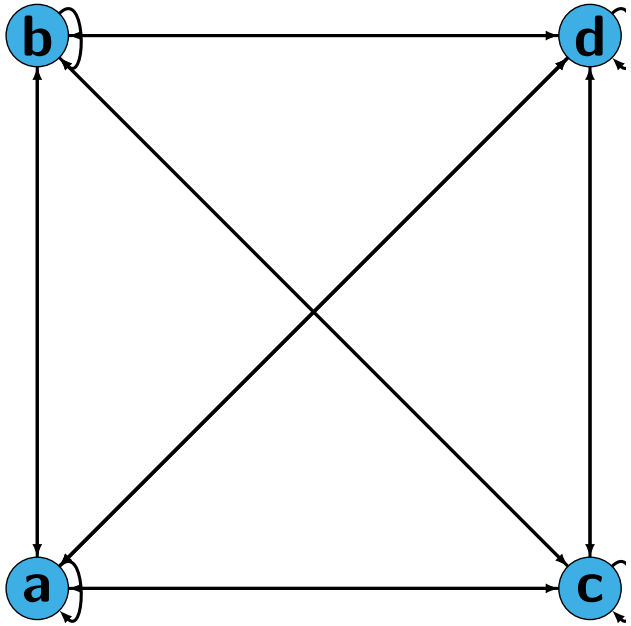
Maximum number of links in **directed** network with n nodes?

without self-loops



$$2 \cdot \binom{n}{2} = n \cdot (n-1)$$

with self-loops



$$n^2$$

Notes:

- For directed networks, we again distinguish whether we do or do not allow self-loops:
- Without self-loops
  - We consider a sampling process with ordering and without replacement
  - We can easily calculate this number based on the formula for undirected networks. Since each undirected link exists in two directions, we just multiply by this number by two
  - For a directed network without self-loops and $n = 4$ we have: $2 \cdot \binom{n}{2} = 12$ possible links
- With self-loops
  - Here we can simply consider all permutations (with repetitions) of length two, with elements chosen from a set of size n
  - In general, we have n^k so permutations of length k, so for pairs of n nodes we get n^2
  - For undirected networks with self-loops and $n = 4$ we thus have: $n^2 = 16$ possible links
- **Important:** in the following study of random graph models, we consider **undirected networks with self-loops**. However, using the discussion above it is easy to generalise this to other cases.

$G(n, m)$ model: microstate probability

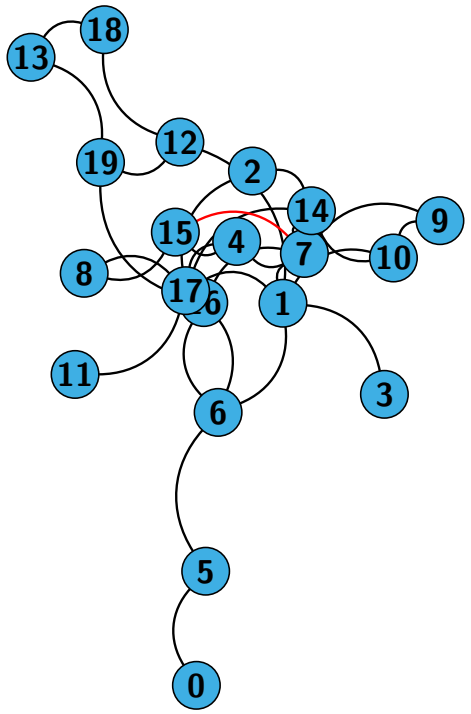
- consider $G(n, m)$ model for **undirected** networks **with** self-loops (no multi-edges)
- number of different microstates that are consistent with the macrostate is

$$Z(n, m) := |\Omega(n, m)| = \binom{\binom{n+1}{2}}{m}$$

- in $G(n, m)$ model all **microstates are equiprobable**, i.e.

$$P(G) = \frac{1}{Z(n, m)} \text{ for all } G \in \Omega(n, m)$$

- what is the probability that a link (v, w) is generated?



random microstate generated using $G(n, m)$ model for $n = 20$ and $m = 30$

Notes:

- Since we know the number of possible links in a network with n nodes, we can easily calculate the number of possible graphs with n nodes and m links. For an undirected network with n nodes and self-loops, we have $\binom{n+1}{2}$ possible links. From this set of possible links, we chose m links at random without replacement (and we don't care about the order in which they are chosen). So the number of networks with n nodes and m links is $\binom{\binom{n+1}{2}}{m}$ (i.e. this is the size of our sample space $\Omega(n, m)$).
- Since we draw links uniformly at random, all microstates are generated with the same probability, so we have $P(G) = \frac{1}{Z(n,m)}$ where $Z(n, m) = \binom{\binom{n+1}{2}}{m}$ gives the size of the sample space $\Omega(n, m)$.
- In statistical mechanics, $Z(n, m)$ is called "partition function" or "Zustandssumme" (because it sums statistical weights over all possible states).
- In other words: in the $G(n, m)$ model, the microstate is chosen uniformly at random from the space of all networks with n nodes and m edges. This assumes that, e.g. in a social network with n users and m friendships, each realization of the network is equally probable.
- There is one complication: consider the probability that a specific pair of nodes (v, w) is connected? Since we sample m links **without** replacement, links are actually not sampled independently, which makes it cumbersome to calculate this probability.

$G(n, p)$ model

- consider different stochastic generative stochastic model → E N Gilbert, 1959
  - create links **independently**
  - relax constraint of having **exactly** m links
- **model parameters**
  - number of nodes n
  - probability p to form links between pairs of nodes
- **microstates**: all networks with n nodes and **any number of links** (incl. self-loops)
- expected number of links $\langle m \rangle$ generated for a microstate G is

$$\langle m \rangle = p \cdot \binom{n+1}{2}$$

Edgar Nelson Gilbert

1923 – 2013

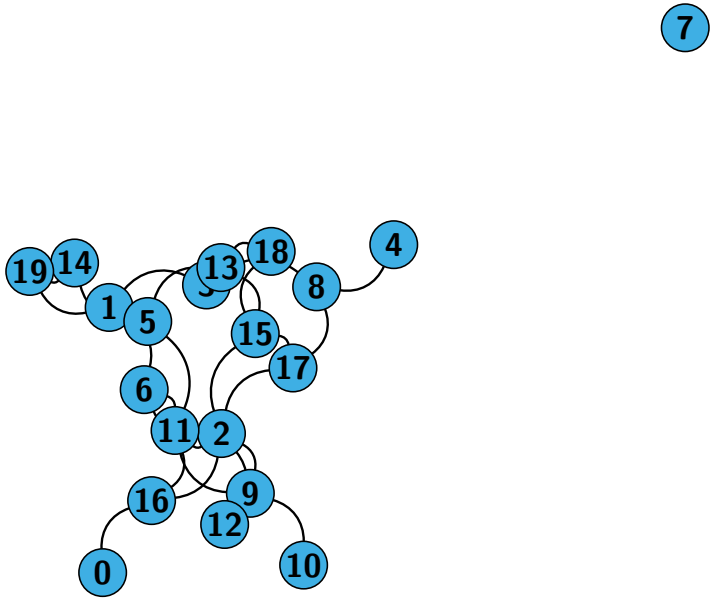


image credit: legacy.com, fair use

Notes:

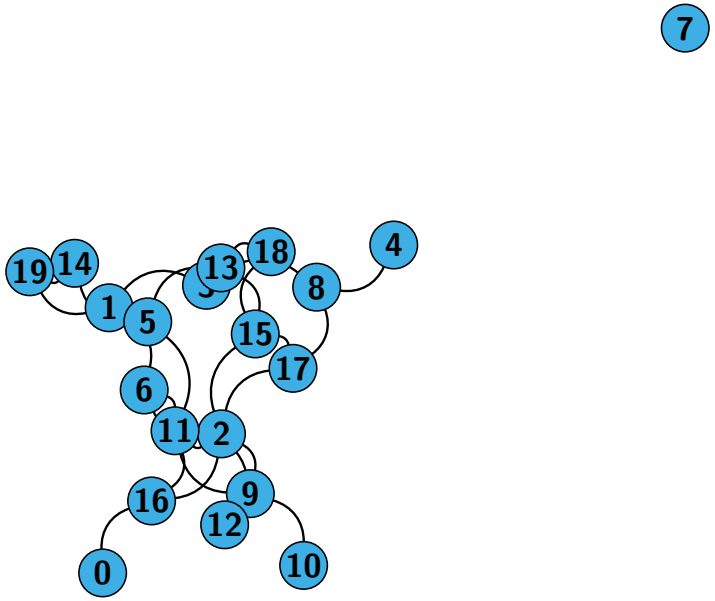
- The $G(n, p)$ model allows for a simpler mathematical treatment of the sampling process. The idea is to connected pairs of nodes independently from each other, by defining a simple single "link creation" probability p. This simple random graph model was introduced in the 1959 paper "Random Graphs" by Edgar Nigel Gilbert.
- It corresponds to a generative model where, for each pair of nodes, we perform a Bernoulli trial and generate a link upon the success of this trial (which happens with probability p).
- importantly, here the sample space consists of all networks with exactly n nodes and any number of links. So, different from the $G(n, m)$ model, every network topology with n nodes is a microstate in the resulting statistical ensemble. Network realizations with different numbers of links have, in general, different probabilities. As an exercise, you can try to find a parameter p where this is not the case.
- If we allow self-loops, we have $\binom{n+1}{2}$ possible pairs of nodes. Since we make $\binom{n+1}{2}$ Bernoulli trials, where each trial yields a link with probability p, the expected number of links generated by the $G(n, p)$ model is $p \cdot \binom{n+1}{2}$.

