

Network Science

Similarity Networks

Lab 06

Lecture 16

Today's topics

- Recap of similarity measures
- Analysing Bipartite Networks

Similarity Based Networks

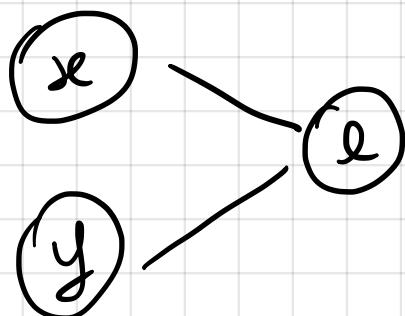
links are not always explicit

We have sometimes implicitly
create links based on
attributes

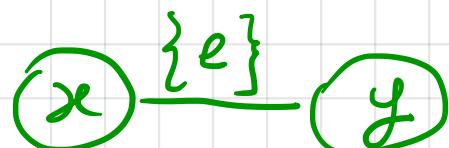


is *i* and *j*
are similar along
some sets of features

In bipartite networks :



both *x* and *y*
"attended" the same
event *e*

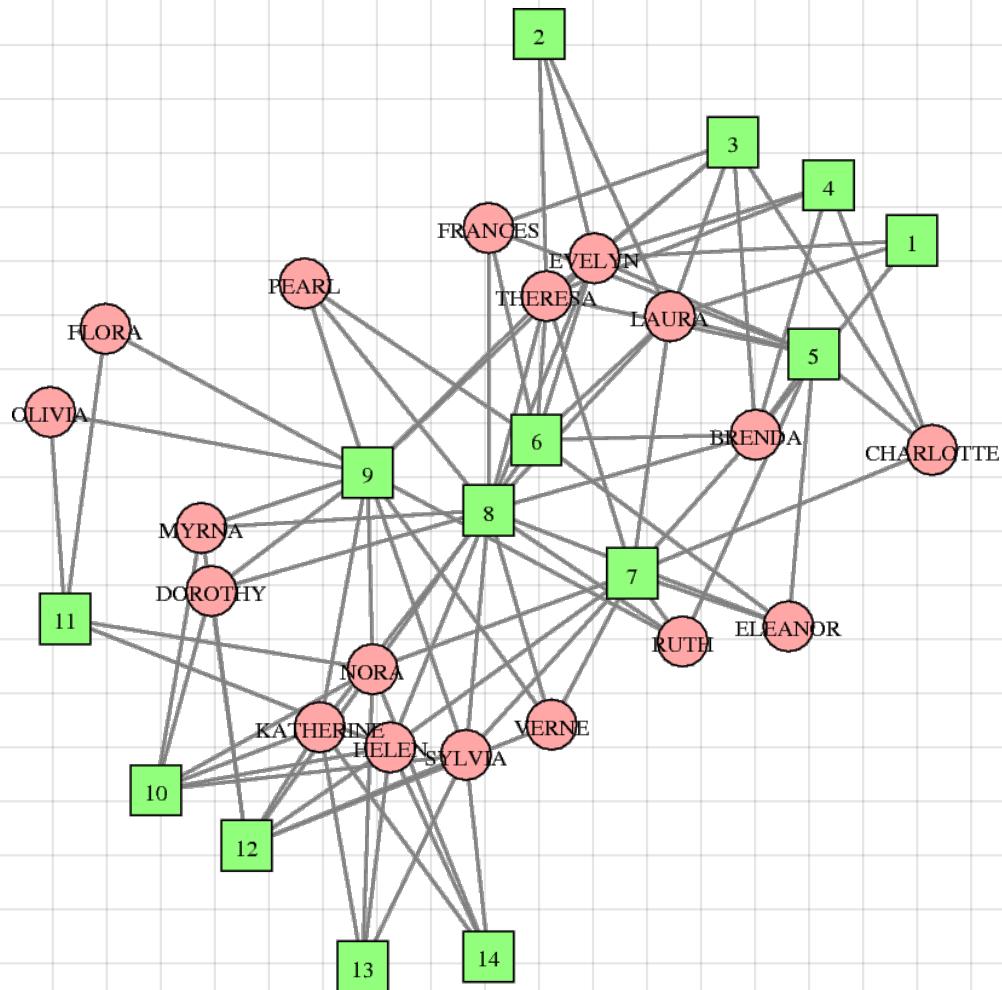


Exercise (homework)

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

$\text{sim}(u, v) \rightarrow$ the higher
the value
the more

u and v
are "aligned"

Hemming Distance

Used To compare strings

- "Karolin" and "K~~e~~thrin" is 3

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

the complementary measure is
a similarity

$$h("Karolin", "K\cancel{e}thrin") =$$

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

It works with binary or
categorical attributes

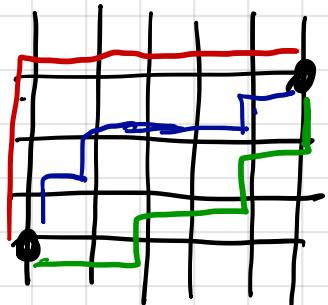
Manhattan Distance (or city block)

it is an extension of the Hamming Distance

it works with non-binary attributes

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

"how many blocks away is a given intersection"

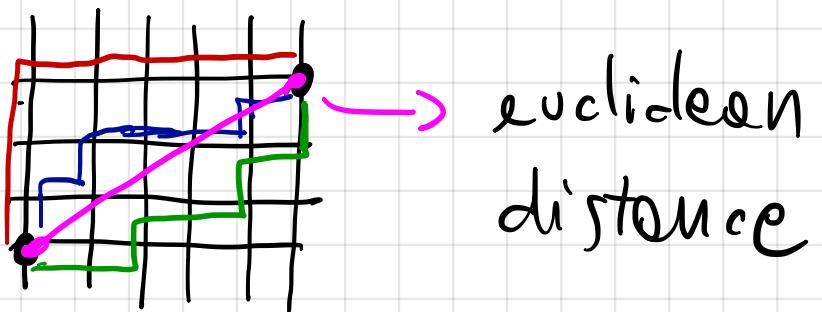


If you normalize and complement your transform the distance to a similarity measure again

Euclidean Distance

The "real" shortest path

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



rarely used
(Why?)

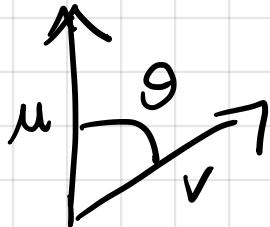
Cosine similarity

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

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

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

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

Pearson Correlation

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

It overestimates 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.

$$\varphi(\mu, v) = \frac{\text{cov}(\mu, v)}{\sigma_{\mu} \sigma_v}$$

covariance
standard deviations
or

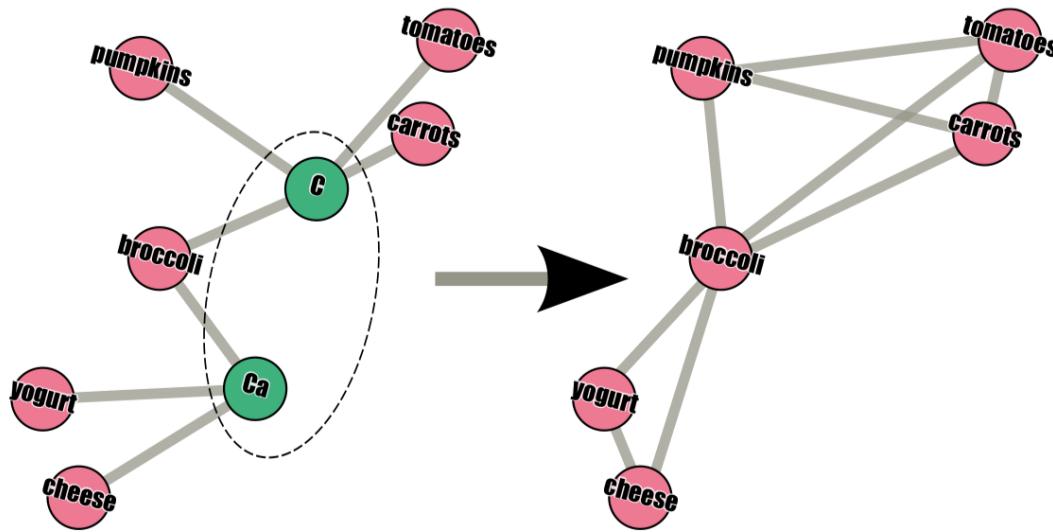
$$\rho(\mu, v) = \frac{\sum_{i=1}^n (\mu_i - \bar{\mu})(v_i - \bar{v}_i)}{\sqrt{\sum_{i=1}^n (\mu_i - \bar{\mu})^2} \sqrt{\sum_{i=1}^n (v_i - \bar{v}_i)^2}}$$

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 ("K_bipartite")
4. "Analyze" its projections
5. Food similarity = it can improve link interpretation then a "pure" projection

Projection :



Hint : assign weights to
the induced edges ...

It's your
turn !