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 05
Random Graphs and Small World Networks

November 16, 2022

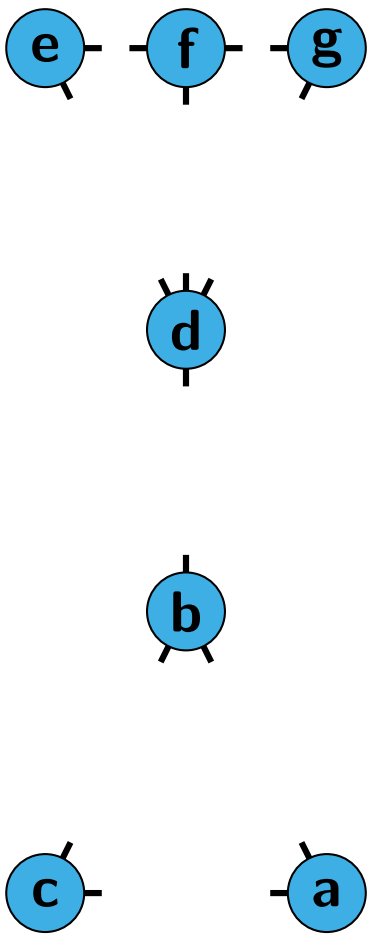
Notes:

- **Lecture L05:** Random Graphs and Small World Networks 16.11.2022
- In this lecture you will learn how we can use random graph models to compute expected properties of empirical networks. You will further see how the small-world property influences the navigability of networks, and how we can synthetically generate random networks that exhibit this property of real social networks.
  - Degree distribution of random graphs
  - Diameter and clustering coefficient in random graphs
  - Small-world networks
  - The Watts-Strogatz model

- **Exercise Sheet 03:** Random graphs and small-worlds due 23.11.2022

Statistical Ensembles

- we introduced **statistical ensembles of networks**
  - ensemble of all networks with given aggregate characteristics $\rightarrow$ **macrostate**
  - macrostate defines sample space of specific realizations $\rightarrow$ **microstates**
- we obtain probability space of possible microstates $\rightarrow$ **statistical ensemble**
- models **lack of knowledge** about topological details of microstate given a macrostate
- helps to reason about **expected properties** of empirical networks



$$V = \{a, b, c, d, e, f, g\}$$
$$S(G) = (2, 3, 2, 4, 2, 3, 2)$$

Notes:

- In the previous lecture, we introduced the **statistical ensemble perspective on networks**, which will keep us busy for three more weeks. Within this perspective, we fix aggregate characteristics like the number of nodes and links, the degree sequence or degree distribution (i.e. the macrostate) and define a probability space (statistical ensemble) that assigns a probability to each possible network realization (the microstate) that is consistent with that macrostate. This probability space can be seen as a simple model that captures our lack of knowledge about the details of a network, given that we only know aggregate statistics.
- We have considered two **random graph models** that fix the number of nodes n and the number of links m or, equivalently, the number of nodes n and the expected degree $\langle k \rangle$ of a node. These models can be used to answer the question which of the properties of a given data set we can expect "at random" given the size of a network and the "density" of connections. Such a random baseline is important if we want to argue which of a network's properties constitute a pattern and which other properties can be expected at random.
- We concluded the last lecture with the microstate probabilities of the $G(n, p)$ and $G(n, m)$ model, which we calculated based on a simple combinatorial argument of how many links can possibly exist in a network with n nodes and how many networks with a given number of links we can possibly have. Naturally, this depends on whether we consider directed or undirected networks with or without self-loops. To simplify matters, in the following we consider the case of undirected networks with self-loops, where the maximum number of links is $\binom{n+1}{2} = \frac{n(n+1)}{2}$.

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

$G(n, m)$ model

"urn model" of random graphs

- macrostate: n nodes and m links
- microstates: networks with m links
- $\langle k \rangle = \frac{2m}{n}$
- equiprobable microstates

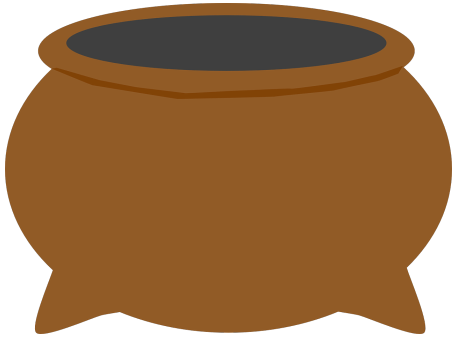
$$P(G) = \left(\binom{n+1}{2} \atop m \right)^{-1}$$

$G(n, p)$ model

"coin tossing model" of random graphs

- macrostate: n nodes and $\langle k \rangle = np$
- microstates: networks with expected number of $p \cdot \binom{n+1}{2}$ links
- microstates G can have **any** number of links m_G (for $p \in (0, 1)$)
- microstate probability of G depends on m_G

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

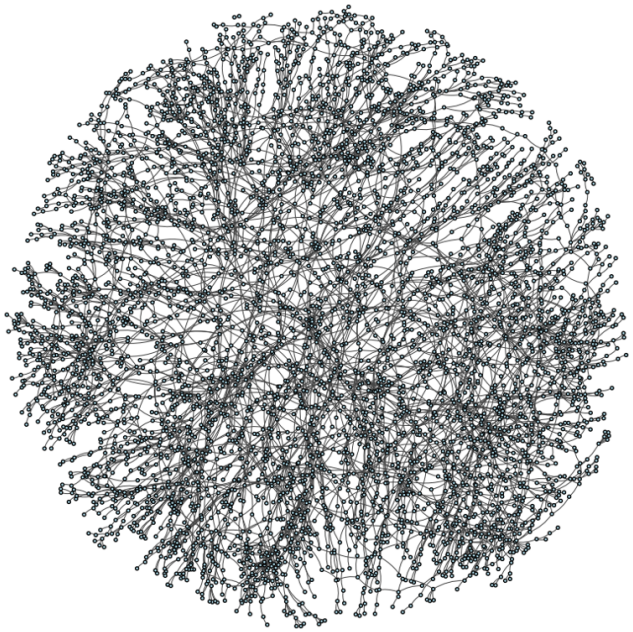


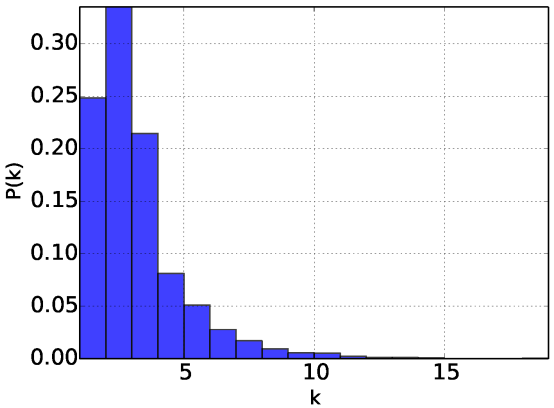
Notes:

- The two random graph models introduced in the previous lecture can be loosely associated with fundamental random experiments considered in probability theory.
- The $G(n, m)$ **model** can be viewed as the graph-theoretic equivalent of an urn model. We start with an empty network with n nodes and draw m links from the set of $\binom{n+1}{2}$ possible links uniformly at random. This model helps us to address the question which properties we can expect in a network, given that we only know the number of nodes and links. It produces equiprobable microstates with n nodes and exactly m links. Since each microstate has exactly n nodes and m links, all generated realizations have a mean degree of $\frac{2m}{n}$, which is also the expected degree of a randomly chosen node.
- The $G(n, p)$ **model** is the graph-theoretic equivalent of a repeated Bernoulli trial. I.e. we perform a coin tossing experiment, where we toss a coin for each pair of nodes and draw an edge with probability p. This model produces microstates with a fixed expected degree of nodes but with variable numbers of edges.

Expected properties?

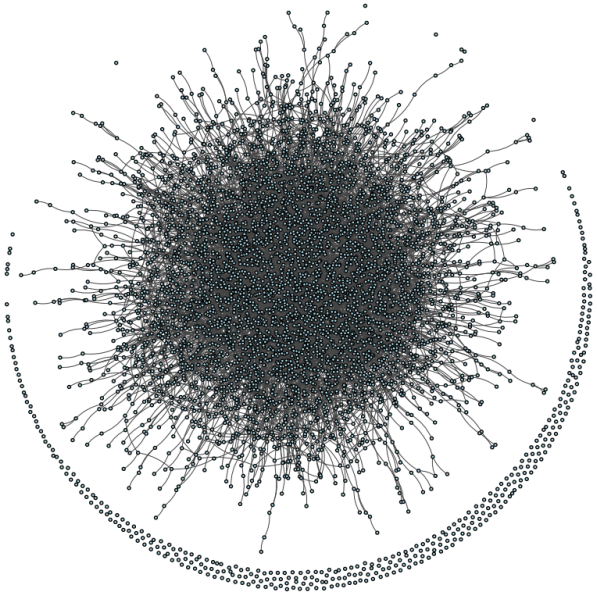
empirical network of transformer stations in US power grid

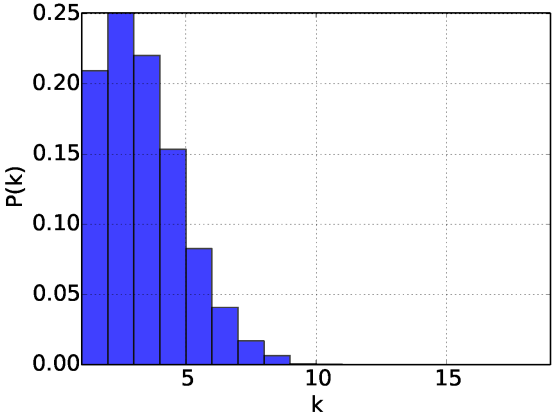




empirical degree distribution

random network with same number of nodes and links





random degree distribution

Notes:

- We ended the previous lecture with an empirical example, the power grid network of the western states in the United States (see lecture L01). For this network, we have reported a number of stylized facts, which included a plot of the degree distribution. We also used a random graph model to generate a random network with the same aggregate statistics, here the number of nodes and links.
- For the resulting random network, the topology, i.e. who is connected to whom and which nodes are connected to how many other nodes, is generated by a simple stochastic model that has no concept of power stations or power lines whatsoever. Nevertheless, we see that we recover some of the stylized facts that we observed in the empirical network, which is probably created by much more complicated processes.
- Which of the empirical network's properties can be explained by a simple random graph model? To answer this question, we first study the expected degree distribution of random networks.

$G(n, p)$ model: degree distribution

- what is the **degree distribution** of a random microstate $G = (V, E)$?
- for $v \in V$ chosen uniformly at random define **discrete random variable** $X : V \to \mathbb{N}^0$ with $X := d_v$
- to calculate probability mass function $P : \mathbb{N}^0 \to [0, 1]$ consider generative process
  - for node v we perform n **Bernoulli trials** with success probability p (i.e. incl. self-loops)
  - probability of exactly k successes given by **binomial distribution** $\mathrm{Binom}(n, k)$
- **degree distribution** of random microstate is

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

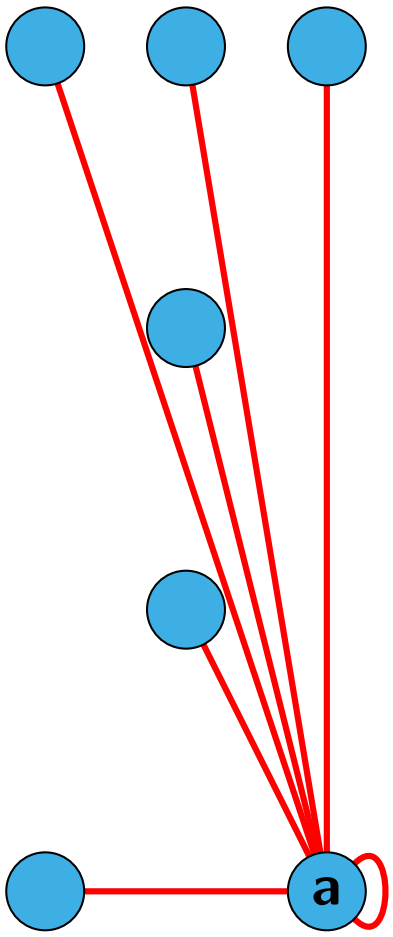


illustration of Bernoulli trials performed for node a in $G(n, p)$ model

Notes:

- Simple combinatorial arguments, and the interpretation of the $G(n, p)$ model as a repeated Bernoulli trial, allow us to calculate the **degree distribution of the microstates generated by the $G(n, p)$ model**. Given parameters n and p we calculate the probability $P(k)$ that a random node in a random microstate has degree k.
- For each of the n nodes, we perform n (with self-loops!) Bernoulli trials, where each trial yields a link with probability p. What is the probability to have exactly k successes in a sequence of n independent Bernoulli trials that succeed with probability p? This probability is given by the Binomial distribution, where the binomial coefficient captures the number of different ways how k successes can occur within the n trials, while the second and third factor capture the probability of k successes and $n - k$ "failures" respectively.
- Note that we obtain the Binomial distribution because we do not differentiate between different microstates, as long as we have k successes. I.e. we do not care to which nodes the k links are connected. This is why (different from the microstate probability above) we must include the binomial factor, which counts in how many different ways a node in a network with n can be connected to k nodes.

First raw moment of $P(k)$

- **moments of degree distributions** capture important properties of networks

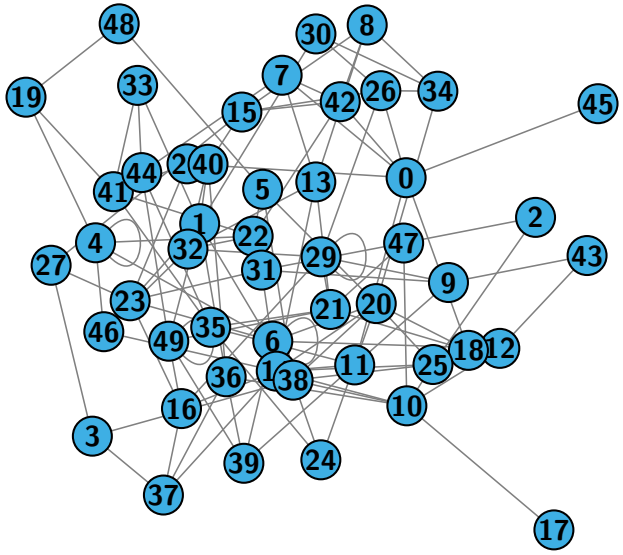
- ***m*-th raw moment** of discrete random variable $X \in \mathbb{N}^0$ with probability mass function P is

$$\langle X^m \rangle := E[X^m] = \sum_{k=0}^{\infty} k^m P(X = k)$$

- first raw moment of $P(k)$ is the **mean degree**

$$\langle k \rangle := \sum_{k=0}^{\infty} k \cdot P(k)$$

- for **degrees k of random nodes in $G(n, p)$ microstates** we have $\langle k \rangle = np$



microstate of $G(n, p)$ model,
$n = 50$, $p = 0.08$
$\langle k \rangle = np = 4$

Notes:

- Moments are important parameters of statistical distributions and we can use them define meaningful quantities like, e.g. the mean, the variance, kurtosis, or skewness.
- The **raw moments of a discrete probability distribution** are a straight-forward generalization of the mean that are calculated based on the powers of a random variable. For a probability mass function $P(X)$, the kth-raw moment is simply defined as the expected value of the random variable X^k. This implies that the first raw moment ($k = 1$) of a distribution is simply the expected value of X.
- For a distribution $P(X = k)$, where X assumes the degree of a randomly chosen node, the **first raw moment** is thus the expected degree of a randomly chosen node.
- For the binomial distribution, the first raw moment is the expected number of successes in n trials with success probability p, i.e. for the $G(n, p)$ model we have a first raw moment of np, which is also the expected degree of nodes in a microstate.

Second raw moment of $P(k)$

- **second raw moment of degree distribution** $P(k)$ is the mean of squared degrees

$$\langle k^2 \rangle := \sum_{k=0}^{\infty} k^2 \cdot P(k)$$

- **MOSSOM mnemonic**: variance of distribution is

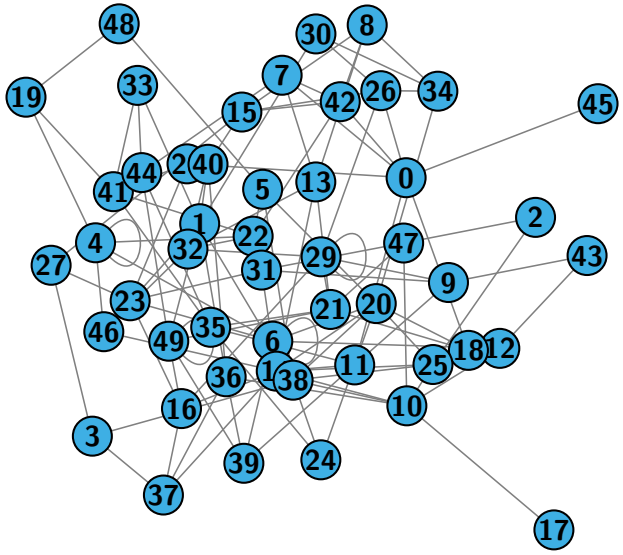
$$\mathrm{Var}(X) := \langle k^2 \rangle - \langle k \rangle^2$$

- for **degrees k of random nodes in $G(n,p)$ microstates** we have

$$\langle k^2 \rangle = np(1-p+np)$$

and thus

$$Var(X) = np(1-p+np) - n^2p^2 = np(1-p)$$



microstate of $G(n,p)$ model
with $n = 50$, $p = 0.08$
$\langle k^2 \rangle = np(1-p+np) = 19.68$
$Var(k) = np(1-p) = 15.26$

Notes:

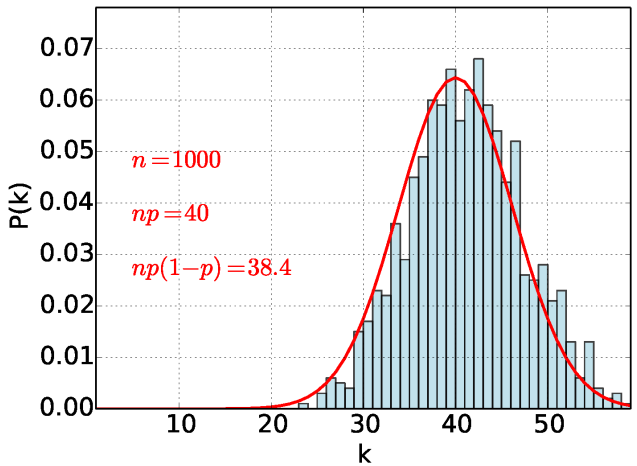
- The second raw moment of a probability mass function $P(X)$ is the expected value of the random variable X^2, which is related but not identical to the variance of the distribution. The variance is the expected value of the squared difference from the mean, i.e. the first raw moment of the distribution. This leads to the fact that the variance $Var(X)$ is related to the second raw moment $\langle k^2 \rangle$ via the relation $\mathrm{Var}(X) = \langle k^2 \rangle - \langle k \rangle^2$. You can remember this relation by the mnemonic: "mean of square minus square of mean" (MOSSOM).
- For the $G(n, p)$ model, we can calculate the second raw moment of the degree distribution based on the binomial distribution. We obtain $\langle np(1 - p + np) \rangle$. With this, for the variance of the degree distribution of the $G(n, p)$ model we have $Var(X) = np(1 - p)$.
- As we will see in the following lectures, the ratio between the first and the second raw moment of a degree distribution has important and surprising consequences for the expected properties of a random network. Intuitively it captures the density of links in a network as well as how heterogeneous the degree distribution is. Interestingly, we will show that this is all we have to know about a degree distribution in order to calculate the expected diameter, the connectedness, etc. of a random network.

$G(n, p)$ model: limiting degree distribution

- for given number of n nodes, degree distribution of $G(n, p)$ microstate is

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- what if n is large?



De Moivre-Laplace Theorem

for $n \to \infty$ the degree distribution of random $G(n, p)$ microstates converges to

$$P(k) \sim \mathcal{N}(\mu, \sigma^2) = \mathcal{N}(np, np(1-p))$$

i.e. **normal distribution** with

$$P(k) = \frac{1}{\sqrt{np(1-p)}\sqrt{2\pi}} e^{-\frac{(k-np)^2}{2np(1-p)}}$$

central limit theorem for node degrees in random graphs

example

degree distribution of random microstate generated by $G(n, p)$ model with $n = 1000$ and $p = 0.04$

Notes:

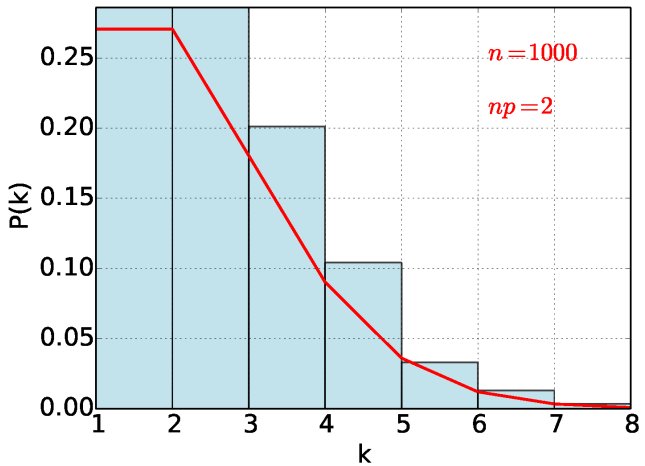
- The ensemble perspective on networks is particularly useful when we want to reason about networks with a large number of nodes. It is thus helpful to study the **limiting degree distribution** as the number of nodes n goes to infinity. Let us consider this for the $G(n, p)$ model: For any network of finite size n, the degree distribution is a binomial distribution but can we approximate this for large n?
- For $n \to \infty$ we can use the *De Moive-Laplace theorem*, which follows as a special case from the *central limit theorem*. It states that for $n \to \infty$ the binomial distribution with success probability p converges to a normal distribution with mean np and variance $np(1-p)$. We confirm this in the example network shown on the right. We use the $G(n, p)$ model to generate a microstate with $n = 1000$ nodes and a link probability of $p = 0.04$ and plot the degree distribution of the microstate. The red curve shows a normal distribution with $np = \mu = 40$ and $np(1-p) = \sigma = 38.4$. Note that the normal distribution has a continuous domain, i.e. we have a probability distribution function rather than a probability mass function.
- The normal distribution closely approximates the degree distribution of a random network even for moderately large networks with a few hundred or a few thousand nodes. This tells us that we should not be surprised if we find a normal-shaped degree distribution in a large network. We expect this at random and all we need to explain it is a random link formation process. If we find a distribution that does not follow this expected shape, we need a different explanation. I.e. we consider this a pattern that we do not expect (at least not for the random baseline of a $G(n, p)$ model) and that justifies further investigation.

$G(n, p)$ model: sparse infinite networks

- degree distribution of $G(n, p)$ model

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- what if n is large and $\langle k \rangle = np$ **is small**



Poisson limit theorem

for $n \to \infty$ and $p \to 0$ the degree distribution of random $G(n, p)$ microstates converges to

$$P(k) \sim \frac{\lambda^k e^{-\lambda}}{k!}$$

i.e. **Poisson distribution** with $\lambda = \langle k \rangle = np$

example

degree distribution of random microstate generated by $G(n, p)$ model with $n = 1000$ and $p = 0.002$

- $p \to 0 \Rightarrow np \equiv$ const, i.e. **sparse** network

Notes:

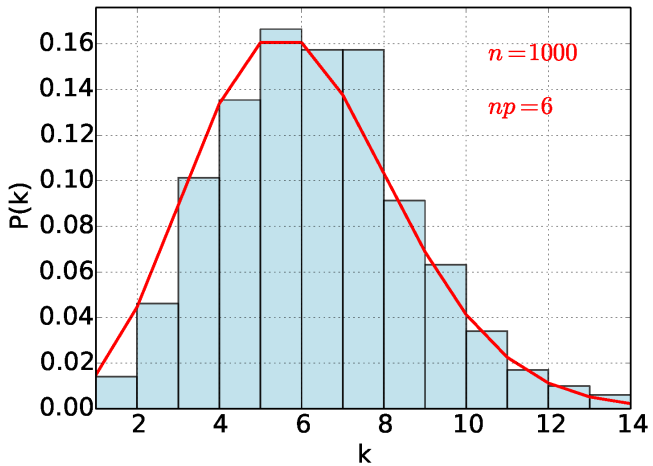
- The approximation of the degree distribution by a normal distribution assumes that the number of nodes n goes to infinity while the link probability p remains constant. If n goes to infinity, the expected degree np goes to infinity as well, i.e. the generated networks are becoming increasingly dense.
- We can instead consider **sparsely connected infinite random networks** with $n \to \infty$ but with a fixed expected degree $np \equiv const \ll n$. For $n \to \infty$ this implies $p \to 0$.
- Under these assumptions, the **Poisson limit theorem** states that the limiting degree distribution is a Poisson distribution with parameter $\lambda = np$. Hence, the degree distribution of sparse networks with a large number of nodes can be approximated by a Poisson distribution. We demonstrate this in the example shown on the right, which shows the degree distribution of a random microstate from the $G(n, p)$ model with $n = 1000$ and $np = 2$, i.e. $p = 0.002$. The red curve is the probability mass function of the limiting Poisson distribution. The limiting distribution closely approximates the empirical degree distribution, even for a moderately large network with one thousand nodes.
- Using Stirling's approximation $k! \approx \sqrt{2\pi k}\frac{k}{e}^k$ we easily see that the tail of a Poisson degree distribution decays faster than exponential, i.e. it is extremely unlikely to observe a node with degrees that are much larger than the expected degree np. This tells us that we do not expect to find "outliers" in terms of the degree in microstates that are generated using the $G(n, p)$ model.

$G(n, p)$ model: sparse infinite networks

- degree distribution of $G(n, p)$ model

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- what if n is large and $\langle k \rangle = np$ **is small**



Poisson limit theorem

for $n \to \infty$ and $p \to 0$ the degree distribution of random $G(n, p)$ microstates converges to

$$P(k) \sim \frac{\lambda^k e^{-\lambda}}{k!}$$

i.e. **Poisson distribution** with $\lambda = \langle k \rangle = np$

example

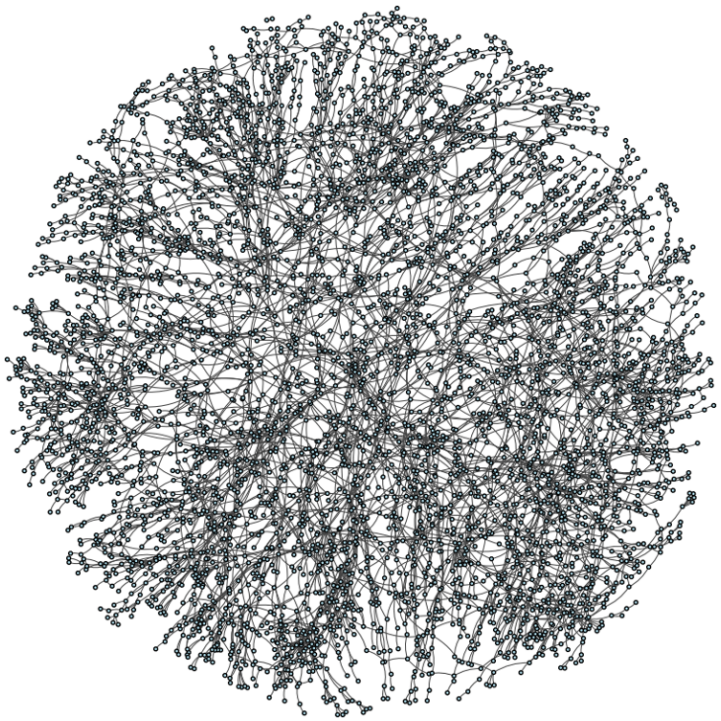
degree distribution of random microstate generated by $G(n, p)$ model with $n = 1000$ and $p = 0.006$

- $p \to 0 \Rightarrow np \equiv$ const, i.e. **sparse** network

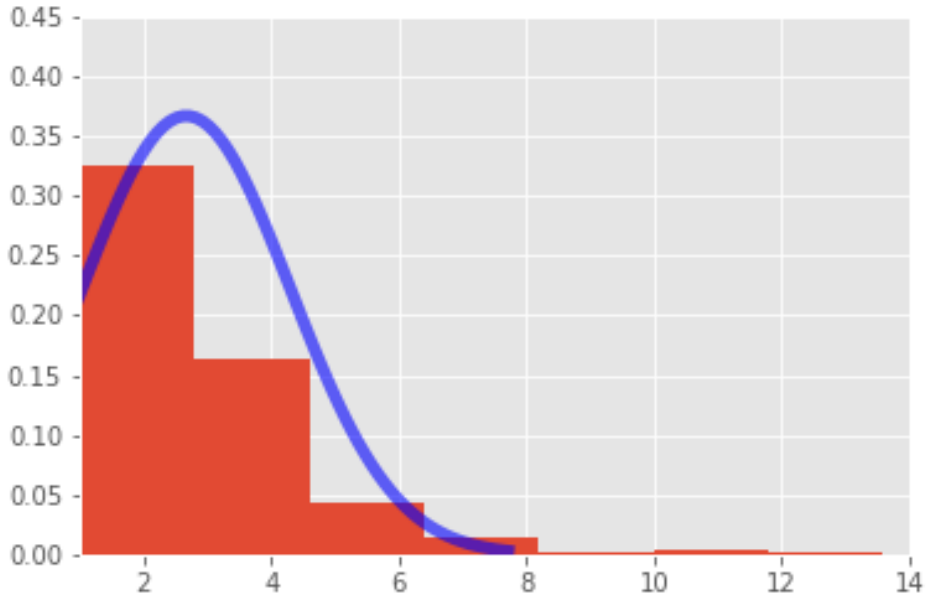
Notes:

- The example on the right shows the degree distribution of a $G(n, p)$ microstate with $n = 1000$ and $p = 0.006$, as well as the probability mass function of the Poisson distribution with $\lambda = np = 6$.
- This shows that, as we increase the expected degree np, for sufficiently large n the Poissonian degree distribution becomes increasingly similar to a Normal distribution.

Empirical 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**
r	**0**
C	**0.11**

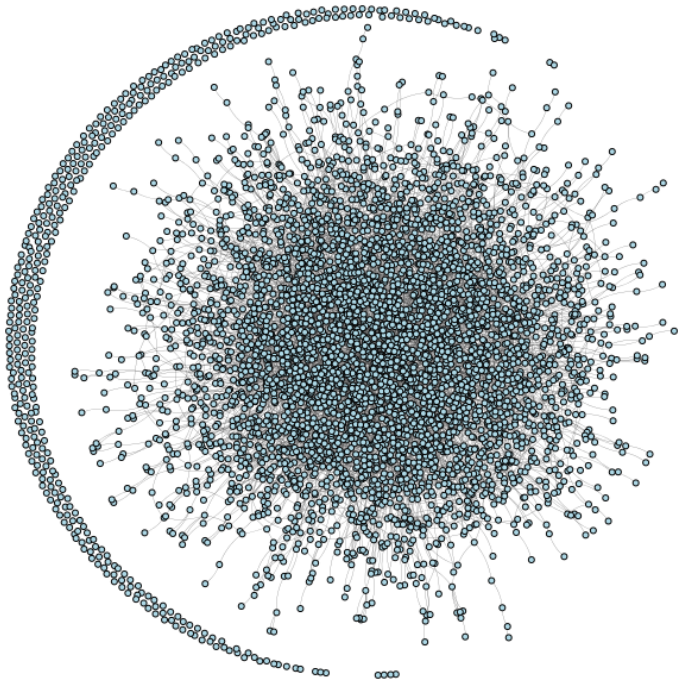


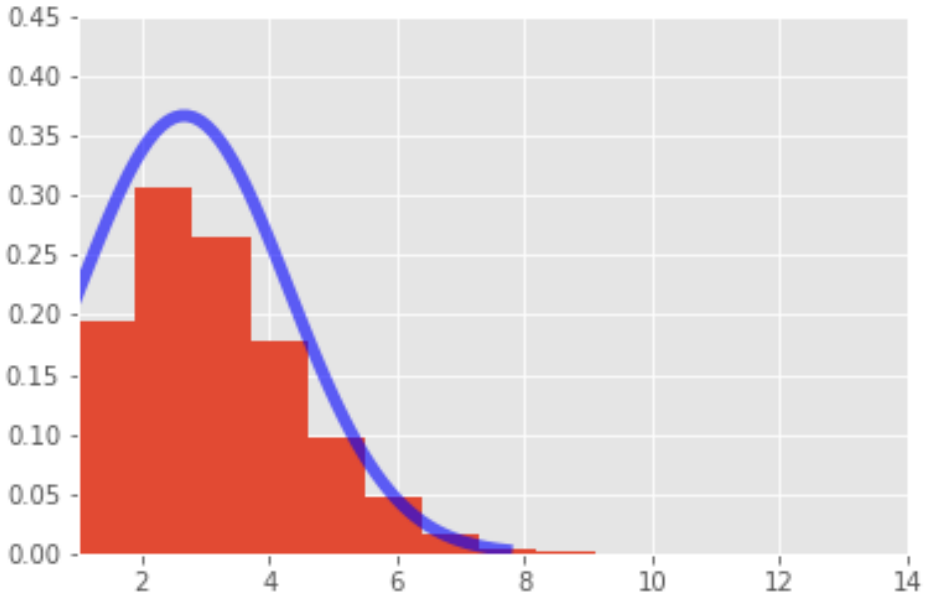
Notes:

- For the power grid network, the shape of the empirical degree distribution resembles the Poissonian shape of a random graph with the same expected degree. The blue line is the probability mass function of the expected Poisson distribution with $\lambda = 2.67$. In the spirit of Occam's razor (which emphasizes the role of **parsimonious** explanations in science), this suggest that it is not justified to consider a more complex model than a simple $G(n, p)$ random graph model – at least not if we only want to explain the observed degree distribution of the power grid.
- But what about other aggregate quantities in the empirical network, like, e.g., its diameter, average shortest path length, clustering coefficient, or degree assortativity)? Can we explain those by a random $G(n, p)$ or $G(n, m)$ model as well? Which of the observed stylized facts deserve further investigation? Which of them constitute a "pattern" and which others are expected in a random network of that size?
- For degree assortativity, we recall that it is the Pearson correlation coefficient of the degrees of nodes connected by links. Since we randomly connect pairs of nodes, we expect the Pearson correlation coefficient to be zero, which is in line with value observed in the real network.
- But what about the other quantities? What do we expect at random?

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

$\lvert V \rvert$	4943
$\lvert E \rvert$	6595
$\langle k \rangle$	2.67
$diam(G)$	**19** (gcc)
$\langle l \rangle$	**8.47** (gcc)
r	**0**
C	**0.00054**





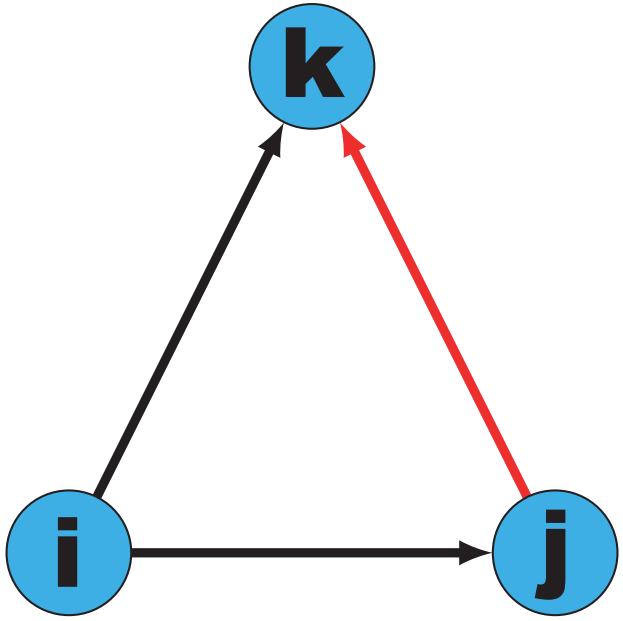
Notes:

- To reason about the expected properties of random networks we can use the $G(n, p)$ and the $G(n, m)$ models to generate a sample of realizations that preserve the number of nodes and the (expected) number of links. We can calculate the mean (or the distribution) of a property of interest in the random sample and then compare this to the empirical network. In the previous practice session, we did this for the clustering coefficient. This approach is one basis for statistical hypothesis testing in networks. In particular, for a given test statistic we can calculate p-values for null hypotheses, i.e. the probability to observe in a random network a value for the test statistic at least as extreme as the value that we observed in an empirical network. (Remember that the p-value is **not** the probability that the null hypothesis is true!)
- In the example above we generate one microstate of the resulting ensemble. Note that – different from the real system – in the random microstate we have multiple connected components. We thus compute the diameter and the average shortest path lengths for the largest connected component only. We will study the size of the largest connected component in random networks in more detail in one of the the next lectures. We see that at random, we expect a much smaller diameter and average shortest path length. The zero degree assortativity coefficient is consistent with that of the empirical network while the clustering coefficient of random microstates is much smaller than that of the real network.
- Rather than generating a large number of microstates to obtain estimates for expected properties, it is preferable to **analytically calculate expected properties**, which in the following we do for diameter and clustering coefficient.

Random graphs: clustering coefficient

- **clustering coefficient** C can be interpreted as conditional probability $P(A|B)$

$$P\left[(j,k) \in E | (i,j) \in E, (i,k) \in E\right]$$



- in $G(n,p)$ model **links are created independently**, i.e.

$$P\left[(j,k) \in E | (i,j) \in E, (i,k) \in E\right]$$

$$= P\left[(j,k) \in E\right] = p$$

- for microstates of $G(n,p)$, for sufficiently large n **we expect** $C \approx p$

Notes:

- In the last lecture, we have defined the local clustering coefficient of nodes, which captures the fraction of triads around a node that is closed. We further introduced the **average or global clustering coefficient**, which is defined as the arithmetic mean of the local clustering coefficients of all nodes. To see what clustering coefficient we can expect for a random microstate in the ensemble of networks generated by the $G(n, p)$ model, we interpret the average clustering coefficient as conditional probability of a link (j, k), conditional on the fact that two links (i, j) and (i, k) exist.
- With this interpretation of the clustering coefficient, we can now calculate the **expected clustering coefficient** in a random microstate generated by the $G(n, p)$ model. For two independent events A and B, we have $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$. Hence, for the $G(n, p)$ model (where links are generated independently) the (conditional) probability of any link (i, k) is simply p (because it is independent of the existence of the other two links).
- Due to the law of large numbers, for sufficiently large n, we thus expect the fraction of closed triads that is generated by the $G(n, p)$ model to converge to p. This holds if we fix the parameter p in the $G(n, p)$ model.

Clustering coefficient: large random networks

- for networks with n nodes and m links, we have

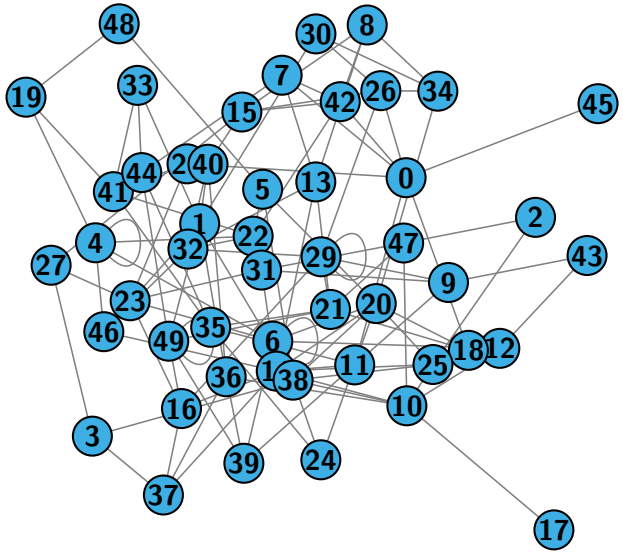
$$\langle k \rangle = \frac{2m}{n}$$

- for the $G(n, p)$ model:

$$\langle k \rangle = np \Leftrightarrow p = \frac{\langle k \rangle}{n} = \frac{2m}{n^2}$$

- for **sparse networks** with $\langle k \rangle = np \equiv const$

$$C \to 0 \qquad (n \to \infty)$$



example

microstate of $G(n, p)$ model with $n = 50$, $p = 0.08$

$C \approx 0.09$

Notes:

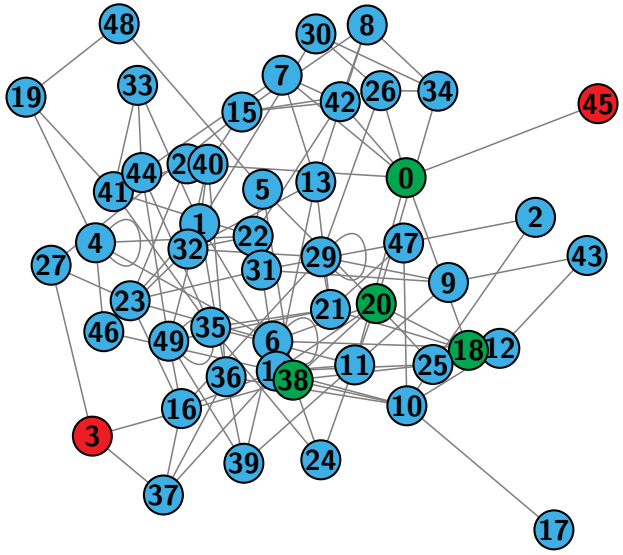
- What if we observe an **empirical network with n nodes and a given number of links m**? We can calculate the mean degree of the network as $\frac{2m}{n}$. We further know that the expected degree for a random node in the $G(n, p)$ model is np. This allows us to calculate the parameter $p = \frac{2m}{n^2}$ for which the $G(n, p)$ model generates networks with an expected number of m links.
- With this we have actually **fitted the parameters of the model to the empirical network**. This ensures that the expected number of links in a random microstate matches the observed number of links, which in turn maximizes the likelihood of the model (more on this in a later chapter on *statistical inference*).
- We find a link probability of $p = \frac{2m}{n^2}$ which implies that, in the limit of infinite n, the clustering coefficient of random networks with finite mean degree (i.e. sparse networks) converges to zero.
- We can apply this to the power grid network, which exhibits a clustering coefficient of $C = 0.11$. We first calculate the parameter $p = 0.00054$ for which a $G(n, p)$ model reproduces the observed mean degree of 2.67. For such a model we expect a clustering coefficient of $C = p = 0.00054$ (which is exactly the clustering coefficient of the random microstate shown on slide 10).
- We find that the clustering coefficient of the power grid network is more than 200 times larger than expected at random (given the number of nodes and links of the network).

Diameter of random graphs

- we can **approximate the expected diameter** of large sparse random graphs
- **consider randomly chosen node** v
  - v has expected number of $\langle k \rangle$ neighbors
  - each neighbor n of v has expected number of $\langle k \rangle$ neighbors → really?!? see Lecture 07
  - how many neighbors of n are neighbors of v?
  - for sufficiently **large sparse networks**: none since $C \to 0$ $(n \to \infty)$
  - we expect $\Rightarrow \langle k \rangle^l$ nodes at distance l to v
- number of nodes w at $dist(v, w) \leq D$

$$\sum_{l=0}^{D} \langle k \rangle^l = \frac{\langle k \rangle^{D+1} - 1}{\langle k \rangle - 1} \propto \langle k \rangle^D$$

- diameter $D \Leftrightarrow \langle k \rangle^D \approx n$ and thus $D \approx \frac{\log n}{\log \langle k \rangle}$



example

microstate of $G(n, p)$ model with $n = 50$, $p = 0.08$

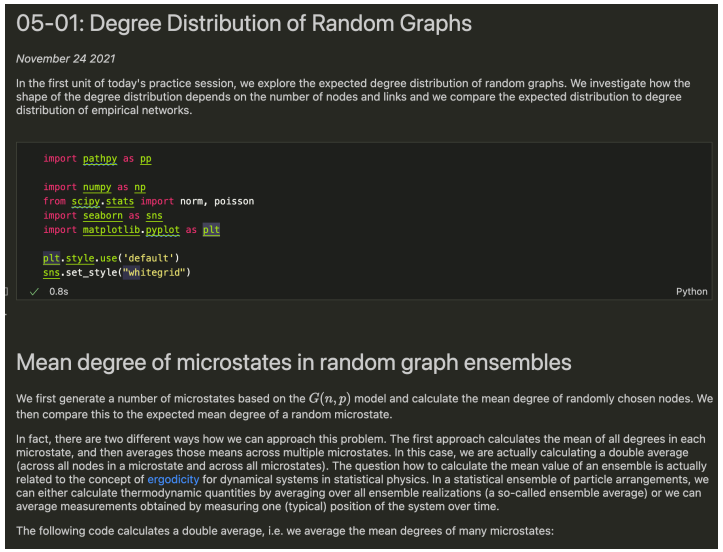
diam $= 5$, $\frac{\log n}{\log np} = 2.82$

Notes:

- We next approximate the expected diameter in random microstates generated by the $G(n, p)$ model. The idea is to pick a random node v in the network and ask how many nodes we can expect to be at a given distance l from v. For $l = 1$ this is easy to answer, because the number of expected neighbors is $\langle k \rangle$. To calculate this number for $l = 2$ we can recursively apply this to each of the $\langle k \rangle$ neighbors of node v. How many neighbors does a a neighbor of v have? This is a non-trivial question because the neighbors of v were not chosen uniformly at random. Moreover, we want to count nodes at distance $l = 2$ i.e. we should exclude the neighbor of node v (which is at distance $l = 1$). Finally, there could even be a link to node v (at distance 0), which we must account for.
- Regarding the last point, a link from a neighbor x of a neighbor w of v back to v would lead to a closed triad (v, w), (w, x), (v, x) around node w. We know that the probability of such a closed triad goes to zero if the network is sufficiently large and the link probability p is fixed, so we can neglect this case in large sparse networks.
- The first point is more difficult to address and we will develop a formal argument for this next week. For now, let me say that for a $G(n, p)$ network the number of neighbors of a node w that is a neighbor of v is simply $\langle k \rangle$, despite the fact that w was not chosen uniformly at random! Through the recursive application of this argument we count $\langle k \rangle^l$ nodes at distance l from node v. We sum this number across all path lengths to get the number of nodes at a distance smaller or equal than l. If this number equals the size of the network, we have found the diameter D, i.e. we solve the equation by D and obtain the approximate expected diameter of a random network.

Practice session

- we use `pathpy` to calculate the **degree distribution, diameter, and clustering coefficients** of random graphs
- we explore the dependence of those properties on the **model parameters**
- we compare **expected properties** to those of empirical networks



05-01: Degree Distribution of Random Graphs

November 24 2021

In the first unit of today's practice session, we explore the expected degree distribution of random graphs. We investigate how the shape of the degree distribution depends on the number of nodes and links and we compare the expected distribution to degree distribution of empirical networks.

```python
import pathpy as pp

import numpy as np
from scipy.stats import norm, poisson
import seaborn as sns
import matplotlib.pyplot as plt

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

✓ 0.8s Python

Mean degree of microstates in random graph ensembles

We first generate a number of microstates based on the $G(n,p)$ model and calculate the mean degree of randomly chosen nodes. We then compare this to the expected mean degree of a random microstate.

In fact, there are two different ways how we can approach this problem. The first approach calculates the mean of all degrees in each microstate, and then averages those means across multiple microstates. In this case, we are actually calculating a double average (across all nodes in a microstate and across all microstates). The question how to calculate the mean value of an ensemble is actually related to the concept of ergodicity for dynamical systems in statistical physics. In a statistical ensemble of particle arrangements, we can either calculate thermodynamic quantities by averaging over all ensemble realizations (a so-called ensemble average) or we can average measurements obtained by measuring one (typical) position of the system over time.

The following code calculates a double average, i.e. we average the mean degrees of many microstates:

practice session

see notebooks 05-01 – 05-02 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 use the implementations of random graph models introduced last week to study their expected properties. We further compare those properties to the properties observed in real networks.

The “small world” effect in social networks

- at random, we expect social networks to have a **small diameter**

  → “six degrees of separation”

Milgram’s small world experiment

- chose 160 random people in Omaha, NE, USA
- asked to forward letter to stockbroker in Boston, MA
- forward to an acquaintance known on first-name basis

result and implications

44 letters reached the target, taking **5.5 steps on average**

- social networks have small diameter
- social networks are **navigable**

Frigyes Karinthy

1887 – 1938

Stanley Milgram

1933 – 1984

image credit: left: public domain, right: Wikimedia Commons, Harvard Department of Psychology, fair use, bottom: user Ageev Andrew, Wikimedia Commons, CC-BY-SA

Notes:

- The small expected diameter of random networks brings us to an interesting experiment conducted in the 1960s, which investigated the **small diameter of social networks**. In the context of social networks, you may have heard of the “six degrees of separation”. This refers to the hypothesis that every person is at most six hops away from any other person on the planet. This idea was used by Hungarian author Frigyes Karinthy, who used it in the 1929 play “chains” long before random graph theory and network science. Interestingly, it was was later confirmed by empirical studies.
- Such a confirmation of the small diameter would not be surprising, since we expect the diameter of a random network to scale as a logarithm of the number of nodes. However, in a famous experiment Stanley Milgram also showed that people are able to navigate the social network, i.e. they are able to find short paths without knowledge of the whole network.
- For this, he chose 160 people in Omaha, Nebraska, asking them to forward a letter to a stock broker in Boston exclusively by passing on the letter to acquaintances known on a first-name basis. 44 letters reached the target person after an average of 5.5 steps in the social network.

Navigability of social networks

- **navigability** is an important characteristic of social networks
- how can we navigate social networks with **local knowledge**?

small-world experiment

conduct a thought experiment in which you are a person receiving a letter in Milgram's small-world experiment

- target person is **Adele**
- **whom** of your friends would you chose and **why**?

Adele

born 1988

image credit: marcen27 - Flickr, CC BY 2.0

Notes:

- The navigability of social networks exposed by Milgram's experiment is an interesting and important feature, because it influences how individuals utilise their social network to efficiently find (or spread) information.
- To better understand this, we make a thought experiment, assuming that we are one of the participants in Milgram's experiment. We further assume that the target person to whom we should forward the letter to Adele. You have to forward a letter to one of your friends, known to you on a first-name basis. What would be your strategy of choosing the *optimal* friend to forward the letter to?

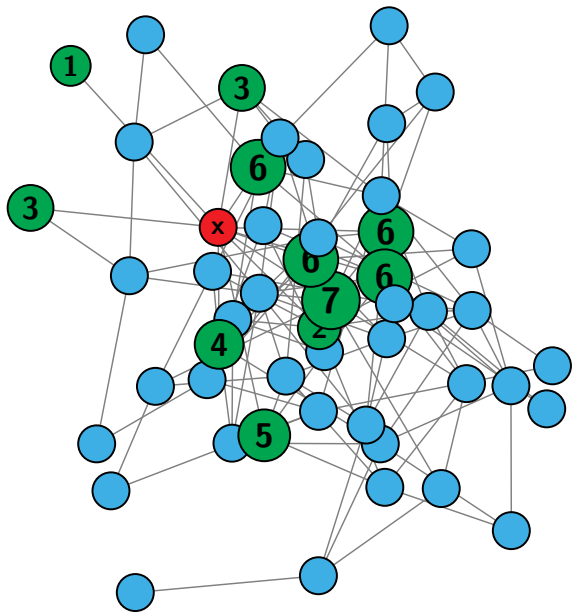
Greedy routing in networks

- we can perform **greedy routing** based on **local heuristic** computed for all neighbours
- example: **geographic distance** to target
- forward message to neighbor that maximizes/minimizes heuristic
- can you think of **topological heuristics**?

shortest path heuristic

- estimate **number of shortest paths** between you and all other nodes that pass through each of your friends
- choose the friend with maximum estimate

- in random networks shortest path heuristic does not yield a clear choice



example: random graph

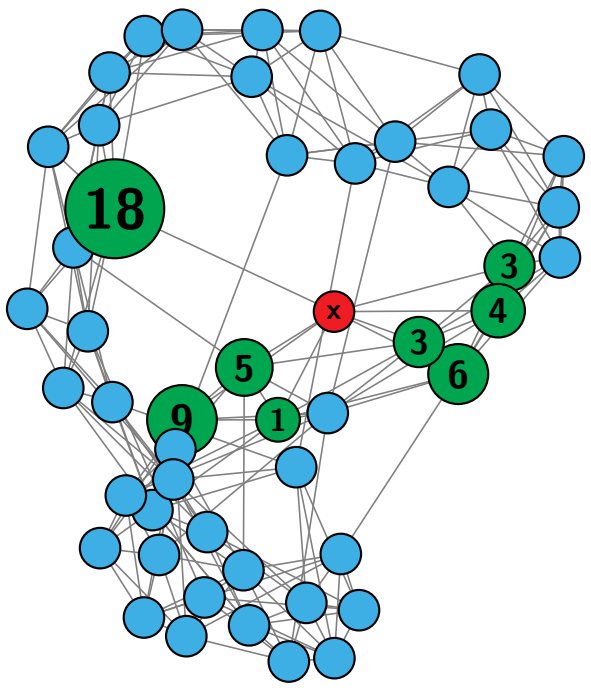
number of shortest paths from node x (red) to all other nodes (blue) passing to all neighbours (green) of node x in a random graph generated using $G(n, p)$ model with $n = 100$, $p = 0.05$

Notes:

- This can be phrased as a problem of routing with local knowledge. A simple strategy is greedy routing, i.e. in each step we chose a neighbor who maximizes or minimizes a heuristic. We could, for instance use geographic information, e.g. sending the letter to a friend living in Great Britain, closest to Adele. What if we do not have any additional information (like the position) available? We can still use topological heuristics to identify an optimal neighbor.
- As an example, we could try to estimate the number of shortest paths to any other node that pass through each of your friends. The friend with the highest number has then the highest chance to be on the (unknown) shortest path between you and the target person. In your own social network, you may have a gut feeling about this number for your friends, because you may know whom of your friends are "well-connected".
- We can quantify this property of your friends in an example network with 100 nodes. A random source node x (i.e. YOU) is colored in red while the neighbors of x (i.e. your friends) are colored green. The numbers indicate the number of shortest paths from you to any other node that pass through a particular friend. We observe some variation among those friends but in a random graph these numbers are generally rather similar.

Path funneling in social networks

- we know that random networks have small diameter but small clustering coefficient
- many social networks have **high clustering coefficient** and **small diameter**
- "funneling": majority of shortest paths to other nodes pass through **few friends**
  - good choices to forward the letter to
  - we often know intuitively who these friends are
- choose a friend whose friends have **minimal overlap** with your own friends!



example: path funneling

number of shortest paths from node x (red) to all other nodes (blue) passing to all neighbours (green) of node x in a network with **high clustering coefficient** and **small diameter**

Notes:

- Social networks often have both a small diameter as well as a high clustering coefficient. In such networks the situation is very different, as shown in the example on the right.
- We now observe that one neighbor lies on 18 shortest paths, while the number for most of the other neighbors is much smaller. For a random target node, this neighbor would be a good choice to the letter to.
- This property of a network is called funneling (funnel in German: “Trichter”). It refers to the fact that most shortest paths around a node are funneled through only a few of its neighbors. What is the reason for this funneling property in social networks? It is related to the concept of **strong and weak ties** in social networks that was proposed in the 1970s. The idea is that those neighbors that are on many shortest paths are typically the ones whose neighbors have the smallest overlap with your own neighbors, i.e. they are so-called weak ties because these are nodes outside your own densely connected cluster of acquaintances.

Strong vs. weak ties

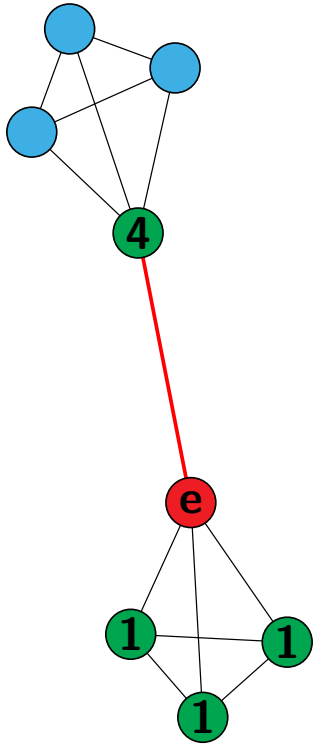
- **strong ties** within clusters or cliques

> *"... likely contacts of two individuals who are closely acquainted tend to be more overlapping than those of two arbitrarily selected individuals."* → Rapoport, 1954

- **weak ties** "bridge" cluster or cliques

> *"... weak ties are more likely to link members of different small groups than strong ones, which tend to be concentrated within particular groups"* → Granovetter, 1973

> *"... whatever is to be diffused can reach a larger number of people, and traverse greater social distance when passed through weak ties rather than strong."* → Granovetter, 1973



Mark Granovetter

born 1943

image credit: wikibooks.org, public domain

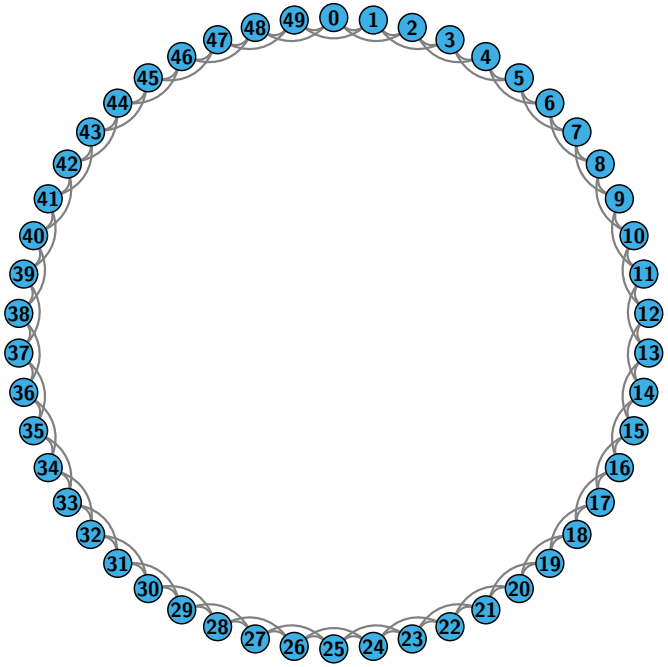
Notes:

- The classification of links in social networks into **strong and weak ties** has been put forth in two influential sociological papers:

    1. Anatol Rapoport: "Spread of information through a population with socio-structural bias: III. Suggested experimental procedures", Bulletin of Mathematical Biophysics, Vol. 16, 1954
    2. Mark Granovetter: "The strength of weak ties", American Journal of Sociology, Vol. 78, 1973

- Another empirical finding has been made based on a study of how people got to know about their current job: *"... there is a structural tendency for those to whom one is only weakly tied to have better access to job information one does not already have."* Mark Granovetter: "The strength of weak ties: A network theory revisited", Sociological theory, Vol. 1, No. 1, 1983. It turns out that it is the presence of weak ties that makes social networks navigable and those can be used to maximize the spread of information. This work was so influential that Mark Granovetter is often listed as a candidate for an Economics Nobel Prize.
- Which measure captures the presence of densely connected clusters in which connected nodes have largely overlapping neighbor lists? The clustering coefficient. We have already seen that (sparse) random networks have a small diameter, but a close to zero clustering coefficient. So how can we construct networks that combine both a large clustering coefficient and a small diameter (thus creating networks with both strong and weak ties)?

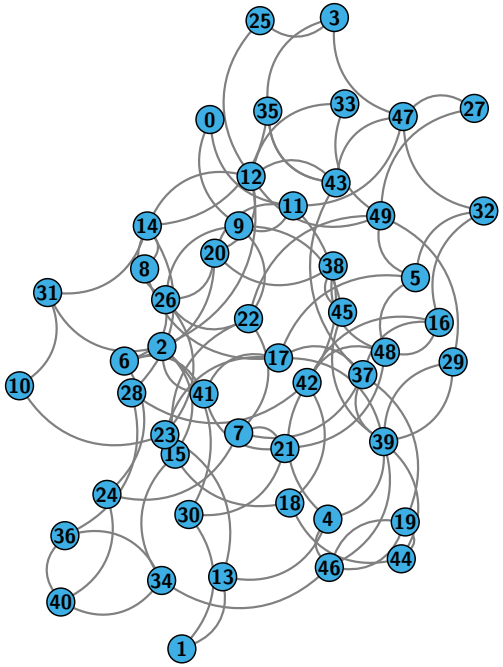
Clustering coefficient vs. diameter



ring lattice with $n = 50$, $m = 100$

large clustering coefficient

large diameter



$G(n, m)$ with $n = 50$, $m = 100$

small clustering coefficient

small diameter

can you think of a model that combines **large clustering coefficient** with **small diameter**?

Notes:

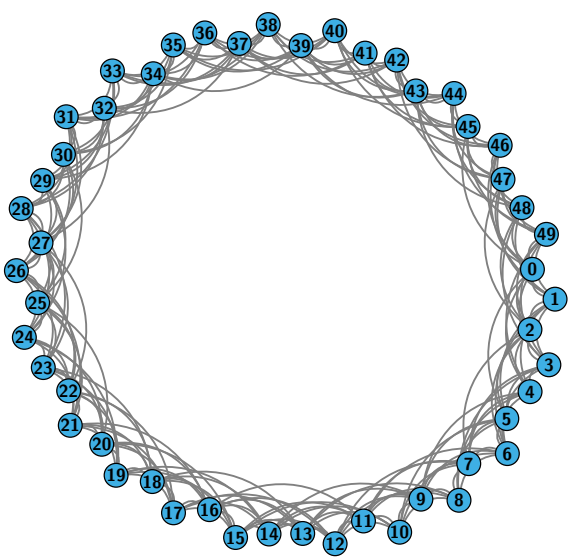
- Before we introduce such a model, we first consider two extreme types of networks:
  - In the **random network** shown on the right, the clustering coefficient is small, meaning that two nodes connected by a link are unlikely to share a common neighbor (if the network is sparse). According to Rapoport and Granovetter, this would correspond to a situation where **strong ties are absent**.
  - In the **ring lattice network** shown on the left, the clustering coefficient is large, because the network is such that nodes share most of their contacts. However, here the diameter is large and nodes are not connected to other nodes whose neighbors do not overlap with their own neighbors. According to the notions of Rapoport and Granovetter, this corresponds to a situation where **weak ties are absent**, Moreover, short paths are absent as well.
- What we need is a middle ground between these two models, i.e. a model which generates networks in which both strong and weak ties exist.

Watts-Strogatz random lattice networks

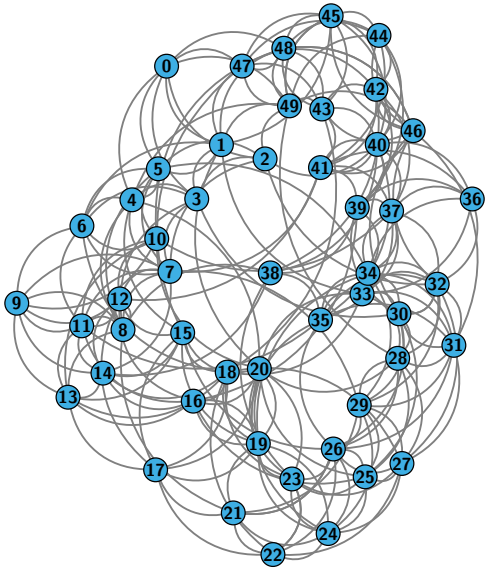
Watts-Strogatz model

Consider regular lattice network $G = (V, E)$ with n nodes and dimension d, where nodes at **lattice distance** smaller than s are connected. For each edge $(v, w) \in E$ "rewire"' node v to randomly chosen node with probability p.

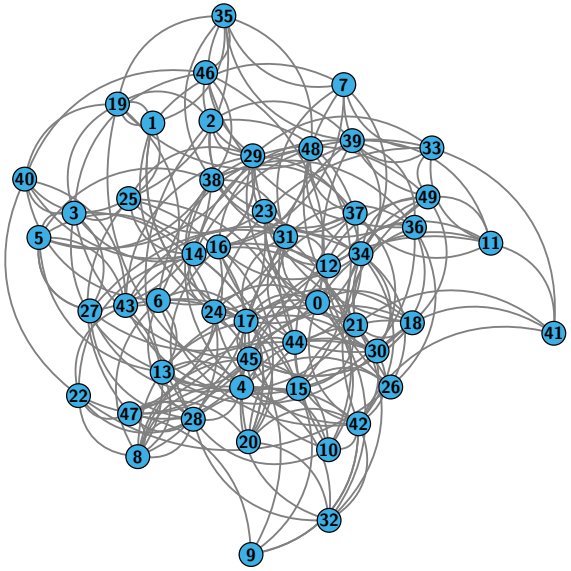
- $d = 1$: **ring lattice**
- $d = 2$: **grid lattice**
- $d = 3$: **cube lattice**



$p = 0$



$p = 0.2$



$p = 1$

Notes:

- Such a model was proposed by Duncan J. Watts and Steven H. Strogatz in their influential paper "Collective dynamics of 'small-world' networks", Nature, Vol. 393, pp. 440-442, June 1998. It starts from a lattice network with dimensionality d and with lattice parameter s. For $d = 1$ we assign one-dimensional coordinates to nodes and connected nodes up to lattice s to each other. We further consider a periodic boundary condition in which we interconnect the endpoints of the lattice to each other (i.e. we wrap-around the coordinates). We then additionally consider a probability p with which each link is rewired to a node chosen uniformly at random.
- For $p = 0$ we end up with a (deterministic) lattice network, i.e. we have large clustering and a large diameter. For $p = 1$ we obtain a network in which all links have been rewired randomly. As a first approximation, we can think of the resulting networks as realizations of the $G(n, m)$ model. However, a more detailed analysis reveals that even for $p = 1$ the rewiring process leaves some "memory" about the initial lattice network. As a result, even for $p = 1$ there are slight deviations compared to a fully random network. For details see: A. Barrat, M. Weight: "On the properties of small-world network models", Eur. Phys. J. B, Vol. 11, pp. 547-560, 2000.
- For intermediate values of p, we obtain networks with small diameter and large clustering coefficient. Note that, for small values of s, there is a chance that the network becomes disconnected due to the rewiring process. To avoid this, there is the so-called Newman-Watts variant of the model, in which - rather than rewiring links - new random links are added to the initial lattice. In the figures above I have applied a force-directed layout algorithm to each network.

Watts-Strogatz networks: $p = 0$ and $p = 1$

ring lattice (n nodes, $d = 1$)

- **clustering coefficient** → exercise sheet 03

$$C = \frac{3s - 3}{4s - 2}$$

- **diameter**

$$D \approx \frac{n}{2s}$$

- **degree distribution**

$$P(k) = \begin{cases} 1 & \text{if } k = 2s \\ 0 & \text{else} \end{cases}$$

random graph (n nodes, $\langle k \rangle = 2s = np \ll n$)

- **clustering coefficient**

$$C = \frac{2s}{n} \to 0$$

- **diameter**

$$D \approx \frac{\log n}{\log 2s}$$

- **degree distribution**

$$P(k) \sim \text{Poisss(2s)} = \frac{(2s)^k e^{-2s}}{k!}$$

Notes:

- What can we say about the expected properties of Watts-Strogatz networks for $p = 0$ and $p = 1$?
- As an exercise, you should consider the definition of the global clustering coefficient and show that the above expression for the clustering coefficient holds for a ring lattice. For a random graph, the expected clustering is simply the expected degree divided by the number of nodes.
- We can easily validate the expression for the diameter for two boundary cases:
  - if each node is connected to exactly one neighbor on both sides ($s = 1$), then the diameter is exactly half the size of the ring
  - if on the other hand the parameter we have $s = \frac{n}{2}$ (i.e. each node is connected to all other nodes), then the diameter is one
- The degree distribution of the ring lattice is a delta function, because all nodes have exactly the same degree $k = 2s$.
- Note that – due to the memory effects mentioned on the previous slide – the expression found for the random network does not hold exactly even for $p = 1$. But ignoring these (small) effects, we can use it as an approximation for the degree distribution of large networks.
- What can we say about clustering coefficient and path lengths in the intermediate range of parameters p?

Watts-Strogatz: clustering coefficient

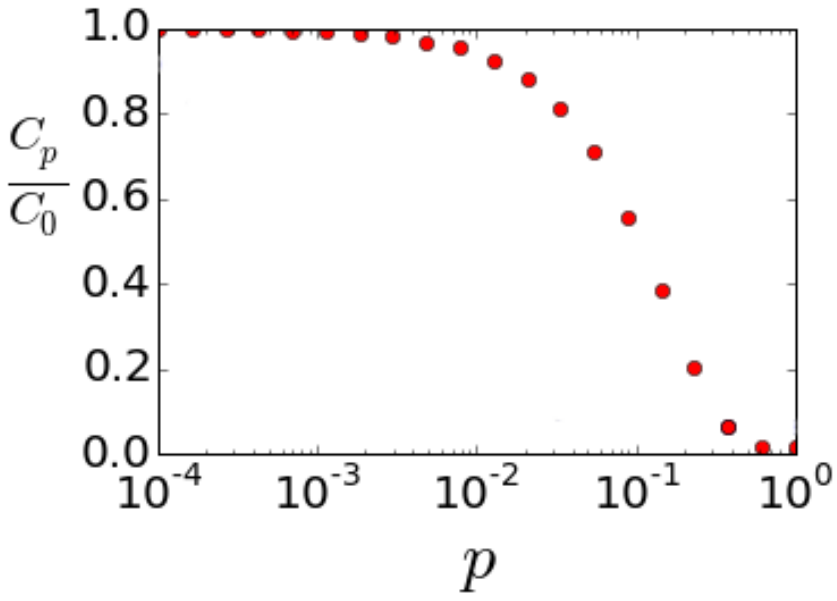
- for $p = 0$ we have an initial **clustering coefficient**

$$C_p = C_0 = \frac{3s - 3}{4s - 2}$$

- how does C_p change compared to initial value C_0 as we increase p?
- for each p we can calculate ratio

$$\frac{C_p}{C_0}$$

- for $p \ll 1$ we have **larger than random clustering coefficient**



example

$\frac{C_p}{C_0}$ for microstates generated by Watts-Strogatz model with parameters $d = 1$, $n = 1000$, $s = 5$ and different rewiring probabilities p

Notes:

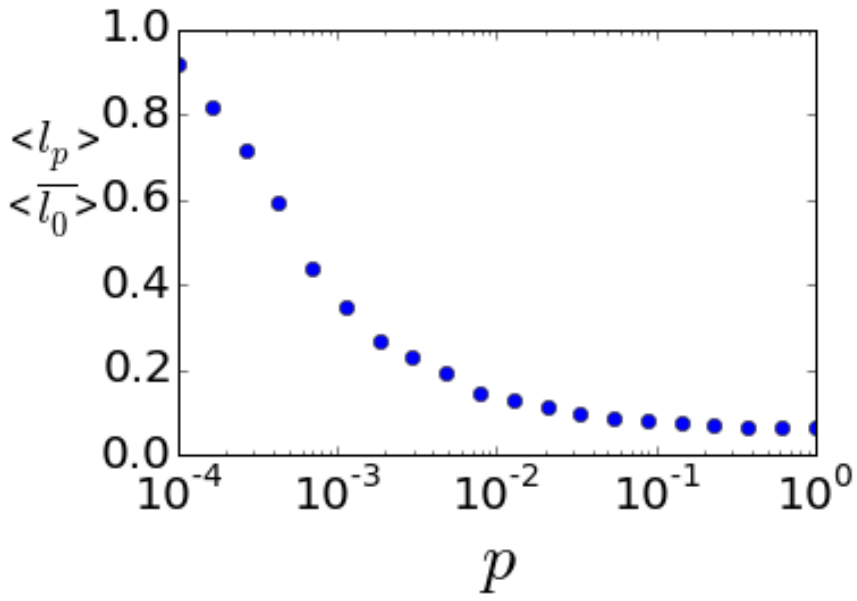
- For $p = 0$ we can calculate the (local) clustering coefficient of a node v in a (perfect) ring topology with parameter s $\rightarrow$ exercise sheet 03 . Unfortunately, this is more difficult for non-zero values of p, so we instead use simulations to generate microstates and calculate the empirical clustering coefficient.
- For any rewiring probability p we calculate the clustering coefficient C_p as well as the ratio $\frac{C_p}{C_0}$, which captures how much of the initial (large) clustering coefficient was preserved under the rewiring.
- The red points in the plot on the right show the average clustering coefficient of 10 microstates generated by the Watts-Strogatz model with different rewiring probabilities p. For each microstate, the model was initialized using a ring lattice with $s = 5$, $n = 1000$.
- We observe a **slow decrease** of the clustering coefficient as the rewiring probability p is increased.

Watts-Strogatz: shortest path lengths

- let l_p be **average shortest path length** in microstate generated by Watts-Strogatz model with parameters n and p
- let l_0 be average shortest path length in ring lattice ($p = 0$)
- for each p we can calculate= ratio

$$\frac{\langle l_p \rangle}{\langle l_0 \rangle} \in [0, 1]$$

- for $p \gg 0$ we have **average shortest path lengths similar to random networks**



example

$\frac{\langle l_p \rangle}{\langle l_0 \rangle}$ for microstates generated by Watts-Strogatz model with parameters $d = 1$, $n = 1000$, $s = 5$ and different rewiring probabilities p

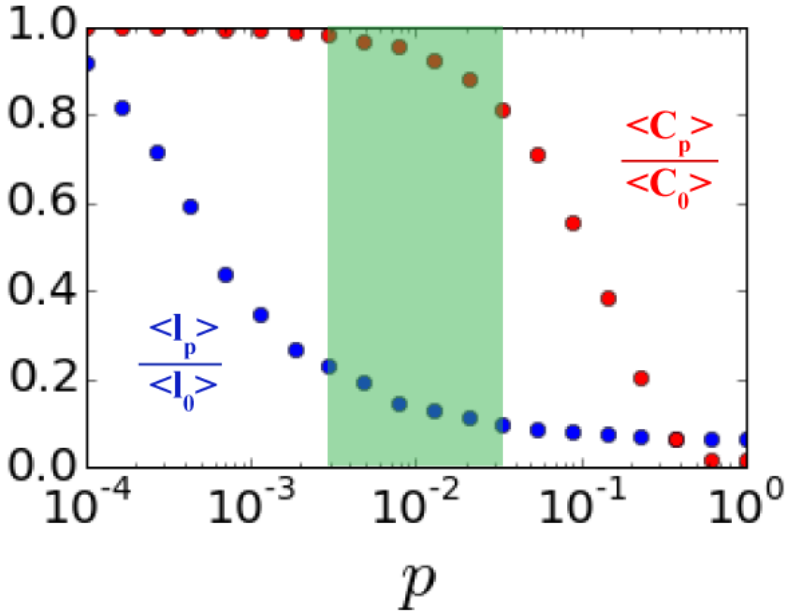
Notes:

- We can use the same approach to calculate the average shortest path lengths for Watts-Strogatz networks with different rewiring probabilities.
- Starting from the initial average shortest path length $\langle l_0 \rangle$ of the ring lattice, we calculate the average shortest path length $\langle l_p \rangle$ for networks generated using a rewiring probability p. The ratio $\frac{\langle l_p \rangle}{\langle l_0 \rangle}$ tells us how much shorter paths are on average compared to the (maximum) average shortest path length in the ring lattice.
- We observe a **fast decrease** of the average shortest path length as the rewiring probability p increases. This is due to the fact that a single rewired link has a large impact on the average shortest path length, as it affects all shortest paths in the network. In contrast, the effect of a rewired link on the global clustering coefficient is much smaller as it only affects local clustering coefficients in the neighborhood of those nodes that are connected.

Watts-Strogatz: the small-world regime

- **clustering coefficient** C_p of WS network with rewiring probability p
- **average shortest path length** $\langle l_p \rangle$ of WS network with parameter p
- **small world regime** (green region in graph)

$$\frac{\langle l_p \rangle}{\langle l_0 \rangle} \ll 1 \text{ and } \frac{C_p}{C_0} \gg 0$$



small-world network

we call a network that combines **small, similar to random average path length** with a **larger than random clustering coefficient** a small-world network

example

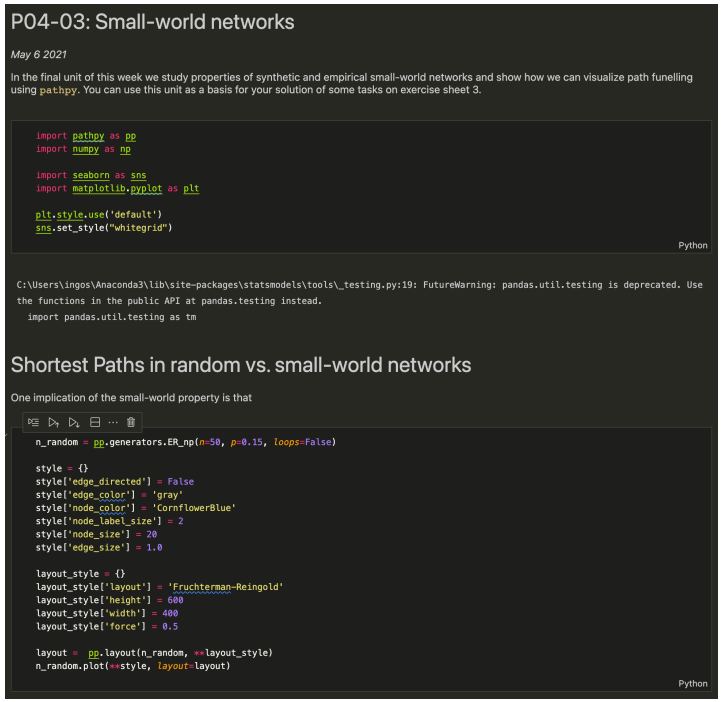
clustering coefficient (red) and average shortest path length (blue) ratio for microstates generated by Watts-Strogatz model with parameters $d = 1$, $n = 1000$, $s = 5$ and different rewiring probabilities p

Notes:

- This (localized) effect of link rewiring on the clustering coefficient, and its (global) effect on shortest path structures generates an interesting effect: the fast drop in the average shortest path length and the much slower drop in the clustering coefficient implies that there is an intermediate regime of p where the average shortest path is already small, while the clustering coefficient is still large
- in this intermediate regime (realized for parameters $0 \ll p \ll 1$), the generated networks reproduce features of social networks which are crucial for efficient navigability.
- We particularly generate network topologies in which nodes have both strong and weak ties, which can be used to efficiently navigate the network. We call this intermediate range of p the small-world regime of the model.

Practice session

- we **implement the Watts-Strogatz model** of small-world networks
- we explore the dependence of average path length and clustering coefficient on the **rewiring probability**
- we visualize path funnelling in small-world networks
- we test the **small-world property** of empirical social networks



P04-03: Small-world networks

May 6 2021

In the final unit of this week we study properties of synthetic and empirical small-world networks and show how we can visualize path funelling using `pathpy`. You can use this unit as a basis for your solution of some tasks on exercise sheet 3.

```python
import pathpy as pp
import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt

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

Python

```
C:\Users\ingos\Anaconda3\lib\site-packages\statsmodels\tools\_testing.py:19: FutureWarning: pandas.util.testing is deprecated. Use
the functions in the public API at pandas.testing instead.
  import pandas.util.testing as tm
```

Shortest Paths in random vs. small-world networks

One implication of the small-world property is that

```python
n_random = pp.generators.ER_np(n=50, p=0.15, loops=False)

style = {}
style['edge_directed'] = False
style['edge_color'] = 'gray'
style['node_color'] = 'CornflowerBlue'
style['node_label_size'] = 2
style['node_size'] = 20
style['edge_size'] = 1.0

layout_style = {}
layout_style['layout'] = 'Fruchterman-Reingold'
layout_style['height'] = 600
layout_style['width'] = 400
layout_style['force'] = 0.5

layout = pp.layout(n_random, **layout_style)
n_random.plot(**style, layout=layout)
```

Python

practice session

see notebooks 05-03 – 05-04 in `gitlab` repository at

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

Notes:

- In the second practice session, we implement the Watts-Strogatz model and explore the small-world property of randomly generated and empirical networks.

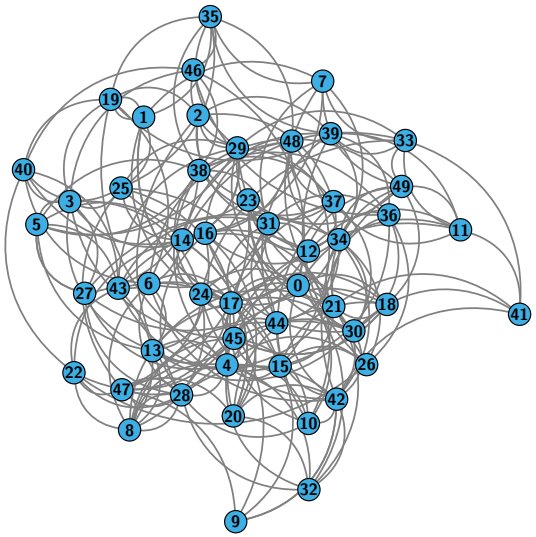
In summary

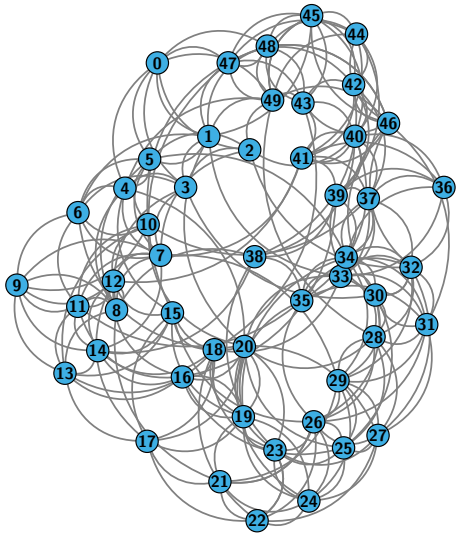
random graphs

- $G(n, p)$ and $G(n, m)$ model are useful to study **expected properties** of complex networks
- for $\langle k \rangle \to \infty$ convergence to **Normal degree distribution**
- for $\langle k \rangle \to$ const convergence to **Poisson degree distribution**
- sparse random networks have **small diameter** and **small clustering coefficient**

small-worlds

- networks with small average shortest path length and large clustering coefficient
- combination of **weak and strong ties** makes networks navigable
- small-world property is **common in social networks**
- small-world networks can be generated by **Watts-Strogatz model**





Notes:

- In summary, we have used random graph models as (maximally) simple reference (or null) models for complex networks. Despite their simplicity, these models allow us to study which of the characteristics of an empirical network we can already expect at random. We learned that, if we find a network with a Poisson- or Normal-shaped degree distribution, we should not be surprised. Random edge formation is enough to explain these degree distributions. On the other hand, this tells us that we need more elaborate models if we want to explain degrees that do **not** follow a Poisson or Normal distribution (such as the broad degree distributions that we have seen in some of th earlier examples).
- These simple reference models can help us to decide which statistical patterns are likely to be the result of more complex organization principles resulting, e.g., from growth processes, stochastic optimization schemes or strategic link formation processes.
- We have further seen that social networks exhibit the small-world property and we have introduced a simple stochastic model (the Watts-Strogatz model) that allows to generate networks which exhibit this property. In an intermediate regime of the model parameter p, the generated microstates reproduce the small-world property of real networks, which we define as the combination of a small average shortest path length and larger than random clustering coefficient.

Exercise sheet 03

- **third exercise sheet** is available on WueCampus
  - prove the limiting degree distribution
  - study the diameter of random graphs
  - explore properties of small-world networks
- solutions are due **November 23nd** (via WueCampus)
- solutions will be discussed on December 8th
- 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 03

Published: November 23, 2021
Due: December 1, 2021
Total points: 12

1. **Limiting degree distribution**
   (a) Consider a function [3P]

$$G_P(x;n,p) = \sum_{k=0}^{n} P_{n,p}(k)x^k \text{ with } P_{n,p}(k) = \binom{n}{k} \cdot p^k(1-p)^{n-k}$$

   as well as a function

$$G_Q(x;\lambda) = \sum_{k=0}^{\infty} Q_\lambda(k)x^k \text{ with } Q_\lambda(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

   For $n \to \infty$ and $np \to \lambda$ show $G_P(x;n,p) \to G_Q(x;\lambda)$.

   Hint: Use the binomial theorem to rewrite $G_P(x;n,p)$.

   (b) In lecture L05 we have shown that the degrees of a random microstate generated by the $G(n,p)$ model follow a Binomial distribution. We further stated that for $n \to \infty$ and $np \to \lambda$ the degree distribution converges to the Poisson distribution with parameter λ. Investigate the definition of a so-called *probability generating function* and use it to explain how your proof from (a) can be used to prove the convergence to a Poisson degree distribution. [2P]

2. **Diameter of random graphs**
   (a) Consider the $G(n,p)$ model of random graphs for a sufficiently large n and variable link probability p. Plot the dependency of the diameter of random realizations on the parameter p. Compare the observed dependency with the expected dependency based on the approximation of the diameter made in lecture L05. Explain possible discrepancies. [2P]

   (b) Use the $G(n,p)$ model to generate a number of random microstates (undirected with self-loops) with a fixed expected degree $\langle k \rangle$. Plot the distribution of degrees of randomly chosen nodes in the generated microstates and compare it to the expected value $\langle k \rangle$. Repeat the same procedure for the degrees of *neighbors* of randomly chosen nodes. How do those distributions differ? Explain the phenomenon and how it might affect the approximation of the diameter made in lecture L05. [2P]

3. **Watts-Strogatz Model**
   (a) Use the Watts-Strogatz model to generate microstates that exhibit the small-world property. Operationalize the definition of weak vs. strong ties from lecture L05, i.e. define an edge-level measure that captures to what extent an edge can be considered a weak tie. Use your measure to highlight weak ties in a plot of generated microstates. [1P]

   (b) Show that the average clustering coefficient of a one-dimensional undirected ring lattice with parameter $s \geq 1$ is $\frac{3s-3}{4s-2}$. [1P]

   (c) Use the Watts-Strogatz model to generate ring lattice networks (rewiring probability $p = 0$) with different lattice parameters $s \geq 1$. Plot the dependency of the average shortest path length of the ring lattice networks on the lattice parameter s. [1P]

Notes:

Self-study questions

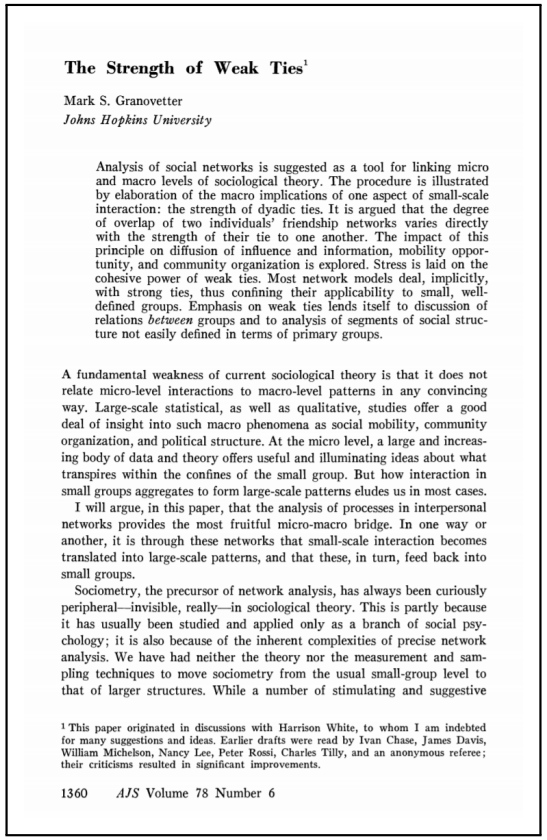
1. What is the limiting degree distribution for networks generated by the $G(n, p)$ model for $n \to \infty$?
2. Under what condition does the degree distribution converge to a Poisson distribution?
3. Show that the variance of a degree distribution is given by $\langle k^2 \rangle - \langle k \rangle^2$.
4. What degree assortativity r do you expect for microstates from the ensemble generated by the $G(n, p)$ model?
5. What is the expected clustering coefficient for $G(n, p)$ microstates?
6. Show that the diameter of Erdős-Rényi networks scales logarithmically with the network size.
7. Explain the difference between weak and strong ties in a social network.
8. Generate a network with small-world properties and pick a node whose neighbors exhibit a "funnelling" of shortest paths.
9. Show that the clustering coefficient of a ring lattice with lattice parameter s is $\frac{3s-3}{4s-s}$.
10. For which parameters of the Watts/Strogatz model do we expect generated networks to exhibit the small-world property?

Notes:

References

reading list

- F Karinthy: **Chains**, 1929
- B Bollobás: **The Diameter of Random Graphs**, Transactions of the American Mathematical Society, 1981
- O Riordan, N Wormald: **The diameter of sparse random graphs**, Combinatorics, Probability and Computing, 2010
- A Rapoport: **Spread of information through a population with sociostructural bias: III. Suggested Experimental Procedures**. Bulletin of Mathematical Biophysics, 1954
- S Milgram: **The Small World Problem**, Psychology Today, 1967
- M Granovetter: **The Strength of Weak Ties: A Network Theory Revisited**. Sociological Theory, 1983
- DJ Watts, SH Strogatz: **Collective dynamics of 'small-world' networks**, Nature, 1998



The Strength of Weak Ties[1]

Mark S. Granovetter
Johns Hopkins University

Analysis of social networks is suggested as a tool for linking micro and macro levels of sociological theory. The procedure is illustrated by elaboration of the macro implications of one aspect of small-scale interaction: the strength of dyadic ties. It is argued that the degree of overlap of two individuals' friendship networks varies directly with the strength of their tie to one another. The impact of this principle on diffusion of influence and information, mobility opportunity, and community organization is explored. Stress is laid on the cohesive power of weak ties. Most network models deal, implicitly, with strong ties, thus confining their applicability to small, well-defined groups. Emphasis on weak ties lends itself to discussion of relations *between* groups and to analysis of segments of social structure not easily defined in terms of primary groups.

A fundamental weakness of current sociological theory is that it does not relate micro-level interactions to macro-level patterns in any convincing way. Large-scale statistical, as well as qualitative, studies offer a good deal of insight into such macro phenomena as social mobility, community organization, and political structure. At the micro level, a large and increasing body of data and theory offers useful and illuminating ideas about what transpires within the confines of the small group. But how interaction in small groups aggregates to form large-scale patterns eludes us in most cases.

I will argue, in this paper, that the analysis of processes in interpersonal networks provides the most fruitful micro-macro bridge. In one way or another, it is through these networks that small-scale interaction becomes translated into large-scale patterns, and that these, in turn, feed back into small groups.

Sociometry, the precursor of network analysis, has always been curiously peripheral—invisible, really—in sociological theory. This is partly because it has usually been studied and applied only as a branch of social psychology; it is also because of the inherent complexities of precise network analysis. We have had neither the theory nor the measurement and sampling techniques to move sociometry from the usual small-group level to that of larger structures. While a number of stimulating and suggestive

[1] This paper originated in discussions with Harrison White, to whom I am indebted for many suggestions and ideas. Earlier drafts were read by Ivan Chase, James Davis, William Michelson, Nancy Lee, Peter Rossi, Charles Tilly, and an anonymous referee; their criticisms resulted in significant improvements.

1360 *AJS* Volume 78 Number 6

Notes: