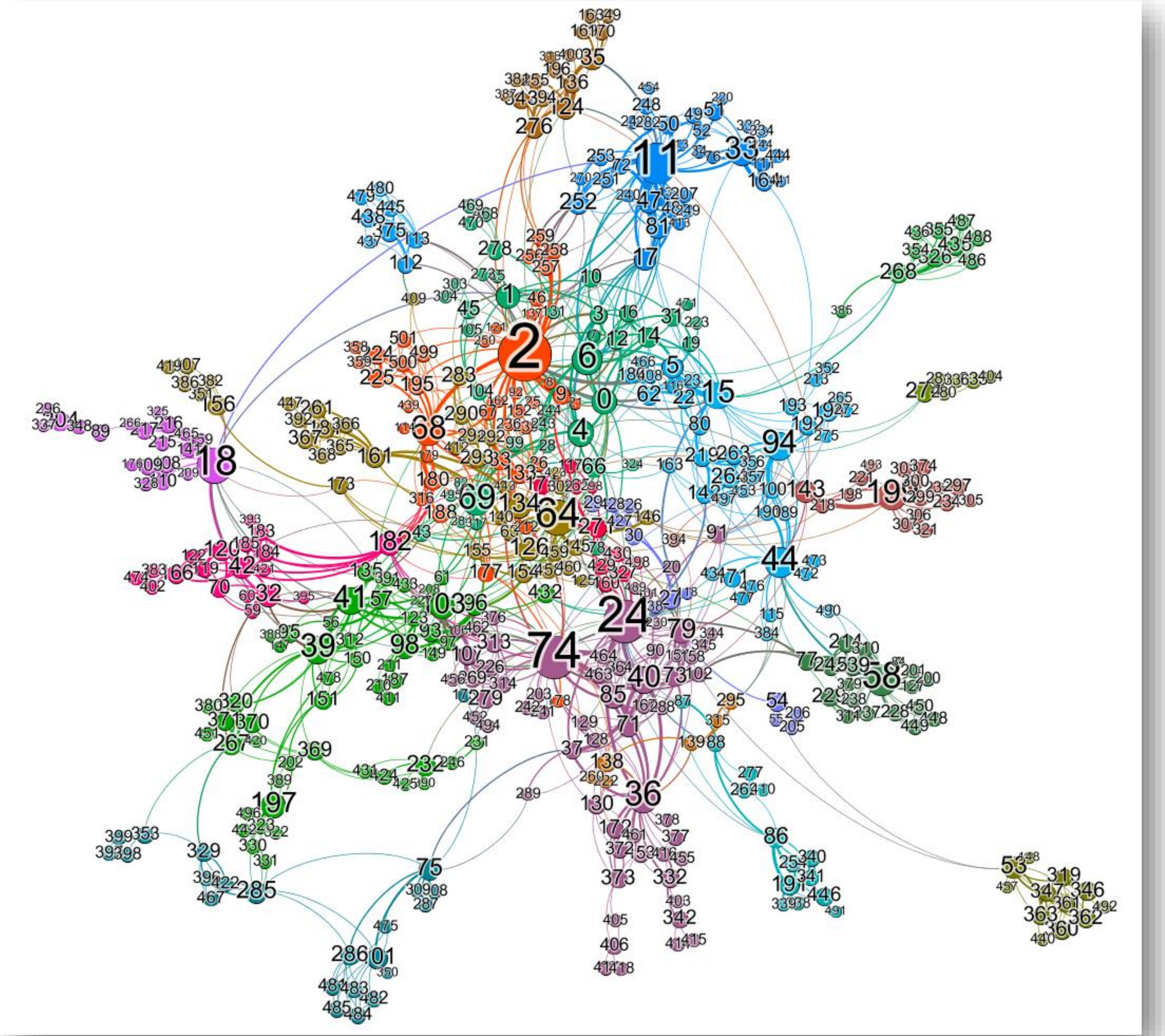
DATA ANALYSIS II

Community Network Models

2021/22



References

- Barabási, L-A. (2016) *Network Science*. Cambridge University Press. [Communities: http://networksciencebook.com/chapter/9]

- Holme, P., Kim, B. J. (2002). *Growing scale-free networks with tunable clustering*. Physical review E, 65(2), 026107. [https://arxiv.org/pdf/cond-mat/0110452.pdf]

- Bianconi, G., Darst, R. K., Iacovacci, J., Fortunato, S. (2014). *Triadic closure as a basic generating mechanism of communities in complex networks*. Physical Review E, 90(4), 042806. [https://arxiv.org/pdf/1407.1664.pdf]

What the community is?

- Connectedness and Density Hypothesis: *A community is a locally dense connected subgraph in a network.*

- Maximum Cliques: *A community is as group of individuals whose members all know each other.* In graph theoretic terms this means that a community is a complete subgraph, or a clique.

  - While triangles are frequent in networks, larger cliques are rare.

  - Requiring a community to be a complete subgraph may be too restrictive, missing many other legitimate communities.

