Installing Gephi

- Current stable version 0.9.2 but you might want to use an older version for compatibility with plugins
- Windows, Mac OS X, and Linux are supported
- Requires Java JRE 7 or 8
- Memory requirements:

<table>
<thead>
<tr>
<th>Network size (nodes + edges)</th>
<th>~Memory suggested</th>
</tr>
</thead>
<tbody>
<tr>
<td>~1000</td>
<td>128 MB</td>
</tr>
<tr>
<td>~10,000</td>
<td>512 MB</td>
</tr>
<tr>
<td>~100,000</td>
<td>2 GB</td>
</tr>
<tr>
<td>~1 M</td>
<td>&gt;8 GB</td>
</tr>
</tbody>
</table>

Sample Gephi datasets are available at: https://github.com/gephi/gephi/wiki/Datasets
Search for Class 1880-1881 and import it into Gephi

By default, Gephi is configured to start with 512 Mb of memory allocated to JVM. This might not be enough for larger graphs. To allocate more memory, increase the value of the `–Xmx` option (e.g., set `-Xmx1024m`) in `gephi.conf` configuration file. Then close and reopen Gephi for the new options to take effect.
Sample Network Analysis in Gephi

Sample Gephi datasets are available at:

https://github.com/gephi/gephi/wiki/Datasets

Search for Class 1880-1881 and import it into Gephi
The origins of preferential attachment
Link selection model -- perhaps the simplest example of a local or random mechanism capable of generating preferential attachment.

**Growth:** at each time step we add a new node to the network.

**Link selection:** we select a link at random and connect the new node to one of nodes at the two ends of the selected link.

To show that this simple mechanism generates linear preferential attachment, we write the probability that the node at the end of a randomly chosen link has degree $k$ as

$$q_k = Ck p_k$$

In (5.26) $C$ can be calculated using the normalization condition $\Sigma q_k = 1$, obtaining $C = 1/\langle k \rangle$. Hence the probability to find a degree-$k$ node at the end of a randomly chosen link is

$$q_k = \frac{k p_k}{\langle k \rangle}.$$
Section 9 Originators of preferential attachments

1953

In An Informal Theory of the Statistical Structure of Languages [26] Benoît Mandelbrot proposes optimization as the origin of power laws.

1955

In On a Class of Skew Distribution Functions Herbert Simon [6] proposes randomness as the origin of power laws and dismisses Mandelbrot's claim that power law are rooted in optimization.

3

Mandelbrot publishes a comment on Simon's paper [27] writing:

Simon's model is analytically circular...

1959

Dr. Mandelbrot's principal and mathematical objections to the model are shown to be unfounded

4

The essence of Simon's lengthy reply a year later is well summarized in its abstract [28].

5

In a 19 page response entitled Final Note, Mandelbrot states [29]:

...Most of Simon's (1960) reply was irrelevant.

1961

This present 'Reply' refutes the almost entirely new arguments introduced by Dr. Mandelbrot in his 'Final Note'...

6

Simon's subsequent Reply to 'Final Note' by Mandelbrot does not concede [30].

7

In the creatively titled Post Scriptum to Final Note Mandelbrot [31] writes

My criticism has not changed since I first had the privilege of commenting upon a draft of Simon.

1961

Dr. Mandelbrot has proposed a new set of objections to my 1955 models of Yule distributions. Like earlier objections, these are invalid.

8

Simon's final reply ends but does not resolve the debate [32].
György Pólya [1887-1985] was a Hungarian mathematician. His work on the 'Pólya process' [2] led to the concept of preferential attachment in 1923. This concept describes how certain elements in a network tend to connect to already well-connected elements, explaining the power-law distribution of degrees in real-world networks.

Robert Gibar [1904-1980] proposed the idea of proportional growth to explain how the size and growth rate of a company are independent. In 1955, he published a paper on proportional growth, which laid the groundwork for the concept of preferential attachment in the context of social networks.

Herbert Alexander Simon [1916-2001] used preferential attachment to explain the fat-tailed nature of the distributions describing city sizes, word frequencies, or the number of papers published by scientists [6].

Geoffrey Pyke [1902-1950] was a British statistician who worked on preferential attachment to explain the distribution of wealth in society [5].

Derek de Solla Price [1922-1983] used preferential attachment to explain the citation statistics of scientific publications, referring to it as cumulative advantage [7].

