Network generation

Applied numerics in systems biology SS2017

Jacob Gora

Content

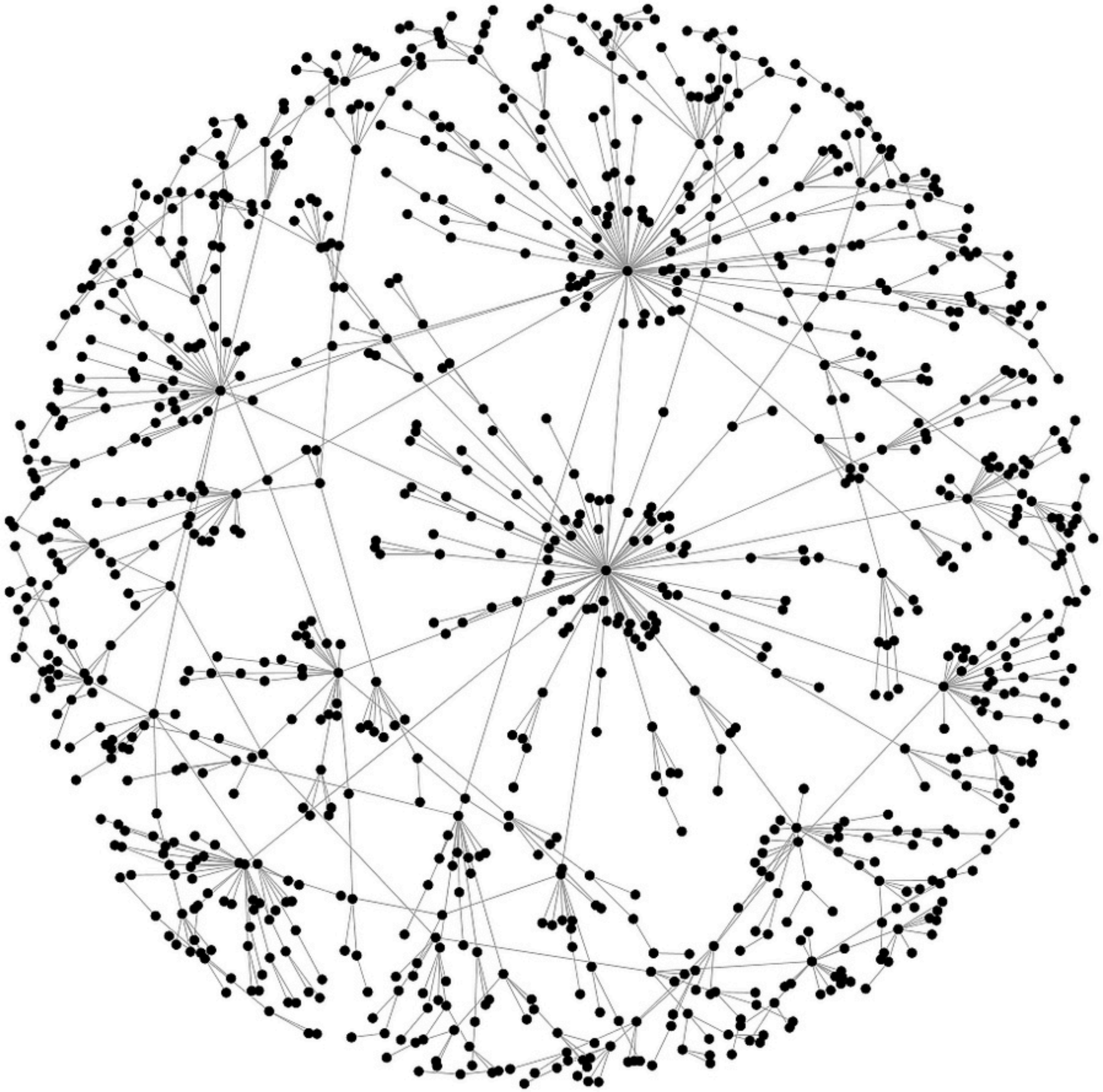
- **Scale-free networks**
- **Configuration model**
  - Sampling from Power-law functions
  - Catanzaro et al. structural cut-off model
  - Results
- **Modeling PPI topology (Schneider et al.)**
  - Validation of model implementation

Scale-free networks

Pertubation resistant (fault tolerant)

(Often) Clustering coefficient distribution decreases with increasing degree

Low mean shortest-path

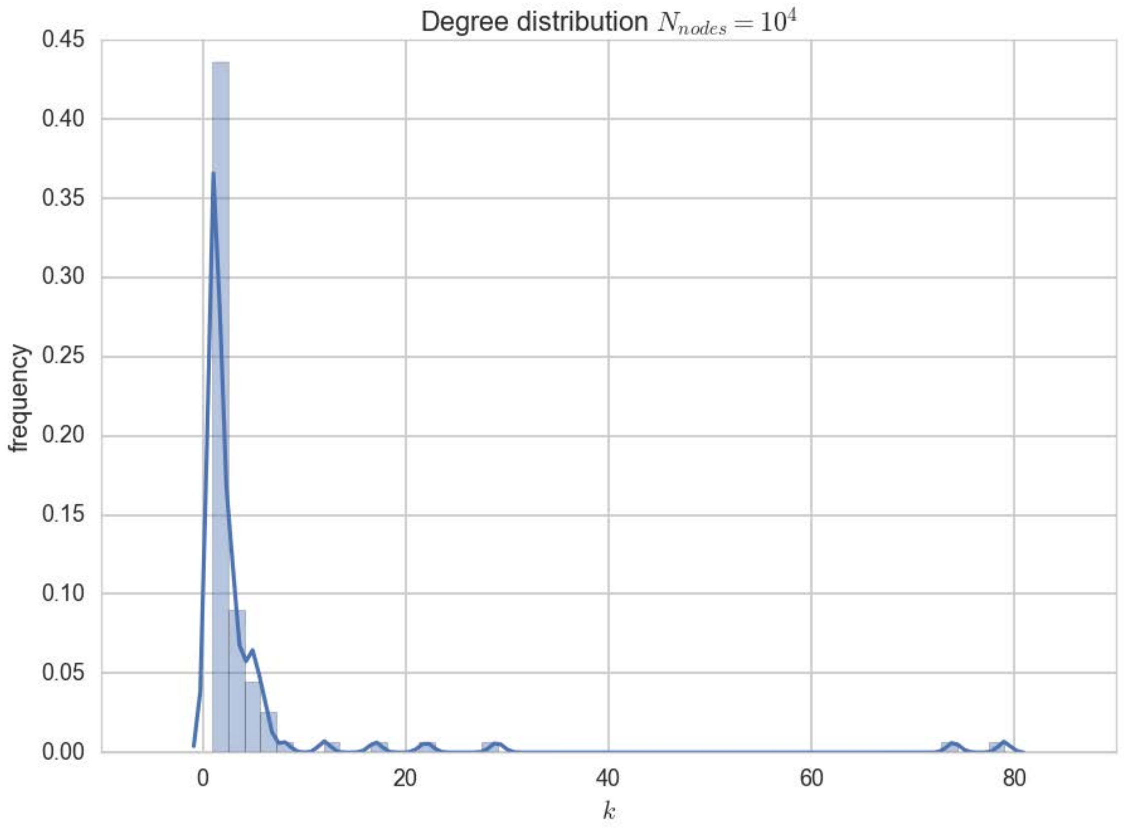


Scale-free networks cont.

Networks whose degree distribution follow a power-law

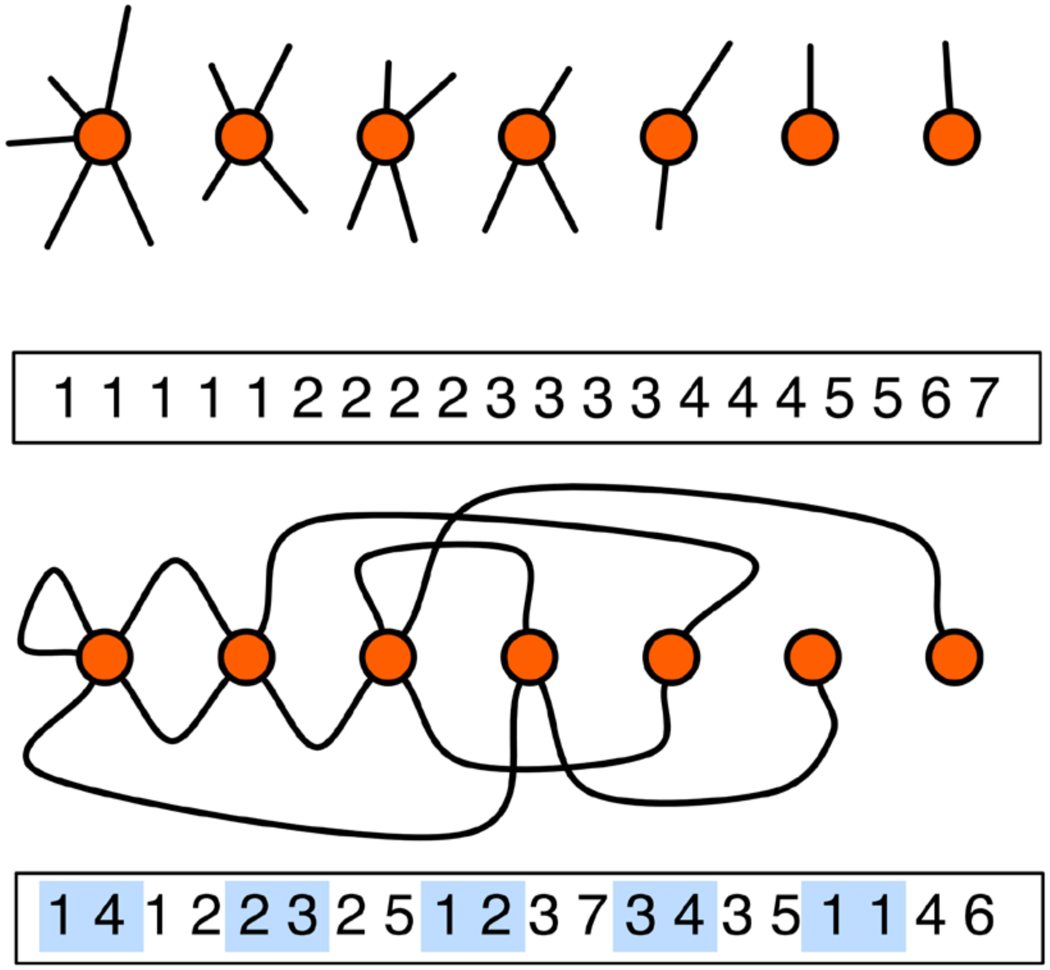
$P(k) \sim k^{-\alpha}, 2 < \alpha < 3$

Variety of methods to generate them (Configuration model, preferential attachment)



Configuration model (CM)

- Sample a power-law distributed (even) degree sequence $(d_0, d_1, \ldots, d_n)$
- For each d_i create $\sum^{d_i} k_i$ stubs where i denotes the node id
- Random shuffle these stubs
- Iterate over them and connect pairs
- Approach can introduce degree correlations



Sampling a power-law distribution within an interval using an uniform distributed variable

$f(k) = k^{\alpha}$, where $k \in [k_0, k_1]$

$\int_{k_0}^{k_1} C f(k) dk = 1$

$F(k) = C\left[\frac{k^{\alpha+1}}{\alpha+1}\right]_{k_0}^{k_1}$

$\Leftrightarrow C\left[\frac{k_1^{\alpha+1} - k_0^{\alpha+1}}{\alpha+1}\right] = 1 \Leftrightarrow C = \frac{\alpha+1}{k_1^{\alpha+1} - k_0^{\alpha+1}}$

Now let

$u \sim U\,[0, 1]$

$\int_{k_0}^{k} F(k) dk = C\frac{k^{\alpha+1} - k_0^{\alpha+1}}{\alpha+1} \equiv u$

We can then solve this for our k.

$\Leftrightarrow k^{\alpha+1} = \left(\frac{\alpha+1}{C} u + k_0^{\alpha+1}\right)$

$\Rightarrow k = \left(\frac{\alpha+1}{C} u + k_0^{\alpha+1}\right)^{\frac{1}{\alpha+1}}$

Insertion of C yields our target function

$degree = \left[\left(k_1^{\alpha+1} - k_0^{\alpha+1}\right) u + k_0^{\alpha+1}\right]^{\frac{1}{\alpha+1}}$

Issue of degree correlations (assortativity/disassortative)

- Assortative mixing: highly connected vertices prefer highly connected vertices
- Disassortative mixing: highly connected vertices prefer the poor
- Cantanzaro et al. show numerically/analytically that naive CM will introduce these degree correlations for $\alpha < 3$
- Problematic when testing numerical behaviour on these system as analytical solutions expect uncorrelated networks

Issue of degree correlations cont.

- Constraining the maximum degree by $K_s(N) \sim N^{\frac{1}{2}}$ solves (reduces) these issues

Implementation and experiments of CM and uncorrelated CM

Generation of networks of size $N = 10^5$

Measuring the clustering coefficient (CF) and average nearest neighbor degree (ANND)

Average nearest neighbor degree $\mathrm{k_{nn}(k)}$

$\mathrm{k_{nn}(k)} = \sum_{k'} k' P(k|k')$, where $P(k|k')$ is the conditional probability that an edge k points to a node of degree k'

- An increasing $\mathrm{k_{nn}(k)}$ indicates that the network is assortative
- Vice-versa a decrease indicates a disassortative network

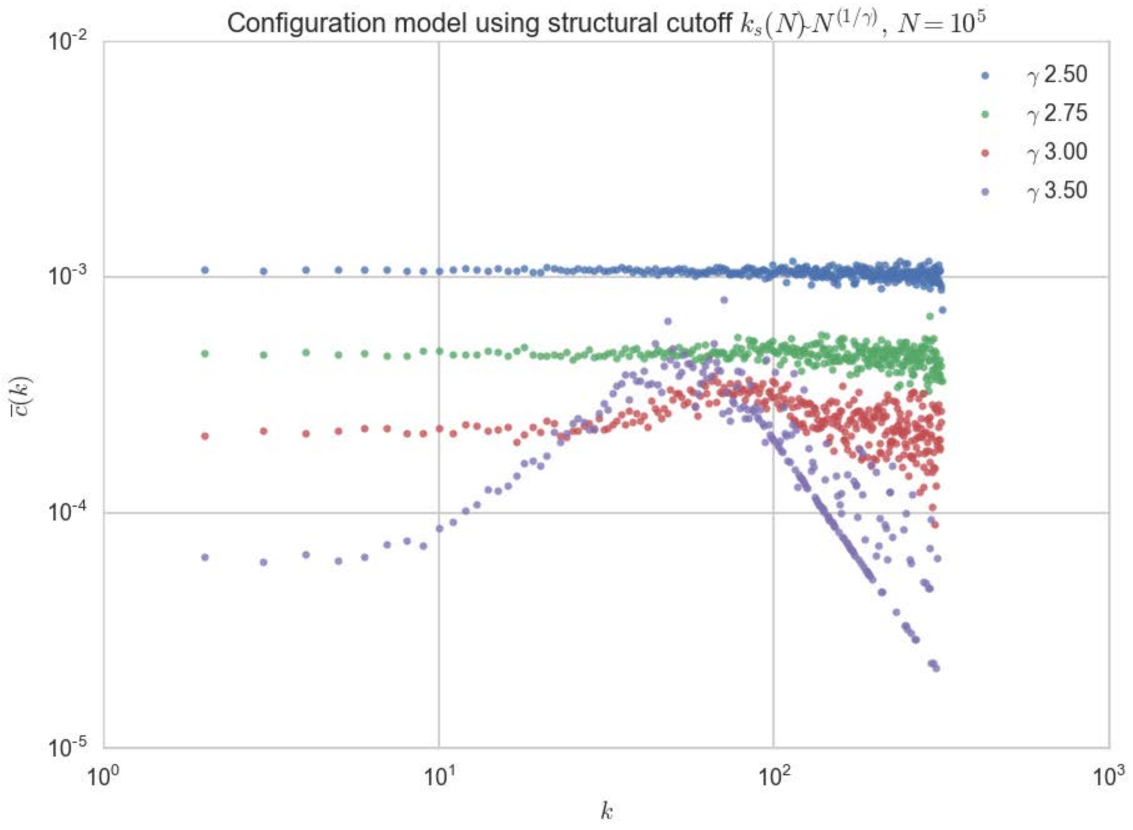
Average clustering coefficient of nodes of degree k (CF)

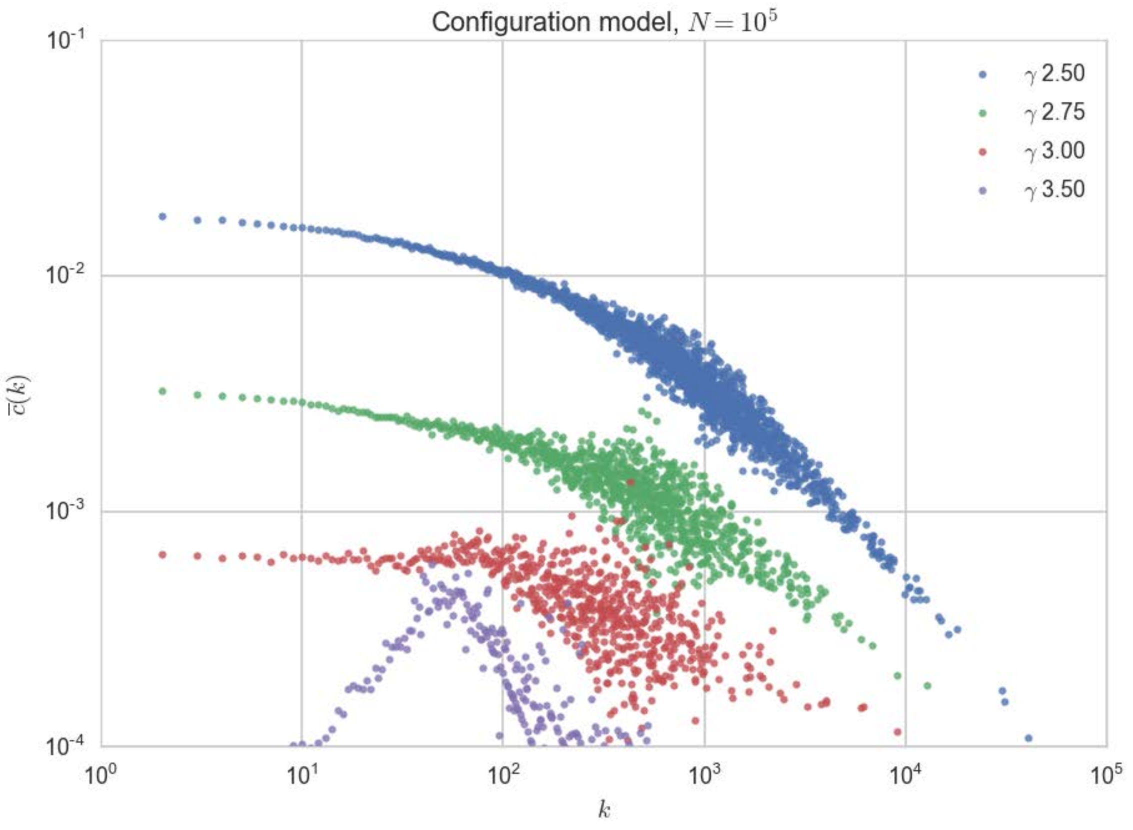
$C_i = \frac{2\ L_i}{k_i(k_i-1)}$, where L_i denotes the number of links between k_i neighbors of node i.

Averaged over nodes of degree k.

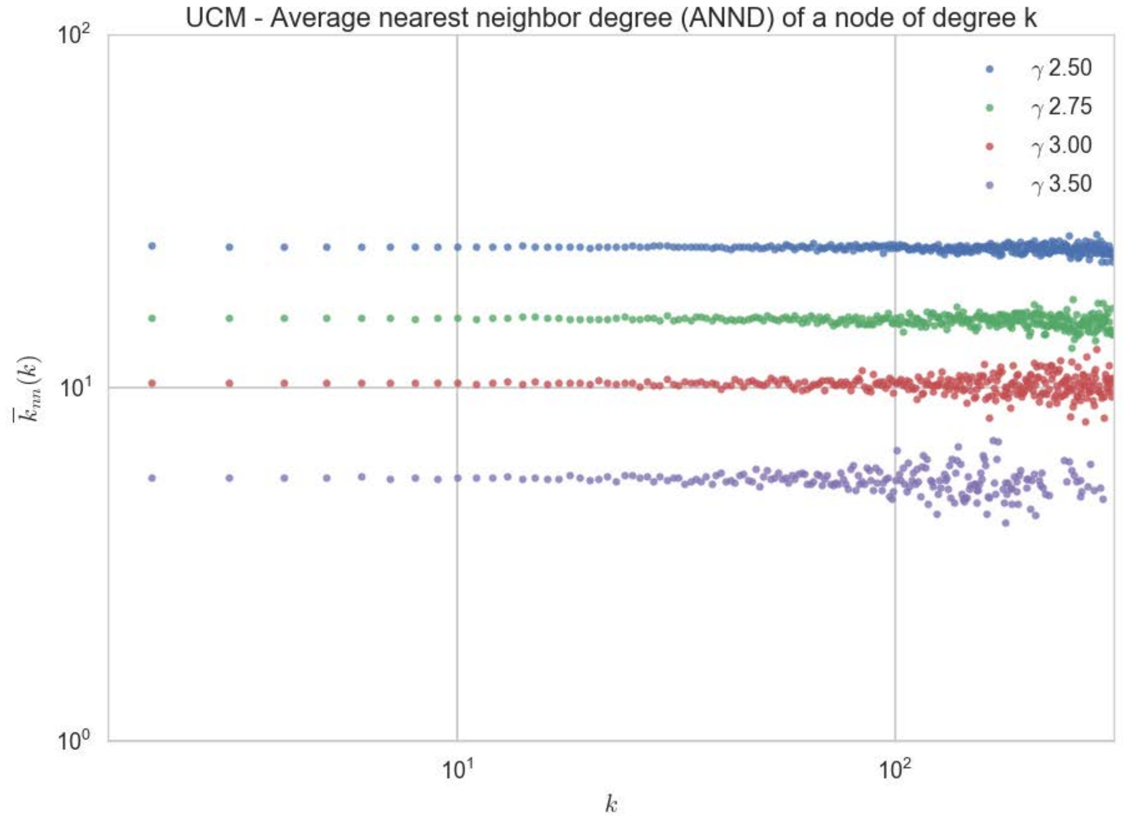
Measures the local link density of nodes of degree k.
The more connected a neighborhood is the higher the CF.

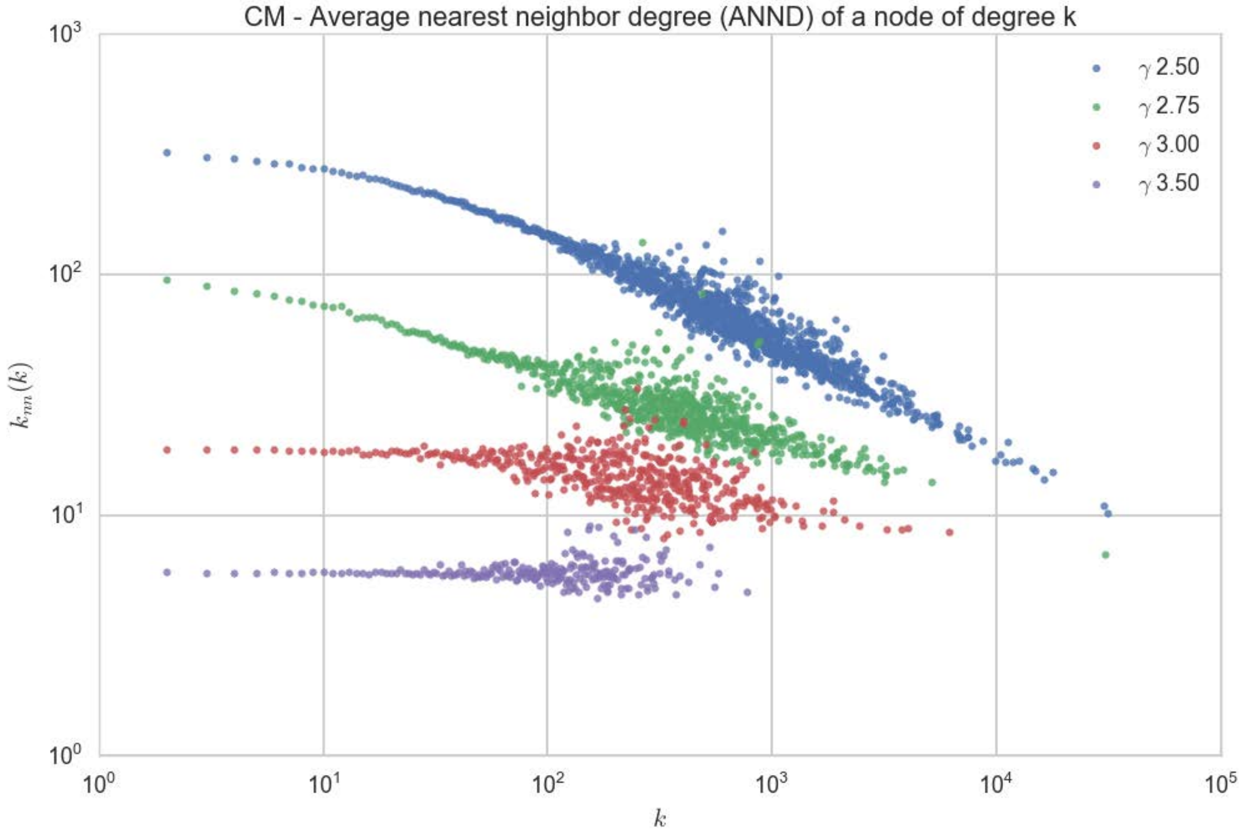
Uncorrelated CM and CM results (CF)





Uncorrelated CM and CM results (ANND)





Modeling PPI topology (Schneider et al.)

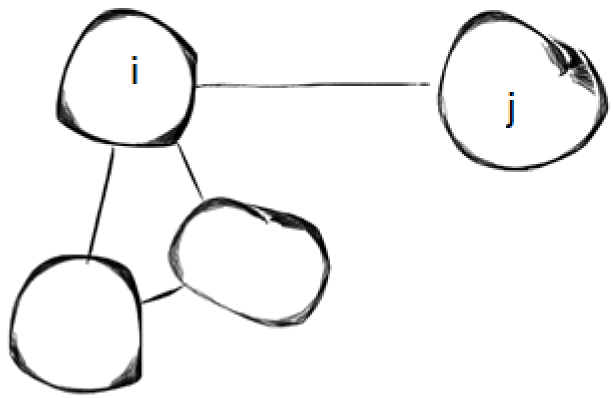
Schneider describes an algorithm to generate PPI like networks given an N_{bio}, $|edge|_{bio}$ and a parameter α.

Schneider’s model - depletion step

(i) Start with a **fully connected** network with N_{bio} nodes

(i) Choose at random a node *i*

(ii) Choose at random an edge $e_{i,j}$ and remove it with a probability $p_{i,j}$ related to the degree k_j of the neighbor *j* of node *i*:



$$p_{i,j} = \frac{1^{-\alpha}}{2^{-\alpha}+2^{-\alpha}+2^{-\alpha}}$$

$$p_{i,j} = \frac{p_j}{N_i}, where\ p_i = \begin{cases} {k_j}^{-\alpha}, k_j > 1 \\ 0, \quad else \end{cases} \quad \text{and } N_i = \sum_{l=1}^{k_i} p_l$$

Repeat until the number of edges is equal to N_{bio}.

Schneider's model - similarity step

(i) Choose at random two nodes *i* and *j*. Add an edge between these nodes with probability

(ii) Connect these nodes based on probability $p_{i,j}$

$p_{i,j} = [N_c(i,j)]^2(k_i k_j)$, where $N_c(i,j)$ is the number of common neighbors of i, j.

Repeat until number of edges is equal to $|edge|_{bio}$

Validation of model implementation

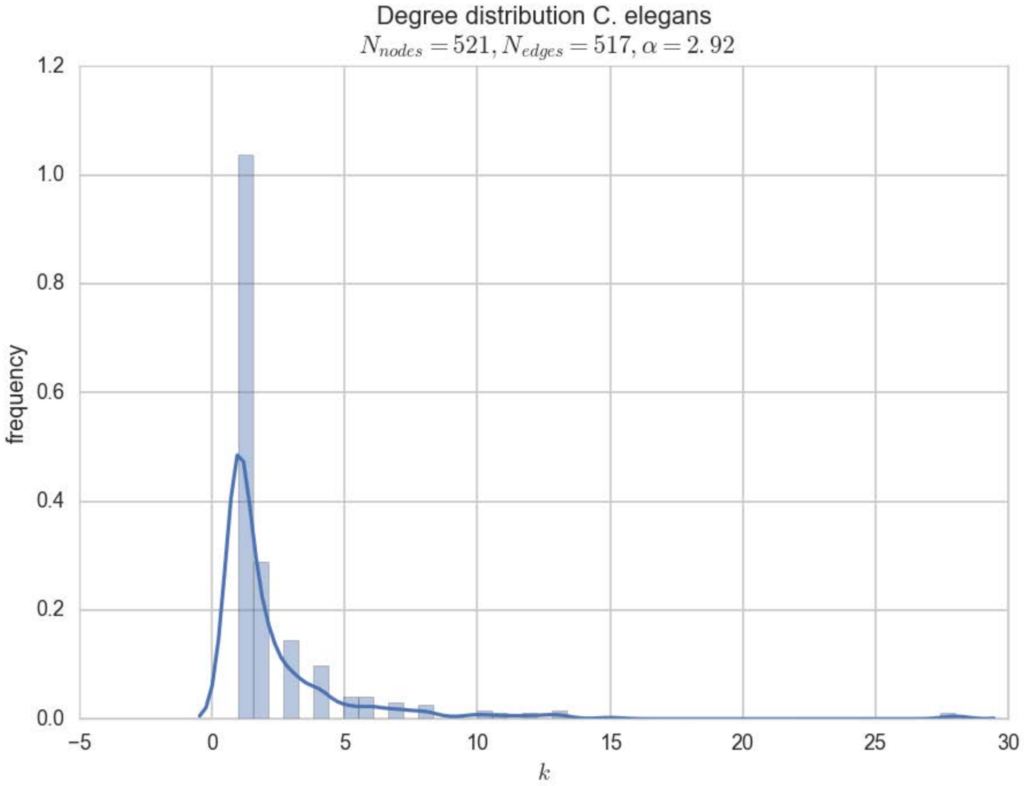
- Data from the Harvard CCSB Interactome database (C. Elegans, S. cerevisiae)

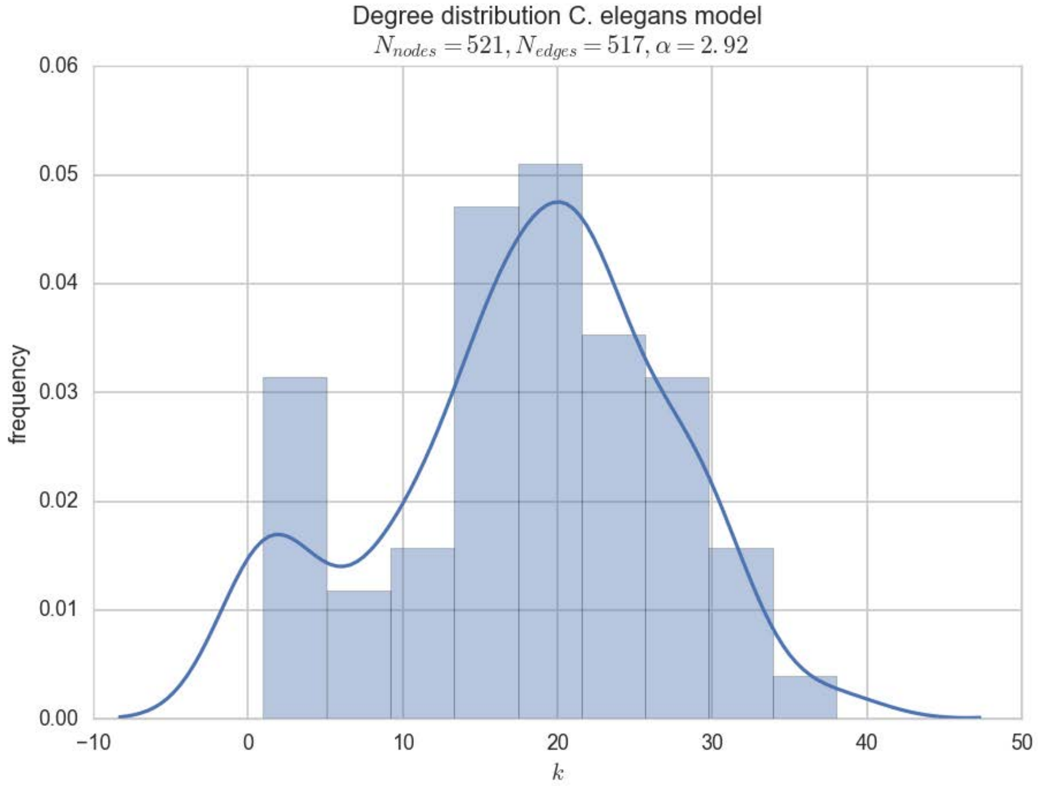
- Fitting the alpha parameter (scipy.optimize.curve_fit)

$$\alpha_{opt} \equiv \underset{\alpha}{\text{argmin}} \sum_{i=1}^{N} \left[y_i - f(x_i, \alpha)\right]^2, \text{ where } f(x, \alpha) = x^{-\alpha}$$

- Run Schneiders algorithm using the corresponding node and edge amounts

Validation of model implementation cont.





$\alpha = 14.91$ in reality

References

- Michele Catanzaro et al. 2005. Generation of uncorrelated random scale-free networks, DOI: 10.1103/PhysRevE.71.027103
- Christian M. Schneider et al. 2011. Modeling the topology of protein interaction networks. DOI: 10.1103/PhysRevE.84.016112
- Dong Yao et al. 2017. Average nearest neighbor degrees in scale-free networks. DOI:10.1140/epjb/e2017-80031-x
- Albert –Laszlo Barabasi. 2016. Network Science
- S. Boccaletti et al. 2015. Complex networks: Structure and dynamics
- http://interactome.dfci.harvard.edu