Network Science

Network Measures

Lab 03

Lecture 08

Today's topics

- Slice Weighted Networks
- Measuring Networks
- Case Study: Panama Papers

Slice Weighted Networks

In weighted networks, some edges are strong, some are weak

Sometimes you cannot keep all the edges

Slicing is the process of eliminating low-strength edges

Simplest form: you choose a cut-off threshold T that controls the density of the resulting network

$\forall$ edge e :
  if weight $(e) < T$
    remove e;

how to select T

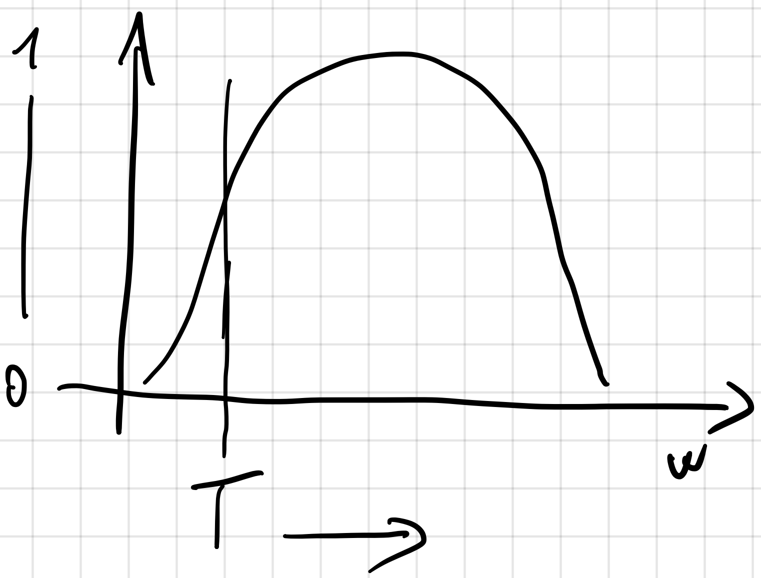
1. Select a T based on edge weight distribution

2. Slice the network

3. Calculate some network properties (# components, density, ...)

4. if you are unsatisfied with results, go back to 1.

run 06_Slicing.ipynb

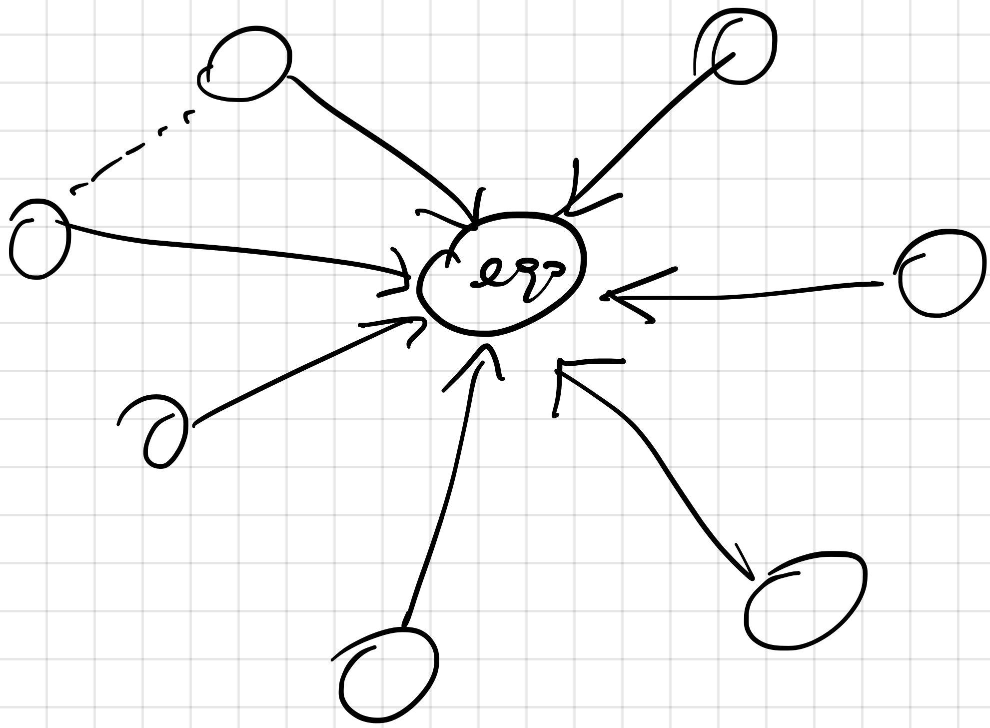


Measuring Networks

with notebook 07_measuring we will learn how to calculate the following properties with networkx:

- density
- clustering coefficient
- transitivity
- paths and distances
- eccentricity, diameter, radius, periphery
- basic centralities
- assortativity
- homophily

the notebook calculates some metrics for the ego network of a previously retrieved Wikipedia page



ego network = ego + alters + connections between the alters

Density

fraction of existing edges
by not existing edges

m: # edges
N: # nodes

G: graph

$$d = \frac{m}{n(n-1)}$$ if G is directed

$$d = \frac{2m}{n(n-1)}$$ if G is undirected

Clustering coefficient

Networkx : only for undirected graph

$$cc(i) = \frac{2|L(i)|}{|N(i)|\,(|N(i)|-1)}$$

i : node

$N(i)$ = Neighbors of i

$L(i)$ = links connecting neighbors of i

$cc(i)$: is the probability that "friends" of i are connected each other

otherwise : the number of triangles closed by i

clustering coefficient **metrics**:

cc (i) : clustering coefficient of a node

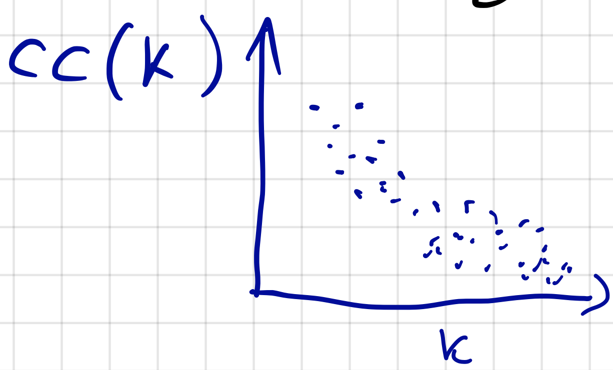
cc (G) : average clustering coefficient of the graph

$$cc(A) = \sum_{i=1}^{n} cc(i)$$

transitivity : number of triangles in the graph

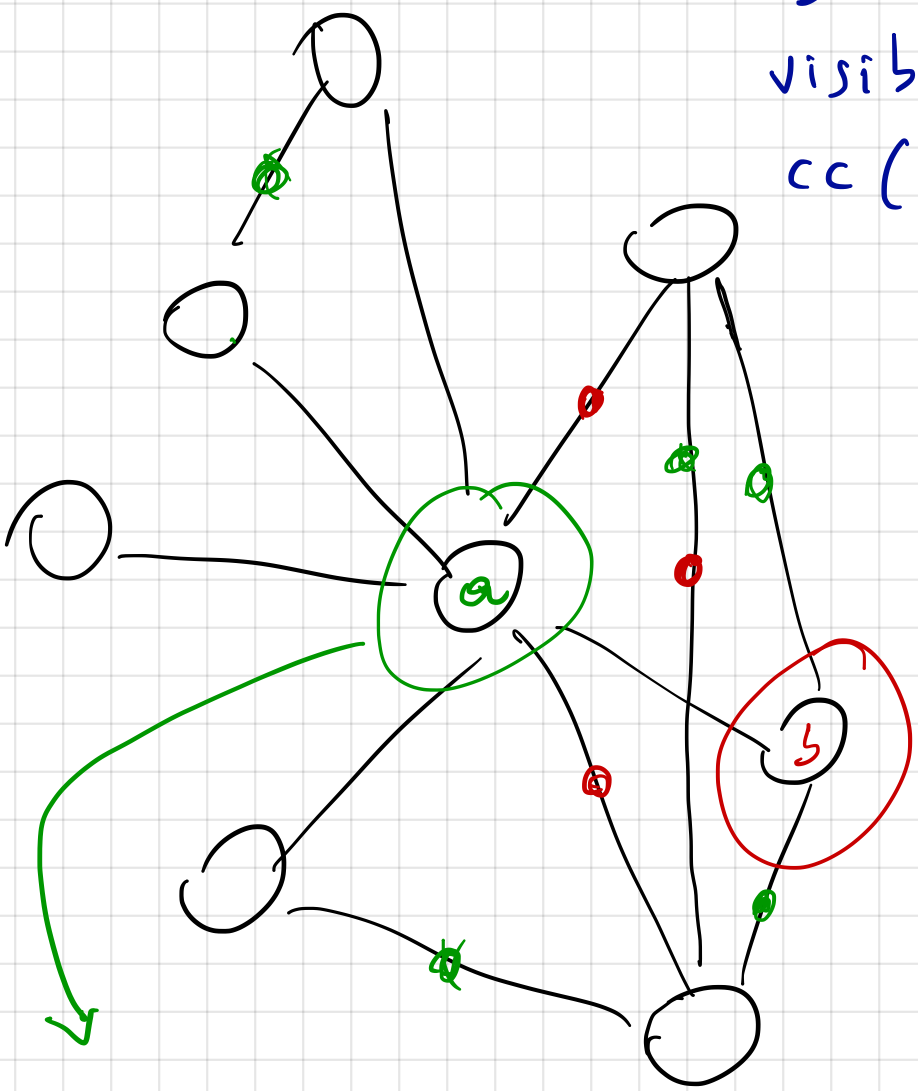
distribution of clustering coefficient:

usually plotted as a function of the node degree



Hubs are more unlikely to have high clustering coefficient w.r.t. low degree nodes.

↳ this is "usually" visible if distribution of cc(k) is plotted



5 links (between friends)

7 neighbors

$$cc(a) = \frac{2 \cdot 5}{7(7-1)} = \frac{10}{42}$$

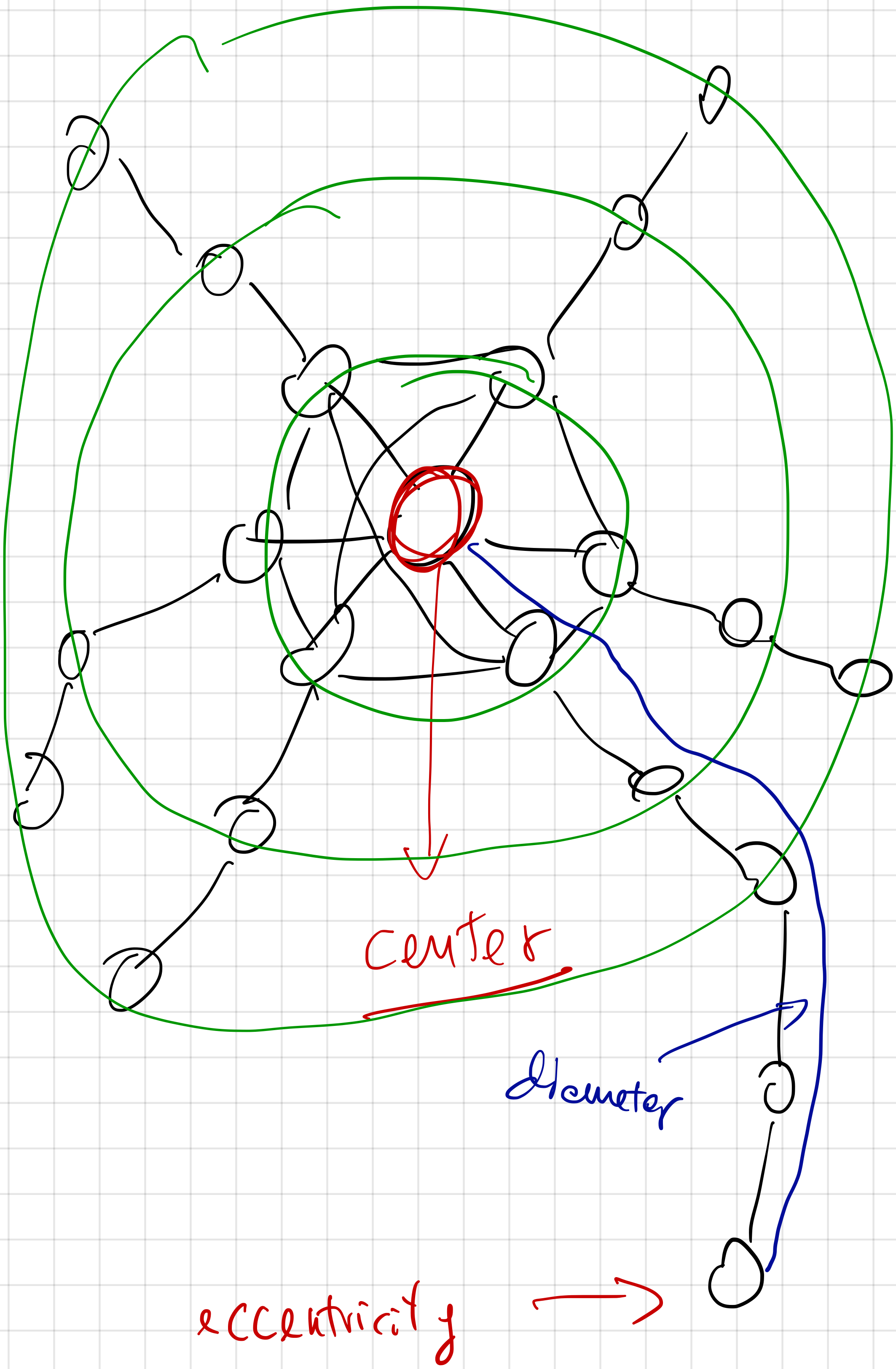
3 links

3 neighb

$$cc(b) = \frac{2 \cdot 3}{3 \cdot 2} = 1$$

$$cc(a) < cc(b)$$

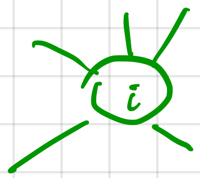
Eccentricity, diameter, radius, and center of a Graph



Degree Centrality

popularity

$$D(i) = \frac{k_i}{N-1}$$



$k_i = 5$

Betweenness Centrality

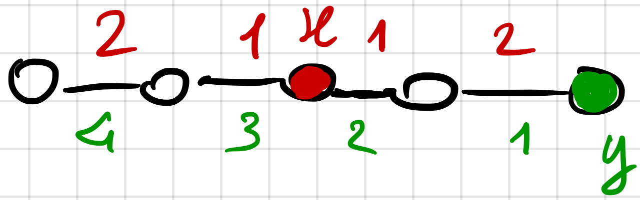
$$B(i) = \sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}$$

another kind of "importance"

- "structural holes"
- gate keepers
- ...

Closeness Centrality

$$C(i) = \left[\sum_{j=1}^{N} d(i,j)\right]^{-1}$$



$= 6 \rightarrow C(x) = \frac{1}{6}$

$= 10 \rightarrow C(y) = \frac{1}{10}$

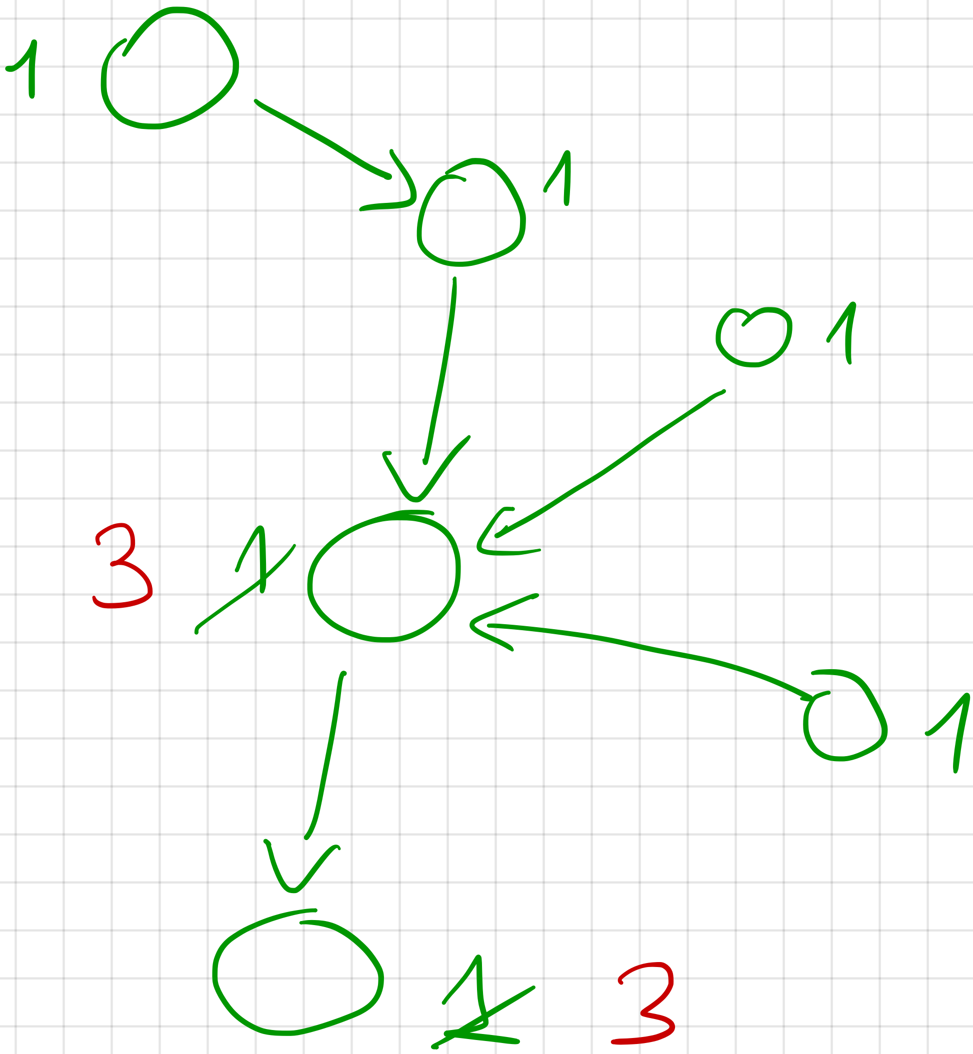
Eigenvector Centrality

(more complicated)

high eigenvector centralities identifies nodes that are surrounded by other nodes with high eigenvector centrality

It's a measure of prestige

if we set initial eigenvector centrality values to 1, after few iterations we may refine our calculations



Page Rank

similar to Eigen Vector centrality

developed by Google (Page, Brin)

the rank of a node is calculated as the probability that a "person" randomly traversing the edges will arrive at the node

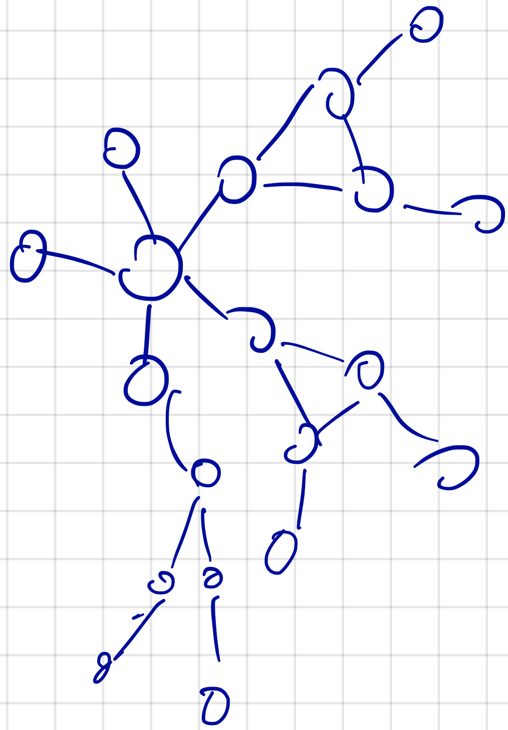
damping factor ($\alpha = 0.85$):

↳ the probability that the user will continue clicking

HITS: HUBS and Authorities

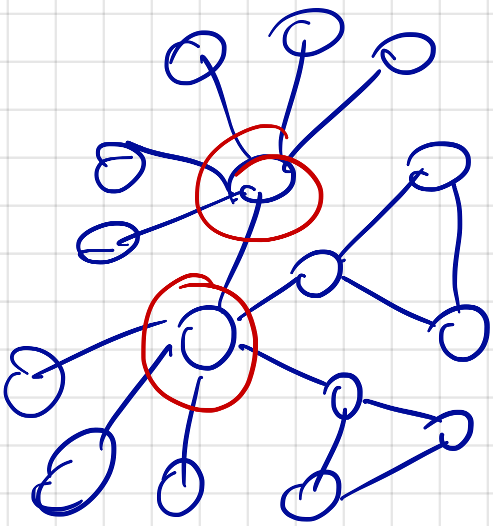
it is an extended version of Page Rank

Assortativeness



hierarchy

disassortative = heterophily by degree

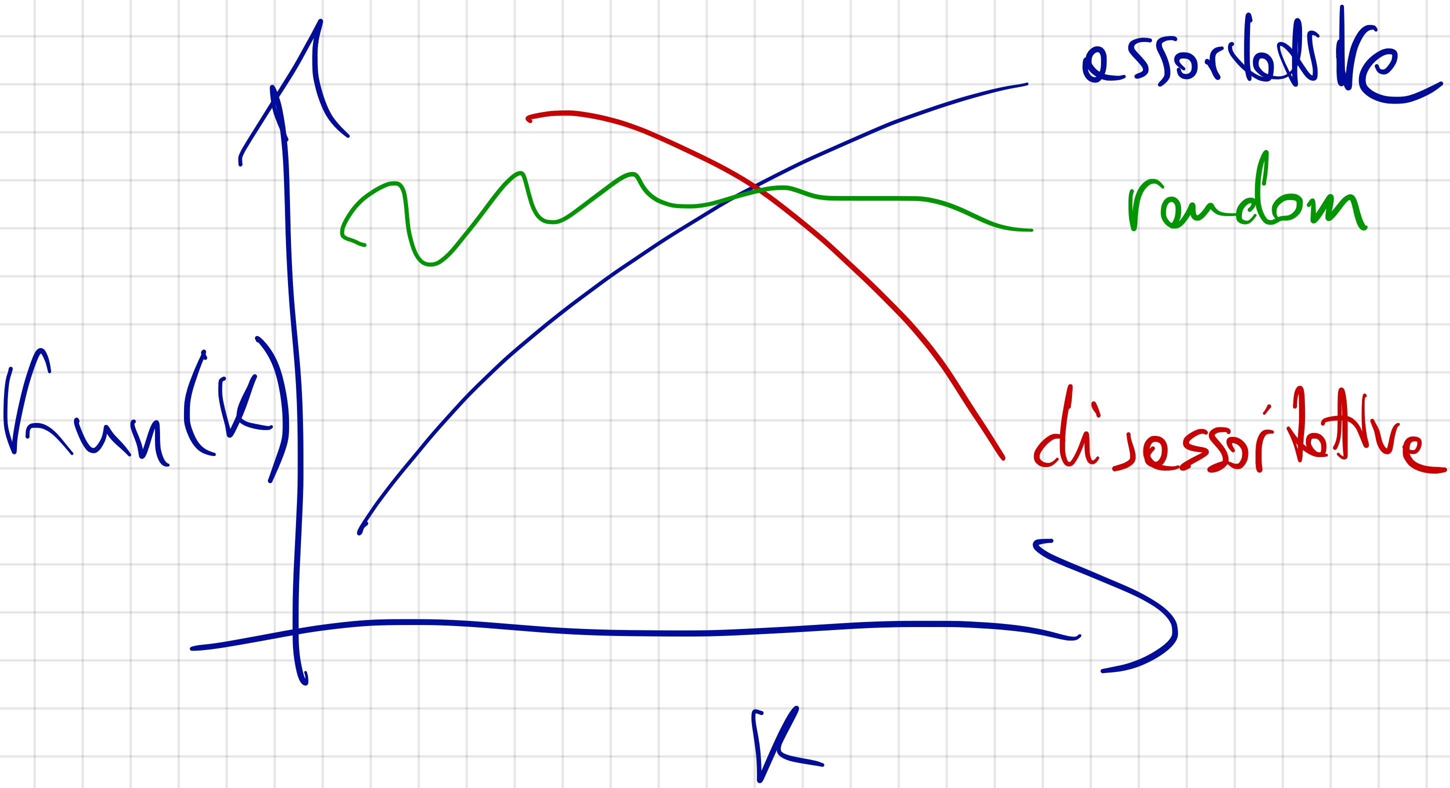


core-periphery

assortative = homophily by degree

$K_{nn}(k)$ =

average degree
for
nearest neighbors
of nodes with
degree = k



Exercise

Go back to your "classmates" network and calculate all the basic properties we learnt to measure today

Also : Focus on Homophily
(gender , study program, ...)

Panama Papers

run 08_ panama

09_ panama-ca

Exercise

modify previous notebooks,
to evaluate the measures
you obtained with
"panama papers" data.
to evaluate = are they
"significant"

Hint: you need some
baseline ...

Hint 2: you may want
to consider random
models ...