Barabási & Albert [1972] introduced the term preferential attachment in the context of networks to explain the origin of their power-law degree distribution.
MECHANISMS RESPONSIBLE FOR PREFERENTIAL ATTACHMENT

1. Copying mechanism
   directed network
   select a node and an edge of this node
   attach to the endpoint of this edge

2. Walking on a network
   directed network
   the new node connects to a node, then to every
   first, second, … neighbor of this node

3. Attaching to edges
   select an edge
   attach to both endpoints of this edge

4. Node duplication
   duplicate a node with all its edges
   randomly prune edges of new node
Section 9

**Copying model**

(a) **Random Connection:** with probability $p$ the new node links to $u$.

(b) **Copying:** with probability $p$ we randomly choose an outgoing link of node $u$ and connect the new node to the selected link's target. Hence the new node “copies” one of the links of an earlier node.

(a) the probability of selecting a node is $1/N$.
(b) is equivalent with selecting a node linked to a randomly selected link. The probability of selecting a degree-$k$ node through the copying process of step (b) is $k/2L$ for undirected networks.

The likelihood that the new node will connect to a degree-$k$ node follows preferential attachment.

**Social networks:** Copy your friend’s friends.

**Citation Networks:** Copy references from papers we read.

**Protein interaction networks:** gene duplication,
Preferential Attachment!

$$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) \sim \frac{\Delta k_i}{\Delta t}$$

For given $\Delta t$: $\Delta k \propto \Pi(k)$

$k$ vs. $\Delta k$: linear increase in the # of links

S. Cerevisiae PIN: proteins classified into 4 age groups

SUMMARY: PROPERTIES OF THE BA MODEL

- Nr. of nodes: \( N = t \)
- Nr. of links: \( L = m \, t \)
- Average degree: \( \langle k \rangle = \frac{2L}{N} \rightarrow 2m \)
- Degree dynamics: \( k_i(t) = m \left( \frac{t}{t_i} \right)^\beta \), \( \beta = \frac{1}{2} \) \( \beta \): dynamical exponent
- Degree distribution: \( P(k) \sim k^{-\gamma} \), \( \gamma = 3 \) \( \gamma \): degree exponent
- Average Path Length: \( l \approx \frac{\ln N}{\ln \ln N} \)
- Clustering Coefficient: \( C \sim \frac{(\ln N)^2}{N} \)

The network grows, but the degree distribution is stationary.
Can we change the degree exponent?
Section 9

Optimization model

\[ C_i = \min_j[\delta d_{ij} + h_j] \]
**Section 9  Optimization model**

\[ C_i = \min_j [\delta d_{ij} + h_j] \]

The vertical boundary of the star configuration is at \( \delta = (1/2)^{1/2} \). This is the inverse of the maximum distance between two nodes on a square lattice with unit length, over which the model is defined. Therefore, if \( \delta < (1/2)^{1/2} \), for any new node \( \delta d_{ij} < 1 \) and the cost (5.28) of connecting to the central node is \( C_i = \delta d_{ij} + 0 \), always lower than connecting to any other node at a cost of \( f(i,j) = \delta d_{ij} + 1 \). Therefore, for \( \delta < (1/2)^{1/2} \) all nodes connect to node 0 (star-and-spoke network (c)).
Section 9

Optimization model

\[ C_i = \min_j[\delta d_{ij} + h_j] \]

The oblique boundary of the scale-free regime is \( \delta = N^{1/2} \). Indeed, if nodes are placed randomly on the unit square, then the typical distance between neighbors decreases as \( N^{-1/2} \). Hence, if \( d_{ij} \sim N^{-1/2} \) then \( \delta d_{ij} \geq h_j \) for most node pairs. Typically the path length to the central node \( h_j \) grows slower than \( N \) (in small-world networks \( h \sim \log N \)). Therefore, \( C_i \) is dominated by the \( \delta d_{ij} \) term and the smallest \( C_i \) is achieved by minimizing the distance-dependent term. Note that, strictly speaking, the transition only occurs in the \( N \to \infty \) limit. In the white regime we lack an analytical form for the degree distribution.
Section 9 Optimization model

\[ C_i = \min_j[\delta d_{ij} + h_j] \]
Diameter and clustering coefficient
$D \sim \frac{\log N}{\log \log N}$
Section 10  

Clustering coefficient

Reminder: for a random graph we have:

\[ C_{\text{rand}} = \frac{\langle k \rangle}{N} \sim N^{-1} \]

What is the functional form of \( C(N) \)?

\[ C = \frac{m}{8} \frac{(\ln N)^2}{N} \]

Konstantin Klemm, Victor M. Eguiluz,
Growing scale-free networks with small-world behavior,
Denote the probability to have a link between node $i$ and $j$ with $P(i,j)$.
The probability that three nodes $i,j,l$ form a triangle is $P(i,j)P(i,l)P(j,l)$.

The expected number of triangles in which a node $l$ with degree $k_l$ participates is thus:

$$Nr_1(\Delta) = \sum_{i=1}^{N} \sum_{j=1}^{N} d_i d_j P(i,j)P(i,l)P(j,l)$$

We need to calculate $P(i,j)$. 

$$C = \frac{Nr(\Delta)}{k(k-1)}$$

$$C = \frac{2}{6}$$
Calculate $P(i,j)$. 

Node $j$ arrives at time $t_j = j$ and the probability that it will link to node $i$ with degree $k_i$ already in the network is determined by preferential attachment:

$$P(i,j) = m \prod_{l=1}^{j-1} k_i(l) = m \frac{k_i(j)}{2mj}$$

Where we used that the arrival time of node $j$ is $t_j = j$ and the arrival time of node is $t_i = i$:

$$P(i,j) = \frac{m}{2} (ij)^{-1/2}$$

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

$$N_{r_{i}}(\Delta) = \int_{i=1}^{N} di \int_{j=1}^{N} dj P(i,j)P(i,j)P(j,l) = \frac{m^3}{8l} \int_{i=1}^{N} di \int_{j=1}^{N} dj (ij)^{-1/2} (il)^{-1/2} (jl)^{-1/2} = \frac{m^3}{8l} \int_{i=1}^{N} di \int_{j=1}^{N} dj = \frac{m^3}{8l} (\ln N)^2$$

$$C = \frac{m}{8l} \frac{(\ln N)^2}{k_i(k_i - 1)/2}$$

$$k_i = m \left( \frac{N}{l} \right)^{1/2}$$

Which is the degree of node $l$ at current time, at time $t = N$.

Let us approximate:

$$k_i(k_i - 1) \approx k_i^2 = m^2 \frac{N}{l}$$

There is a factor of two difference... Where does it come from?
Evolving network models
The BA model is only a minimal model. Makes the simplest assumptions:

- linear growth
- linear preferential attachment

Does not capture
- variations in the shape of the degree distribution
- variations in the degree exponent
- the size-independent clustering coefficient

Hypothesis:
The BA model can be adapted to describe most features of real networks.

We need to incorporate mechanisms that are known to take place in real networks: addition of links without new nodes, link rewiring, link removal; node removal, constraints or optimization.
BA ALGORITHM WITH DIRECTED EDGES

(the simplest way to change the degree exponent)

Undirected BA network: $\beta=1/2; \quad \gamma=3$

Directed BA network: $\beta=1; \quad \gamma=2$

Undirected BA network: $P_{\text{out}}(k_{\text{out}}) = \delta(k_{\text{out}}-m)$

Directed BA network: $P_{\text{in}}(k_{\text{in}}) \sim k^{-2}$

$\frac{\partial k_i}{\partial t} \propto \Pi(k_i) = A \frac{k_i}{\sum_j k_j} = \frac{k_i}{t}$

$\sum_j k_j = 2t$

$\sum_j k_j = t$
EXTENDED MODEL: Other ways to change the exponent

- prob. $p$: internal links
- prob. $q$: link deletion
- prob. $1-p-q$: add node

$$P(k) \sim (k + \kappa(p,q,m))^{-\gamma(p,q,m)}$$

$\gamma \in [1, \infty)$
$P(k) \sim (k + \kappa(p,q,m))^{-\gamma(p,q,m)} \quad \gamma \in [1, \infty)$

→ **Predicts a small-k cutoff**

→ a correct model should predict all aspects of the degree distribution, not only the degree exponent.

→ Degree exponent is a continuous function of $p, q, m$

**Actor network**

- prob. $p$ : internal links
- prob. $q$ : link deletion
- prob. $1-p-q$ : add node

$p = 0.937$

$m = 1$

$\kappa = 31.68$

$\gamma = 3.07$
\[ P(k) \sim (k + \kappa(p,q,m))^{-\gamma(p,q,m)} \quad \gamma \in [1, \infty) \]

\rightarrow \text{Predicts a small-k cutoff}
\rightarrow \text{a correct model should predict all aspects of the degree distribution, not only the degree exponent.}
\rightarrow \text{Degree exponent is a continuous function of } p, q, m

\begin{itemize}
  \item prob. p : internal links
  \item prob. q : link deletion
  \item prob. 1-p-q : add node
\end{itemize}
• Non-linear preferential attachment:

\[
\Pi(k) = \frac{k^\alpha}{\sum_i k_i^\alpha}
\]

→ \(P(k)\) does not follow a power law for \(\alpha \neq 1\)

\(\Rightarrow \alpha < 1\) : stretch-exponential

\[P(k) \approx \exp\left(-\left(\frac{k}{k_0}\right)^\beta\right)\]

\(\Rightarrow \alpha > 1\) : no-scaling (\(\alpha > 2\) : “gelation”)

BA model: $k=0$ nodes cannot acquire links, as $\Pi(k=0)=0$ (the probability that a new node will attach to it is zero)

$$\Pi(k) \approx A + k^\alpha, \quad \alpha \leq 1$$

$A$ - initial attractiveness

Initial attractiveness shifts the degree exponent:

$$\gamma_{in} = 2 + \frac{A}{m}$$

Note: the parameter $A$ can be measured from real data, being the rate at which $k=0$ nodes acquire links, i.e. $\Pi(k=0)=A$

• Finite lifetime to acquire new edges

L. A. N. Amaral et al., PNAS 97, 11149 (2000)

• Gradual aging:

$$\Pi(k_i) \propto k_i(t - t_i)^{-\nu}$$

$\nu$ increases with $\nu$

THE LAST PROBLEM: HIGH, SYSTEM-SIZE INDEPENDENT C(N)

Pathlength Clustering Degree Distr.

\[ P(k) \sim k^{-\gamma} \]

\[ C(k) = \text{const} \]

Regular network

\[ l \approx N^{1/L} \]

Erdos-Renyi

\[ l_{\text{rand}} \approx \frac{\log N}{\log \langle k \rangle} \]

\[ C_{\text{rand}} = p = \frac{\langle k \rangle}{N} \]

\[ P(k) = \delta(k-k_d) \]

Watts-Strogatz

\[ l_{\text{rand}} \approx \frac{\log N}{\log \langle k \rangle} \]

\[ C \sim \text{const} \]

Exponential

Barabasi-Albert

\[ l \approx \frac{\ln N}{\ln \ln N} \]

\[ C \sim \frac{(\ln N)^2}{N} \]

\[ P(k) \sim k^{-\gamma} \]
A MODEL WITH HIGH CLUSTERING COEFFICIENT

- Each node of the network can be either active or inactive.
- There are $m$ active nodes in the network in any moment.
  1. Start with $m$ active, completely connected nodes.
  2. Each timestep add a new node (active) that connects to $m$ active nodes.
  3. Deactivate one active node with probability: $P_d(k_i) \propto (a + k_j)^{-1}$

$\Pi(k) \approx a + k$

$P(k) \approx k^{-2-a/m}$

$C \to C^* \text{ when } N \to \infty$

Linear growth, linear pref. attachment  
\[ \gamma = 3 \]

Nonlinear preferential attachment  
\[ \Pi(k_i) \sim k_i^\alpha \]
no scaling for \( \alpha \neq 1 \)

Asymptotically linear pref. attachment  
\[ \Pi(k_i) \sim a_k k_i \text{ as } k_i \to \infty \]
\[ \gamma \to 2 \text{ if } a_m \to 0 \]
\[ \gamma \to \infty \text{ if } a_m \to \infty \]

Initial attractiveness  
\[ \Pi(k_i) \sim A + k_i \]
\[ \gamma = 2 \text{ if } A = 0 \]
\[ \gamma \to \infty \text{ if } A \to \infty \]

Accelerating growth  
\[ \langle k \rangle \sim t^\theta \]
constant initial attractiveness  
\[ \gamma = 1.5 \text{ if } \theta \to 1 \]
\[ \gamma \to 2 \text{ if } \theta \to 0 \]

Internal edges with probab. \( p \)  
\[ \gamma = 2 \text{ if } \]
\[ q = \frac{1-p+m}{1+2m} \]

Rewiring of edges with probab. \( q \)  
\[ \gamma \to \infty \text{ if } p, q, m \to 0 \]

\( c \) internal edges  
\[ \gamma = 2 \text{ if } c \to \infty \]
\[ \gamma \to \infty \text{ if } c \to -1 \]

Gradual aging  
\[ \Pi(k_i) \sim k_i (t-t_i)^{-\nu} \]
\[ \gamma \to 2 \text{ if } \nu \to -\infty \]
\[ \gamma \to \infty \text{ if } \nu \to 0 \]

Multiplicative node fitness  
\[ \Pi_i \sim \eta_i k_i \]
\[ P(k) \sim \frac{k^{-1-c}}{\ln(k)} \]

Edge inheritance  
\[ P(k_{in}) = \frac{d}{k_{in}^2 \ln(ak_{in})} \]

Copying with probab. \( p \)  
\[ \gamma = (2-p)/(1-p) \]

Redirection with probab. \( r \)  
\[ \gamma = 1 + 1/r \]

Walking with probab. \( r \)  
\[ \gamma = 2 \text{ for } p > p_c \]

Attaching to edges
\[ \gamma = 3 \]
\[ \gamma_{in} = 2 + p \lambda \]
\[ \gamma_{out} = 1 + (1-p)^{-1} + \mu p/(1-p) \]

Barabási and Albert, 1999
Krapivsky, Redner, and Leyvraz, 2000
Krapivsky, Redner, and Leyvraz, 2000
Dorogovtsev, Mendes, and Samukhin, 2000a, 2000b
Dorogovtsev and Mendes, 2001a
Dorogovtsev and Mendes, 2000c
Dorogovtsev and Mendes, 2000b
Albert and Barabási, 2000
Dorogovtsev and Mendes, 2000c
Bianconi and Barabási, 2001a
Kumar et al., 2000a, 2000b
Krapivsky and Redner, 2001
Vázquez, 2000
Dorogovtsev, Mendes, and Samukhin, 2001a
Krapivsky, Rodgers, and Redner, 2001
The network grows, but the degree distribution is stationary.
Consequently, the modeling philosophy behind the model is simple: *to understand the topology of a complex system, we need to describe how it came into being.*

The network grows, but the degree distribution is stationary.
### Section 11: Summary

<table>
<thead>
<tr>
<th>Number of Nodes</th>
<th>( N = t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Links</td>
<td>( N = mt )</td>
</tr>
<tr>
<td>Average Degree</td>
<td>( \langle k \rangle = 2m )</td>
</tr>
<tr>
<td>Degree Dynamics</td>
<td>( k(t) = m \left( \frac{t}{t_i} \right)^\beta )</td>
</tr>
<tr>
<td>Dynamical Exponent</td>
<td>( \beta = 1/2 )</td>
</tr>
<tr>
<td>Degree Distribution</td>
<td>( p_k \sim k^\gamma )</td>
</tr>
<tr>
<td>Degree Exponent</td>
<td>( \gamma = 3 )</td>
</tr>
<tr>
<td>Average Distance</td>
<td>( \langle d \rangle \sim \log N / \log \log N )</td>
</tr>
<tr>
<td>Clustering Coefficient</td>
<td>( \langle C \rangle \sim (\ln N)^2 / N )</td>
</tr>
</tbody>
</table>

- The model predicts \( \gamma = 3 \) while the degree exponent of real networks varies between 2 and 5 (Table 4.2).

- Many networks, like the WWW or citation networks, are directed, while the model generates undirected networks.

- Many processes observed in networks, from linking to already existing nodes to the disappearance of links and nodes, are absent from the model.

- The model does not allow us to distinguish between nodes based on some intrinsic characteristics, like the novelty of a research paper or the utility of a webpage.

- While the Barabási-Albert model is occasionally used as a model of the Internet or the cell, in reality it is not designed to capture the details of any particular real network. It is a minimal, proof of principle model whose main purpose is to capture the basic mechanisms responsible for the emergence of the scale-free property. Therefore, if we want to understand the evolution of systems like the Internet, the cell or the WWW, we need to incorporate the important details that contribute to the time evolution of these systems, like the directed nature of the WWW, the possibility of internal links and node and link removal.
1. There is no universal exponent characterizing all networks.

2. Growth and preferential attachment are responsible for the emergence of the scale-free property.

3. The origins of the preferential attachment is system-dependent.

4. Modeling real networks:
   - identify the microscopic processes that take place in the system
   - measure their frequency from real data
   - develop dynamical models that capture these processes.

5. If the model is correct, it should correctly predict not only the degree exponent, but both small and large k-cutoffs.
Philosophical change in network modeling:

ER, WS models are static models – the role of the network modeler is to cleverly place the links between a fixed number of nodes to that the network topology mimic the networks seen in real systems.

BA and evolving network models are dynamical models: they aim to reproduce how the network was built and evolved. Thus their goal is to capture the network dynamics, not the structure. → as a byproduct, you get the topology correctly.

LESSONS LEARNED: evolving network models
Nodes: proteins
Links: physical interactions (binding)

Puzzling pattern:
Hubs tend to link to small degree nodes. Why is this puzzling?

In a random network, the probability that a node with degree \( k \) links to a node with degree \( k' \) is:

\[
p_{k,k'} = \frac{kk'}{2L}
\]

\( k=50, k'=13, N=1,458, L=1746 \)

\( p_{50,13} = 0.15 \)  Yet, we see many links between degree 2 and 1 links, and no links between the hubs.

\( p_{2,1} = 0.0004 \)