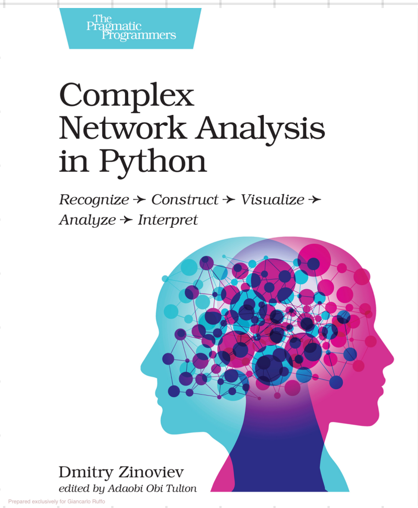
Lecture 10

Complex Network Analysis

Network Measures and Structure

Today's Topics

- Measuring Networks
- Case Study: Panama Papers
- Beyond Social Network Analysis
- Analysing the structure of a network



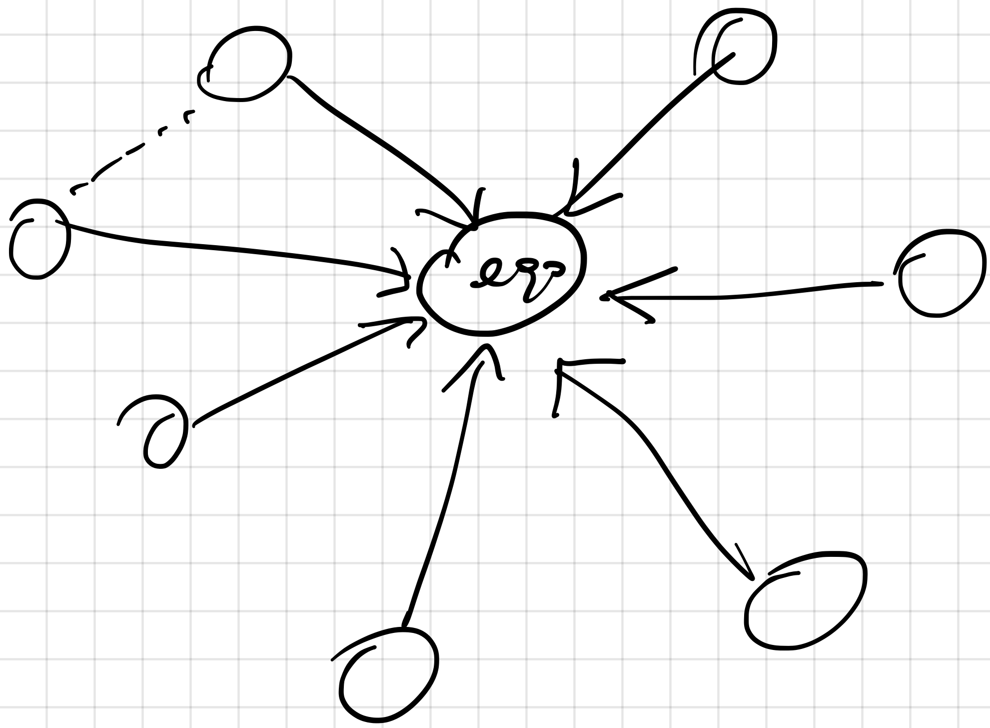
Chapters: 8 – 13

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

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

i : node

$N(i)$ = Neighbors of i

$\mathcal{L}(i)$ = 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

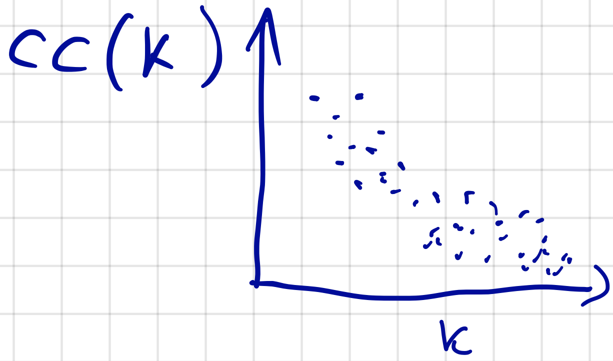
cc (G) : average clustering coefficient of the graph

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

transitivity : number of triangles in the graph

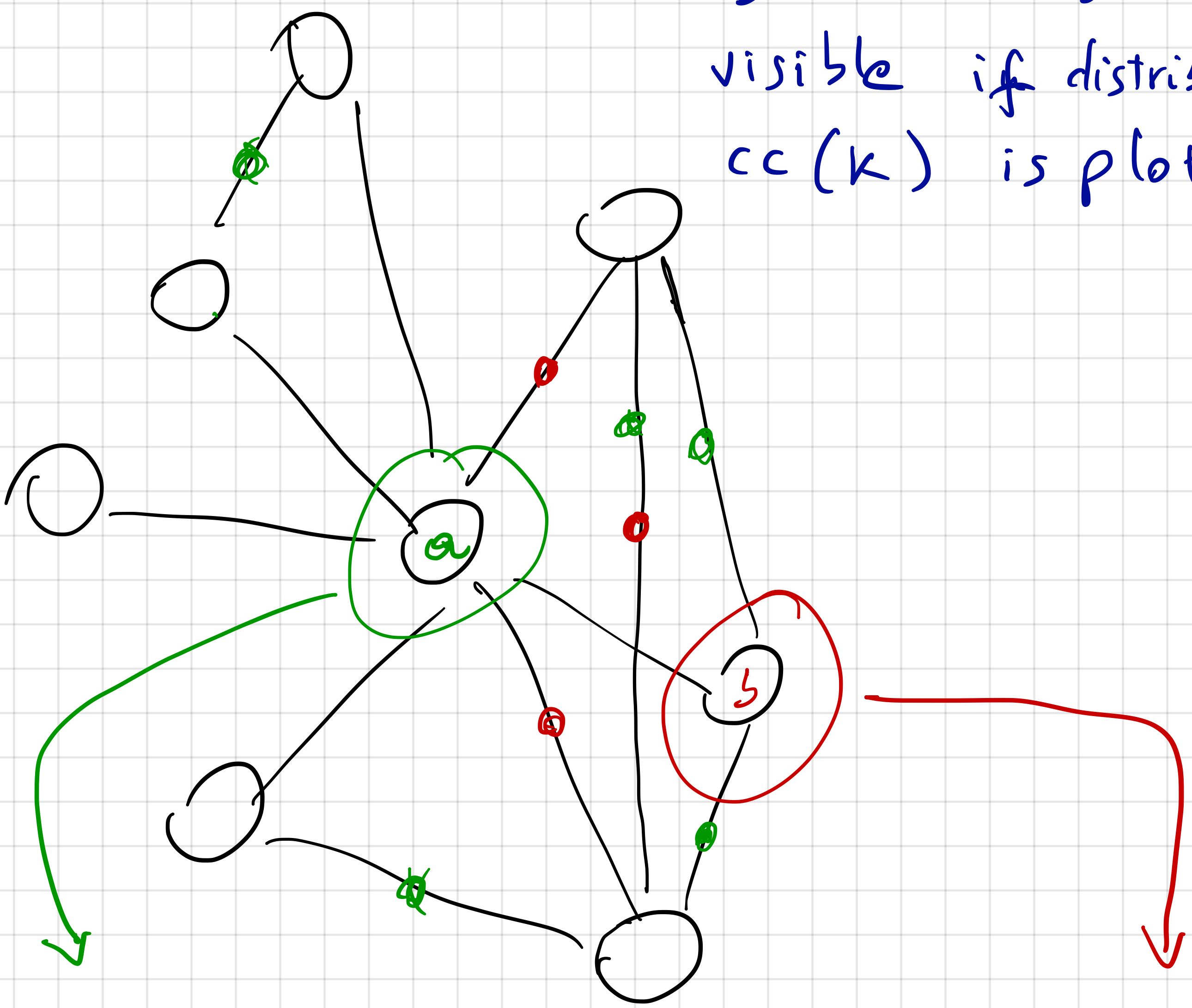
distribution of clustering coefficient:

usually plotted as a function of the node degree



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

$\hookrightarrow$ this is "usually" visible if distribution of cc(k) is plotted



5 links (between friends)

7 neighbors

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

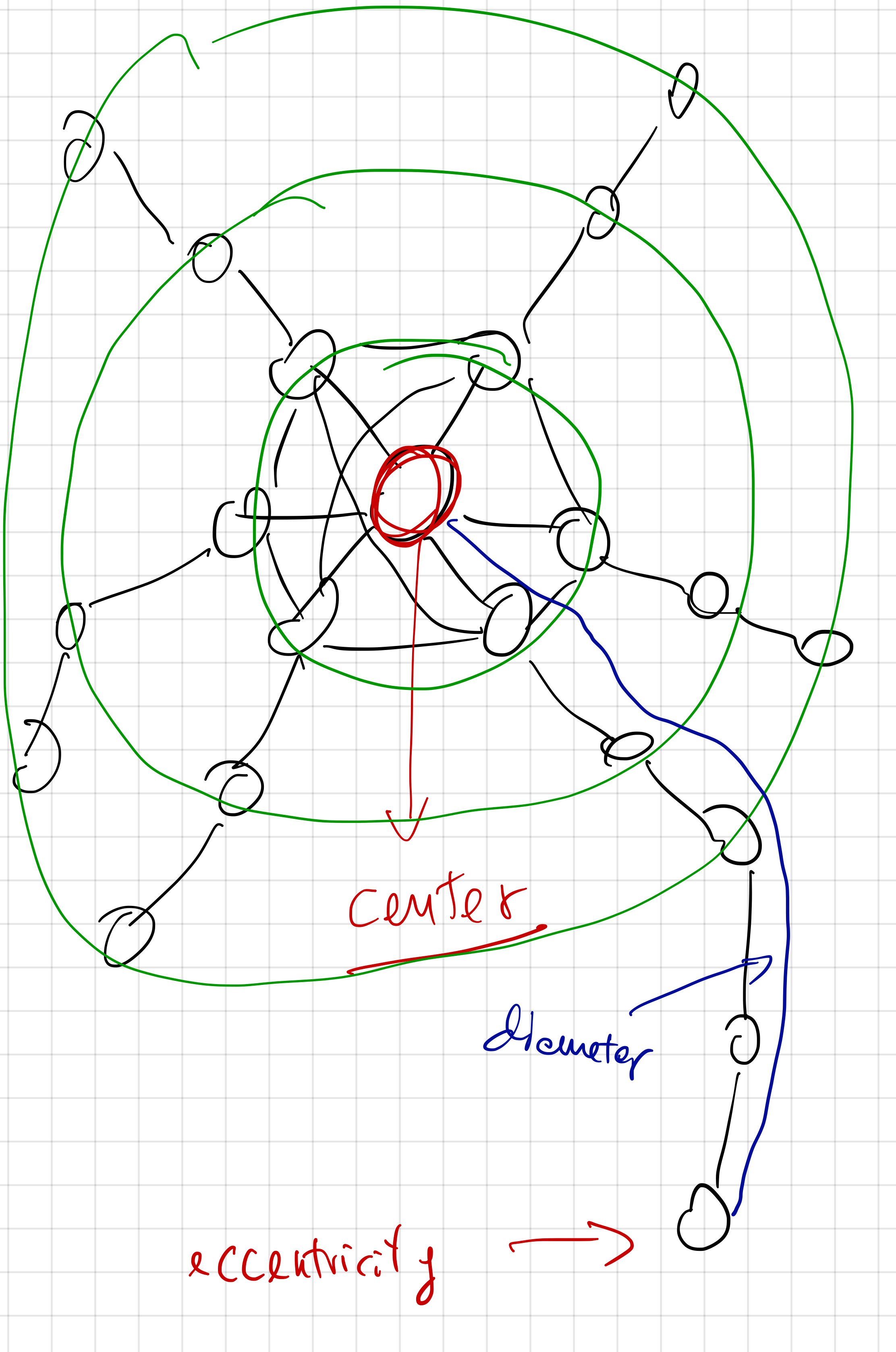
3 links

3 neighb

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

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

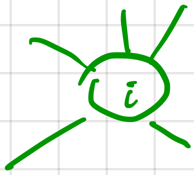
Eccentricity, diameter, radius, and center of a Graph



Degree Centrality

popularity

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



$k_i = 5$

Betweenness Centrality

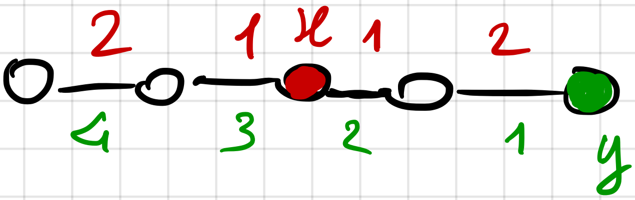
$$B(i) = \sum_{j<k} \frac{d_{jk}(i)}{d_{jk}}$$

another kind of "importance"

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

Closeness Centrality

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



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

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

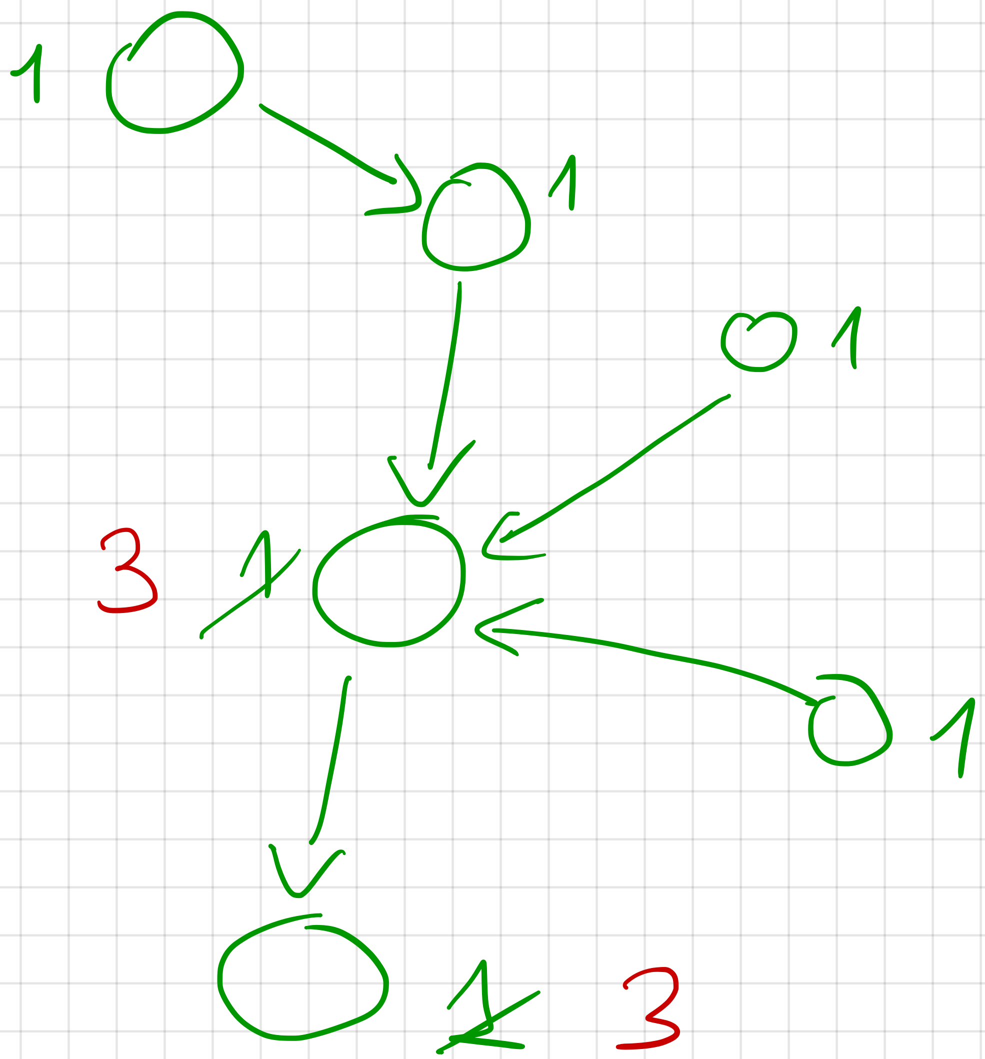
Eigenvector Centrality

(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

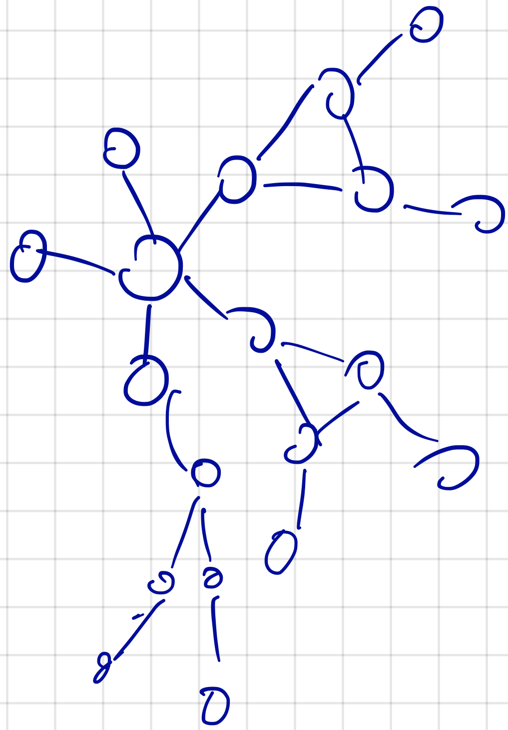
damping factor ($\alpha = 0.85$):

↳ the probability that the user will continue clicking

HITS: HUBS and Authorities

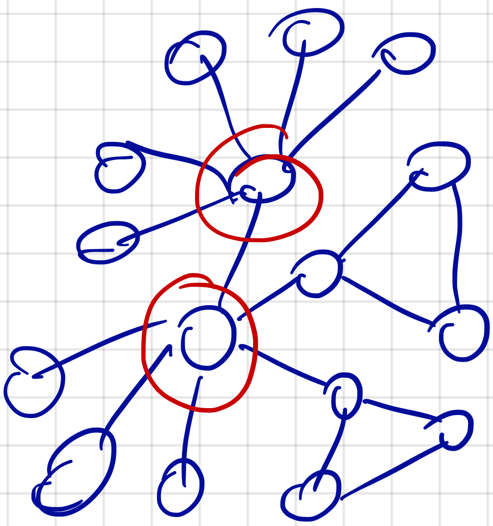
it is an extended version of Page Rank

Assortativeness



hierarchy

disassortative = heterophily by degree

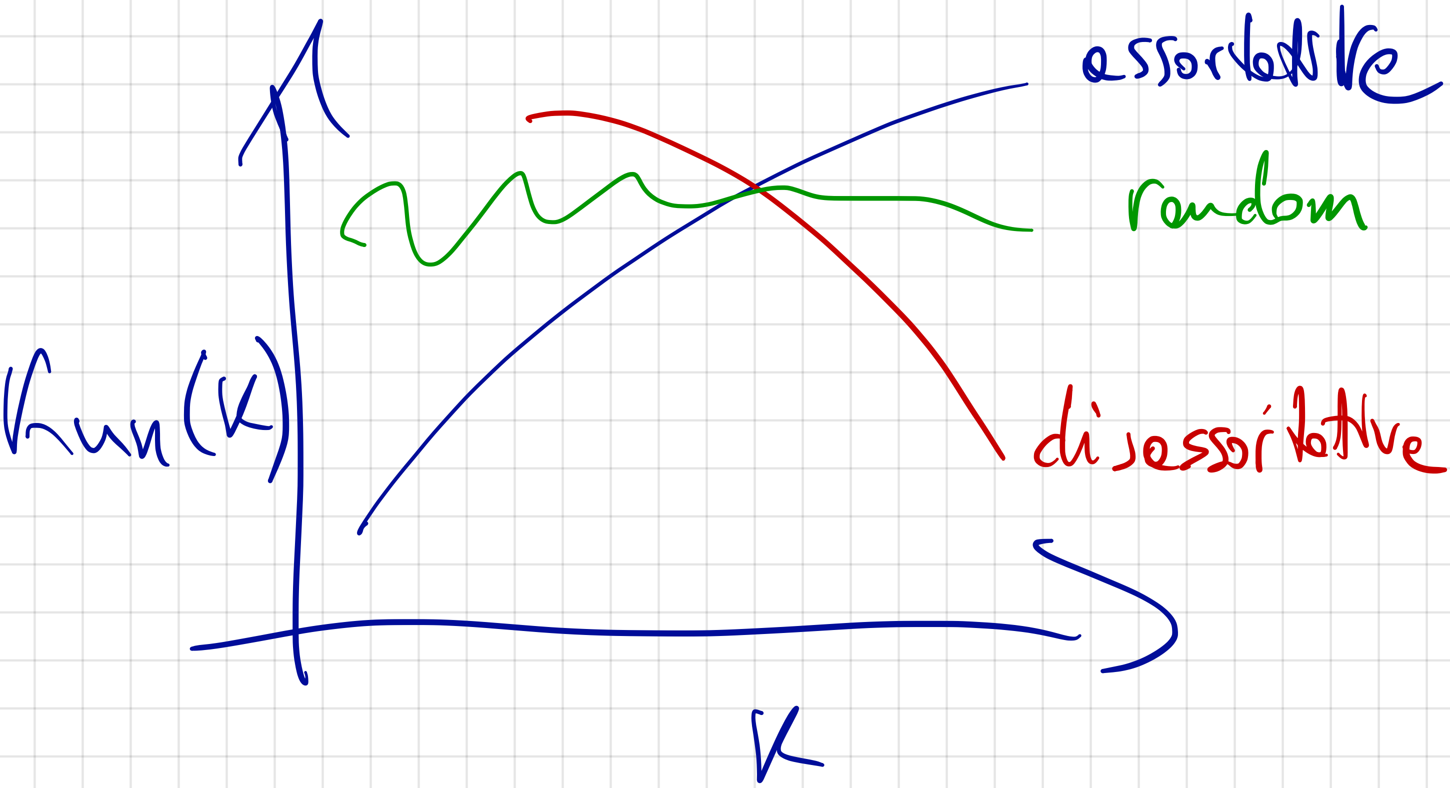


core-periphery

assortative = homophily by degree

$K_{nn}(k)$ =

average degree
for
nearest neighbors
of nodes with
degree = k



Panama Papers

run 08_ panama

09_ panama-cc

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 ...

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 ...

Beyond Social Network Analysis

- Constructing Semantic and Product Networks
- Discovering the Network Structure
- Information Networks (the Web) : <u>next lecture</u>

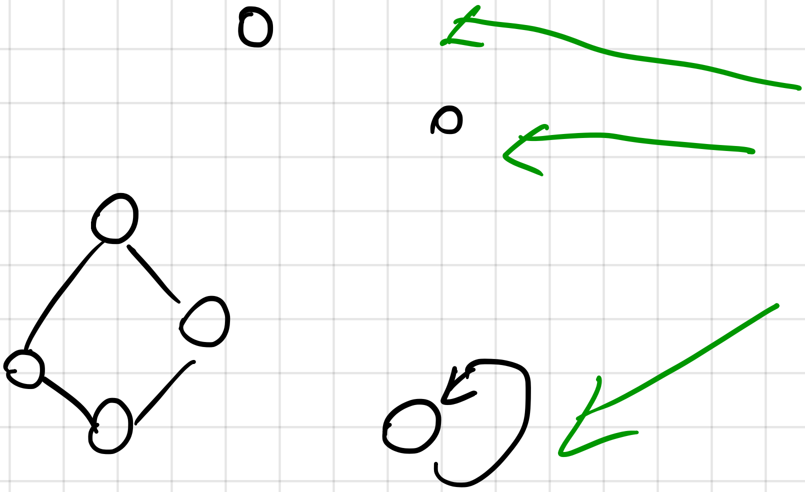
Discovering the network Structure

The Web is (or was...) a bow-tie.

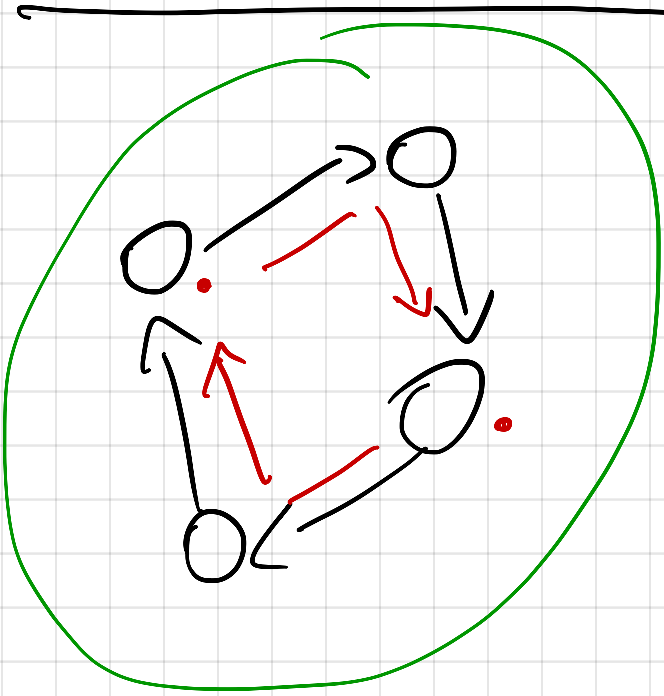
How can we replicate such an analysis on other networks?

For example: we want to find the Giant Connected Component and also:
isolates, components, cliques, communities, K-cores, ...

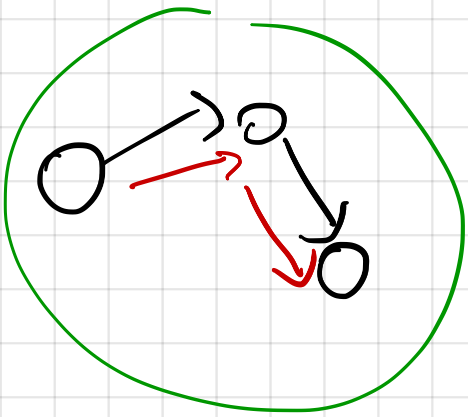
open 10_makeFigures.ipynb and learn how to do that with networkx



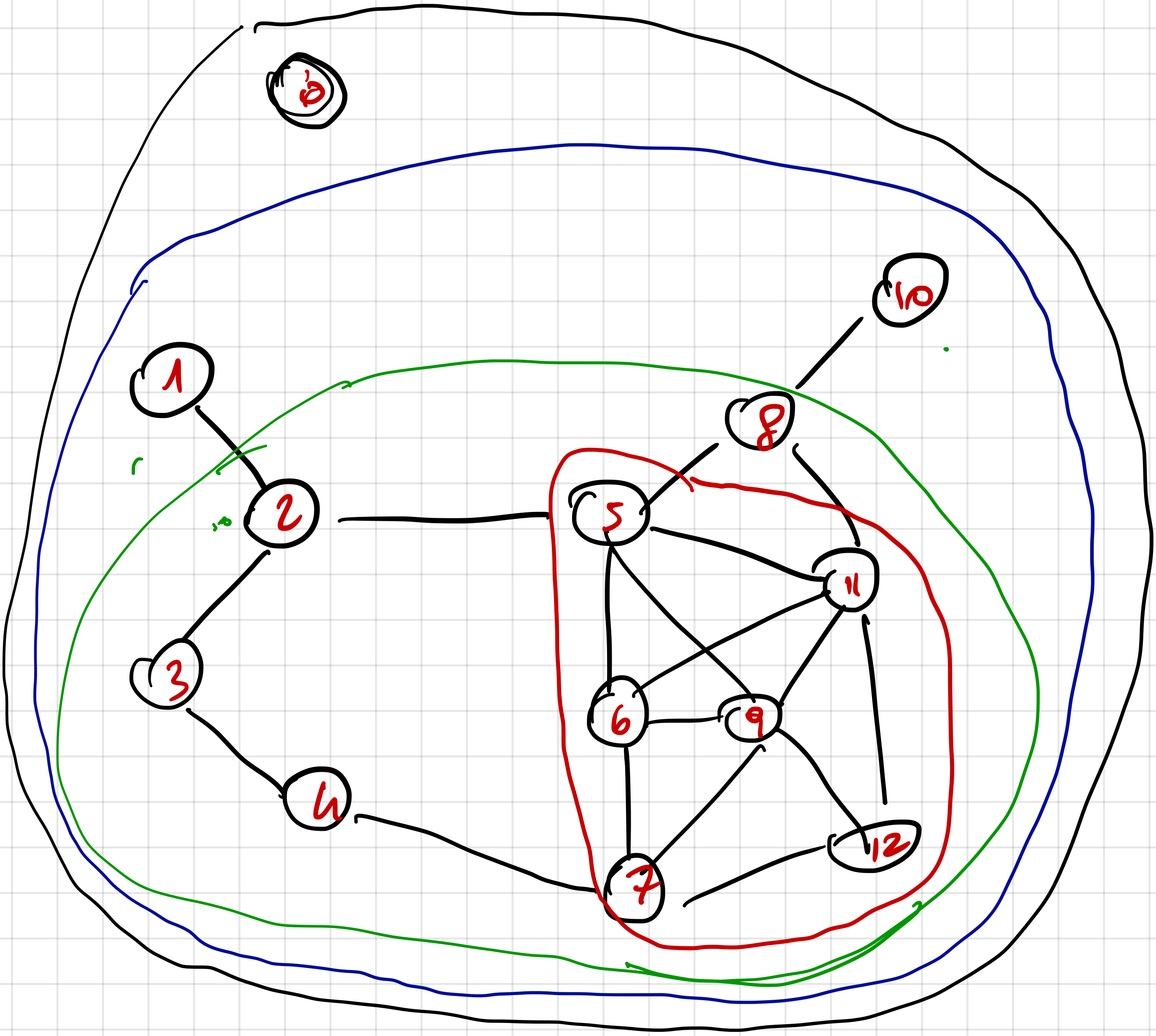
isolates



Strongly Connected Component (SCC)



Weakly CC (WCC)



k -core:

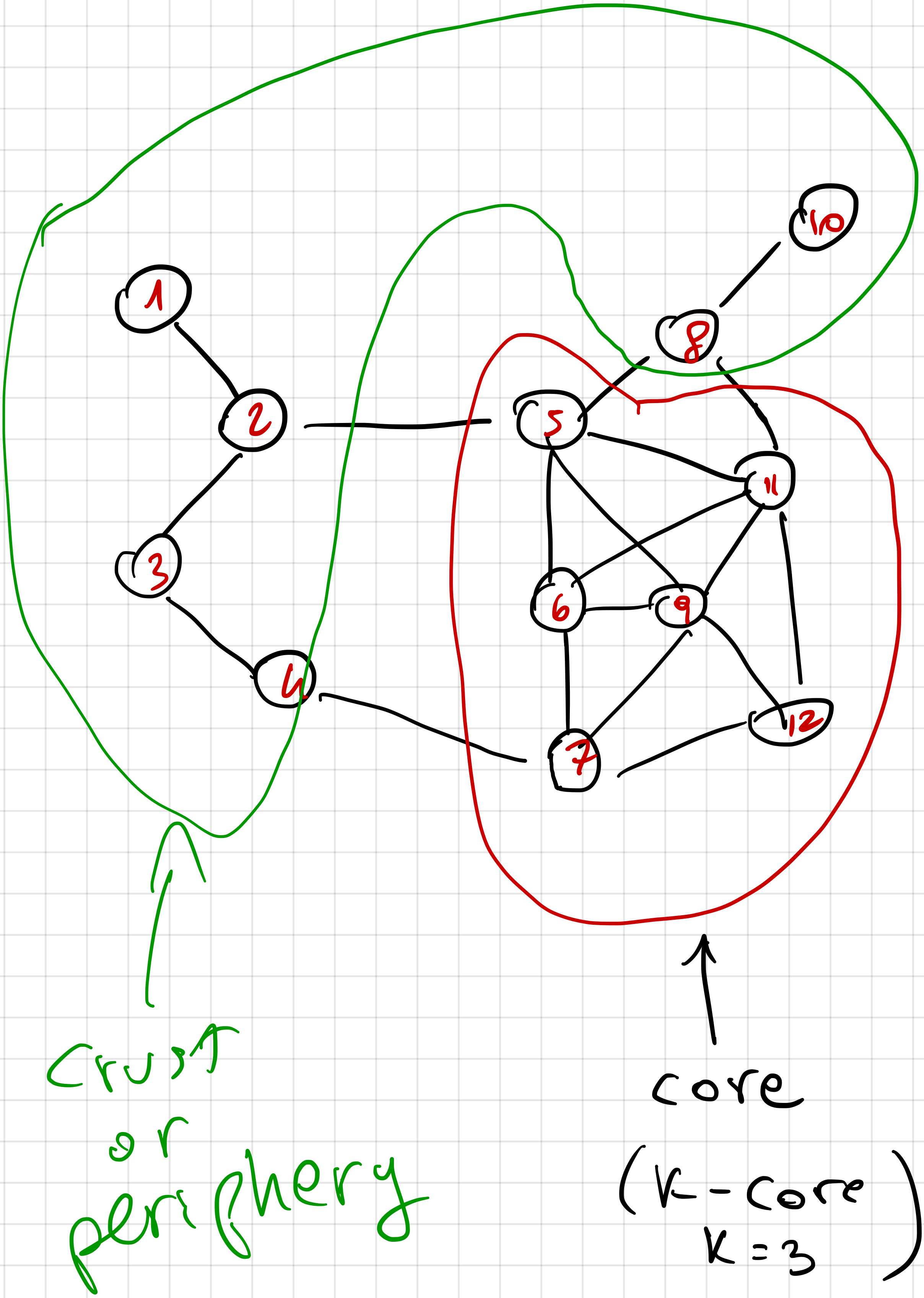
k

0 - core

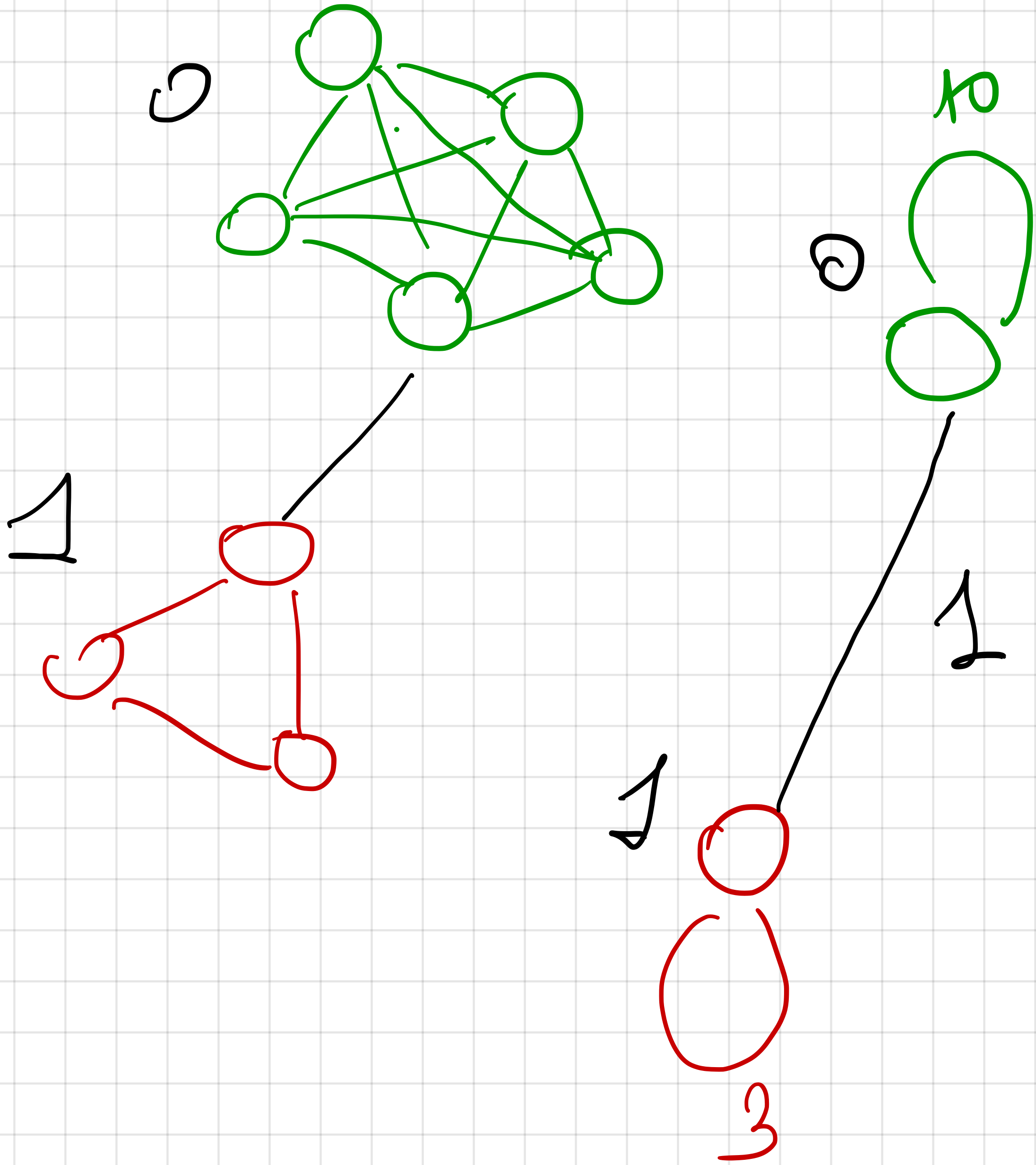
1 - core

2 - core

3 - core


1
2
3
4
5
6
7
8
9
10
11
12
crust
or
periphery
core
(k-core
k=3)

BLOCK MODELING



graph $\Longrightarrow$ induced graph

Semantic and Product Networks

co-occurrence: the property of items being in the same place at the same time

here edges are implicit: you have to deduce, extract, calculate them from other data

We will discover the structure of a network: components, cores, coronas, communities, ...

two examples:

semantic networks

product networks

Semantic Networks

nodes are terms: words, word stems, word groups or concepts

links connect terms that:

i) are commonly used together ("complex" — "networks")

ii) describe the same property ("red" — "blue")

iii) are semantically comparable (synonyms: "program" — "app" hypernyms: "pet" — "cat" antonyms: "erase" — "restore")

Semantic Networks are used by knowledge specialists for semantic domain analysis

Example: Cultural Domain Case Study

notebook: 11_CulturalDomainAnalysis.ipynb

Product Networks

Retail Networks:

nodes are items purchased by individuals

links represent co-occurrence of items in customers in their "shopping baskets"

items are complements

weights : frequency of co-purchasing

Example: from products to projects Case Study

Notebook: 12_Products.ipynb