Strong and weak communities

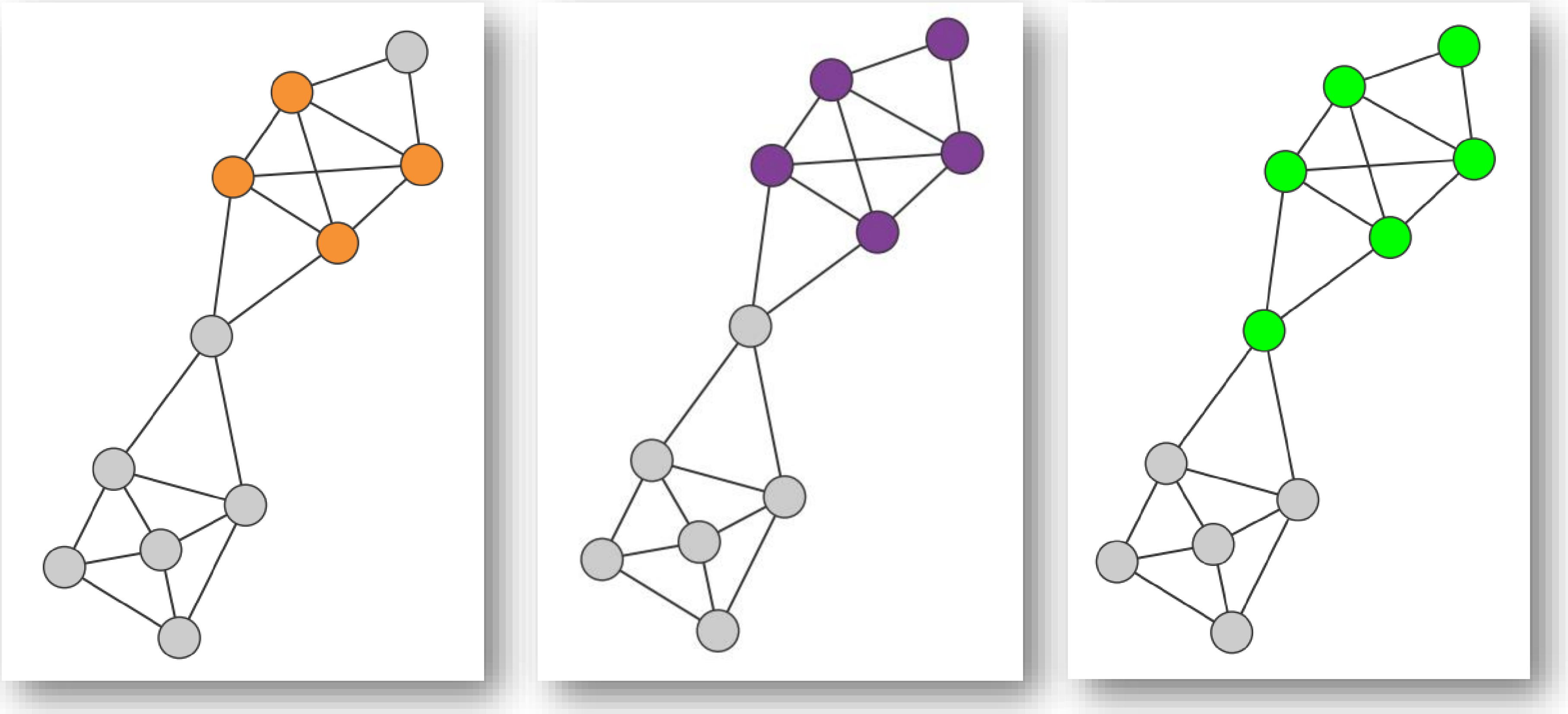
- C is a *strong community* if each node within C has more links within the community than with the rest of the graph. Specifically, a subgraph C forms a strong community if for each node $i \in C$,

$$k_i^{\mathrm{int}}(C) > k_i^{\mathrm{ext}}(C).$$

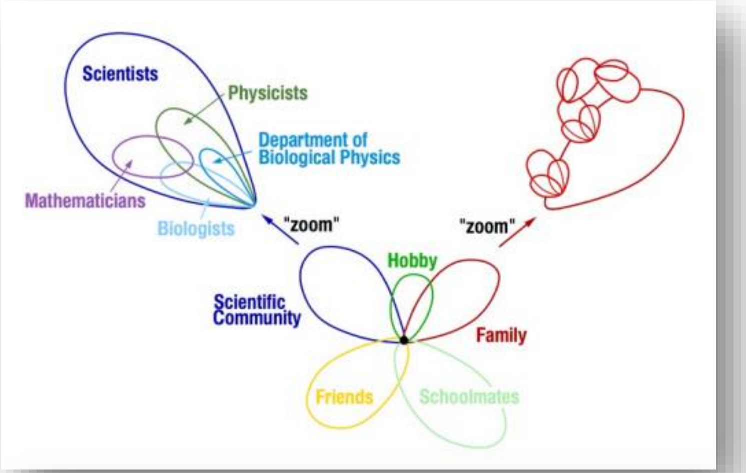
- C is a *weak community* if the total internal degree of a subgraph exceeds its total external degree. Specifically, a subgraph C forms a weak community if

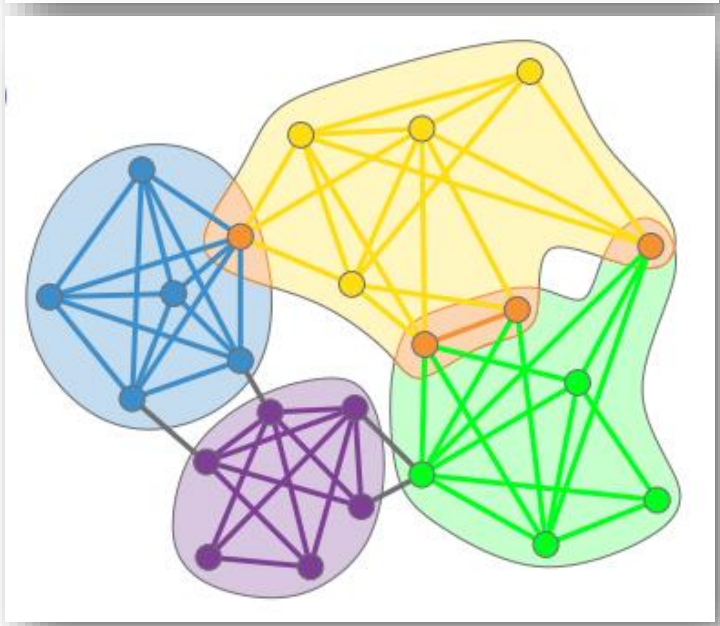
$$\sum_{i \in C} k_i^{\mathrm{int}}(C) > \sum_{i \in C} k_i^{\mathrm{ext}}(C).$$

