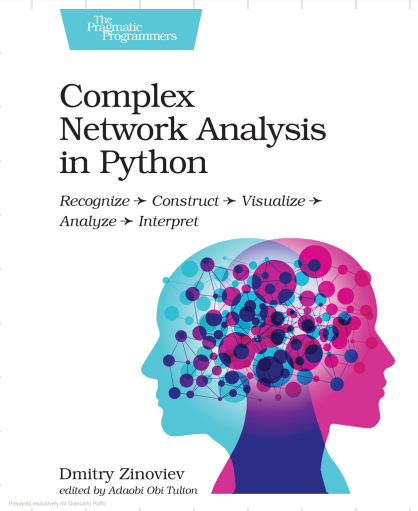
Lecture 15b

Complex Network Analysis

Similarity Networks

Today's topics

- Recap of similarity measures
- Analysing Bipartite Networks



Chapters:
14 - 16

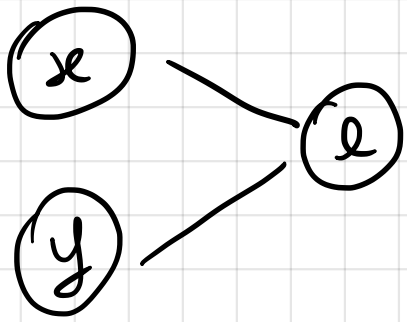
Similarity Based Networks

links are not always explicit

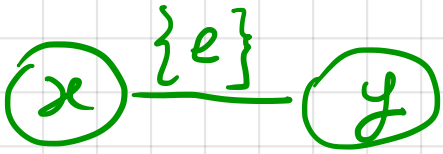
We have sometimes implicitly create links based on attributes

is i and j are similar along some sets of features

In bipartite networks:



both x and y "attended" the same event e

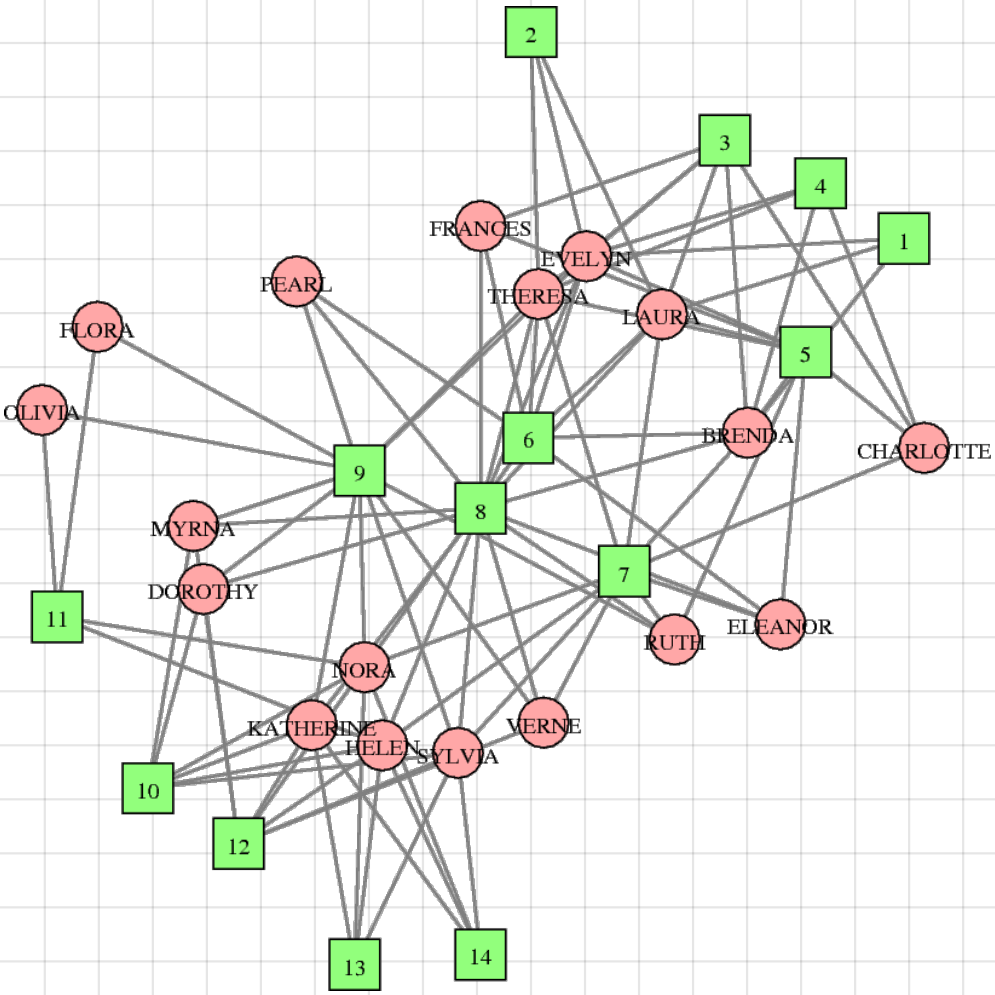


Exercise (homework)

1. Generate the famous "davis southern women" network

   (networkx function: davis_southern_women_graph())

2. "Analyse" both projections



Similarities

Sim. measures are numeric on the scale from -1 to 1 or 0 to 1

Quantify some qualitative measures to calculate similarities

You deal with objects in a multidimensional space: a sequence of (attribute, value) pairs

u, v : objects

$\underline{sim(u,v)}$ → the higher the value the more u and v are "aligned"

Hamming Distance

Used To compare strings

- "Karolin" and "Kathrin" is 3

the distance can be normalized by the length of the string (or vector)

the complementory measure is a similarity

$$h(\text{"Karolin"}, \text{"Kathrin"}) =$$

$$= \frac{4}{7} = 0.57$$

It works with binary or categorical attributes

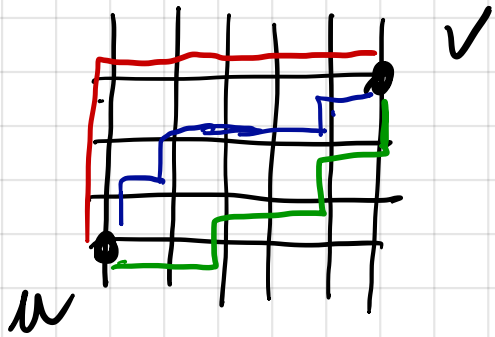
Manhattan Distance (or city block)

it is an extension of the Hamming Distance

it works with non-binary attributes

$$d(u,v) = \|u-v\|_1$$
$$= \sum_{i=1}^{n} |u_i - v_i|$$

"how many blocks away is a given intersection"

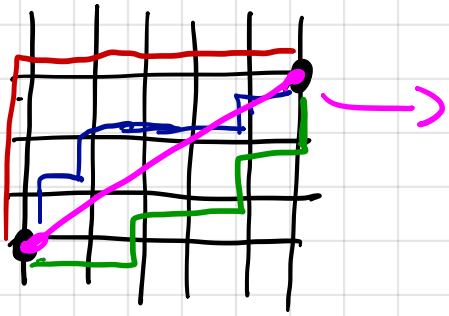


If you normalize and complement you transform the distance to a similarity measure again

Euclidean Distance

The "real" shortest path

$$d(u,v) = \sqrt{\sum_{i=1}^{n} (u_i - v_i)^2}$$



euclidean distance

rarely used

(Why?)

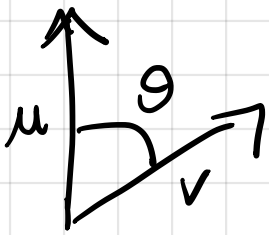
Cosine similerity

Sometimes it is useful to treat attributes as "directions"

It is a measure of angular distance

the cosine of the "angle" is the measure

$$-1 < \cos(\theta) < 1$$



remember that:

$$u \cdot v = \|u\| \cdot \|v\| \cos\theta$$

then

$$\text{sim}(u,v) = \cos\theta = \frac{u \cdot v}{\|u\| \cdot \|v\|}$$

$$= \frac{\sum_{i=1}^{n} u_i v_i}{\sqrt{\sum_{i=1}^{n} u_i^2}\sqrt{\sum_{i=1}^{n} v_i^2}}$$

Pearson Correlation

cosine sim is not invariant to shifts : it fails to detect small variations of attributes

It overstimates similarities!

Pearson is another angular measure ; it is a "correlation" that is not affected by shifts

It is like the cosine but attribute vectors subtract the mean

Usually you need to calculate the p-value together with pearson correlation.

p-value < 0.01

$$\rho(u,v) = \frac{\mathrm{cov}(u,v)}{\sigma_u \sigma_v}$$

covariance

standard deviations

or

$$\rho(u,v) = \frac{\sum_{i=1}^{n} (u_i - \bar{u}) \sum_{i=1}^{n} (v_i - \bar{v})}{\sqrt{\sum_{i=1}^{n} (u_i - \bar{u})^2} \sqrt{\sum_{i=1}^{n} (v_i - \bar{v})^2}}$$

---

notebook : "13_similarities"

Bipartite Networks

1. go back to "Nutrients" notebook

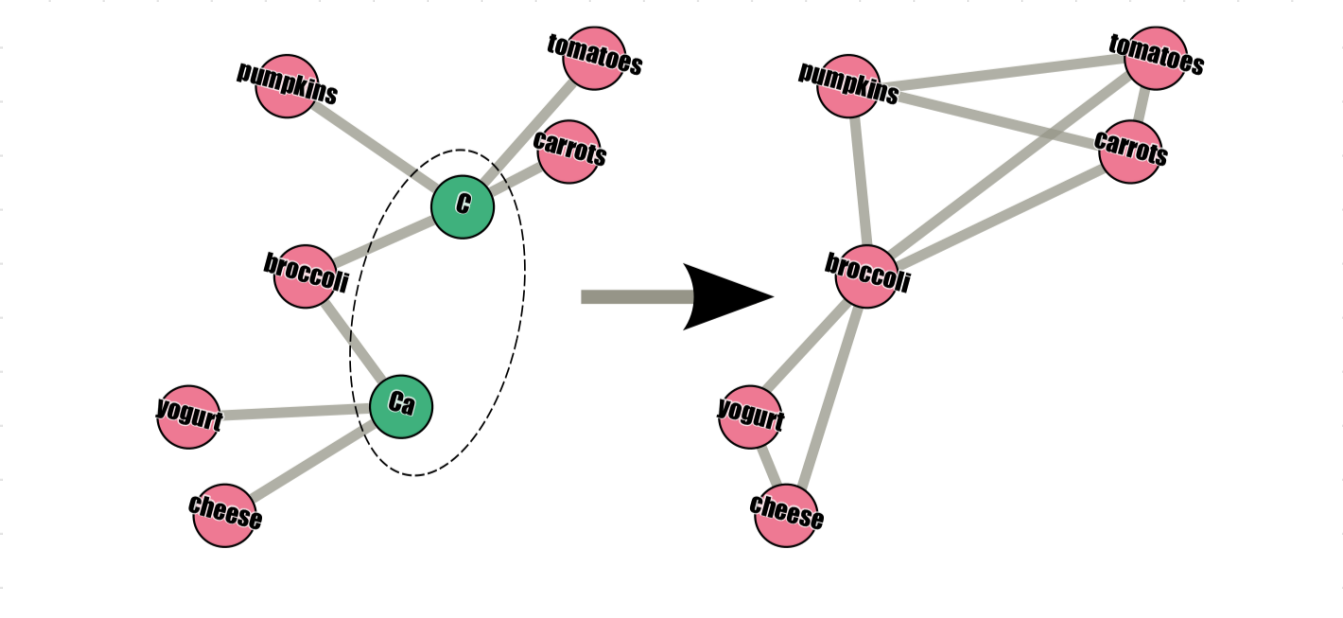
2. "pickle" the resulting graph

3. load the pickle file in a new notebook ("K1_bipartite")

4. "Analyze" its projections

5. Food similarity: it can improve link interpretation than a "pure" projection

Projection :



Hint : assign weights to the induced edges ...

It's your turn !