$G(n, p)$ model: microstate probability

- **microstate probability** $P(G)$ depends on number of links in microstate
- for microstate G with n nodes and m_G links

$$P(G) = p^{m_G} \cdot (1-p)^{\binom{n+1}{2} - m_G}$$

- how does the microstate probability change if we change p?
- is there a value for which we recover equiprobable microstates?
- what is the degree distribution of microstates?



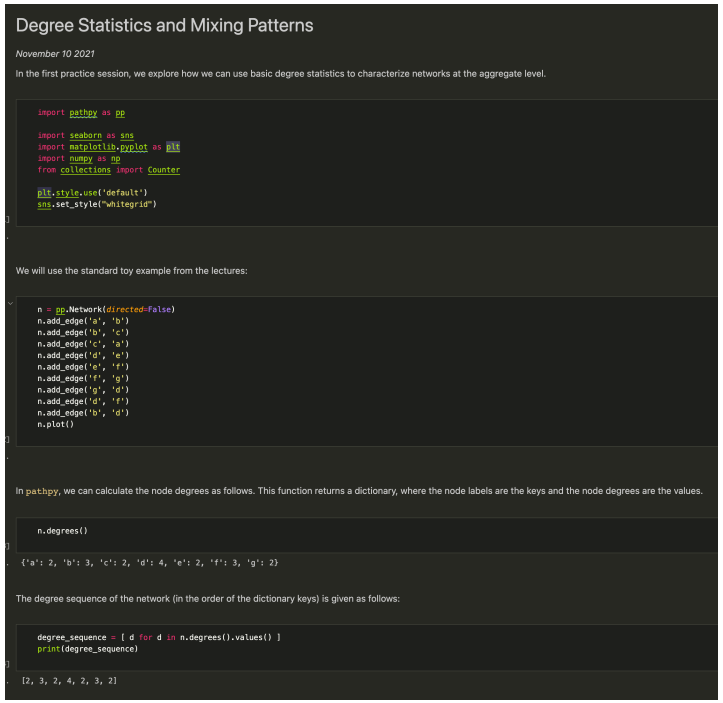
random microstate generated using $G(n, p)$ model with $n = 20$ and $p = 0.18$

Notes:

- In the $G(n, m)$ model, we could easily calculate the microstate probability based on (i) the known total number of microstates, and (ii) the fact that all microstates are equally likely. But what is the **probability of a microstate in the $G(n, p)$ model**?
- Here, not all microstates are equally likely because the probability to generate a particular microstate depends on the number of links of this microstate.
- What we can do is to calculate the probability of a microstate with exactly m links: this corresponds to having m successes in a sequence of $\binom{n+1}{2}$ Bernoulli trials that succeed with probability p, while all other $\binom{n+1}{2} - m$ trials are not successful (which happens with probability $1 - p$). This allows us to calculate the probability of a microstate with m links.
- Note that in the $G(n, p)$ model, the probabilities of all microstates that have the same number of links are the same (i.e. the model does not differentiate between microstates as long as they have the same number of links).

Practice session

- we use `pathpy` to implement two simple **random graph models**
- we study **properties of random graphs** and compare them to empirical networks



Degree Statistics and Mixing Patterns

November 10 2021

In the first practice session, we explore how we can use basic degree statistics to characterize networks at the aggregate level.

```
import pathpy as pp

import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
from collections import Counter

plt.style.use('default')
sns.set_style("whitegrid")
```

We will use the standard toy example from the lectures:

```
n = pp.Network(directed=False)
n.add_edge('a', 'b')
n.add_edge('b', 'c')
n.add_edge('c', 'a')
n.add_edge('d', 'e')
n.add_edge('e', 'f')
n.add_edge('f', 'g')
n.add_edge('g', 'd')
n.add_edge('d', 'f')
n.add_edge('b', 'd')
n.plot()
```