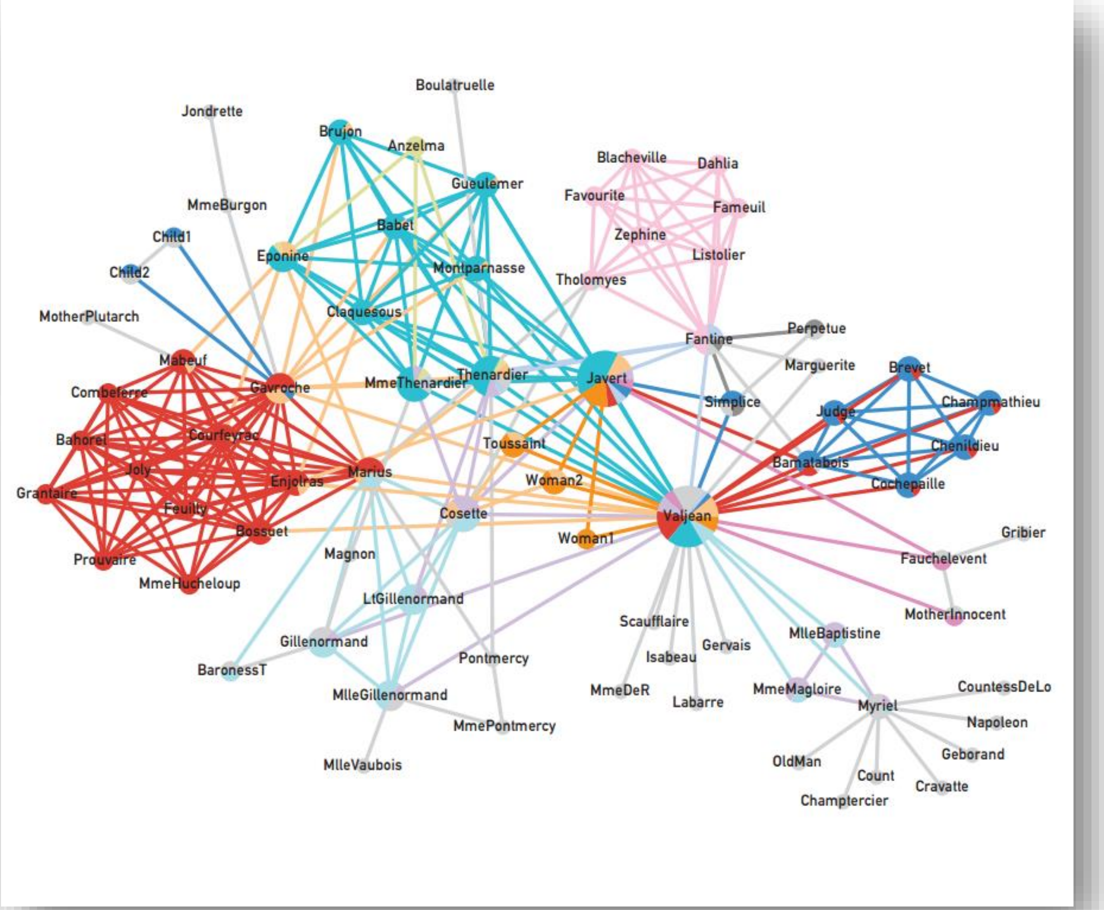
Examples



Overlapping communities







Problem definition

- Most of the complex social, technological and biological networks have a significant community structure.
- The community structure of complex networks has to be considered as *a universal property*, together with the much explored small-world and scale-free properties of these networks.
- Models of network growth based on simple *triadic closure* naturally lead to the emergence of community structure, together with fat-tailed distributions of node degree, high clustering coefficients.

Triadic closure

- *Triadic closure* is a concept in social network theory, first suggested by German sociologist Georg Simmel in his 1908 book *Soziologie*.

- Triadic closure is the property among three nodes *A, B,* and *C,* such that if a strong tie exists between *A*-*B* and *A*-*C,* there is a weak or strong tie between *B*-*C*.

Wikipedia

Preferential Attachment?

- With preferential attachment, it is the node's degree that determine the probability of linking, implying that each new node knows this information about all other nodes (B-A model). *Is it realistic?*

- The triadic closure induces an effective preferential attachment: getting linked to a neighbor A of a node corresponds to choosing A with a probability increasing with the degree k_A of that node. *This well known principle is at the basis of several generative network models.*

Model (Holme & Kim)

- Extension of the standard scale-free network (Barabasi-Albert) model to include a "triad formation step."
  - **Initial condition:** To start with, the network consists of m_0 vertices and no edges.
  - **Growth:** One vertex v with m edges is added at every time step. Time t is identified as the number of time steps.
  - **Preferential attachment (PA):** Each edge of v is then attached to an existing vertex with the probability proportional to its degree, i.e. the probability for a vertex w to be attached to v is:

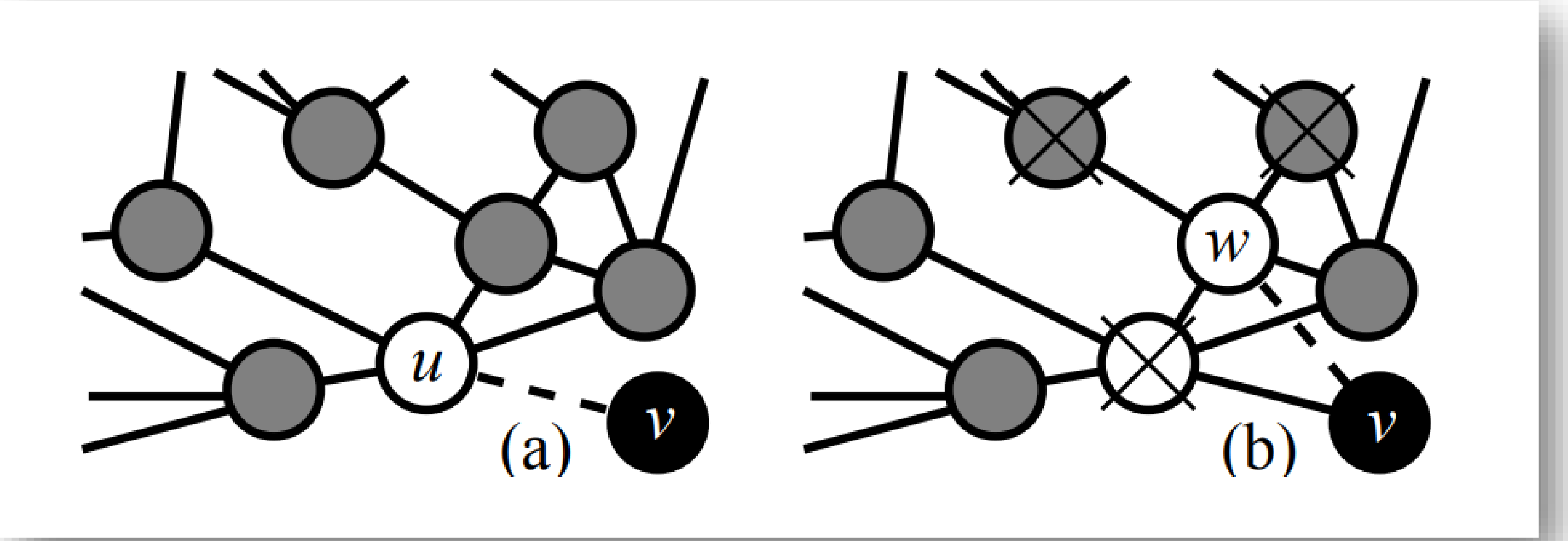
$$P_w = \frac{k_w}{\sum_{v \in \mathcal{V}} k_v}$$

Triad formation step

- **Triad formation (TF):** If an edge between *v* and *w* was added in the previous PA step, then add one more edge from *v* to a randomly chosen neighbor of *w*. If there remains no pair to connect, i.e., if all neighbors of *w* were already connected to *v*, do a PA step instead.

- First, one PA step is performed. Then, a TF step with the probability P_t or a PA step with the probability $1 - P_t$ is performed.

PA x TF



Basic model (Bianconi et al.)

- The starting point is a small connected network of n_0 nodes and $m_0 \geq m$ links.

- *Growth.* At each time t, a new node is added to the network with $m \geq 2$ links.

- *Proximity bias.* The probability to attach the new node to node i depends on the order in which the links are added.

The first link

- The first link of the new node is attached to a random node i_1 of the network. The probability that the new node is attached to node i_1 is then given by

$$\Pi^{[0]}(i_1) = \frac{1}{n_0 + t}.$$

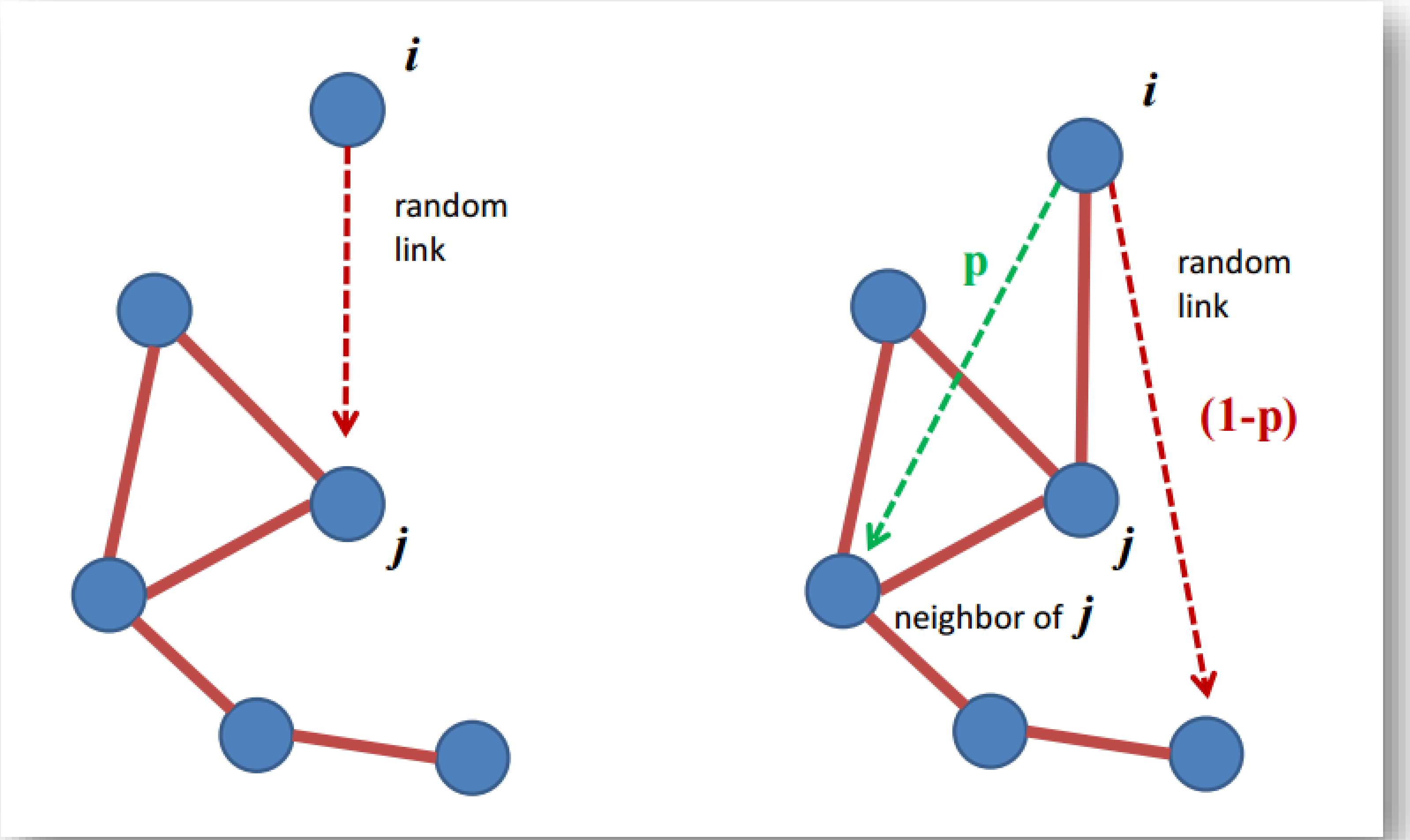
The second link

- The second link is attached to a random node of the network with probability $1-p$, while with probability p it is attached to a node chosen randomly among the neighbors of node i_1.
- Therefore, the probability to attach to a node i_2 <> i_1 is given by

$$\Pi^{[0]}(i_2) = \frac{(1-\delta_{i_1,,i_2})}{n_0+t-1},$$

$$\Pi^{[1]}(i_2) = \frac{a_{i_1,i_2}}{k_{i_1}},$$

Example



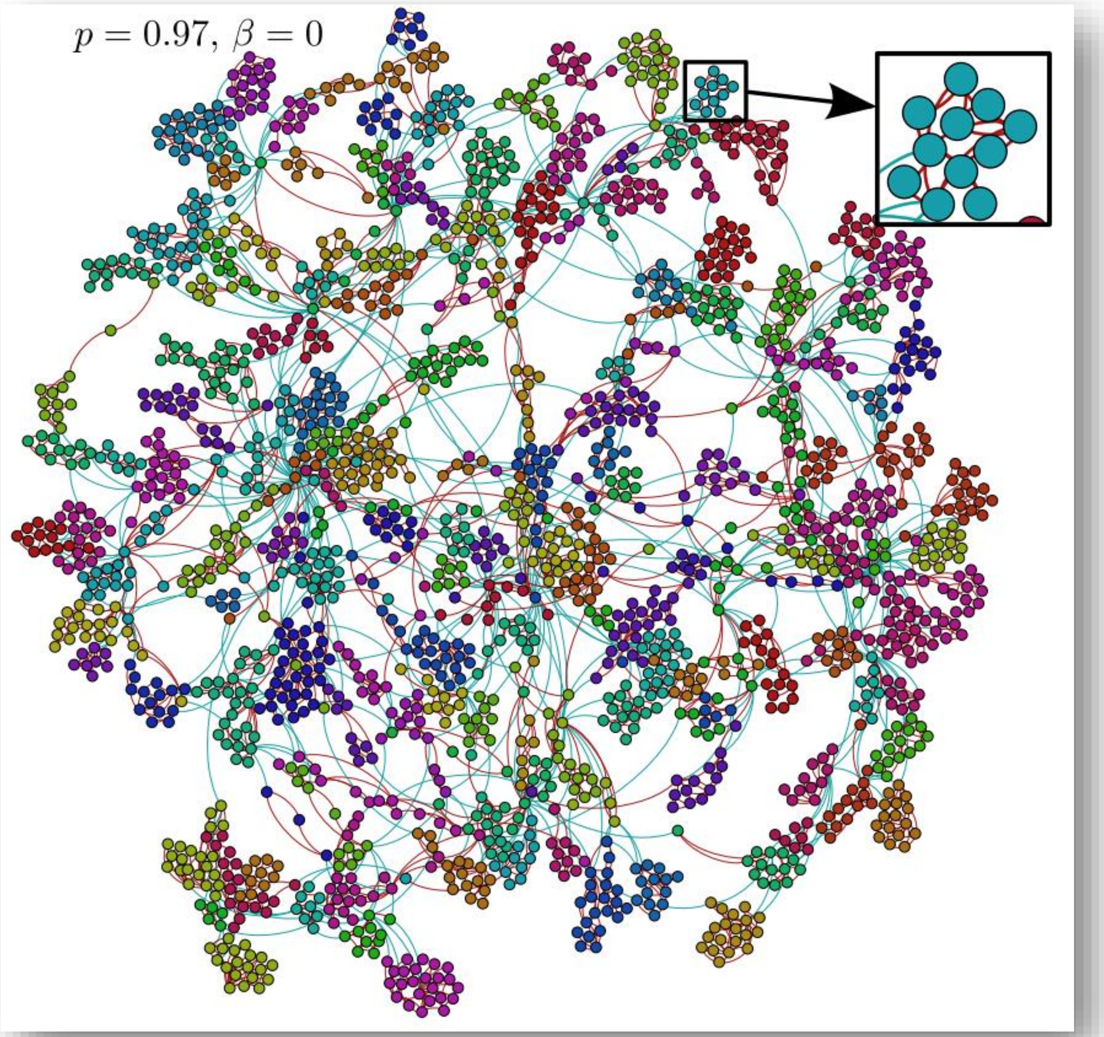
Further edges

- For the model with $m > 2$, further edges are added according to the "*second link*" rule in the previous point.

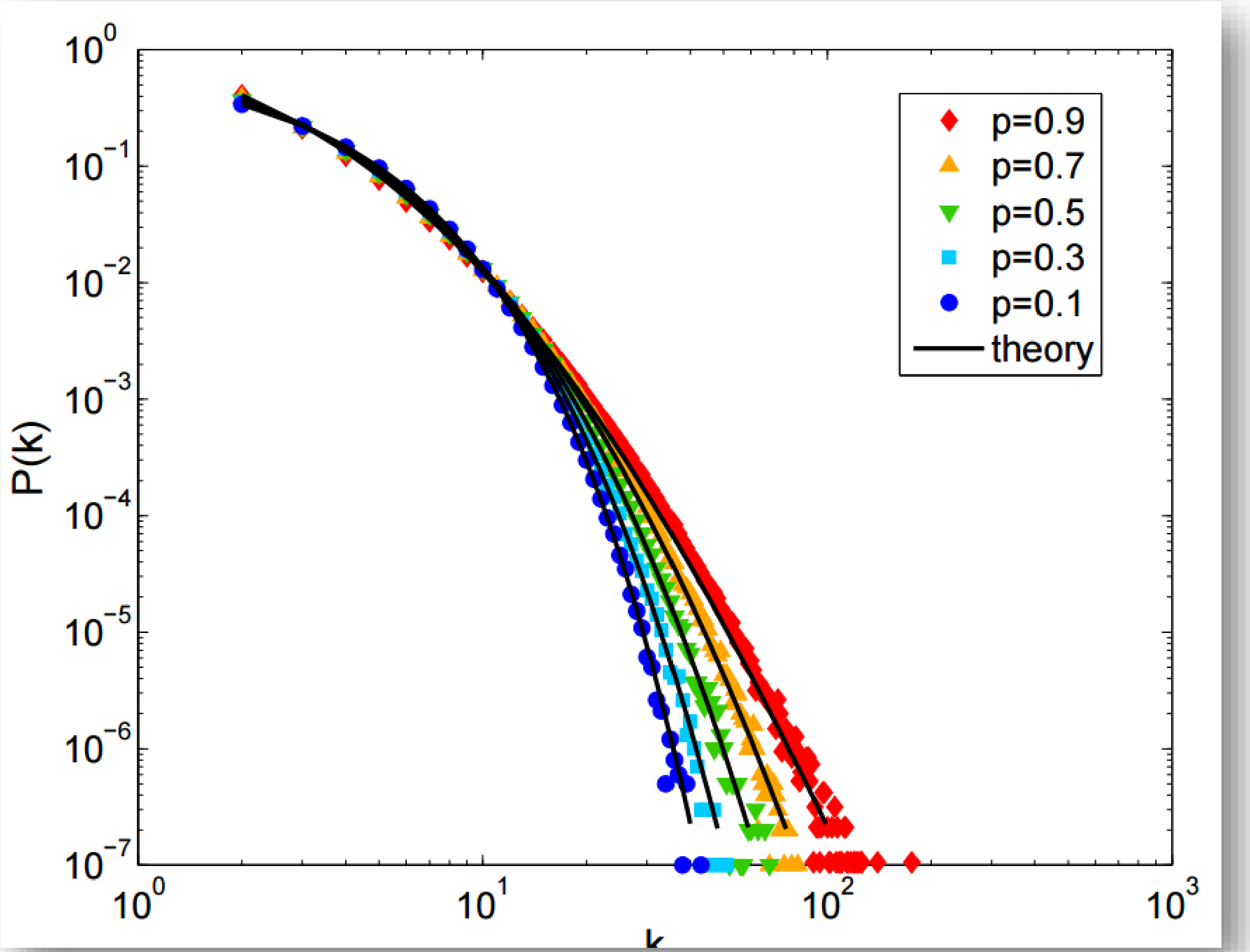
- With probability p, and edge is added to a random neighbor without a link of the first node i_1.

- With probability $1 - p$, a link is attached to a random node in the network without a link already. A total of m edges are added, 1 initial random edge and $m - 1$ involving triadic closure or random attachment.

Example

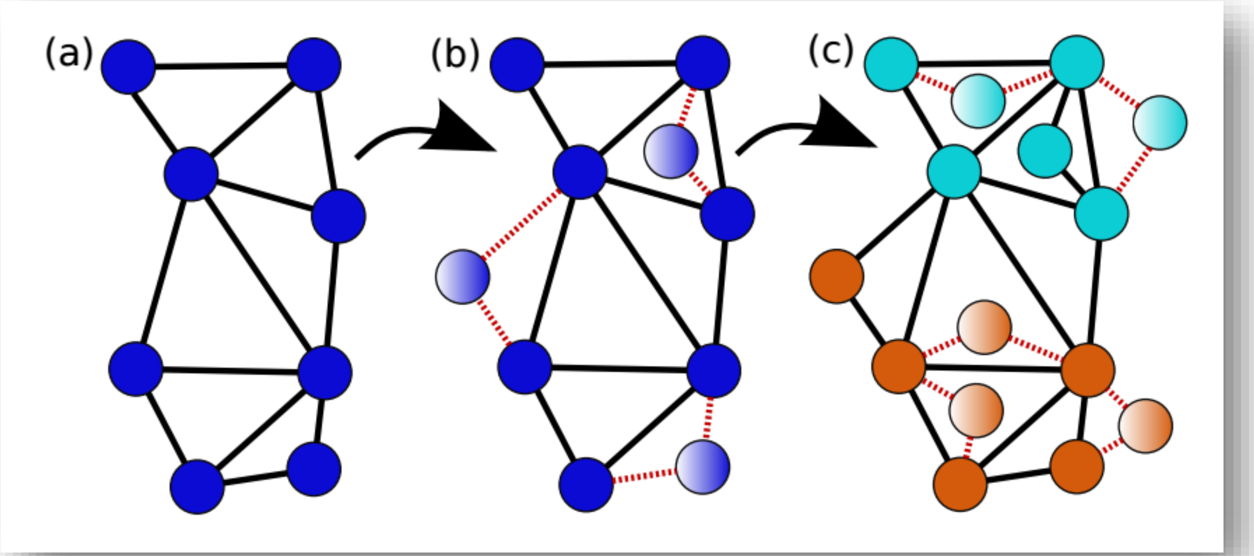


Degree distributions



Community evolution

- Initial inhomogeneities in the link density make more likely the closure of triads in the denser parts, that keep growing until they become themselves inhomogeous, leading to a split into smaller communities.



Fitness

- *Bianconi-Barabási model*, a new concept of preferential attachment (B-A model) called the *fitness*.

- The probability that a new node connects to one of the existing links to a node *i* in the network depends on the number of edges and on the fitness of node *i*, such that,

$$\Pi_i = \frac{\eta_i k_i}{\sum_j \eta_j k_j}.$$

Community model with fitness

- Let us assume that the nodes are not all equal and assign to each node *i* a fitness η_i representing the ability of a node to attract new links.

- We can parametrize the fitness with a parameter $\beta > 0$ by setting with *eps* chosen from a distribution *g* and β representing a tuning parameter of the model ($\nu = 6$).

$$\eta_i = e^{-\beta\epsilon_i} ,$$ $$g(\epsilon) = (1+\nu)\epsilon^{\nu} ,$$ $$\epsilon \in (0,1).$$

Fitness model (no difference)

- The starting point is a small connected network of n_0 nodes and $m_0 \geq m$ links.
- *Growth.* At each time t, a new node is added to the network with $m \geq 2$ links.
- *Proximity bias.* The probability to attach the new node to node i depends on the order in which the links are added.

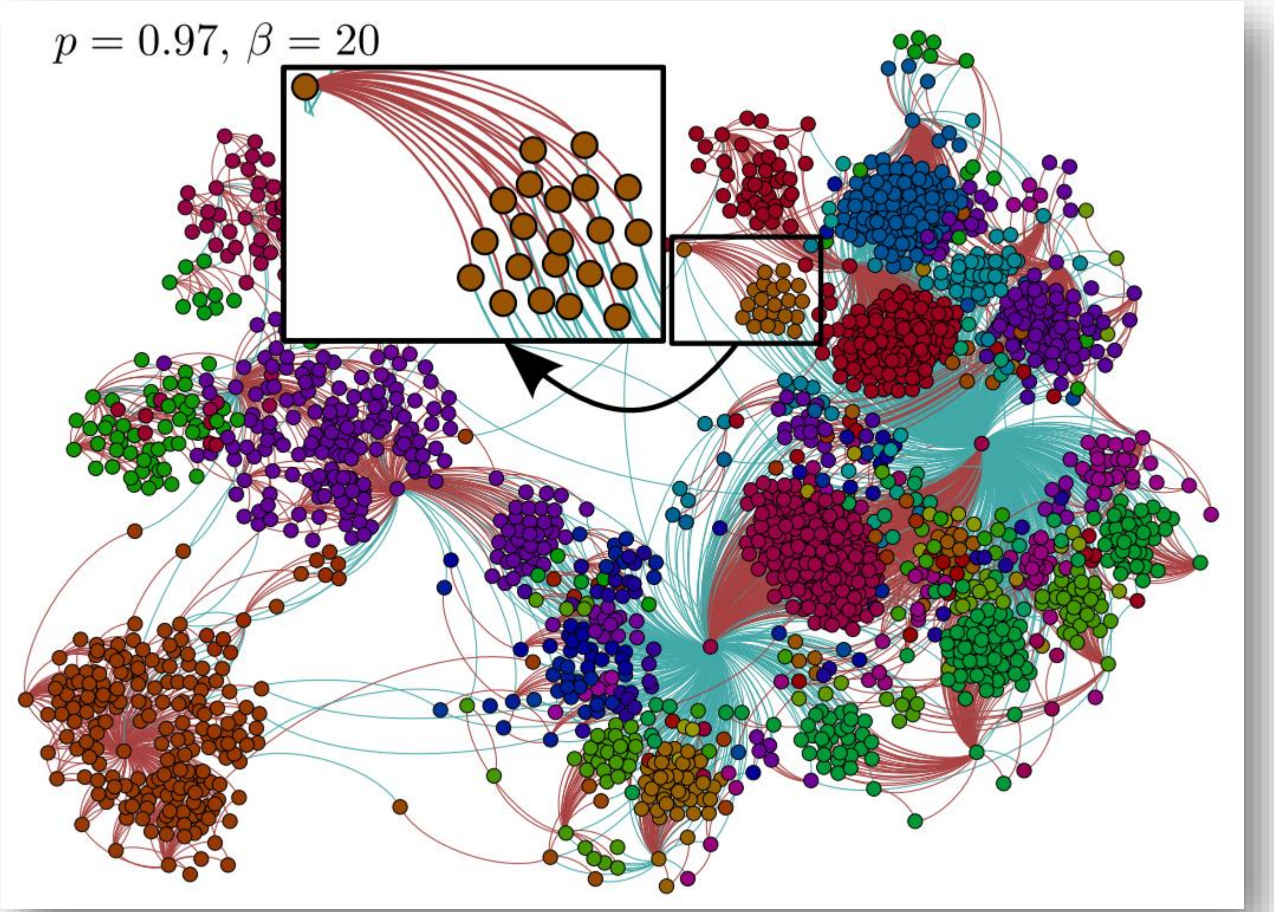
Links (probabilities)

$$\Pi^{[0]}(i_1) = \frac{\eta_{i_1}}{\sum_j \eta_j}$$

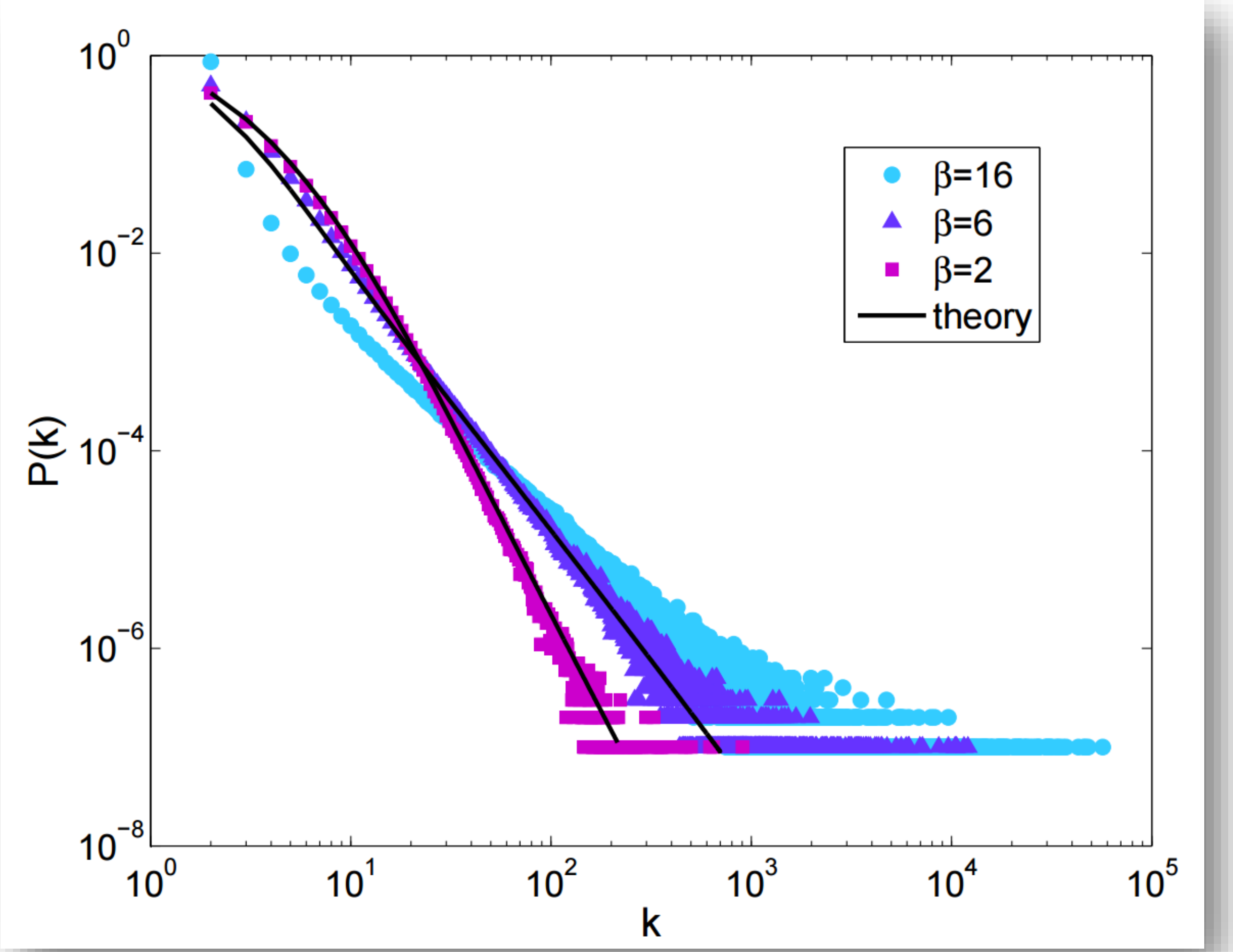
$$\Pi^{[0]}(i_2) = \frac{\eta_{i_2}(1 - \delta_{i_1,i_2})}{\sum_{j \neq i_1} \eta_j}$$

$$\Pi^{[1]}(i_2) = \frac{\eta_{i_2} a_{i_1,i_2}}{\sum_j \eta_j a_{i_1,j}}$$

Example



Degree distribution



Summary

- For $\beta = 0$ all nodes have identical fitness, and the model reduces itself to the basic model.

- With the emergence of communities for sufficiently large values of the probability of triadic closure p, there is a large density of triangles in the network.

- When β is sufficiently large, communities disappear, despite the high density of triangles.

Assignment

- Implement algorithms generating networks with community structure according to Holme & Kim, and Bianconi et al. Use different settings (probability values) and generate networks with one thousand vertices. Visualize these networks and determine their properties (e.g., average degree and distribution, average clustering coefficient).

- Generate networks with one million vertices and compare their properties with the properties of small generated networks.