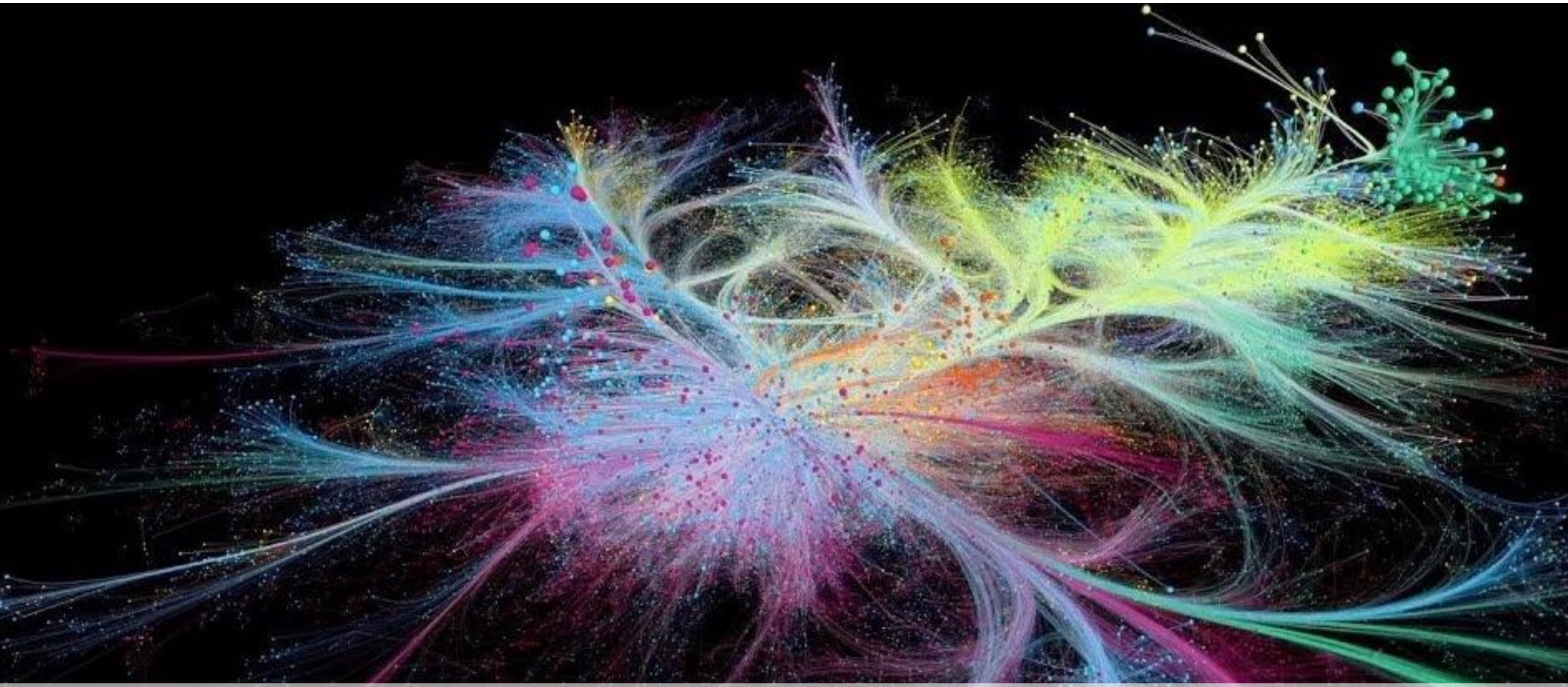
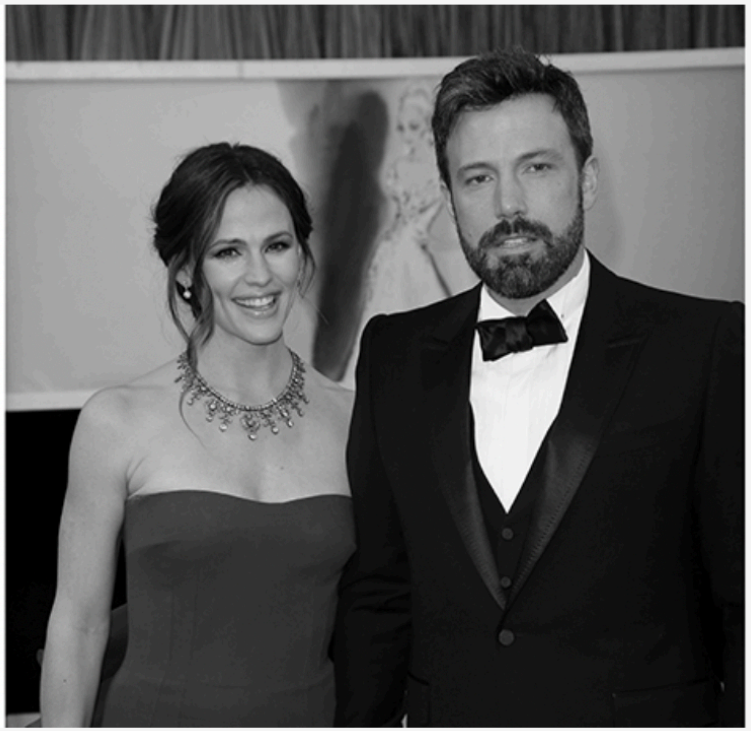