In `pathpy`, we can calculate the node degrees as follows. This function returns a dictionary, where the node labels are the keys and the node degrees are the values.

```
n.degrees()
```

```
{'a': 2, 'b': 3, 'c': 2, 'd': 4, 'e': 2, 'f': 3, 'g': 2}
```

The degree sequence of the network (in the order of the dictionary keys) is given as follows:

```
degree_sequence = [ d for d in n.degrees().values() ]
print(degree_sequence)
```

```
[2, 3, 2, 4, 2, 3, 2]
```

practice session

see notebooks 04-04 – 04-05 in `gitlab` repository at

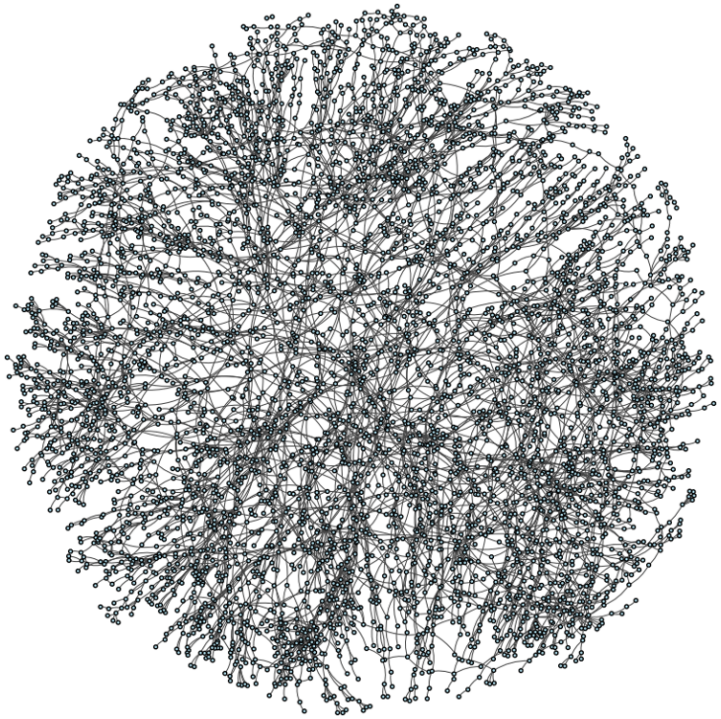
→ `https://gitlab.informatik.uni-wuerzburg.de/ml4nets_notebooks/2022_wise_sna_notebooks`

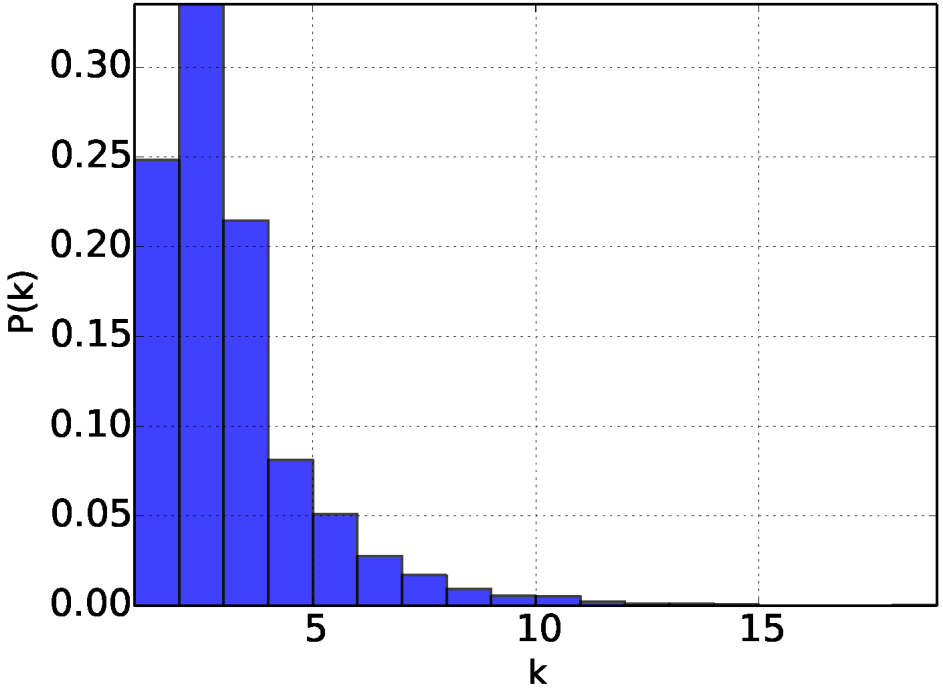
Notes:

- In the second practice session, we implement functions that generate random graphs according to the $G(n, m)$ and $G(n, p)$ model and use them study properties of the generated random graphs.

Example: power grid network

$\lvert V \rvert$	4943
$\lvert E \rvert$	6595
$\langle k \rangle$	2.67
$diam(G)$	46
$\langle l \rangle$	18.99
C	0.11
r	0



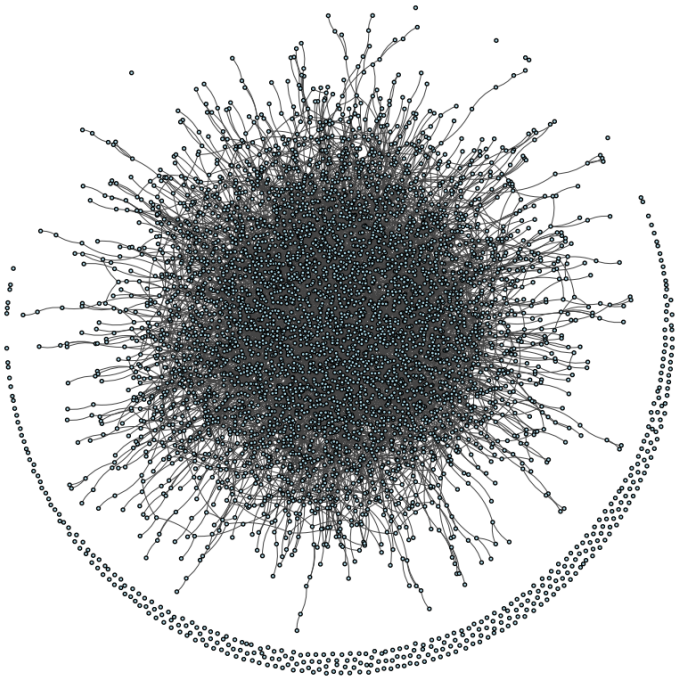


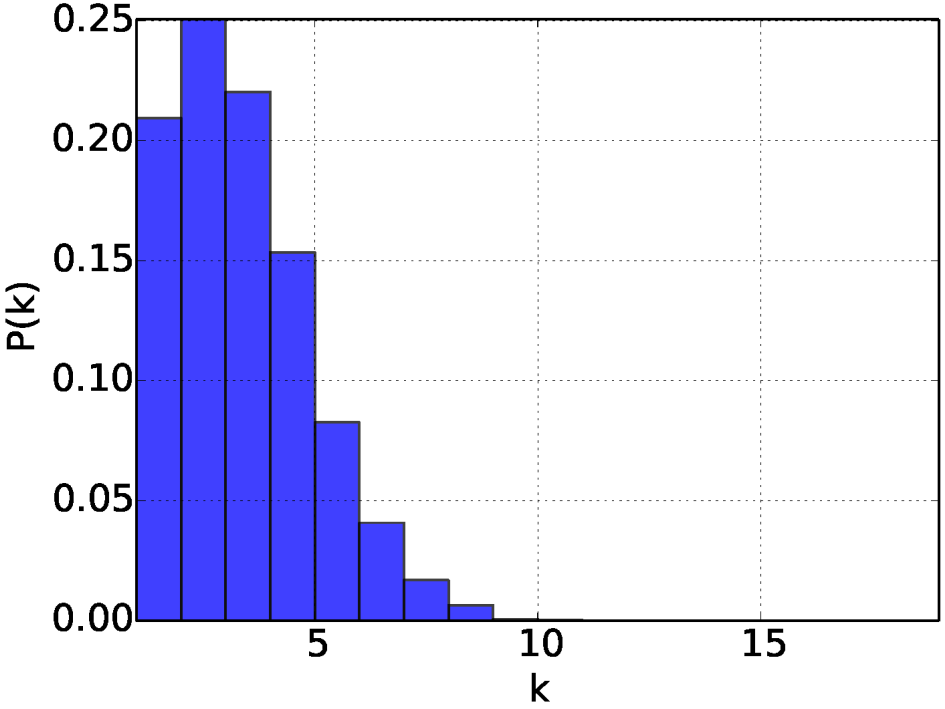
Notes:

- We now reconsider consider the power grid network of the western states in the United States, which we have introduced in lecture L01.
- How do the stylized facts in this network compare to a random network with the same number of nodes and links?

Example: random network

$\lvert V \rvert$	4943
$\lvert E \rvert$	6595
$\langle k \rangle$	2.67
$diam(G)$	19
$\langle l \rangle$	8.5
C	0
r	0





which properties of the power grid network can we expect "at random"?

Notes:

- Let us compare the power grid to a random network. The figure on the left shows a randomly generated network that has the same number of nodes and links as the power grid. This network is one particular microstate generated by the $G(n, m)$ model.
- We see that this microstate is not connected but we can extract a giant connected component that comprises the majority of nodes. All values (except for the mean degree) in the box above were calculated for this giant connected component. We make some interesting observations:
    1. First, the network topology visually looks different from the power grid topology.
    2. Just like the power grid, the random network is not assortative (so we may "expect" this at random)
    3. The degree distribution (shown in the histogram) looks almost indistinguishable from the power grid
    4. The diameter and average shortest path length are similar to those in some of the other examples, and certainly much smaller than in our power grid example. So the large observed values in the power grid network could be considered an interesting finding that is not expected at random in a network with that size.
- We currently lack a formal way to argue about which characteristics of a complex network can be explained by such a simple random model, which limits our ability to reason about interesting patterns. Addressing this issue, in the next lectures we use simple random graph models to establish baselines for the aggregate characterizations of networks.

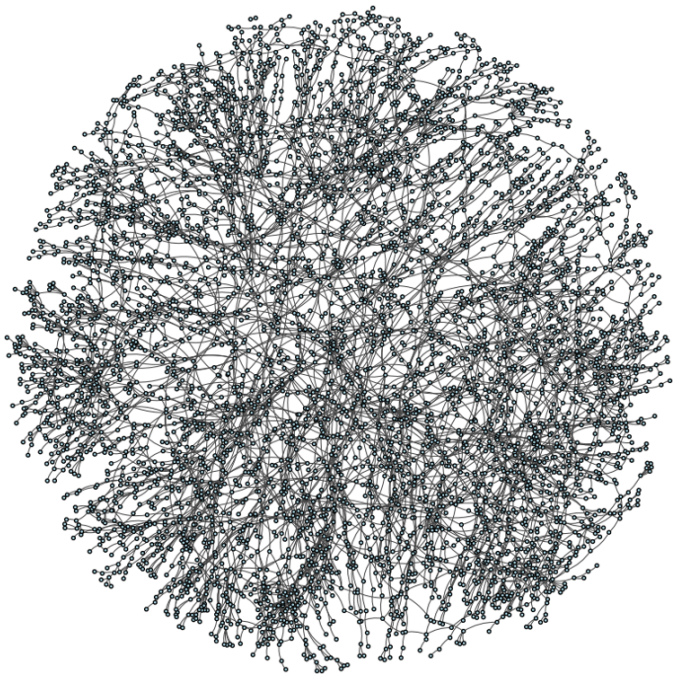
In summary

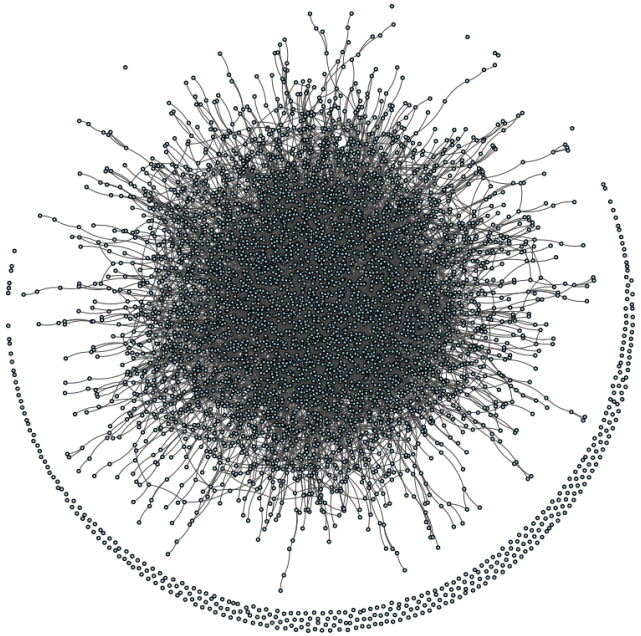
aggregate statistics of networks

- we can characterize networks based on **aggregate statistical quantities**
- **degree assortativity** captures preference of nodes to be connected to nodes with similar degrees
- **clustering coefficient** can be used to assess the fraction of closed triads around nodes

random network ensembles

- we introduced two simple **generative models** that define **random graph ensembles**
- basis to study **expected properties** in statistical ensembles of microstates consistent with a given **macrostate**
- important baseline for **statistical network analysis** and basis for **pattern recognition**





Notes:

- In summary, we have introduced methods that can be used for a macroscopic characterization of complex networks. The degree sequence and distribution are independent of topological details, such as which nodes are connected to each other (as long as the node degrees are preserved). Degree assortativity captures the preference of nodes to connect to other nodes with similar degrees. This is an other example for a measure that is independent of details of the network topology (apart from nodes with certain degrees being connected to each other).
- We can use those measures to identify stylized facts like, e.g., whether networks have broad or narrow degree distributions, whether they exhibit assortative or disassortative mixing, or whether triads are – on average – closed or not. Those facts can be used to reason about similarities between different networked system. However, we additionally need methods to decide which of those stylized facts can be expected at random.
- We introduced two simple random graph models, which can serve as (maximally) simple reference (or null) models for complex networks. We note that the random graph models introduced in this lecture only preserve the number of nodes and the (expected) number of links of a given network, while the topology of links is randomized.
- Despite their simplicity, these models allow us to study which of the characteristics of an empirical network we can already expect at random. These simple reference models help us to decide which statistical patterns are likely to be the result of more

Exercise sheet 02

- **second exercise sheet** is available on WueCampus
  - deepen your understanding of mixing patterns and clustering coefficient
  - take first steps in the analysis of random graphs
  - reason about expected properties of random graphs
- solutions are due by **November 16th** (via WueCampus)
- present your solution to earn bonus points

Statistical Network Analysis
WiSe 2021/2022

Prof. Dr. Ingo Scholtes
Chair of Informatics XV
University of Würzburg

Exercise Sheet 02

Published: November 09, 2021
Due: November 24, 2021
Total points: 10

Please upload your solutions to WueCampus as a scanned document (image format or pdf), a typesetted PDF document, and/or as a `jupyter` notebook.

1. **Mixing Patterns in Networks**
   (a) Prove that the degree assortativity coefficient corresponds to the fraction of the covariance $cov(d_i, d_j)$ of degrees of connected node pairs i, j and the variance $var(d_i)$ of the degree of nodes i, which are endpoints of randomly chosen edges. 3P
   Hint: Start your proof from a definition of the average degree of nodes i, which are endpoints of randomly chosen edges (i, j) in an undirected network.
   (b) Give examples for networks with positive, negative, and zero degree assortativity. Demonstrate the correctness of your solution in `pathpy`. 1P
   (c) Consider one of the empirical networks used in notebook 04-03 of the practice sessions[1]. Use the code from the practice session to generate a set of random microstates generated by the $G(n, p)$ random graph model, where the model parameters are fitted to the empirical network. Compute the degree assortativity for each microstate and plot the distribution of degree assortativity across all random microstates. Use this random graph ensemble to argue whether the degree assortativity that has been observed in the empirical network is significantly different from that in a corresponding random graph. 2P
2. **Random Graph Models**
   (a) Derive an expression for the degree distribution of a microstate generated by the $G(n, p)$ model for random undirected networks with self loops. 2P
   (b) In the lecture you have seen that the *triadic closure* of a network can be computed by averaging the local clustering coefficients of all nodes. Another, non-equivalent, way to compute the triadic closure of a network is given by the so called *global clustering coefficient*. This measure is defined as the fraction between the number T of closed triads and the total number of triads M (closed or not) in a network, i.e. $C = \frac{T}{M}$. 2P
   Consider a $G(n, p)$ network with mean degree $c = np$. Using the independence of edges in a $G(n, p)$ compute the expected global clustering coefficient in the limit of large n.

[1] see https://gitlab.informatik.uni-wuerzburg.de/ml4nets_notebooks/2021_wise_sna_notebooks

Notes:

Self-study questions

1. For an undirected network, define the degree distribution and the mean degree.
2. Define the degree assortativity coefficient. What do positive and negative values indicate, what does a zero value mean?
3. Calculate the local clustering coefficient for all nodes in the network shown on slide 6.
4. What is a stylised fact? Give an example for a stylized fact in a real-world network.
5. Write down an algorithmic formulation (in "pseudocode") of the network construction process in the $G(n, p)$ and the $G(n, m)$ model?
6. What is a statistical ensemble in the context of complex networks? What is a macrostate? What is a microstate?
7. What is the probability of the empty network in the statistical ensemble defined by the $G(n, m)$ model with $n = 5$ and $m = 10$.
8. Can you find a parameter of the $G(n, p)$ model such that all microstates in the resulting statistical ensemble have the same probability?

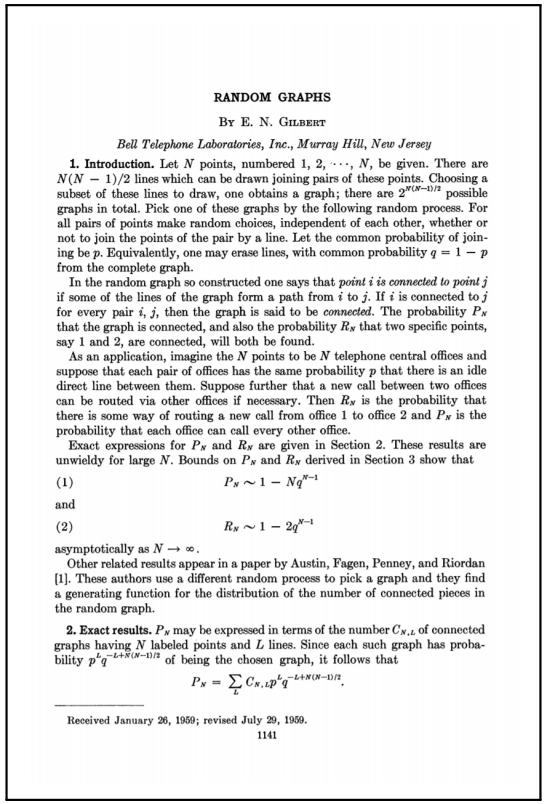
Notes:

References

reading list

- MEJ Newman: **Mixing patterns in networks**, Physical Review E, 2003
- G Simmel: **Soziologie. Untersuchungen über die Formen der Vergesellschaftung**, 1908
- PW Holland, S Leinhardt: **Transitivity in structural models of small groups**, Comparative Group Studies, 1971
- N Kaldor: **Capital Accumulation and Economic Growth**, The Theory of Capital, 1961
- P Erdős, A Rényi: **On Random Graphs. I**, Publicationes Mathematicae, 1959
- EN Gilbert: **Random Graphs**, Annals of Mathematical Statistics, 1959
- S Havlin and R Cohen: **Complex Networks: Structure, Robustness and Function**, Cambridge University Press, 2010

→ Chapter 2



RANDOM GRAPHS

By E. N. Gilbert

Bell Telephone Laboratories, Inc., Murray Hill, New Jersey

1. Introduction. Let N points, numbered $1, 2, \cdots, N$, be given. There are $N(N-1)/2$ lines which can be drawn joining pairs of these points. Choosing a subset of these lines to draw, one obtains a graph; there are $2^{N(N-1)/2}$ possible graphs in total. Pick one of these graphs by the following random process. For all pairs of points make random choices, independent of each other, whether or not to join the points of the pair by a line. Let the common probability of joining be p. Equivalently, one may erase lines, with common probability $q = 1 - p$ from the complete graph.

In the random graph so constructed one says that *point i is connected to point j* if some of the lines of the graph form a path from i to j. If i is connected to j for every pair i, j, then the graph is said to be *connected*. The probability P_N that the graph is connected, and also the probability R_N that two specific points, say 1 and 2, are connected, will both be found.

As an application, imagine the N points to be N telephone central offices and suppose that each pair of offices has the same probability p that there is an idle direct line between them. Suppose further that a new call between two offices can be routed via other offices if necessary. Then R_N is the probability that there is some way of routing a new call from office 1 to office 2 and P_N is the probability that each office can call every other office.

Exact expressions for P_N and R_N are given in Section 2. These results are unwieldy for large N. Bounds on P_N and R_N derived in Section 3 show that

$$P_N \sim 1 - Nq^{N-1} \tag{1}$$

and

$$R_N \sim 1 - 2q^{N-1} \tag{2}$$

asymptotically as $N \to \infty$.

Other related results appear in a paper by Austin, Fagen, Penney, and Riordan [1]. These authors use a different random process to pick a graph and they find a generating function for the distribution of the number of connected pieces in the random graph.

2. Exact results. P_N may be expressed in terms of the number $C_{N,L}$ of connected graphs having N labeled points and L lines. Since each such graph has probability $p^L q^{-L+N(N-1)/2}$ of being the chosen graph, it follows that

$$P_N = \sum_L C_{N,L} p^L q^{-L+N(N-1)/2}.$$

Received January 26, 1959; revised July 29, 1959.

1141

Notes: