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

Bonus Lecture
Modelling Growing Networks

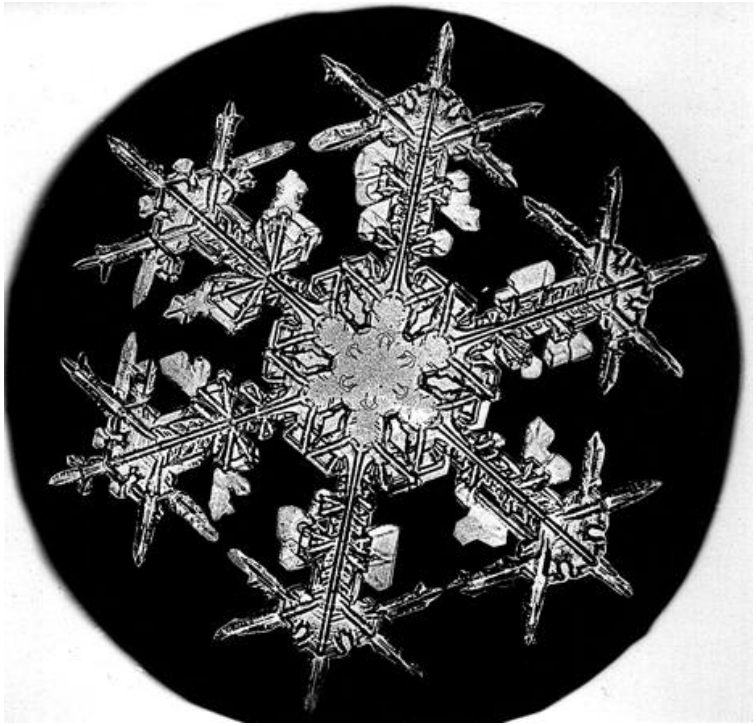
December 21, 2022

Notes:

- Bonus Lecture: Modelling Growing Networks 21.12.2022
- Educational Objective: In this lecture, students will learn that feedback phenomena in the growth of networks can lead to the formation of complex structures
  - Growth models for complex networks
  - Growth of random networks: uniform attachment
  - Feedback in network growth: preferential attachment
  - Preferential attachment and node fitness

- **I wish all of you merry christmas, happy holidays and a good start into a healthy and successful new year!**

Structure formation in complex systems



conservative process

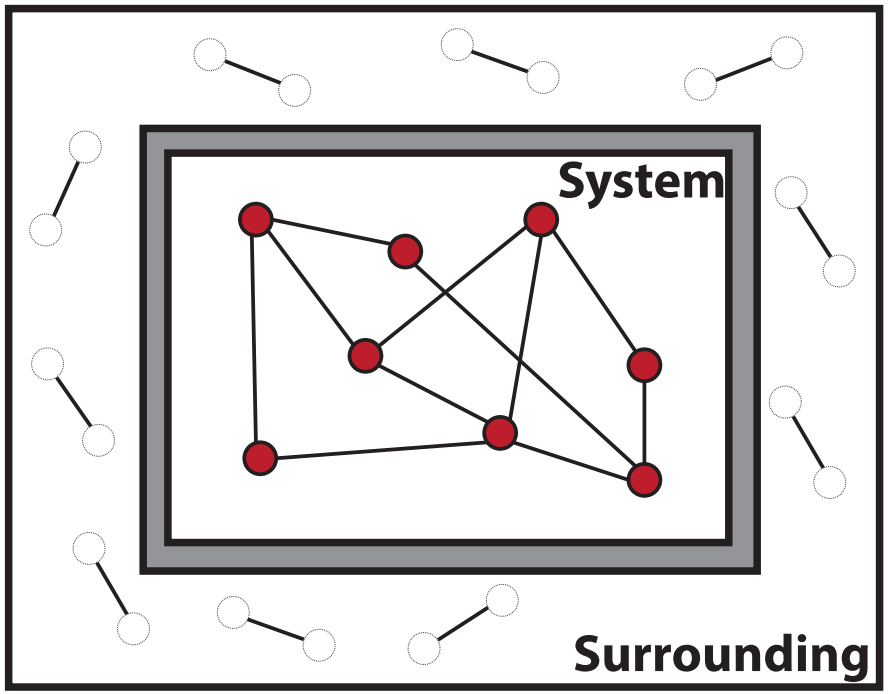
dissipative process

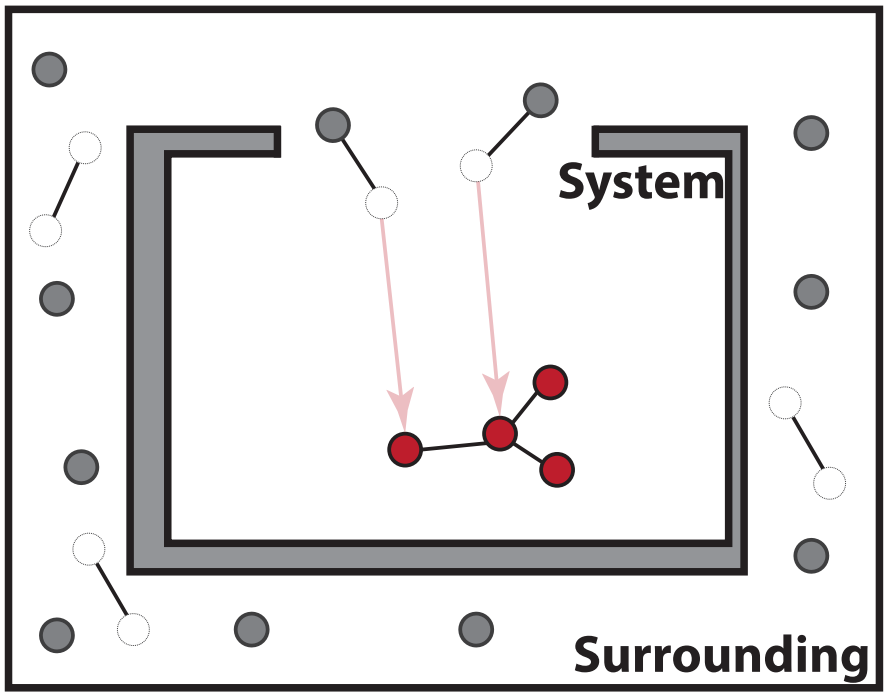
complex structures can form in **equilibrium** and **non-equilibrium** systems.

Notes:

- We can view structure formation in networks in analogy to equilibrium and non-equilibrium systems in physics
- the situation that we are studying today corresponds to the right-hand image, i.e. the formation of complex structures based on a growth process

Equilibrium and non-equilibrium networks





equilibrium network with fixed number of nodes and links

macrostate is stable

non-equilibrium network with net inflow of nodes/links

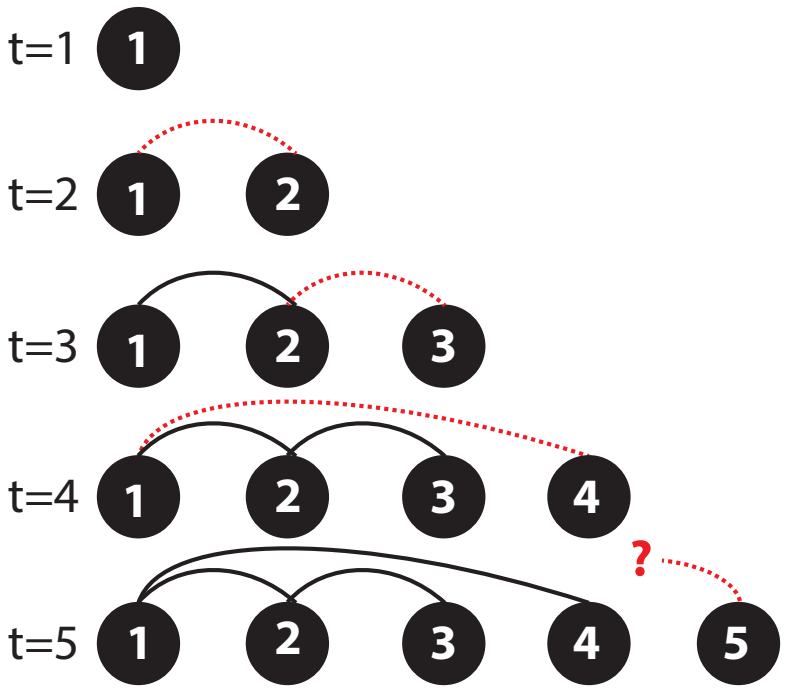
macrostate changes over time

Notes:

- the non-equilibrium case studied today corresponds to a situation where there is a **net inflow** of particles and/or energy
- what we want to understand is how the macrostate description changes over time (and to what stationary state it possible converges)

Modeling network growth

- we study a **homogeneous growth process**
- at each time step a single node is added to the network
- newly added node forms one or more links to existing nodes according to some **stochastic attachment rule**
- any ideas for attachment rules?