CPSC 572/672: Fundamentals of Social Network Analysis and Data Mining

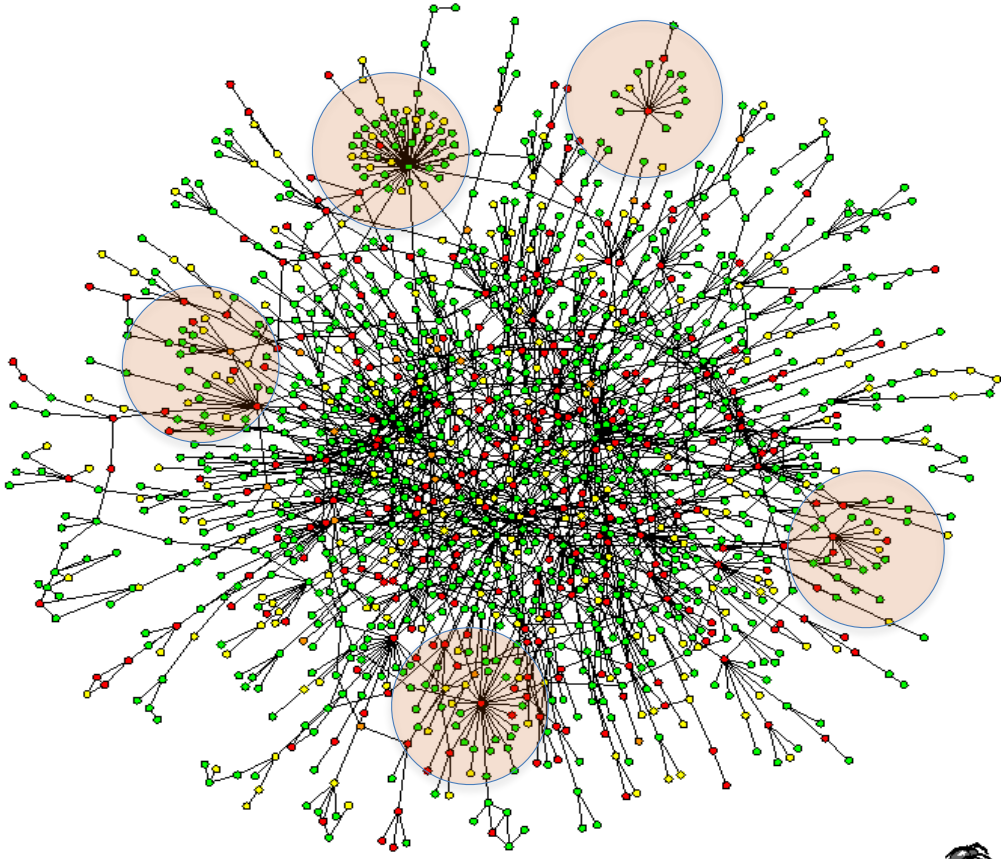
Real Networks:
Degree Correlations



TOPOLOGY OF THE PROTEIN NETWORK

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:

$$p_{kk'} = \frac{kk'}{2L}$$

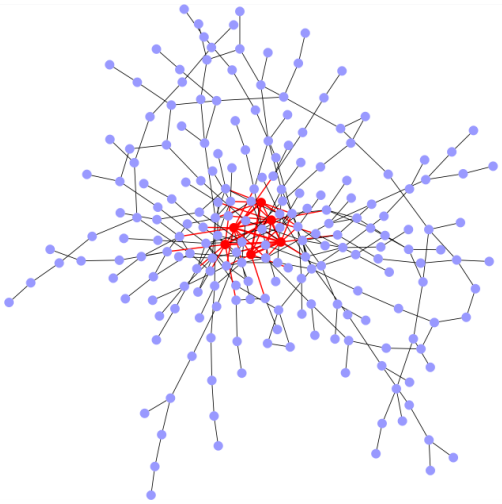
k≅50, k'=13, N=1,458, L=1746

$p_{50,13} = 0.15$
$p_{2,1} = 0.0004$

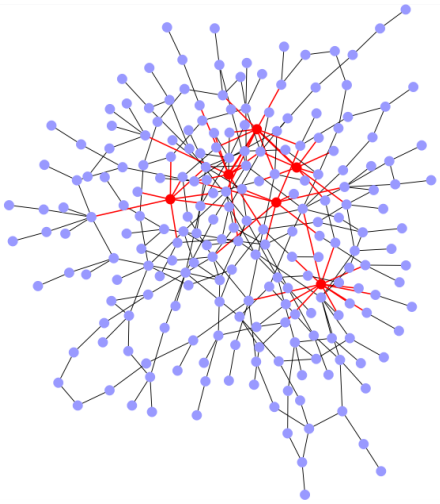
Yet, we see many links between degree 2 and 1 links, and no links between the hubs.

H. Jeong, S.P. Mason, A.-L. Barabasi, Z.N. Oltvai, Nature 411, 41-42 (2001)

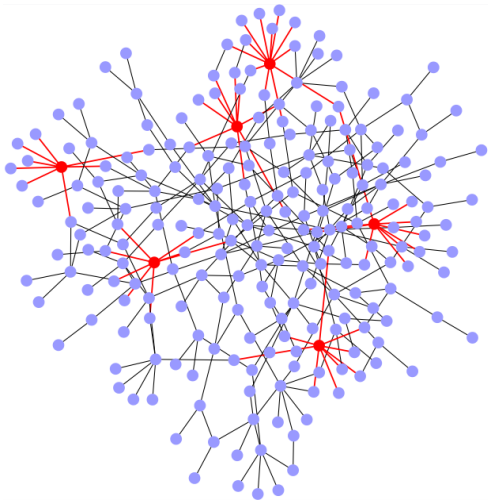
DEGREE CORRELATIONS IN NETWORKS



Assortative:
hubs show a tendency to link to each other.



Neutral:
nodes connect to each other with the expected random probabilities.



Disassortative:
Hubs tend to avoid linking to each other.

Quantifying degree correlations (three approaches):

→ full statistical description (Maslov and Sneppen, Science 2001)
→ degree correlation function (Pastor Satorras and Vespignani, PRL 2001)
→ correlation coefficient (Newman, PRL 2002)

STATISTICAL DESCRIPTION

e_{jk}: probability to find a node with degree j and degree k at the two ends of a randomly selected edge

$$\sum_{j,k} e_{jk} = 1 \qquad \sum_{j} e_{jk} = q_k$$

q_k: the probability to have a degree k node at the end of a link.

M. E. J. Newman, Phys. Rev. Lett. **89,** 208701 (2002)

STATISTICAL DESCRIPTION

e_{jk}: probability to find a node with degree j and degree k at the two ends of a randomly selected edge

$$\sum_{j,k} e_{jk} = 1 \qquad \sum_{j} e_{jk} = q_k$$

q_k: the probability to have a degree k node at the end of a link.

Where: $$q_k = \frac{kp_k}{\langle k \rangle}$$

Probability to find a node at the end of a link is biased towards the more connected nodes, i.e. $q_k=Ckp_k$, where C is a normalization constant . After normalization we find $C=1/\langle k \rangle$, or $q_k=kp_k/\langle k \rangle$

M. E. J. Newman, Phys. Rev. Lett. **89,** 208701 (2002)

STATISTICAL DESCRIPTION

e_{jk}: probability to find a node with degree j and degree k at the two ends of a randomly selected edge

$$\sum_{j,k} e_{jk} = 1 \qquad \sum_{j} e_{jk} = q_k$$

q_k: the probability to have a degree k node at the end of a link.

Where: $$q_k = \frac{kp_k}{\langle k \rangle}$$

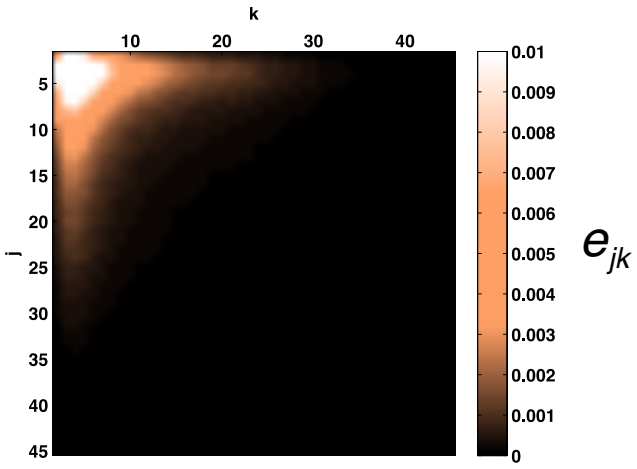
Probability to find a node at the end of a link is biased towards the more connected nodes, i.e. $q_k=Ckp_k$, where C is a normalization constant . After normalization we find $C=1/\langle k \rangle$, or $q_k=kp_k/\langle k \rangle$

If the network has no degree correlations: $$e_{jk} = q_j q_k$$

Deviations from this prediction are a signature of *degree correlations.*

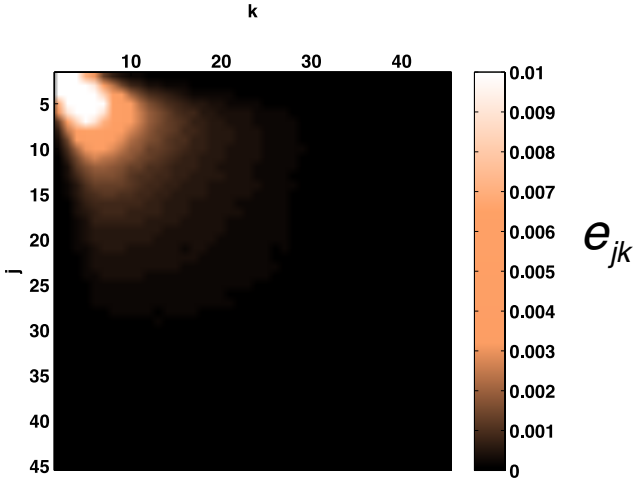
M. E. J. Newman, Phys. Rev. Lett. **89,** 208701 (2002)

EXAMPLE: e_{jk} FOR A SCALE-FREE NETWORK

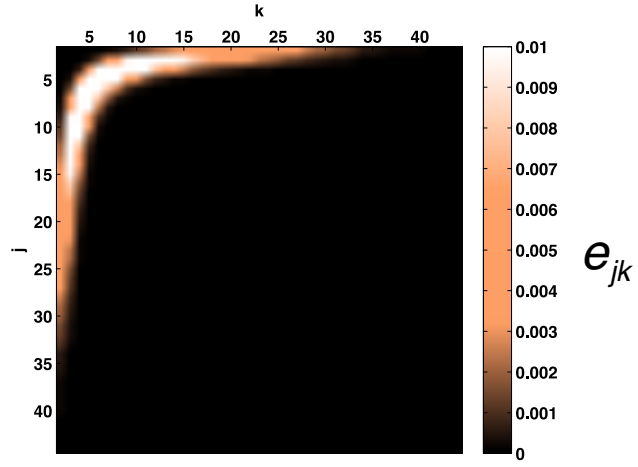


Neutral

Assortative:
More strength in the diagonal, hubs tend to link to each other.



Disassortative:
Hubs tend to connect to small nodes.



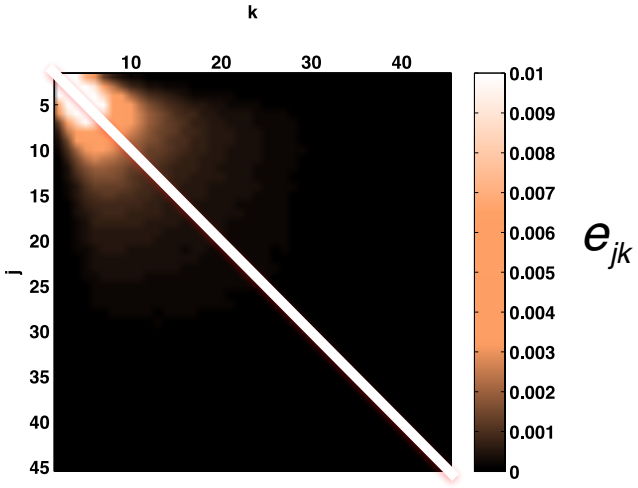
Each matrix is the average of a 100 independent scale-free networks, generated using the static model with $N=10^4$, $\gamma=2.5$ and $\langle k \rangle=3$.

EXAMPLE: e_{jk} FOR A SCALE-FREE NETWORK

Perfectly assortative network:

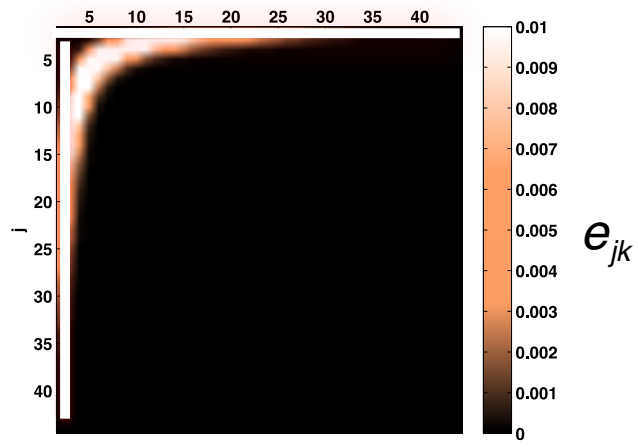
$e_{jk}=q_k\delta_{jk}$

Assortative:
More strength in the diagonal, hubs tend to link to each other.



Perfectly disassortative network:

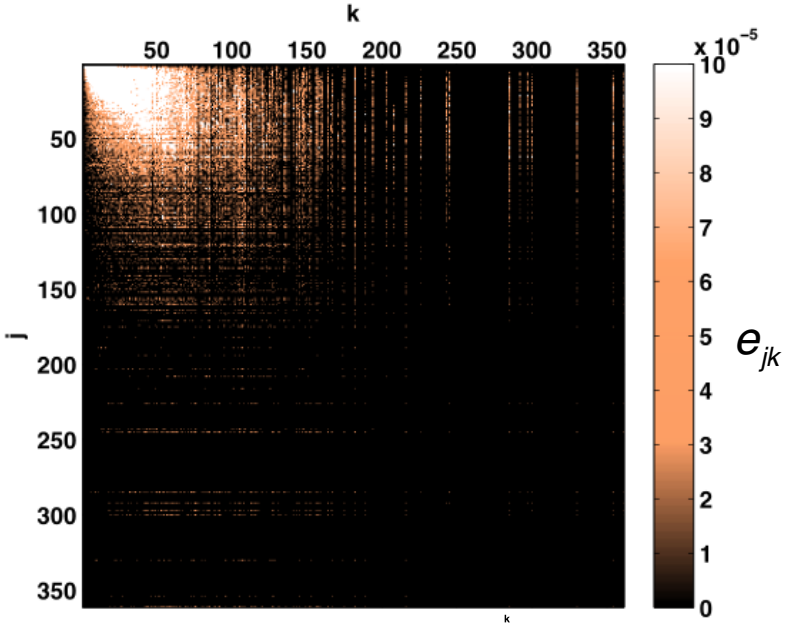
Disassortative:
Hubs tend to connect to small nodes.



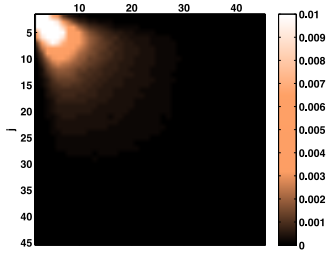
Each matrix is the average of a 100 independent scale-free networks, generated using the static model with $N=10^4$, $\gamma=2.5$ and $\langle k\rangle=3$.

REAL-WORLD EXAMPLES

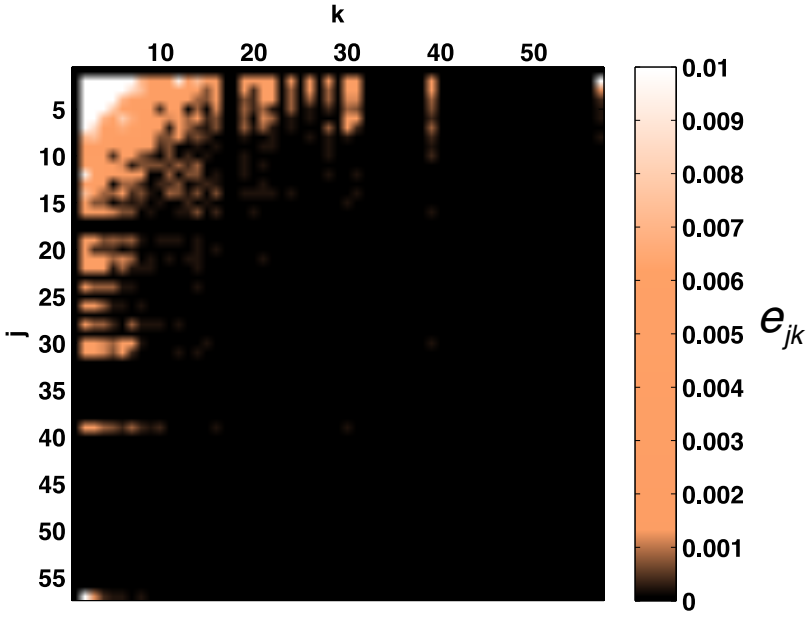
Astrophysics co-authorship network



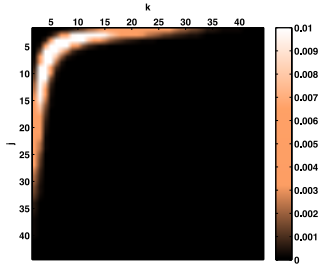
Assortative:
More strength in the diagonal, hubs tend to link to each other.



Yeast PPI

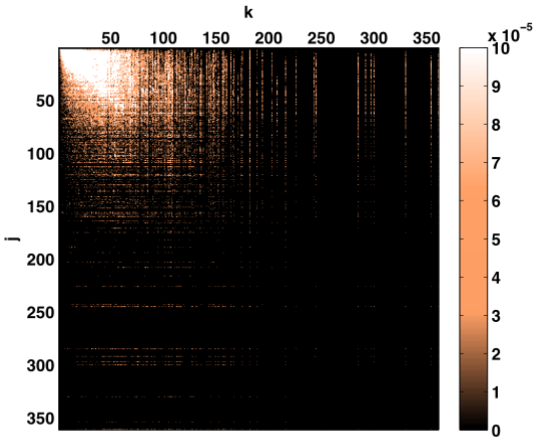


Disassortative:
Hubs tend to connect to small nodes.



PROBLEM WITH THE FULL STATISTICAL DESCRIPTION

(1) Difficult to extract information from a visual inspection of a matrix.



(2) Based on e_{jk} and hence requires a large number of elements to inspect:

$$\frac{k_{\max}\left(k_{\max}-1\right)}{2}-1-k_{\max}$$

Nr. of independent elements

Undirected network: k_{max} x k_{max} matrix

$$\sum_{j,k} e_{jk}=1 \qquad \sum_{j=1,k_{\max}} e_{jk}=q_k$$

Constraints

We need to find a way to reduce the information contained in e_{jk}

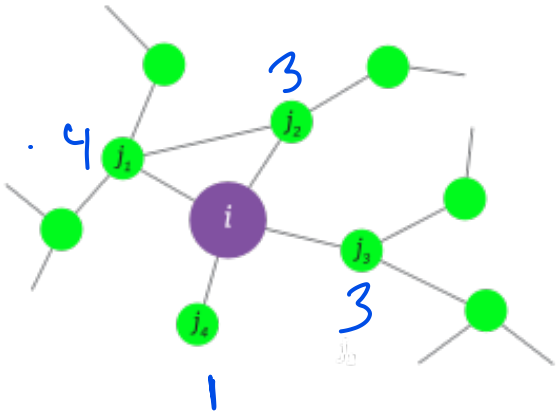
Measuring Degree Correlations

Average next neighbor degree

$k_{annd}(k)$: average degree of the first neighbors of nodes with degree k.

$$k_{nn}(k_i) = \frac{1}{k_i}\sum_{j=1}^{N} A_{ij}k_j$$

$$k_{nn}(k) \equiv \sum_{k'} k'P(k'|k)$$



$$k^{v}_{annd} = \frac{4+3+3+1}{4}$$

average nearest degree

then compare with i node.

R. Pastor-Satorras, A. Vázquez, A. Vespignani, Phys. Rev. E **65**, 066130 (2001)

Average next neighbor degree

$k_{annd}(k)$: average degree of the first neighbors of nodes with degree k.

$$k_{nn}(k) \equiv \sum_{k'} k' P(k' \mid k)$$

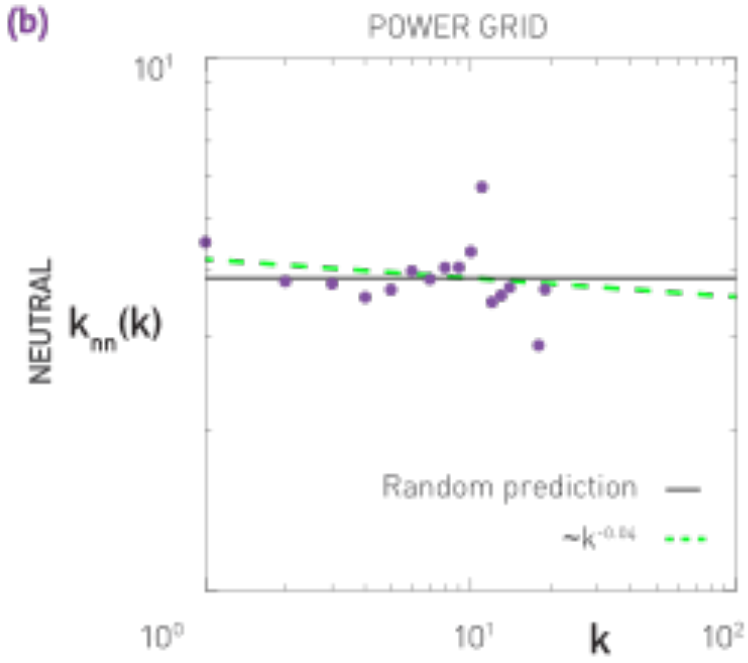
- **Neutral Network**

  For a neutral network (7.3)-(7.5) predict

$$P(k' \mid k) = \frac{e_{kk'}}{\sum_{k'} e_{kk'}} = \frac{e_{kk'}}{q_k} = \frac{q_{k'} q_k}{q_k} = q_{k'}.$$

  This allows us to express $k_{nn}(k)$ as

$$k_{nn}(k) = \sum_{k'} k' q_{k'} = \sum_{k'} k' \frac{k' p(k')}{\langle k \rangle} = \frac{\langle k^2 \rangle}{\langle k \rangle}.$$



R. Pastor-Satorras, A. Vázquez, A. Vespignani, Phys. Rev. E **65**, 066130 (2001)

Box 7.1

Friendship Paradox

The friendship paradox makes a suprising statement: *On average my friends are more popular than I am* [6,7]. This claim is rooted in (7.9), telling us that the average degree of a node's neighbors is not simply $\langle k \rangle$, but depends on $\langle k^2 \rangle$ as well.

Consider a random network, for which $\langle k^2 \rangle = \langle k \rangle (1 + \langle k \rangle)$. According to (7.9) $k_{nn}(k) = 1 + \langle k \rangle$. Therefore the average degree of a node's neighbors is always higher than the average degree of a randomly chosen node, which is $\langle k \rangle$.

The gap between $\langle k \rangle$ and our friends' degree can be particularly large in scale-free networks, for which $\langle k^2 \rangle / \langle k \rangle$ significantly exceeds $\langle k \rangle$ (Image 4.8). Consider for example the actor network, for which $\langle k^2 \rangle / \langle k \rangle = 565$ (Table 4.1). In this network the average degree of a node's friends is hundreds of times the degree of the node itself.

The friendship paradox has a simple origin: We are more likely to be friends with hubs than with small-degree nodes, simply because hubs have more friends than the small nodes.

Average next neighbor degree

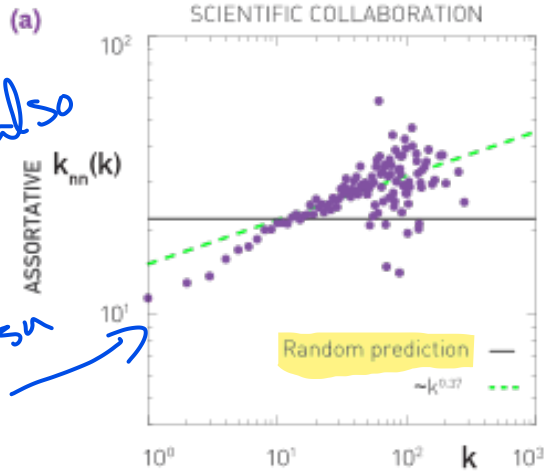
$k_{annd}(k)$: average degree of the first neighbors of nodes with degree k.

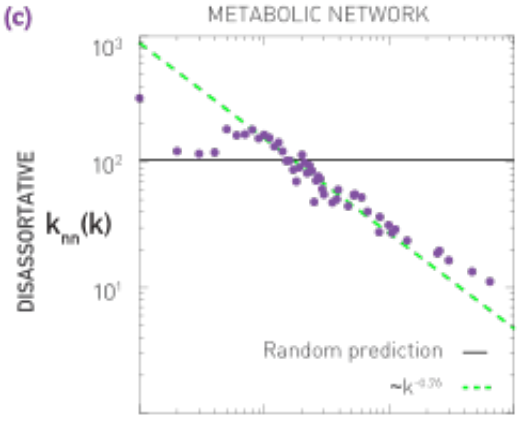
$$k_{nn}(k) \equiv \sum_{k'} k' P(k' \mid k)$$

- **Assortative Network**
  In assortative networks hubs tend to connect to other hubs, hence the higher is the degree k of a node, the higher is the average degree of its nearest neighbors. Consequently for assortative networks $k_{nn}(k)$ *increases with* k, as observed for scientific collaboration networks (Figure 7.6a).

- **Disassortative Network**
  In disassortative network hubs prefer to link to low-degree nodes. Consequently $k_{nn}(k)$ *decreases with* k, as observed for the metabolic network (Figure 7.6c).

if you have ↑ degree neighbours also have ↑ degree and vice versa





↑ d for nodes
↓ d for neighbour nodes

Average next neighbor degree

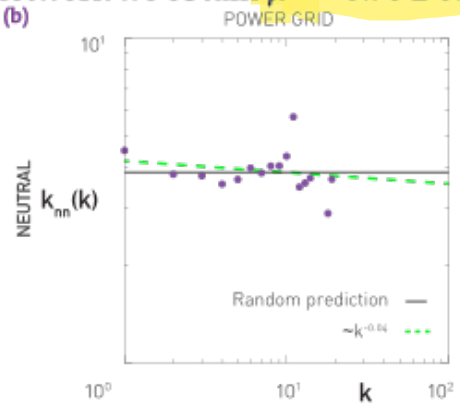
$$k_{nn}(k) = ak^{\mu}$$

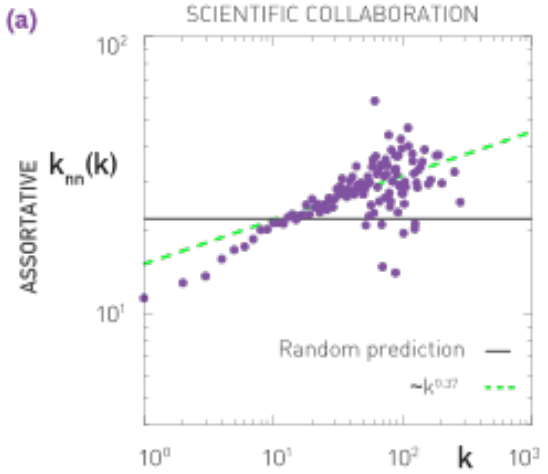
- **Assortative Networks: $\mu > 0$**
  A fit to $k_{nn}(k)$ for the science collaboration network provides $\mu = 0.37 \pm 0.11$ (Figure 7.6a).

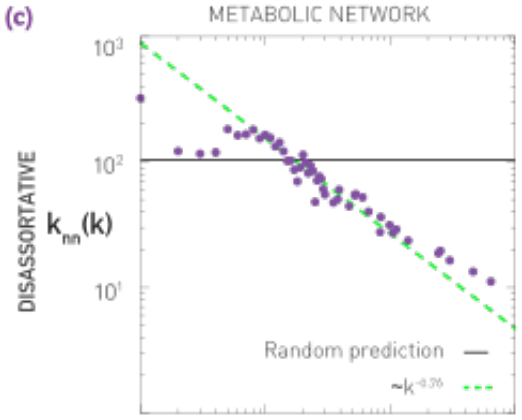
- **Neutral Networks: $\mu = 0$**
  According to (7.9) $k_{nn}(k)$ is independent of k. Indeed, for the power grid we obtain $\mu = 0.04 \pm 0.05$, which is indistinguishable from zero (Figure 7.6b).

  $\therefore \mu = 0$

- **Disassortative Networks: $\mu < 0$**
  For the metabolic network we obtain $\mu = -0.76 \pm 0.04$ (Figure 7.6c).







Average next neighbor degree

$k_{annd}(k)$: average degree of the first neighbors of nodes with degree k.

constraint: $\sum_k k_{annd}(k) \cdot kNp_k = \sum_k k^2 \cdot Np_k$

$\langle k_{annd}(k) k \rangle = \langle k^2 \rangle$ ⟶ k_{max}-1 independent elements

$k_{annd}(k)$ is a k-dependent function, hence it has much fewer parameters,

and it is easier to interpret/read.

R. Pastor-Satorras, A. Vázquez, A. Vespignani, Phys. Rev. E **65**, 066130 (2001)

Degree Correlation Coefficient

If there are degree correlations, e_{jk} will differ from $q_j q_k$. The magnitude of the correlation is captured by <jk>-<j><k> difference, which is:

$$\sum_{jk} jk(e_{jk} - q_j q_k)$$

normalize with the maximum value

<jk>-<j><k> is expected to be:

positive for *assortative* networks,
zero for *neutral* networks,
negative for *dissasortative* networks

To compare different networks, we should normalize it with its maximum value; the maximum is reached for a *perfectly assortative network*, i.e. $e_{jk}=q_k\delta_{jk}$

normalization: $\sigma_r^2 = \max \sum_{jk} jk(e_{jk} - q_j q_k) = \sum_{jk} jk(q_k \delta_{jk} - q_j q_k)$

$$r = \frac{\sum_{jk} jk(e_{jk} - q_j q_k)}{\sigma_r^2} \qquad -1 \le r \le 1$$

$r \le 0$ *disassortative*
$r = 0$ *neutral*
$r \ge 0$ *assortative*

M. E. J. Newman, Phys. Rev. Lett. **89,** 208701 (2002)

REAL NETWORKS

Network	n	r
Physics coauthorship (a)	52 909	0.363
Biology coauthorship (a)	1 520 251	0.127
Mathematics coauthorship (b)	253 339	0.120
Film actor collaborations (c)	449 913	0.208
Company directors (d)	7 673	0.276
Internet (e)	10 697	−0.189
World-Wide Web (f)	269 504	−0.065
Protein interactions (g)	2 115	−0.156
Neural network (h)	307	−0.163
Marine food web (i)	134	−0.247
Freshwater food web (j)	92	−0.276
Random graph (u)		0
Callaway *et al.* (v)		$\delta/(1+2\delta)$
Barabási and Albert (w)		0

Social networks are *assortative*

Biological, technological networks are *disassortative*

r>0: assortative network:
Hubs tend to connect to other hubs.

r<0: disassortative network:
Hubs tend to connect to small nodes.

Relationship between knn and r

$k_{nn}(k) = rk.$ → assumes linear relation

Which one is right?

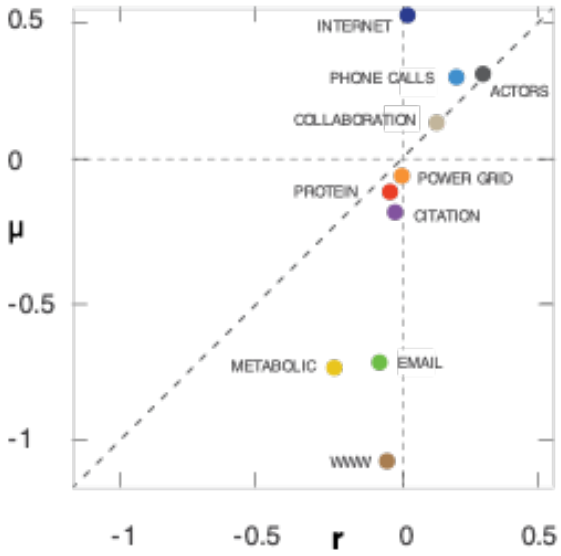
$$k_{nn}(k) = ak^{\mu}$$

Relationship between knn and r

$$k_{nn}(k) - rk.$$

$$k_{nn}(k) = ak^{\mu}$$

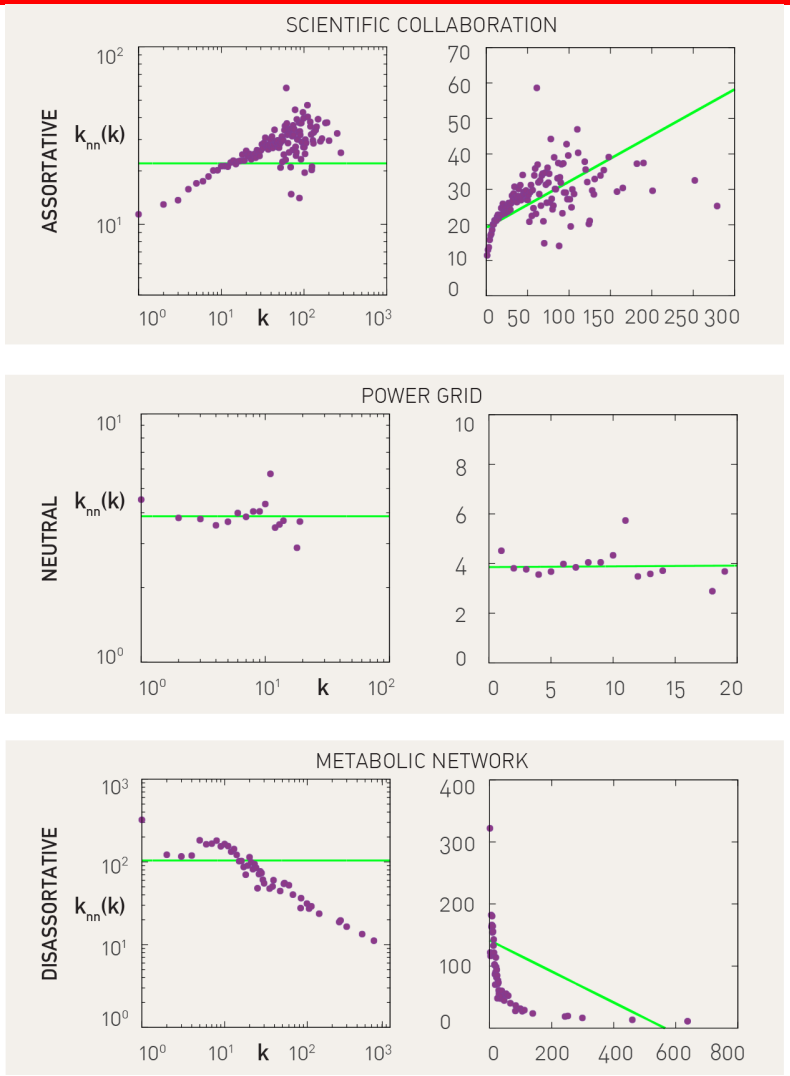
NETWORK	N	r	μ
Internet	192,244	0.02	0.56
WWW	325,729	-0.05	-1.11
Power Grid	4,941	0.003	0.0
Mobile Phone Calls	36,595	0.21	0.33
Email	57,194	-0.08	-0.74
Science Collaboration	23,133	0.13	0.16
Actor Network	702,388	0.31	0.34
Citation Network	449,673	-0.02	-0.18
E. Coli Metabolism	1,039	-0.25	-0.76
Protein Interactions	2,018	0.04	-0.1



Relationship between knn and r

$$k_{nn}(k) - rk.$$

$$k_{nn}(k) = ak^{\mu}$$



even though :

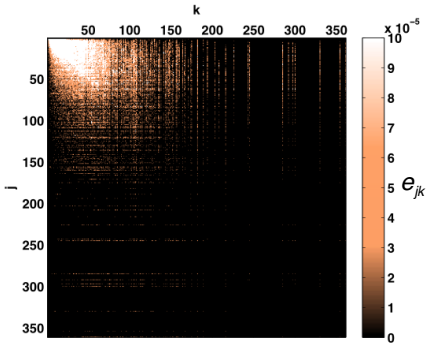
$\mu > 0 \therefore r > 0$

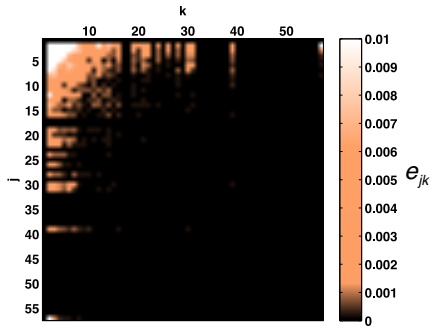
if

$\mu < 0 \therefore r < 0$ → even if its not a perfect fit.

DEGREE CORRELATIONS IN NETWORKS

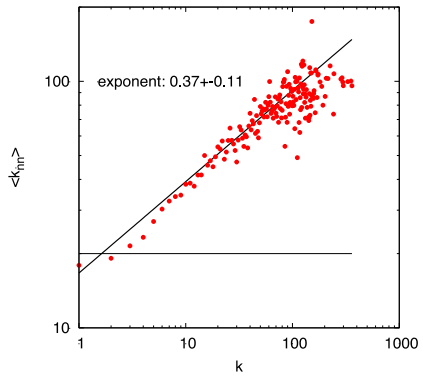
e_{jk}

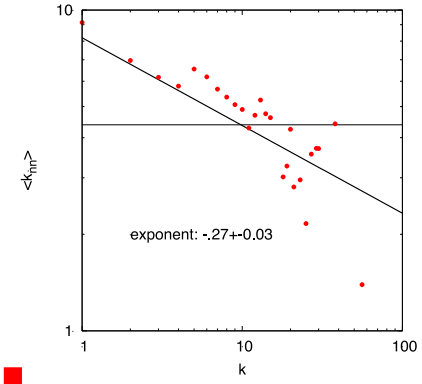




$$\frac{k_{\max}\left(k_{\max}-1\right)}{2}-k_{\max}-1$$

$k_{annd}(k)$





$k_{\max}-1$

r

0.31

-0.16

1

Real Networks:
Null Models

"A deep and insightful book that is a joy to read. There are new ideas on every page, and none of them is obvious!"
—DANIEL GILBERT, Professor of Psychology at Harvard University and author of *Stumbling on Happiness*

Everything Is Obvious*

How Common Sense Fails Us

DUNCAN J. WATTS

*Once You Know the Answer

Good science means making comparisons with respect to an expectation

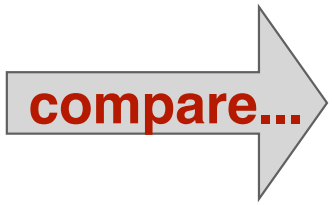
The sad case of Sally Clark

PEOPLE HAVE DIED FROM LACK OF NULL MODELS

null models should tell you what you should expect.

Build an expectation for your measures

Measurement on your network data.



...to networks with same *N* and *<k>* as your data (ER networks)

...to networks with same degree distribution (degree-preserving randomizations, configuration model, ...)

similar edges too

Build an expectation for your measures

Measure, for example, the clustering coefficient C

Comparison to ER network

- Generate many Erdös-Rényi networks with same *N* and <*k*>
- In each of them measure the clustering coefficient, C_i^{ER}
- Average the clustering coefficient over all realizations to get a single number

$$C^{ER} = \sum_i C_i^{ER}$$

Build an expectation for your measures

Measure, for example, the clustering coefficient C

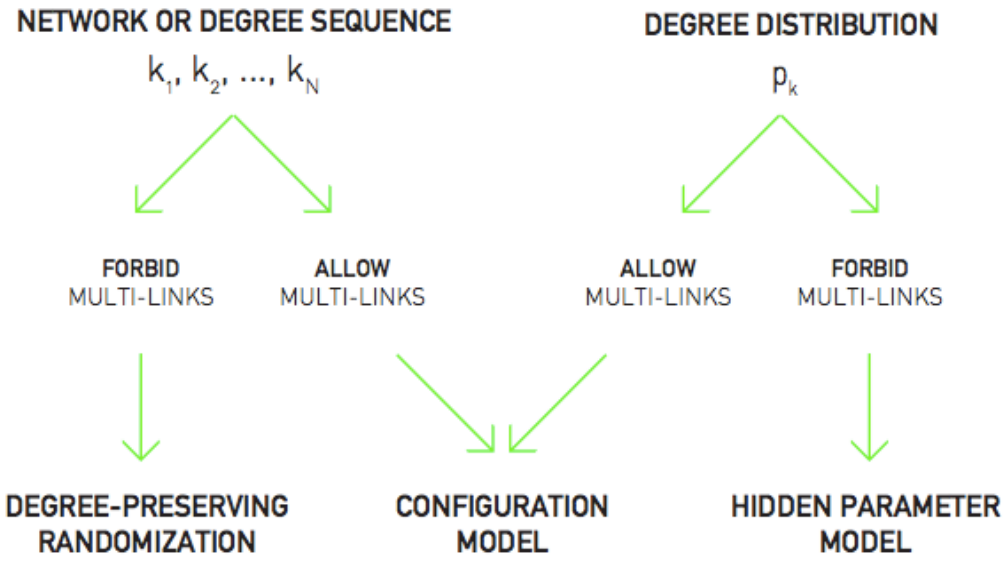
Comparison with networks with same degree sequence

→ Generate many degree-preserved randomizations (same *degree sequence)*

→ In each of them measure the clustering coefficient, C_i^{DR}

→ Average the clustering coefficient over all realizations to get a single number $C^{DR} = \sum_i C_i^{DR}$

Degree-preserving randomization, configuration model, hidden parameter model?



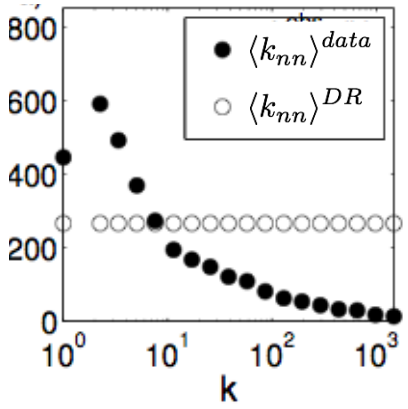
A.-L. Barabási. (2014). *Network Science, Chapter 4.*

For your final project

Have a table for global quantities, like clustering coefficient, average path length (one network --> one number)

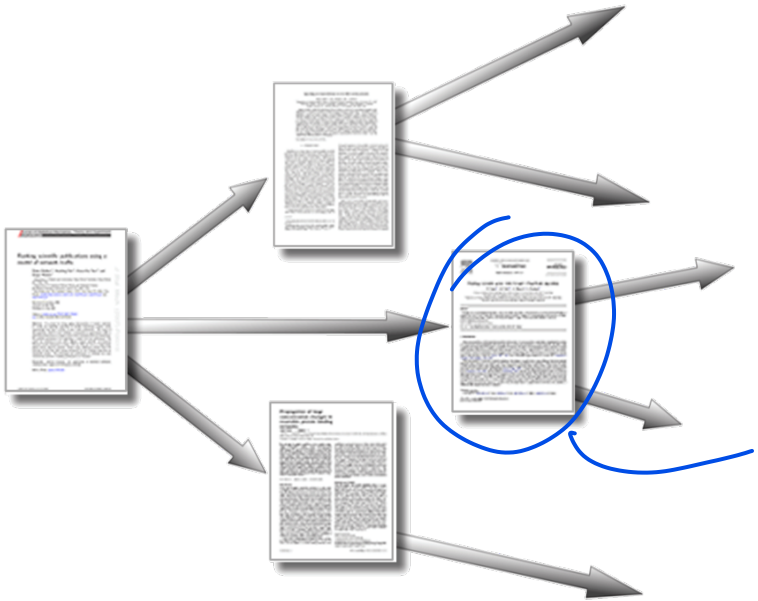
Data	ER	DP

Have plots to compare distributions, scaling, etc



In some cases rewiring (configuration model) is not a good starting point

TIME



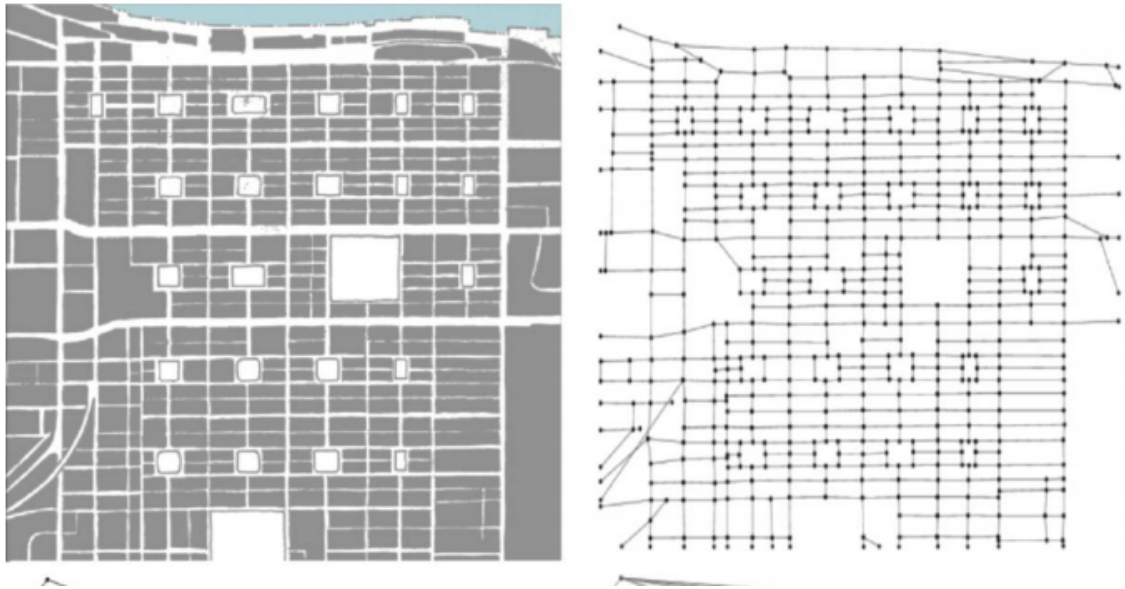
rewiring needs to be taken care of carefully

This paper cannot be referenced by an older paper

citation network

In some cases rewiring (configuration model) is not a good starting point

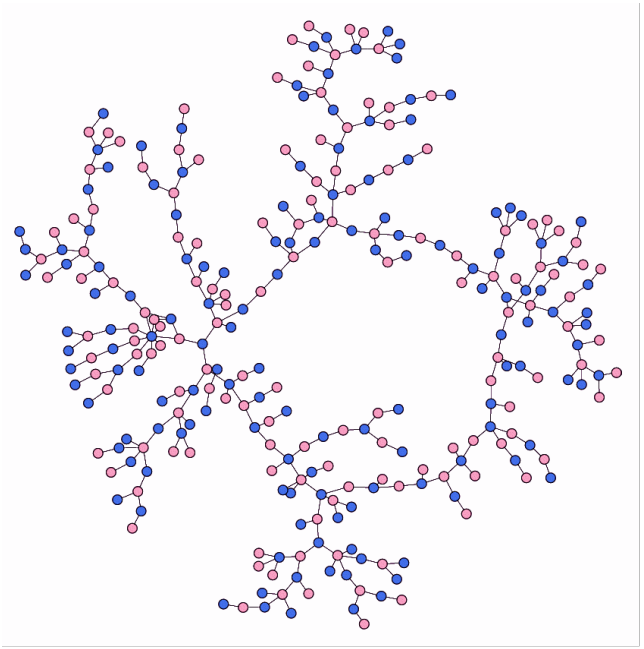
SPACE



Urban infrastructure

In some cases rewiring (configuration model) is not a good starting point

OTHER NODAL PROPERTIES

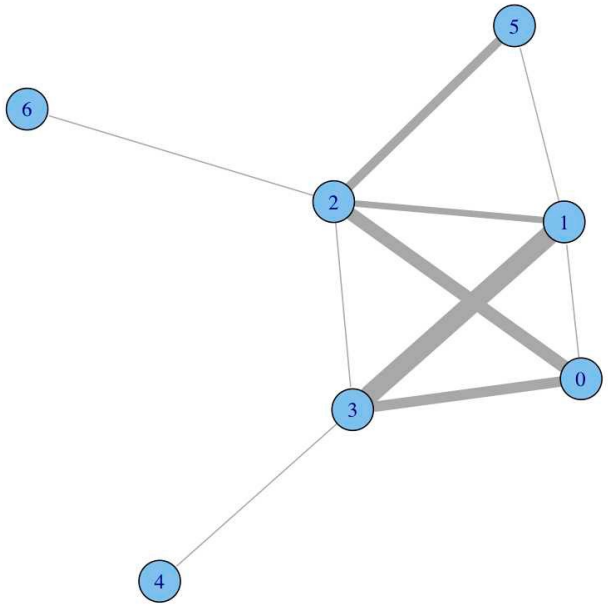


have to wire it based on your network

e.g. sexual relationships

In some cases rewiring (configuration model) is not a good starting point

LINK PROPERTIES



e.g. weighted networks, bipartite networks

Assortativity coefficient

$$r = \frac{\sum_{jk} jk \left(e_{jk} - q_j q_k\right)}{\sigma^2}$$

data

null model

with $\sigma^2 = \sum_k k^2 q_k - \left[\sum_k k q_k\right]^2$

In some cases, the null model is built in, but you still have to be aware of which null model that is!