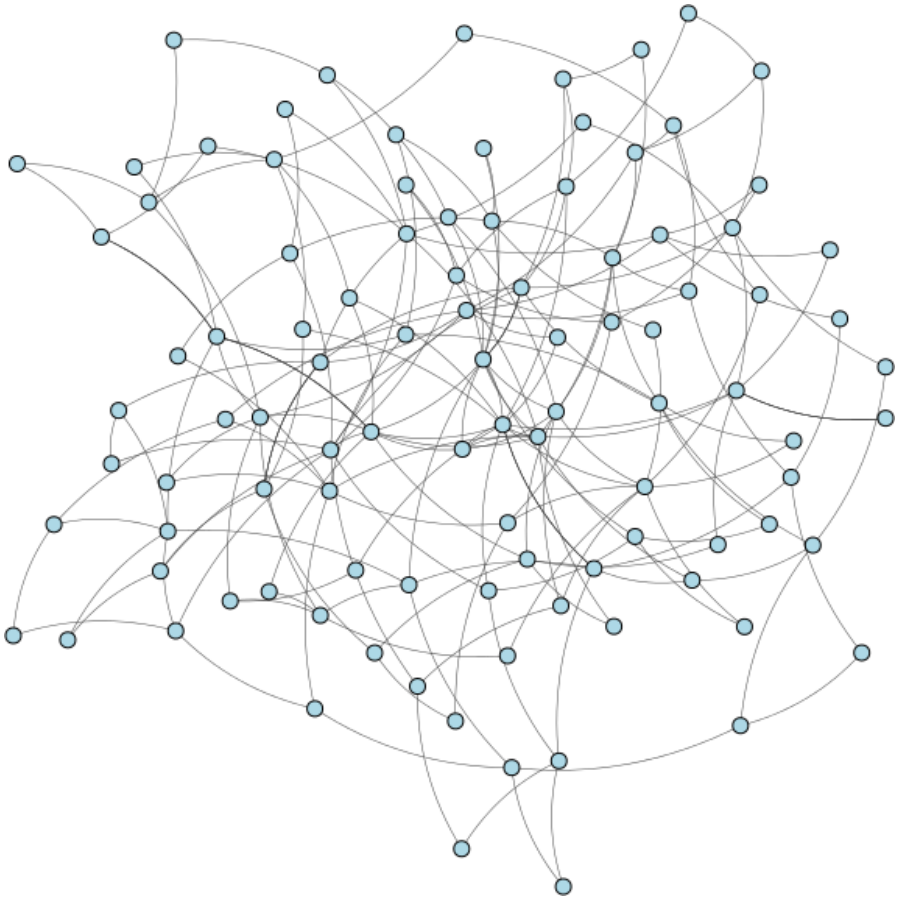


example: at time $t = 5$, node 5 forms links to one or more of the four nodes $1, 2, 3, 4$

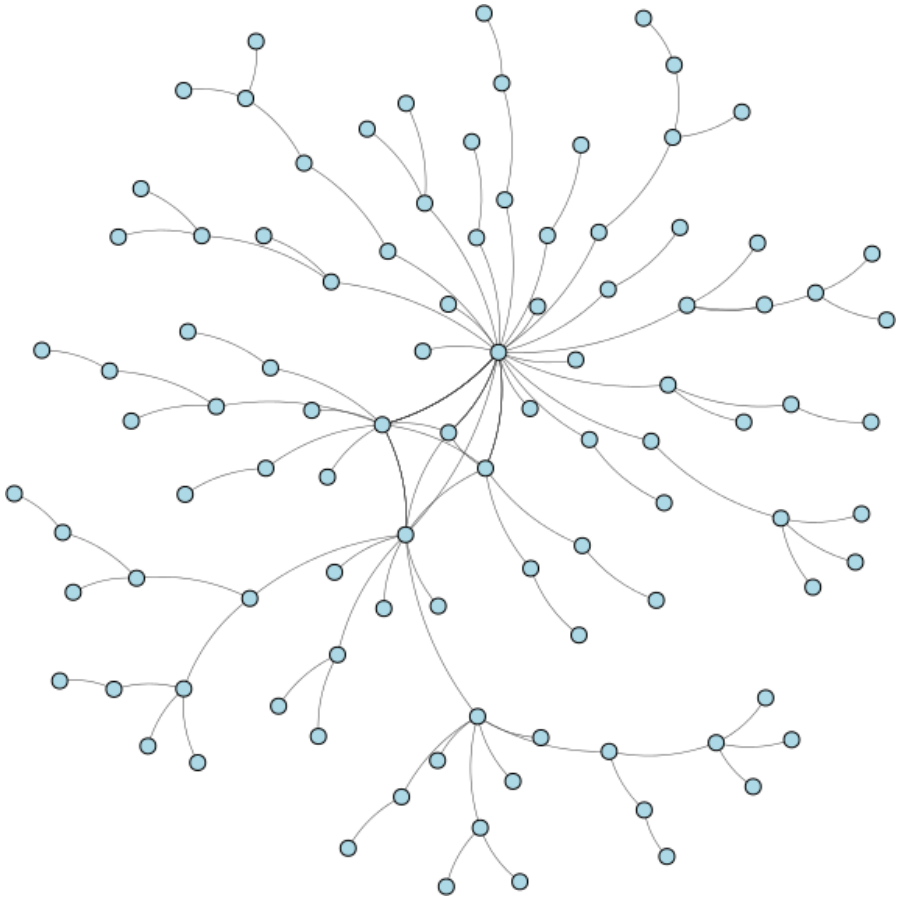
Notes:

- in the following, we will study a simple family of network growth models
- starting from a small initial network, in these models nodes are added to the network one by one
- in the first simple case, we start with a single disconnected node at time $t = 0$
- many real-world networks are formed by growth processes like, e.g., citation networks, many collaboration networks, software structures, possibly protein networks, etc.
- the networks that we observe at a given point in time is the result of such a growth process, so in order to understand why networks look the way they do, we should consider a model for their growth
- we will see that we can analytically study the resulting networks for different types of **stochastic attachment rules**, which model how new nodes connect to the existing network

Two attachment rules



uniform attachment: new nodes form links to existing nodes chosen uniformly at random

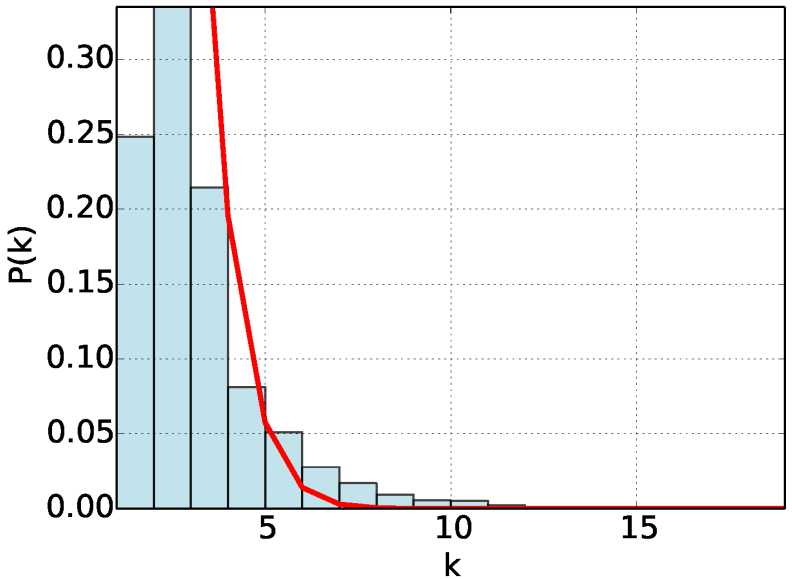


preferential attachment: new nodes preferentially form links to highly connected nodes

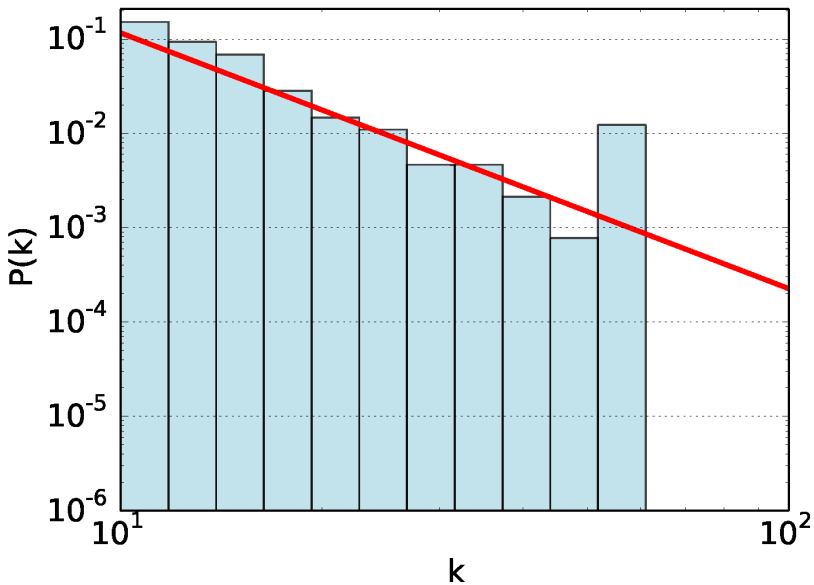
Notes:

- we will study two different attachment rules in more detail
    1. uniform attachment, in which new nodes form links uniformly at random
    2. preferential attachment, in which nodes are more likely to form links to nodes with high degree

Exponential vs. scale-free degree distribution



uniform attachment: gives rise to **exponentially shaped** degree distribution

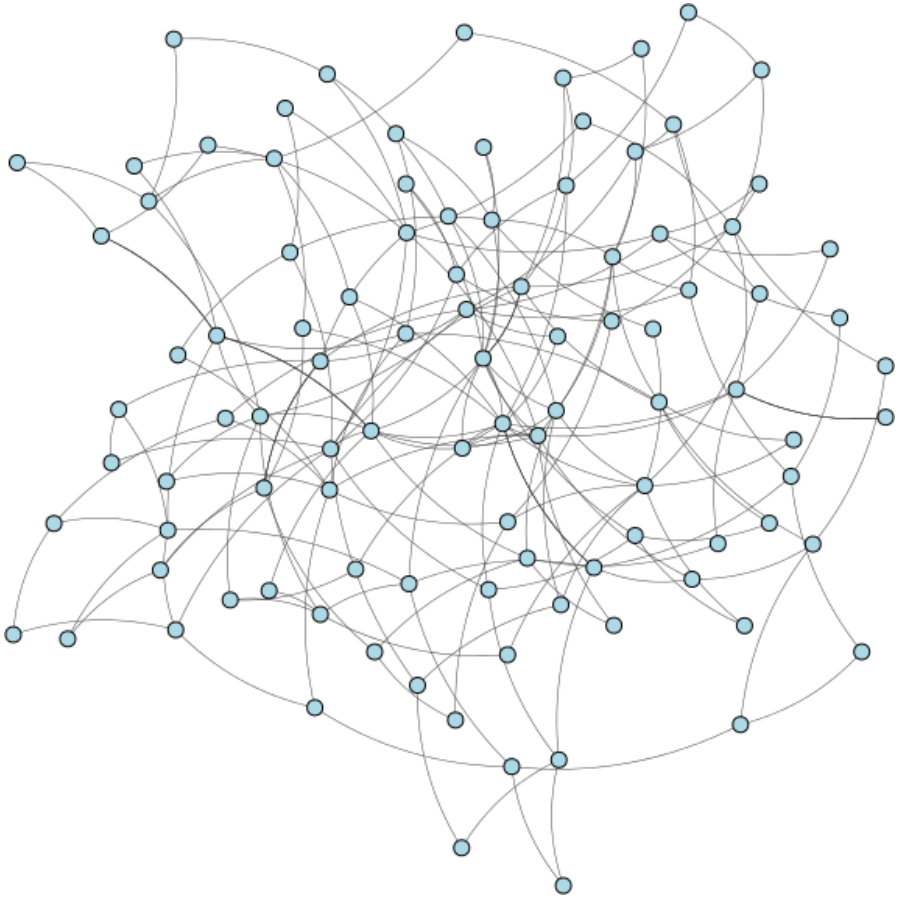


preferential attachment: gives rise to **scale-free** degree distribution

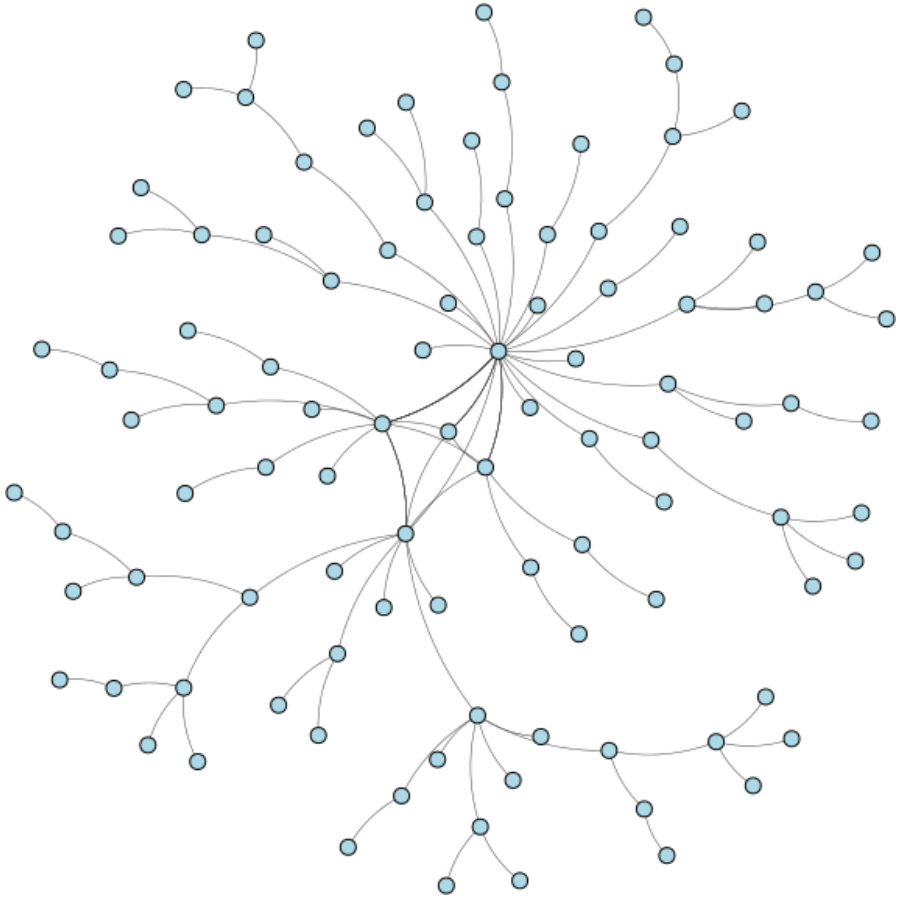
Notes:

- we chose these two attachment rules because they reproduce two broad classes of networks that we have seen before
- the first are networks with exponentially shaped degree distributions, much like the ones generated by simple random network models studied before
- the second class are networks with broad, or even scale-free, degree distributions, which we also have considered in the last lecture

Two analytical approaches



master equation: powerful tool to analyse non-equilibrium systems in statistical physics



continuum limit approximation: allows analysis in terms of differential equation

Notes:

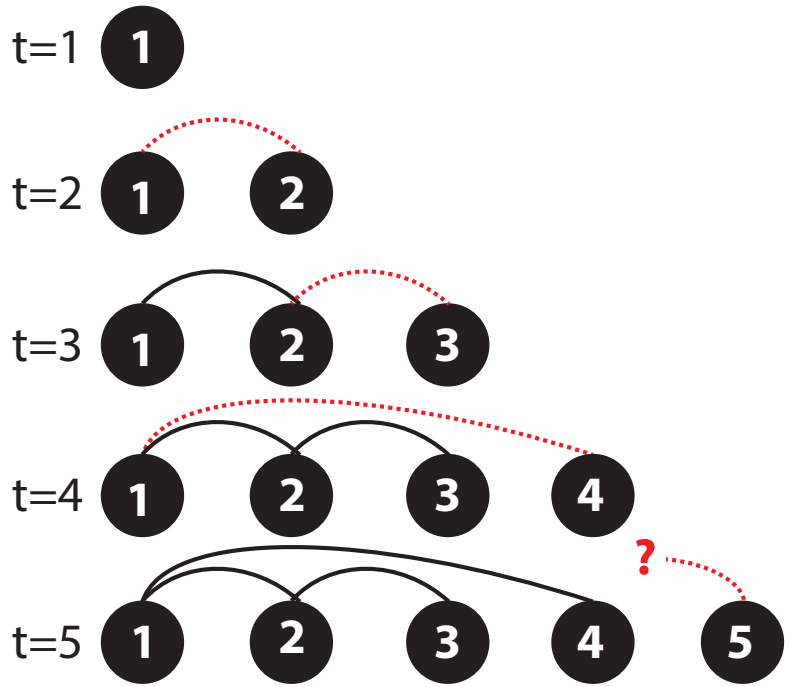
- another reason why we discuss these two attachment rules in more detail, is that it allows us to introduce two powerful analytical approaches, that you may useful in your own studies of networked systems
- the first is based on a discrete perspective (which is the natural one for discrete objects like networks). Here we will use master equations, which some of you may know from physics
- the second makes a simplifying assumption of a continuum limit approximation, which will allow us to write down an ordinary differential equation for the evolution of the macrostate (in this case, the degree distribution). Solving this differential equation will then give the stationary degree distribution of the growth model

Network growth: uniform attachment

- we consider the simplest possible rule
- new nodes form a single link to an existing node chosen uniformly at random
- consider node v_{t+1} added at time $t+1$
- probability p to form link to any one of the existing nodes $v_1, \ldots, v_t$ is

$$p = \frac{1}{t}$$

- observation: probability to receive links decays over time



at time $t = 5$, node 5 chooses one of the four nodes $1, 2, 3, 4$ uniformly at random

Notes:

- let us start with the simplest possible stochastic attachment rule: uniform attachment
- in the following, we assume again a homogeneous network growth, i.e. a single node is added at each time step
- this new node picks a single random existing node and adds an undirected link to this node
- a first observation that we can make is that - from the perspective of existing nodes - the probability to receive links linearly decays over time (as the choices for new nodes grow linearly)
- at the same time, nodes added to the system early, have more chances to get links
- we will see that these two aspects of uniform growth represent “balanced forces” that shape the degree distribution

Analysis with a master equation

- for growing networks, we have a **time-dependent degree distribution**
- let $P(k, s, t)$ be the probability that node added at time s has degree k at time t
- we can write down a **master equation** for the **evolution of** $P(k, s, t)$

$$P(k, s, t+1) = \underbrace{\frac{1}{t} P(k-1, s, t)}_{\text{node } s \text{ with degree } k-1 \text{ drawn}} + \underbrace{\left(1 - \frac{1}{t}\right) P(k, s, t)}_{\text{node } s \text{ with degree } k \text{ not drawn}}$$

- with the initial condition $P(k, s = 1, t = 1) = \delta_{k,0}$

Notes:

- let us have a close look at this master equation
- consider the uniform attachment, i.e. at time t each of the existing nodes is chosen with probability $\frac{1}{t}$
- so what is the probability that node s has degree k at time $t+1$?
- there can be only two cases in which this happens
    1. The node had degree $k-1$ at time t and was chosen to form a link, which happens with probability $\frac{1}{t}$. This case generates the left term in the equation.
    2. The node had degree k at time t and was not chosen (which happens with probability $1-\frac{1}{t}$). This case generates the right term in the equation.
- remember: the delta function $\delta_{i,j}$ is just a function that is zero whenever $i \neq j$ and 1 for $i = j$
- the initial condition just ensures that at time zero, the first node has degree zero

Time-dependent degree distribution

- in addition to the initial condition for $s = t = 1$, we get a **boundary condition** for newly added nodes, i.e. for $s = t$ and $t > 1$

$$P(k, s = t, t) = \delta_{k,1}$$

- **time-dependent degree distribution** $P(k, t)$ is $P(k_s = k, t)$, where k_s is degree of node s chosen uniformly at random

$$P(k, t) = \sum_{s=1}^{t} \frac{1}{t} P(k, s, t)$$

Notes:

- we can now define the time-dependent degree distribution
- $P(k, t)$ is the probability mass function function of a random variable that assumes the degree of a node uniformly chosen from all nodes existing at time t
- we can just define $P(k, t)$ by summing the probabilities of all nodes s having degree k, multiplied by the uniform probability $\frac{1}{t}$ of choosing node s at time t
- the boundary condition for newly added nodes just ensures that - at the time when they are added - every node has exactly degree 1

How does $P(k,t)$ change over time?

- by computing $P(k,t+1)$, we can learn how the degree distribution evolves over time

$$P(k,t+1) = \frac{1}{t+1}\sum_{s=1}^{t+1} P(k,s,t+1)$$

- using the boundary condition $P(k,s=t+1,t+1) = \delta_{k,1}$ we then have

$$P(k,t+1) = \frac{1}{t+1}\left(\sum_{s=1}^{t} P(k,s,t+1) + \delta_{k,1}\right)$$

$\rightarrow$ **and now?**

Notes:

- in the last line (within the sum) we have $P(k, s, t+1)$, which we have encountered before in the master equation describing the change of the probability that a node added at time s has degree k

Using the master equation for $P(k, s, t+1)$

we can substitute $P(k, s, t+1)$ by the **master equation**

$$P(k, t+1) = \frac{1}{t+1}\left(\sum_{s=1}^{t}\left[\frac{1}{t}P(k-1, s, t) + \left(1 - \frac{1}{t}\right)P(k, s, t)\right] + \delta_{k,1}\right)$$

$$= \frac{1}{t+1}\left(\underbrace{\frac{1}{t}\sum_{s=1}^{t}P(k-1, s, t)}_{=P(k-1,t)} + \underbrace{\left(1 - \frac{1}{t}\right)\sum_{s=1}^{t}P(k, s, t)}_{tP(k,t)-P(k,t)} + \delta_{k,1}\right)$$

Notes:

- in the formula for $P(k, t+1)$, we can just substitute $P(k, s, t+1)$ by the master equation
- we then obtain a formula, in which we can identify two terms $P(k-1, t)$ and $tP(k, t) - P(k, t)$ that we can again substitute

A master equation for $P(k, t+1)$

- with this we get a **master equation for** $P(k,t)$

$$(t+1)P(k,t+1) = P(k-1,t) + \underbrace{tP(k,t) - P(k,t)}_{=(t-1)P(k,t)} + \delta_{k,1}$$

- which we can write as

$$(t+1)P(k,t+1) - (t-1)P(k,t) = P(k-1,t) + \delta_{k,1}$$

- can we say something about the **stationary degree distribution** $P(k)$, i.e.

$$P(k) = \lim_{t\to\infty} P(k,t)$$

Notes:

- with this, we obtain a master equation for the time-dependent degree distribution, which tells us how the degree distribution at the time $t + 1$ relates to the degree distribution in the previous step t
- we are interested in the long-term behavior of the system, i.e. we take study the infinite-time limit for the degree distribution
- note that - in general - one has to be careful, because this approach may neglect finite-time effects that may be significant even for large t if the convergence is slow

Recurrence equation for $P(k)$

- with $P(k) = \lim_{t\to\infty} P(k,t)$ we have

$$\lim_{t\to\infty} (t+1)P(k,t+1) - (t-1)P(k,t) = 2P(k)$$

- this gives us a recurrence equation for the stationary degree distribution

$$P(k) = \frac{1}{2}\left(P(k-1) + \delta_{k,1}\right)$$

- we can solve this recurrence equation iteratively and get

$$P(k) = 2^{-k}$$

Notes:

- by taking the limit for infinite t, we get a recurrence equation, which we can easily solve ietratively
- first we have $P(0) = 0$ since every node has at least degree 1
- from this we can compute $P(1) = \frac{1}{2}(P(0) + 1) = \frac{1}{2}$
- for every $k > 1$, since $\delta_{k,1} = 0$ we have $P(k) = \frac{1}{2}P(k-1)$ which gives us 2^{-k}

Stationary degree distribution

- for $t \to \infty$ we obtain

$$P(k,t) \to 2^{-k} = e^{-k \ln(2)}$$

- degree distribution converges to an **exponentially-shaped geometric distribution**
- for the mean degree $\langle k \rangle$ we obtain

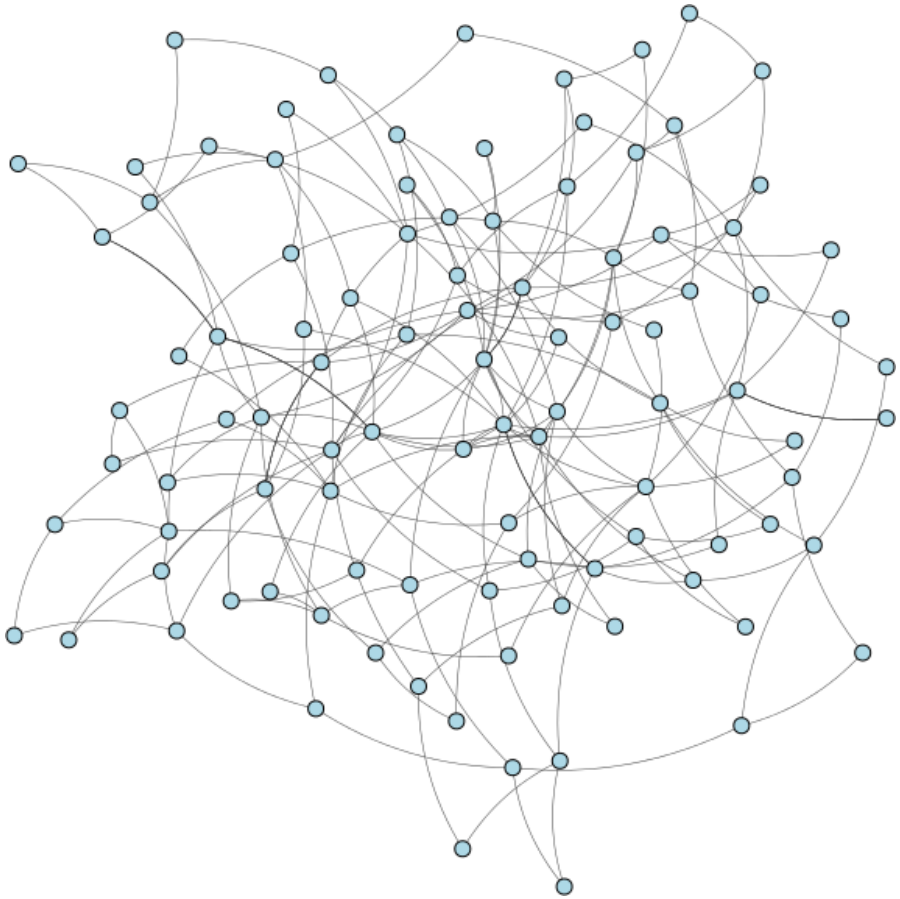
$$\langle k \rangle = \sum_{k=1}^{\infty} k \cdot 2^{-k} = 2$$

Notes:

- we obtain a geometric distribution, which is a member of the family of exponentially-shaped distributions, just like the Poissonian distribution we have derived for the Erdő-Rényi network model

Summary: uniform attachment

- we can **analyse network growth using master equations** for the time-evolution of node degrees
- network growth with uniform attachment gives rise to an **exponentially-shaped degree distribution**
- do you have any intuition about differences to **Erdős-Rényi networks**?



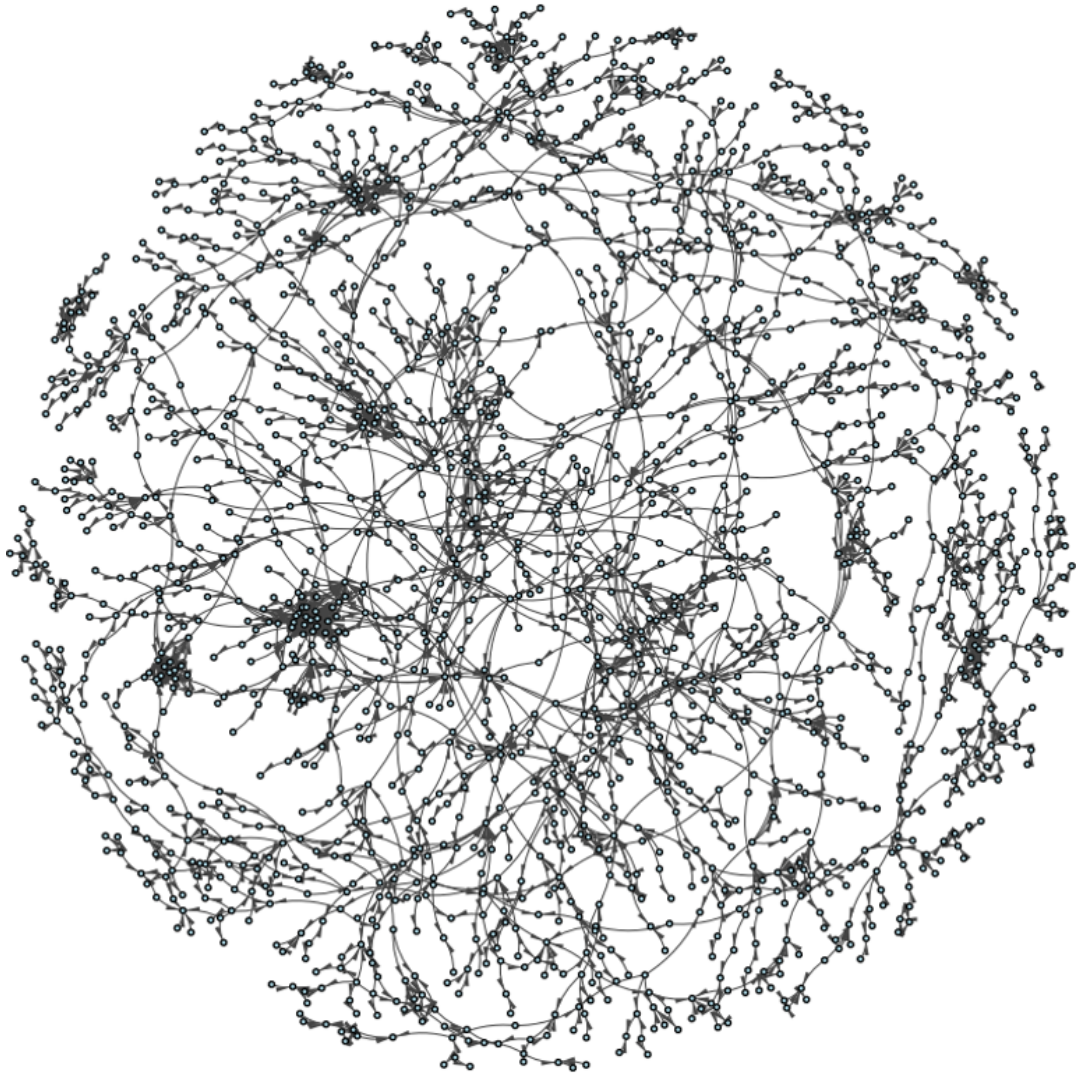
uniform attachment

Notes:

- one additional remark is important: the **uniform attachment model does not generate random Erdős-Rényi networks**
- one can actually show that the average degree of a node added at time s evolves over time as $1 - \ln(\frac{s}{t})$
- this means that - different from the Erdős-Rényi model, there is a dependence of the average degree on the birth time of a node
- a non-equilibrium model that is actually equivalent to the $G(n, p)$ model is one, in which all nodes exist (as a disconnected network without any edges) right from the beginning, and then nodes are activated, adding links to any of the nodes at random. By the very same approach using a master equation, for this model one can show that the network's degree distribution evolves into a Poissonian distribution.

A real growing network: scholarly citations

- we can identify some real-world networks that are formed by growth
- **example:** citation networks
- new articles are published over time and can only cite existing ones
- due to **broad degree distribution** we can rule out uniform attachment model



Notes:

- in citation networks, we observe **broad degree distributions** which cannot be reproduced by the simple uniform attachment model discussed above
- we need another ingredient in the model
- in reality, it is unlikely that papers cite existing paper uniformly at random

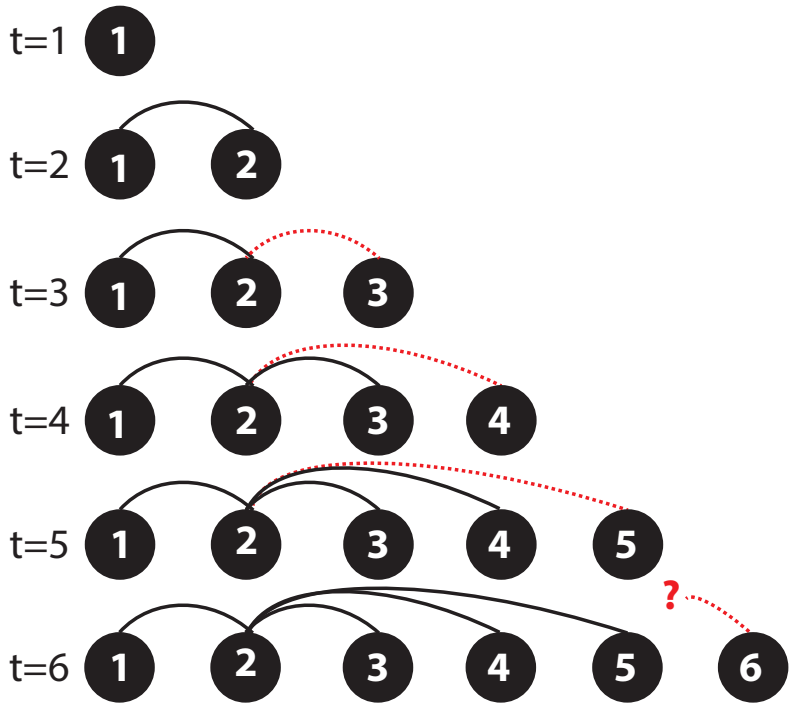
Network growth: preferential attachment

- we extend the attachment rule by a **preference to well-connected nodes**
- node v_{t+1} added at time $t+1$ forms m links to existing nodes
- probability $P(v_i)$ to form link to one of the existing nodes $v_1, \ldots, v_t$

$$P(v_i) = \frac{k_i(t)}{\sum_{j=1}^{t} k_j(t)}$$

where $k_i(t)$ is the degree of node i at time t

- observation: **positive feedback**



at time $t = 6$, node 6 preferentially chooses m well-connected nodes

Notes:

- note that we need formally need to start at time step 2, since at time 1 node 1 will have zero probability to connect to node 0
- in the following, we thus use a simple network consisting of two connected nodes as initial condition
- the model is frequently studied using an initial condition of n_0 fully connected nodes, which is the generalisation of the case studied by us
- this preferential attachment mechanism was first proposed in the context of citation networks by
- Derek J. de Solla Price: “Networks of Scientific Papers”, Science, Vol. 149, pp. 510 - 515, July 1965
- Derek J. de Solla Price: “A general theory of bibliometric and other cumulative advantage processes”, Journal of the Association for Information Science and Technology, Vol. 27, No. 5, pp. 292–306, September 1976
- more than 20 years later it was popularised under the name “preferential attachment” in
- Albert-László Barabási and Reka Albert: “Emergence of Scaling in Random Networks”, Science, Vol. 286, October 1999

Analysis with a differential equation

- we can make the simplifying assumption of continuous time and degrees
- this allows us to give a **differential equation** for the evolution of the degree $k_i(t)$ of a node i at time t

$$\frac{dk_i(t)}{dt} = m \cdot \frac{k_i(t)}{\underbrace{\sum_{j=1}^{t} k_j(t)}_{=2mt}} = \frac{k_i(t)}{2t}$$

- let t_i be the "birth time" of node i, i.e. the time at which the node was added
- since each node initially forms m links, we have the **initial condition**

$$k_i(t_i) = m$$

- we get a **first-order ODE** for every function $k_i(t)$

Notes:

- we could again analyse the stationary degree distribution using a master equation, just like in uniform attachment model
- on purpose, in the following we take a different approach that is common in the physics literature (and the same approach can actually be used to derive the exponential degree distribution of the uniform attachment model)
- we first make the assumption that both time and degrees are continuous
- this continuum limit assumption works in our case, but is generally dangerous
- in the continuum limit, we can write down a differential equation that captures the evolution of node degrees

Evolution of node degrees

- solution of this ODE yields the functions $k_i(t)$ which give degrees of nodes i at time t

$$k_i(t) = m\left(\frac{t}{t_i}\right)^{\frac{1}{2}}$$

- **two observations**
  1. node degrees grow as a square root of time
  2. node degrees are **age-dependent**

- we are interested in $P(k,t)$, i.e. the time-dependent degree distribution, which can be defined as

$$P(k,t) := P(k_i(t) = k)$$

- for a random variable $k_i(t)$ which takes the degree of a randomly chosen node i

Notes:

Time-dependent degree distribution

- consider the probability that a randomly chosen node i has **degree smaller than** k, i.e. $P(k_i(t) < k)$
- using the expression for $k_i(t)$ we can substitute this to

$$P(k_i(t) < k) = P\left(m\left(\frac{t}{t_i}\right)^{\frac{1}{2}} < k\right)$$

- we can solve the expression for t_i as

$$m\left(\frac{t}{t_i}\right)^{\frac{1}{2}} < k \Leftrightarrow m^2\frac{t}{t_i} < k^2 \Leftrightarrow \frac{1}{t_i} < \frac{k^2}{m^2 t} \Leftrightarrow t_i > \frac{m^2 t}{k^2}$$

- with this we have

$$P(k_i(t) < k) = P\left(t_i > \frac{m^2 t}{k^2}\right) = 1 - P\left(t_i \leq \frac{m^2 t}{k^2}\right)$$

Notes:

Distribution of birth times

- we found a relation between the time-dependent degree distribution $P(k_i(t))$ and the **distribution of birth times** $P(t_i)$
- homogeneous growth process $\Rightarrow$ **birth times are uniformly distributed**, i.e. $\forall t' \leq t$

$$P(t_i = t') = \frac{1}{t}$$

- we then get

$$\begin{aligned} P(k_i(t) < k) &= 1 - P\left(t_i \leq \frac{m^2 t}{k^2}\right) \\ &= 1 - \frac{1}{t}\frac{m^2 t}{k^2} = 1 - m^2 k^{-2} = P(k_i < k) \end{aligned}$$

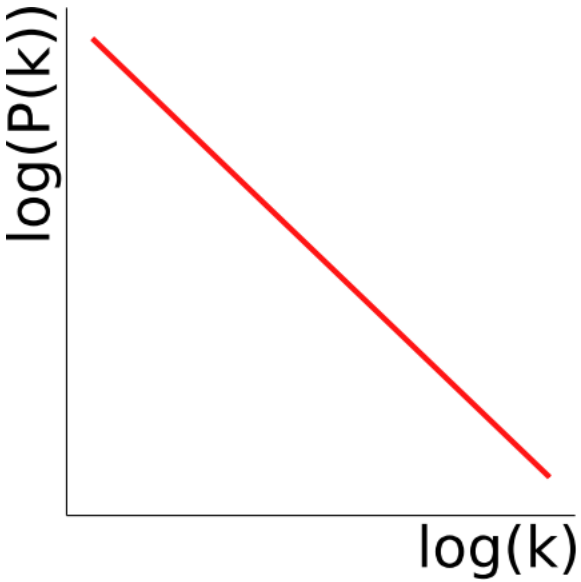
- can we derive $P(k_i = k)$ from $P(k_i < k)$?

Notes:

Stationary degree distribution

- for the stationary degree distribution $P(k) = P(k_i = k)$ we then get

$$P(k_i = k) = \frac{dP(k_i < k)}{dk} = \frac{d}{dk}\left[1 - m^2 k^{-2}\right]$$
$$= 2m^2 k^{-3} \propto k^{-3}$$



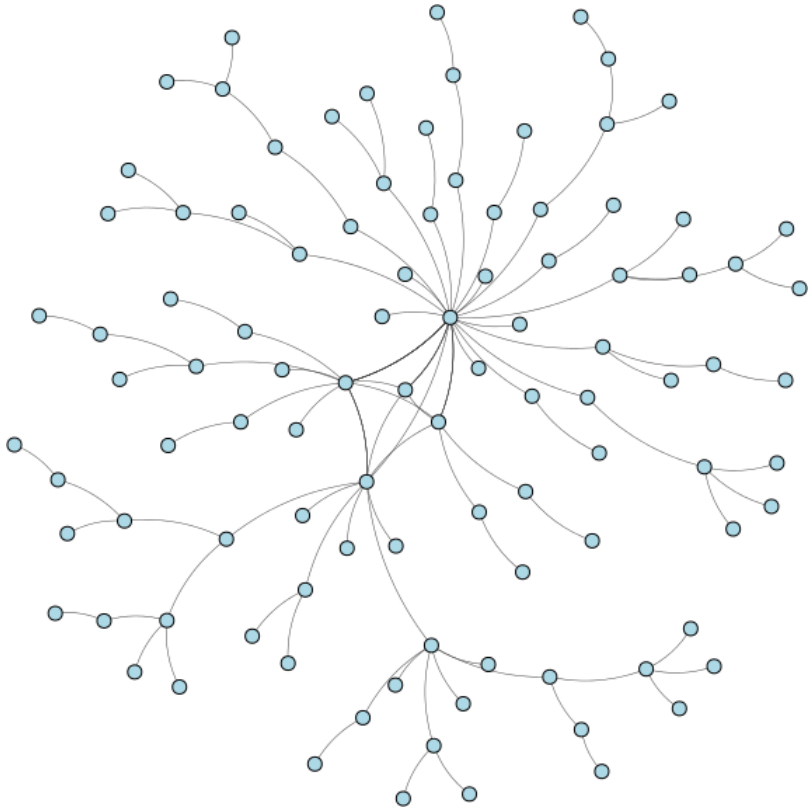
- **degree distribution follows a power law with exponent $\gamma = 3$**
- number of links m added per step is scaling factor of distribution

Notes:

Summary: preferential attachment

- we can **analyse network growth using differential equations** for the time-evolution of node degrees
- network growth with preferential attachment gives rise to a **scale-free degree distribution** network growth

  leads to **first-mover advantage**, enforced via a **Matthew effect**
- what are the **limitations** of this model?



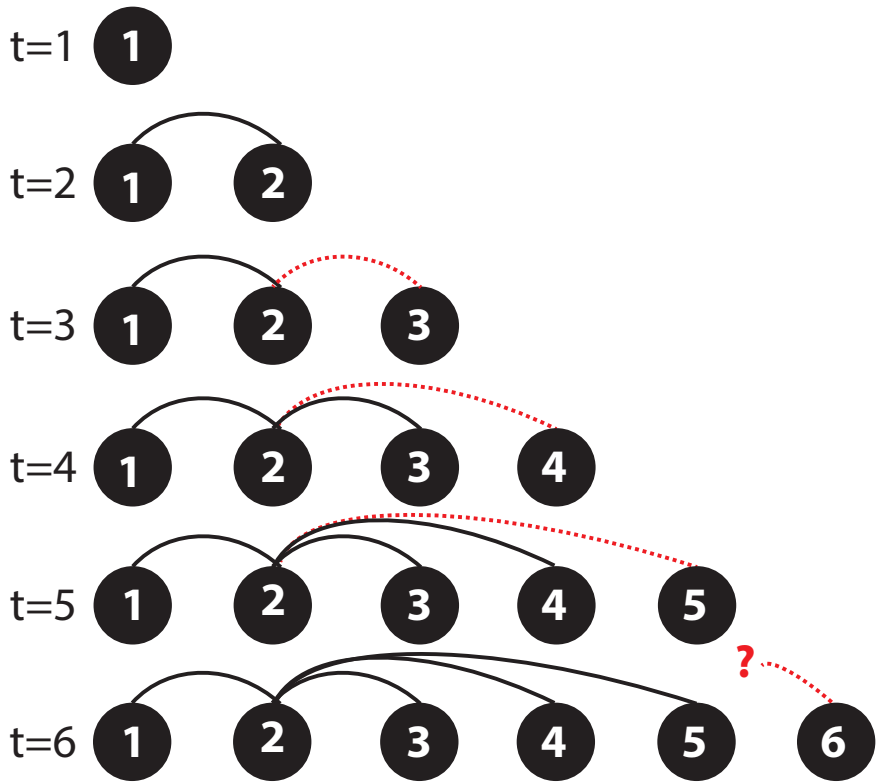
preferential attachment

Notes:

- Why a first-mover advantage? Because the oldest nodes have a high chance to get a larger number of links at the beginning of the growth process.
- those having a slightly higher degree at the beginning of the evolution, are likely to get even more links due to the preferential attachment rule
- the “Matthew effect” refers to a dynamics by which “the rich get richer”, i.e. the degree of those nodes that already have the most links will grow fastest
- the very same mechanism has been introduced in
- G. U. Yule: "A Mathematical Theory of Evolution, based on the Conclusions of Dr. J. C. Willis, F.R.S". Philosophical Transactions of the Royal Society B 213 (402–410): 21–87, 1925
- as well as - in the context of wealth accumulation and city size distribution - in
- Herbert A. Simon: "On a class of skew distribution functions". Biometrika 42 (3–4): 425–440, 1955

Preferential attachment: global knowledge?

- how do new nodes find highly connected nodes?
- new nodes need **global knowledge** of existing network
- preferential attachment is an **observation** rather than a **mechanism**
- do you have any ideas for **local mechanisms** that can explain preferential attachment?

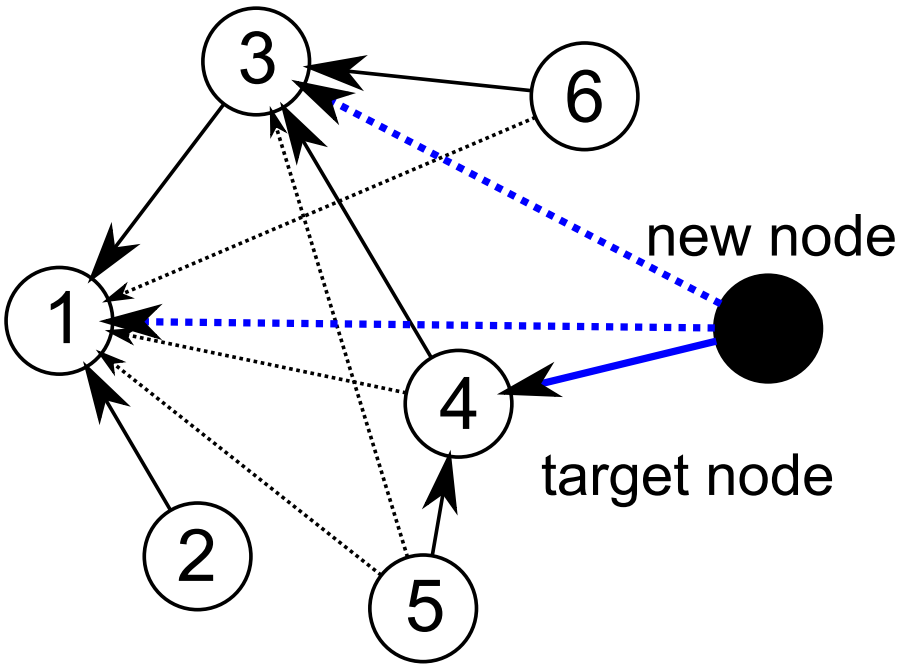


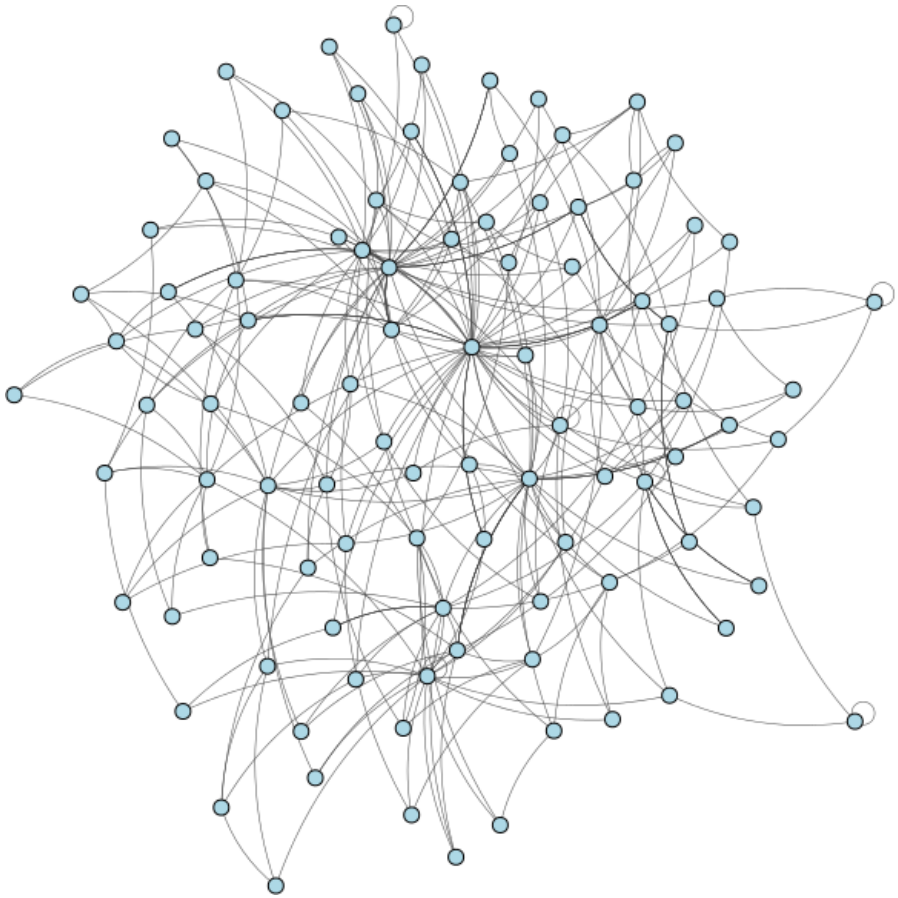
Notes:

- as a model for network formation, simple preferential attachment has the problem that it requires **global knowledge**, because new nodes need knowledge about the degree of all existing nodes in order to make a decision to whom to form a link
- as such, it is not really a good model in terms of describing the actual process of network formation
- can you think of any actual **local mechanisms** that could lead to preferential attachment?

Local mechanism I: copying model

copying model: new nodes form a link to a random node as well as to some (or all) of its neighbours



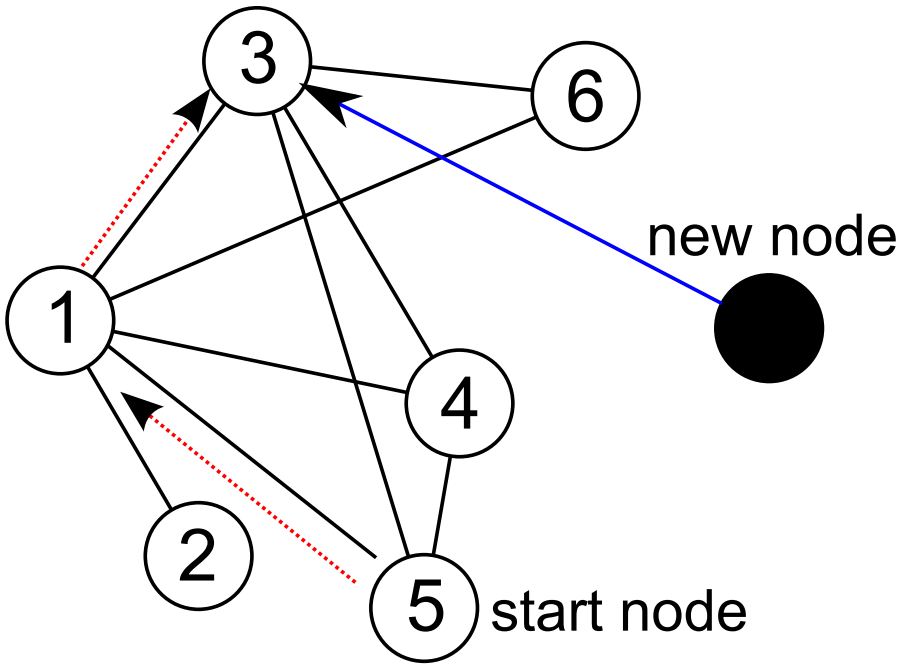


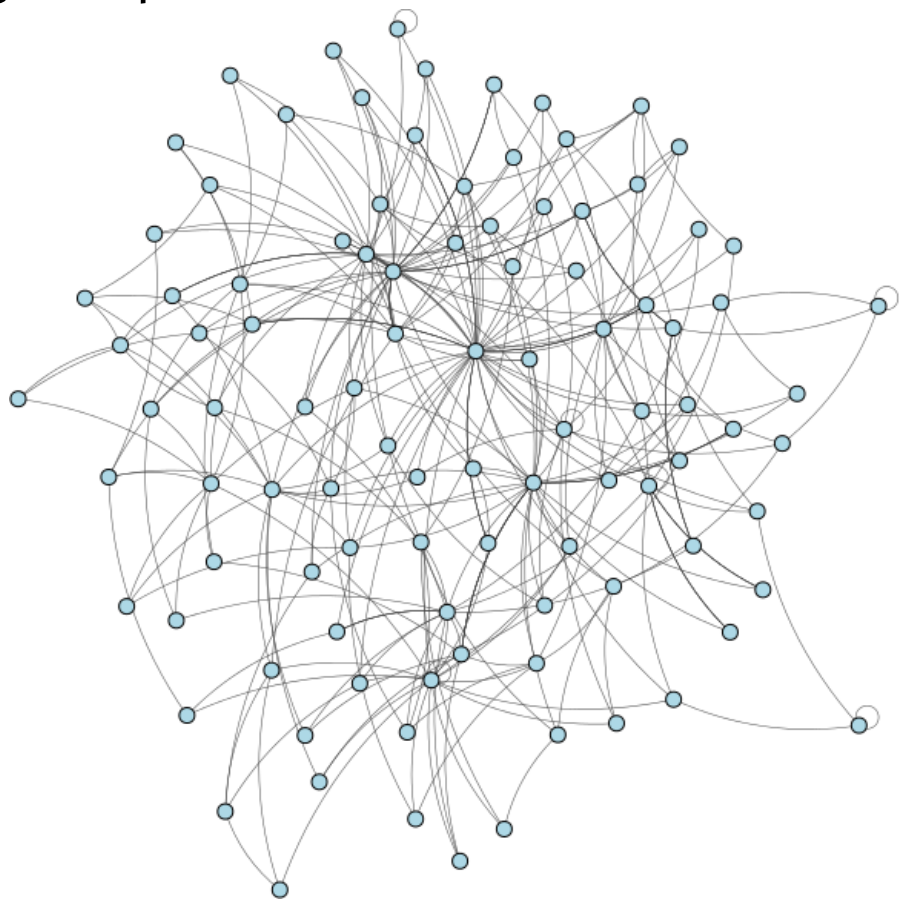
Notes:

- this copying model has been proposed in the context of modeling citation networks in
- P. L. Krapivsky and S. Redner: “Network growth by copying”, Phys. Rev. E 71, 036118, March 2005
- this mechanism leads to preferential attachment: **Why?** ($\Rightarrow$ see Questions slide)

Local mechanism II: random walk model

random walk model: new nodes start random walk at random node and form a link every l steps

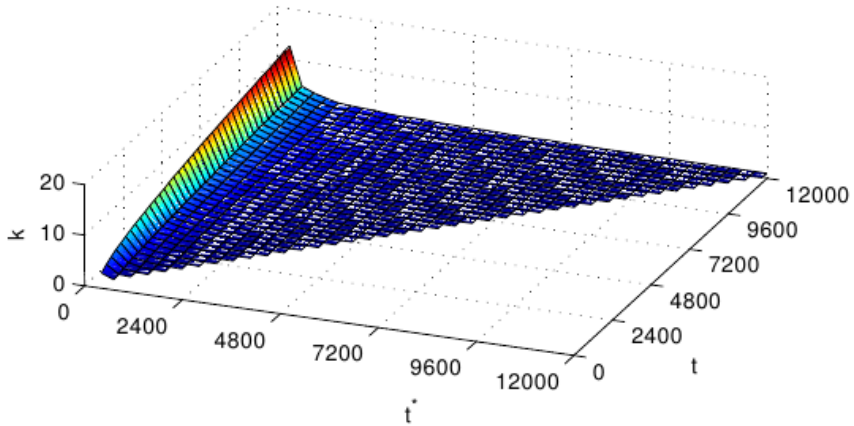




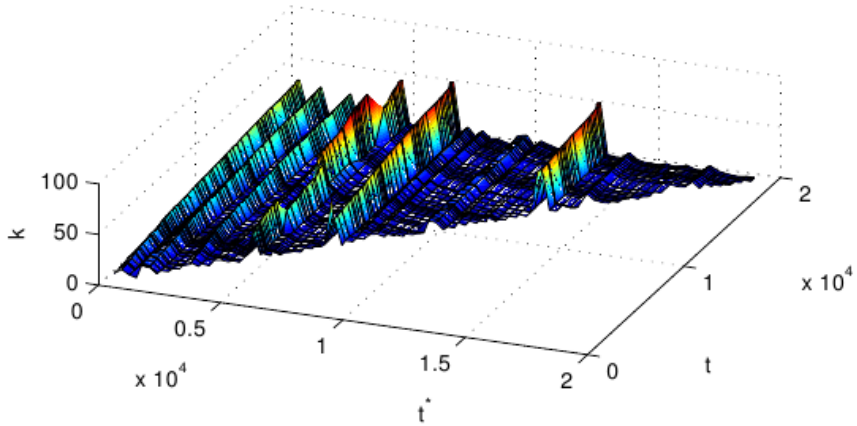
Notes:

- this random walk model has been proposed in
- J. Saramäki and K. Kaski: “Scale-free networks generated by random walkers”, Physica A, Vol. 341, October 2004
- this mechanism leads to preferential attachment: **Why?** ($\Rightarrow$ see Questions slide)
- it is worth noting that this random walk model can produce different clusterings depending on the distance l: small l will lead to higher clustering

Preferential attachment: age vs. degree?



preferential attachment model



empirical software network

- in real-world networks, it matters **which nodes** are most connected
- preferential attachment leads to first-mover advantage: strong correlation between **age** of a node and its **degree**
- do you have an idea for an ingredient that allows late nodes to become highly connected?

Notes:

- the growth of software dependency networks, and the relation between the age of a node and its final degree has been studied in
- Markus M. Geipel: “Dynamics of communities and code in open source software”, ETH Dissertation, Nr. 18480, 2009

Preferential attachment and fitness

- we can include **heterogeneous node fitness**
- fitness $\eta_s \in [0, 1]$ of node s drawn from distribution Q
- for each new node, probability $P(v_i)$ to form link to one of the existing nodes $v_0, \ldots, v_{t-1}$:

$$P(v_i) = \frac{\eta_i k_i(t)}{\sum_{j=0}^{t-1} \eta_j k_j(t)}$$

- **uniform fitness distribution** $Q(\eta) \equiv p \Rightarrow$ **scale-free network** with exponent $\gamma = 2.255$
- degrees of nodes with high fitness shows less pronounced time-dependence
- **exponential fitness distribution** $Q(\eta) \propto e^{-\eta} \Rightarrow$ network with **stretched exponential degree distribution**

Notes:

- this model that combines node fitness with preferential attachment has been introduced in
- G. Bianconi and A.-L. Barabási: “Competition and multiscaling in evolving networks”, Europhysics Letters, Vol. 54, No. 4, 2001

Scale-free network $\neq$ preferential attachment

- **preferential attachment $\Rightarrow$ scale-free networks**

NewScientist

- **scale-free networks $\not\Rightarrow$ preferential attachment**
  - we already discussed three other equilibrium mechanisms
  - preferential attachment is **one of several mechanisms** leading to broad degree distributions

---

All the world's a net: What do the proteins in your body the Internet, a cool collection of atoms and sexual networks have in common? One man thinks he has the answer, and it's going to transform the way we view the world.(Brief Article)

by David Cohen

article header from *New Scientist*, April 2002

Notes:

- today, we have seen that (different variants of) preferential attachmentcan explain the formation of scale-free networks or - in general - the formation of networks with broad degree distributions
- this means that **preferential attachment is a sufficient condition for power law degree distributions**
- it may seem trivial, but it is important to highlight that **preferential attachment is not a necessary condition**, i.e. not all scale-free networks result from preferential attachment
- already in the last lecture, we have seen alternative mechanisms that result in scale-free structures
- especially, in the early literature of the early 2000's, preferential attachment has been **THE** dominant explanation for broad degree distributions
- today, network scientists have a more differentiated perspective on the importance of this particular mechanism (as well as about how wide-spread it actually is in reality).

Self-Study Questions

1. What is a master equation? How can it be used to analyse network growth models?
2. Consider the uniform attachment model and write down the master equation for $P(k, s, t)$ and $P(k, t)$
3. Consider the preferential attachment model and write down the differential equation for the evolution of $k_i(t)$
4. What are the final degree distributions for a growth model with uniform attachment and preferential attachment?
5. In the preferential attachment model, how does the number of links m added in each step affect the final degree distribution?
6. Explain why the random walk model gives rise to preferential attachment. Consider lecture 11 and think about the stationary distribution of a random walker in undirected networks.
7. Explain why the copying model gives rise to preferential attachment. Consider the friendship paradox.

Notes: