# SC435
# Introduction to Complex Networks

## Ch. 6

6.1, 6.2, 6.3, 6.4, 6.4.1, 6.4.2, 6.6, 6.9

6.10, 6.10-1, 6.11, 6.11.1, 6.14.

## Problems

6.1, 6.2, 6.3, 6.4, 6.8

## Ch. 7

7.1, 7.2, 7.3, 7.4, 7.5, 7.6, 7.7

## Problems

7.1, 7.2, 7.3, 7.4

# Similarity

## Similarity in networks

- Nodes that are connected to each other in *(social)* networks tend to similar in their features.
  - friends may have similar features.
  - similar webpages may link to similar others
  - recommendation systems
  - circle of friends tell us about the person
- Assortativity: technical name of measuring similarity. We can make predictions about a person's qualities by inspecting their neighbors.
  - homophily: like attracts likes **or** due to social influence??
  - segregation and polarization of online communities on social media (Echo Chamber)
  - Degree assortativity: core-periphery structure.
- Disassortative: converse of assortative

## questions

- In what way can the vertices in a network be similar?
- How can we quantify similarity?

- Random Variable
- Probability Distribution
- $\langle x^n \rangle = \sum_x x^n \, p\{X = x\}$.
- Jointly distributed Random Variables.
  Covariance, Independence, Correlation

Mean
Variance
Skewness
Kurtosis

$$X_1 \ldots\ldots X_n$$

$$E[X] = \frac{1}{n} \sum_i X_i$$

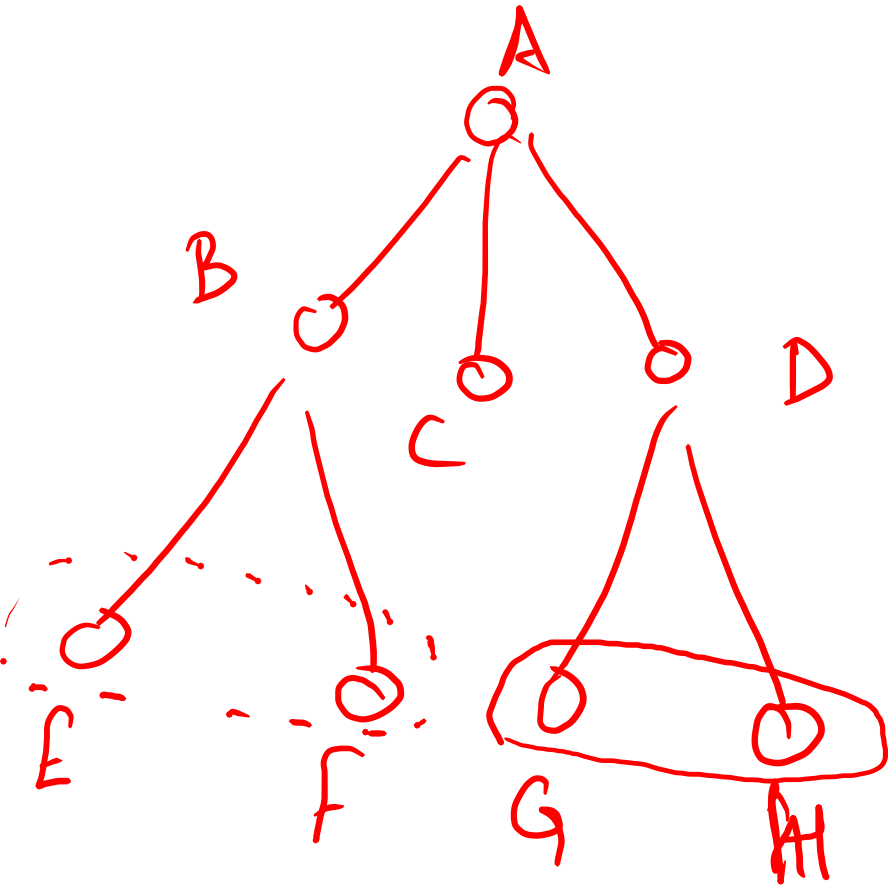
$$Var[X] = E\left[(X - E[X])^2\right].$$

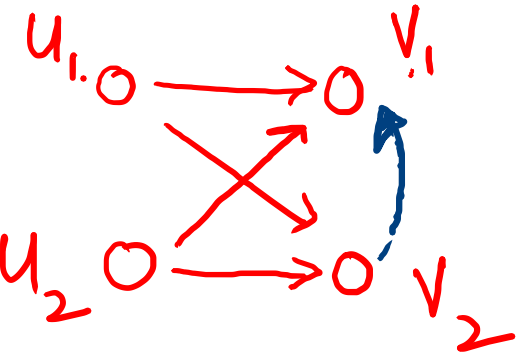
$$Cov(X, Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

# Similarity

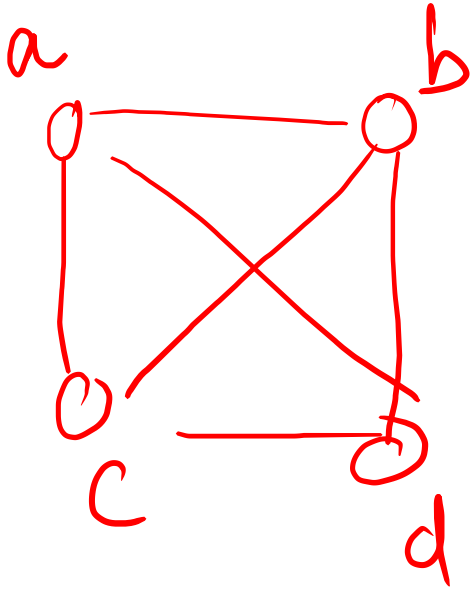
## Constructing measures of similarity

- Structural equivalence: Two vertices of a network are structurally equivalent if they share many of the same neighbors ⇒ in social networks two nodes are similar if they share many of the same neighbors
- Regular equivalence: Two vertices are regularly equivalent if they are equally related to equivalent others ⇒ People with similar roles have same local neighborhood.


A
B
C
D
E
F
G
H

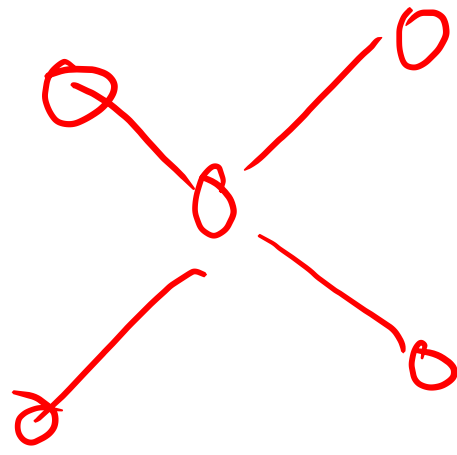



$$A = \begin{array}{c} \\ u_1 \\ u_2 \\ v_1 \\ v_2 \end{array} \begin{array}{c} \begin{array}{cccc} u_1 & u_2 & v_1 & v_2 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{pmatrix} \end{array} \leftarrow 1$$

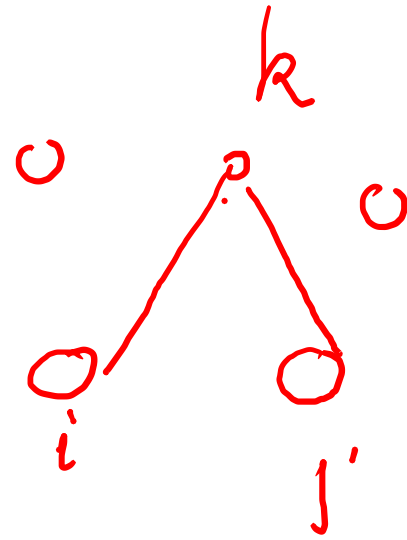



$\{c, d, b\}$

$\{a, c, d\}$.



$\sigma_{ij} = \Big[\quad\Big]$




$n_{ij}$ = number of common neighbors of vertex $i$ & $j$

$= \sum_k A_{ik} A_{kj}$

## Jaccard Similarity.

$$\frac{|N(v_i) \cap N(v_j)|}{|N(v_i) \cup N(v_j)|}$$

$N$ : neighbors.

$$\frac{n_{ij}}{\sum_k A_{ik} + \sum_k A_{jk} - n_{ij}} = \frac{n_{ij}}{k_i + k_j - n_{ij}} = \sigma_{ij}$$

# Cosine Similarity

$$\bar{x} \cdot \bar{y} = |\bar{x}||\bar{y}| \cos\theta.$$

$A_{ik}$ [ ]

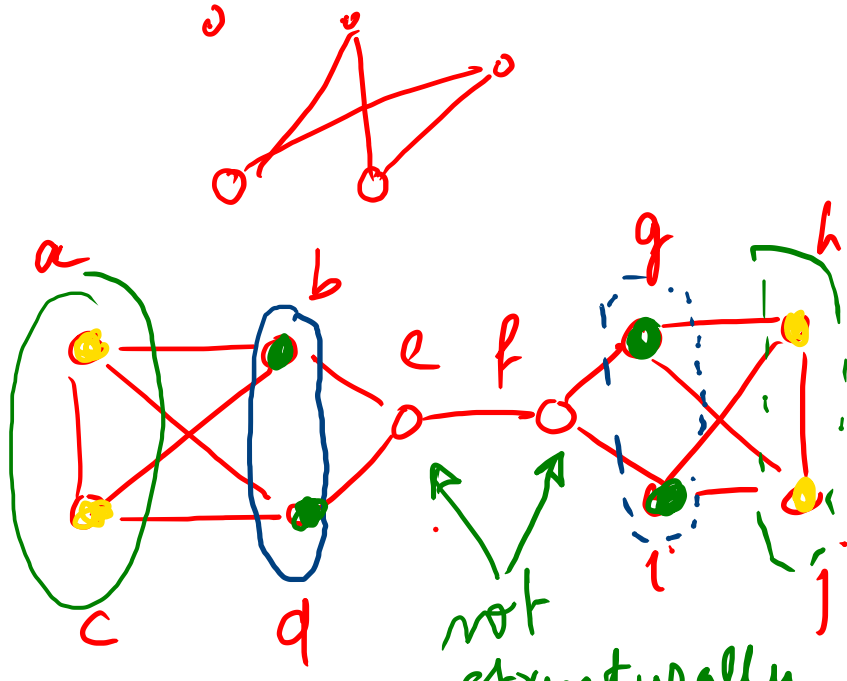
$A_{jk}$ [ ]

$$\cos\theta = \frac{n_{ij}}{\sqrt{\sum_k A_{ik}^2}\sqrt{\sum_k A_{jk}^2}} = \frac{n_{ij}}{\sqrt{k_i k_j}}.$$

Expected overlap if the connection is purely at random

# Similarity

- Structural Equivalence
  - overlap amongst neighbors
    - partial overlap
      - Jaccard Similarity
      - Cosine
      - Pearson
    - Hamming dist.
- Regular Equivalence
  - Social roles

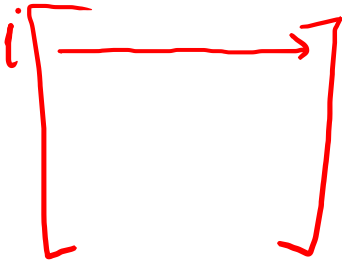


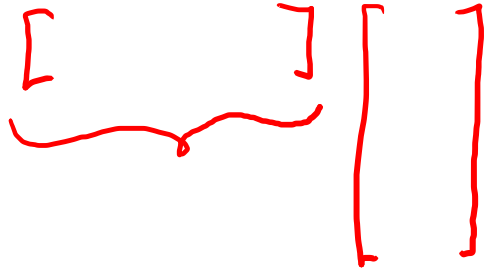

$f \rightarrow \{e, g, i\}$

$e \rightarrow \{b, d, f\}$.

Jaccard $$n_{ij} = \frac{\sum_k A_{ik} A_{kj}}{k_i + k_j - n_{ij}}$$

Cosine $$= \frac{n_{ij}}{\sqrt{\sum_k A_{ik}^2 \sum_k A_{jk}^2}}$$

$A =$ 

# Statistical methods (comparison)

## Covariance

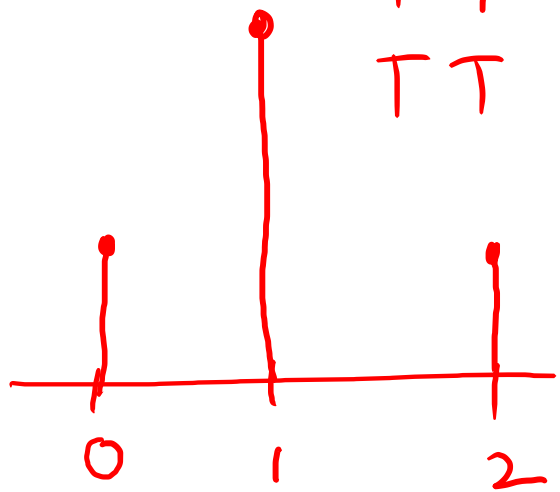
$X$ : number of heads.

$P\{X=0\} = \frac{1}{4}$

$P\{X=1\} = 2/4$

$P\{X=2\} = 1/4$

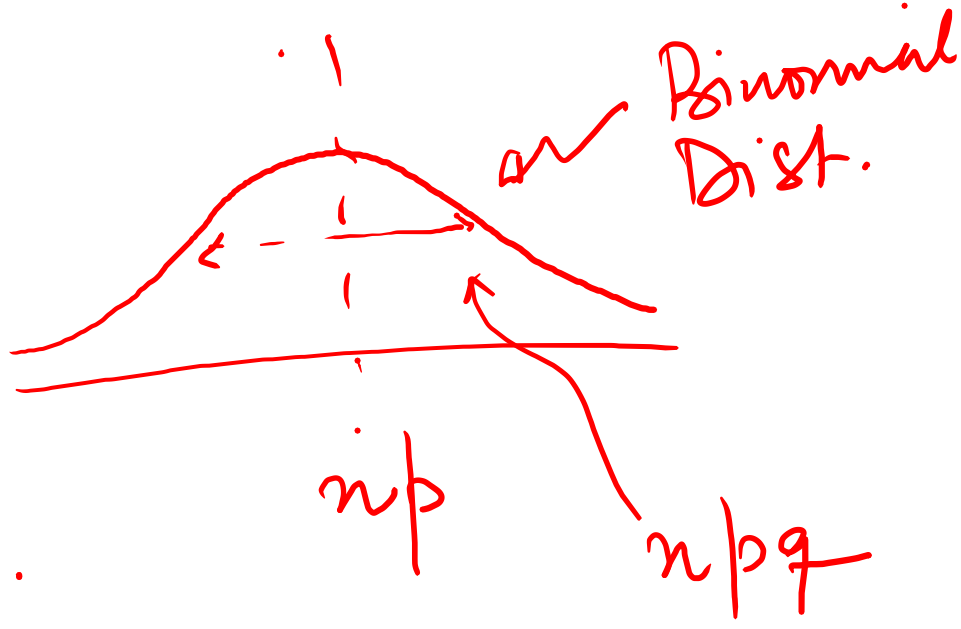
HH
HT
TH
TT




distribution

$n$ flips  $x$ heads

$$\binom{n}{x} p^x (1-p)^{n-x}$$




$$E[X] = \sum_x x \, p\{X = x\}.$$

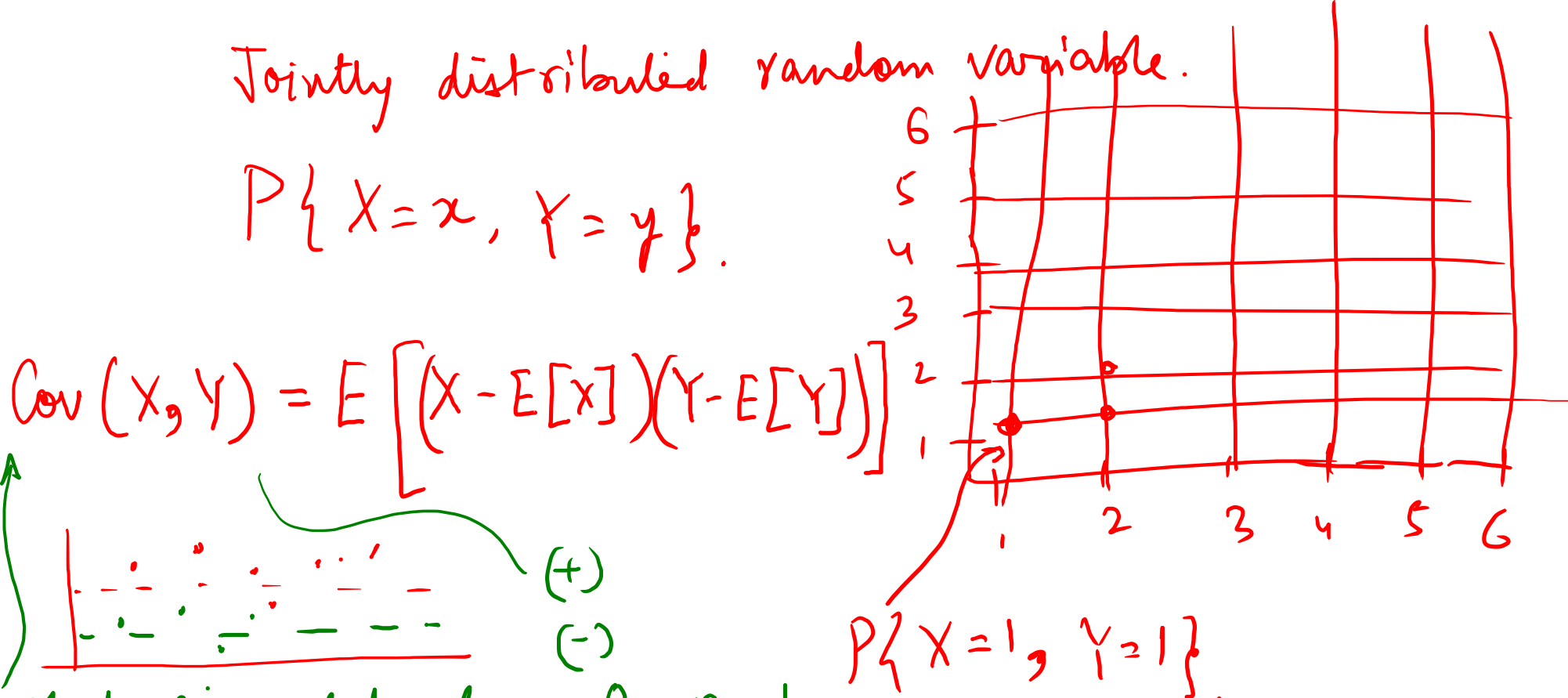
$$E[X^2] = \sum_x x^2 \, p\{X = x\}$$

$$\mathrm{Var}(X) = E\left[(X - E[X])^2\right] = E[X^2] - E[X]^2$$

# Jointly distributed random variable.

$P\{X=x, Y=y\}$.

$$Cov(X,Y) = E\left[(X-E[X])(Y-E[Y])\right]$$



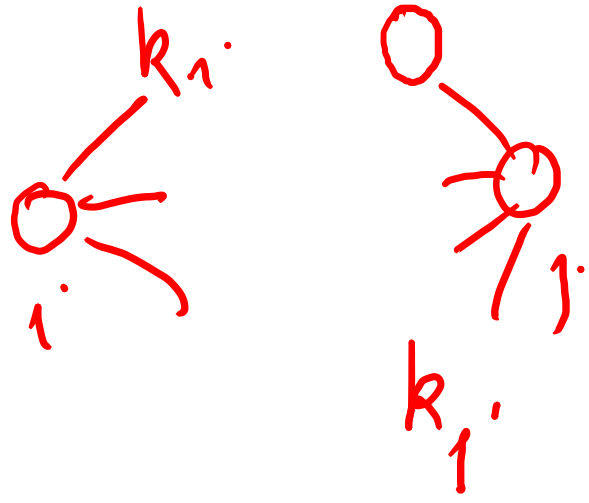

Statistical dependence 0 - random.
between X & Y.

$$\rho = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{Var(X)\ Var(Y)}}$$

Pearson Correlation coefficient

Expected number of common neighbors if the connections are at random

$$\frac{k_j}{n-1} \simeq \frac{k_j}{n}$$




$$\frac{k_i k_j}{n}$$

$$n_{ij} - \frac{k_i k_j}{n}$$

$$= \sum_k A_{ik} A_{kj} - \frac{k_i k_j}{n} \leftarrow \sum_k \left(A_{ik} - \langle A_i \rangle\right)\left(A_{jk} - \langle A_j \rangle\right)$$

$A_{ik} \rightarrow [\ 0 \quad ] \ldots \langle A_i \rangle$

$A_{jk} \rightarrow [\quad\quad ] \ \langle A_j \rangle$

$$\langle A_i \rangle = \frac{1}{n} \sum_k A_{ik}$$

# Similarity

## Measures of Structural equivalence (extent)

- **Count of the number of common neighbors**

$$n_{ij} = \sum_k A_{ik} A_{kj} = [A^2]_{ij}$$

- **Cosine Similarity:**

$$\sigma_{ij} = \frac{\sum_k A_{ik} A_{kj}}{\sqrt{\sum_k A_{ik}^2} \sqrt{\sum_k A_{jk}^2}}$$

$$\sigma_{ij} = \frac{n_{ij}}{\sqrt{k_i k_j}} \qquad \textbf{(simple, unweighted)}$$

- **Jaccard Similarity:**

$$J_{ij} = \frac{n_{ij}}{k_i + k_j - n_{ij}}$$

# Pearson similarity

$$\sigma_{ij} = n_{ij} - \frac{k_i k_j}{n}$$

Expected common neighbors if the connections are drawn at random.

$$= \sum_k A_{ik} A_{kj} - \frac{k_i k_j}{n}$$

$$= \sum_k A_{ik} A_{kj} - n\langle A_i \rangle \langle A_j \rangle$$

$$= \sum_k \left[ A_{ik} A_{kj} - \langle A_i \rangle \langle A_j \rangle \right]$$

$$A_{ik} = A_{ki}$$

$$A^2_{ik} = A_{ik}$$

$$k_i = \sum_k A_{ik}$$

$$\langle A_i \rangle = \frac{1}{n} \sum_k A_{ik}$$

$$= \sum_k \left[ A_{ik} A_{kj} - \langle A_i \rangle \langle A_j \rangle + \langle A_i \rangle \langle A_j \rangle - \langle A_i \rangle \langle A_j \rangle \right]$$

$$\uparrow \; n \langle A_i \rangle \langle A_j \rangle = n \cdot \frac{1}{n} \sum_k A_{ik} \langle A_j \rangle$$

$$\sum_k A_{ik} \langle A_j \rangle$$

$$\sum_k A_{jk} \langle A_i \rangle$$

$$= \sum_k \left[ A_{ik} A_{kj} - A_{ik} \langle A_j \rangle + \langle A_i \rangle \langle A_j \rangle - A_{jk} \langle A_i \rangle \right]$$

$$= \sum_k \left( A_{ik} - \langle A_i \rangle \right) \left( A_{jk} - \langle A_j \rangle \right)$$ ⇝ Covariance

$A = [\quad \cdots \quad]$ ← $A_{1k}$, ← $A_{2k}$

$\sigma_{12}$

$\sigma = [\quad]$ heat maps.

$$\frac{n_{ij}}{\frac{k_i k_j}{n}}$$

Hamming distance

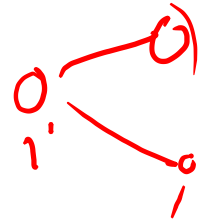
$$d_{ij} = \sum_k \left(A_{ik} - A_{jk}\right)^2$$

$\delta_{ij}$ — Kronecker delta fct.

$\delta_{ij} = 1$ if $i = j$

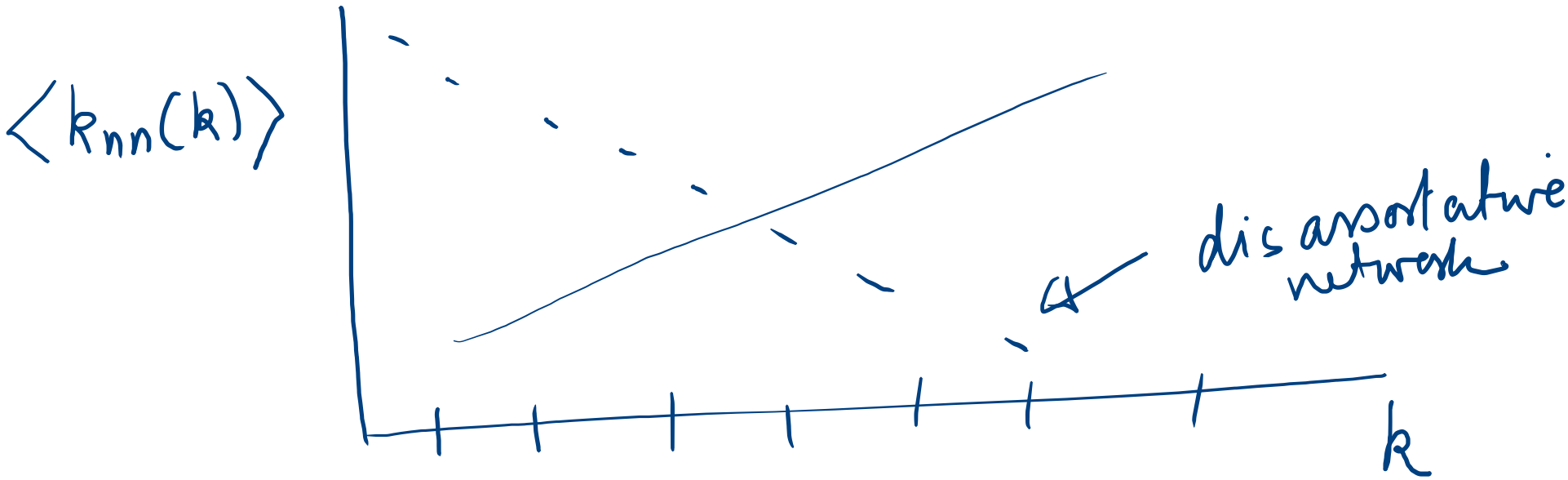
$0$ otherwise

$$k_{nn}(i) = \frac{1}{k_i} \sum_j A_{ij} k_j$$

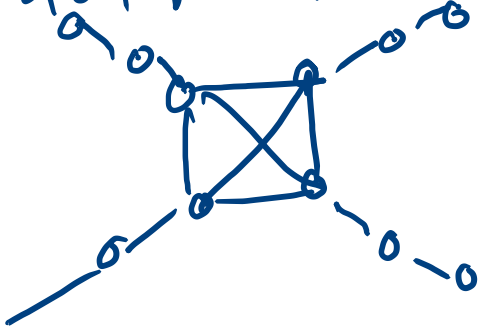



$k_{nn}$ ... nearest neighbor

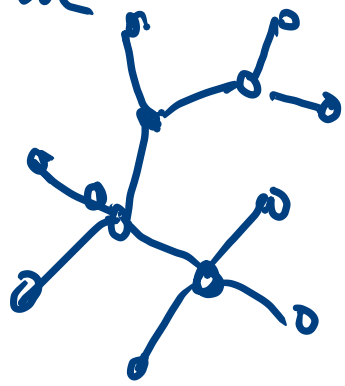
$$\langle k_{nn}(k) \rangle = \frac{1}{N_k} \sum_{i=1}^{n} k_{nn}(i)\, \delta_{k_i,k}$$



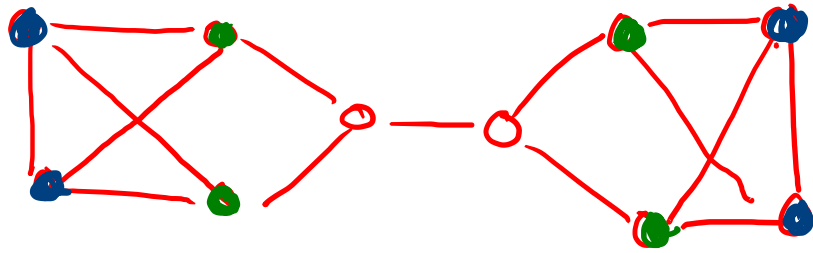

Assortative network — Core - periphery structure



Disassortative network

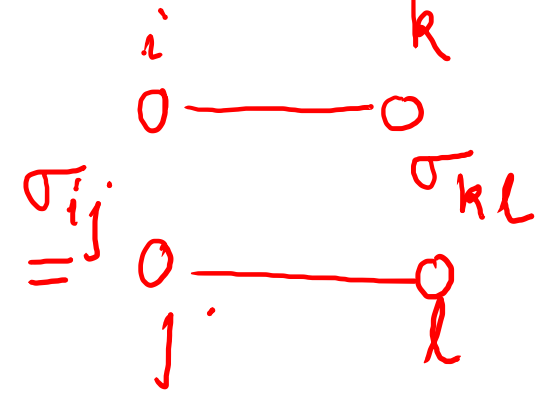


# Regular equivalence



Two vertices are regularly equivalent if they are connected to vertices that are regularly equivalent.

$$\sigma_{ij} = \alpha \sum_{k,l} A_{ik} A_{jl} \sigma_{kl} + \delta_{ij}$$




$$\boxed{\sigma = \alpha A \sigma A}$$

$$\sigma = \alpha A \sigma A + I$$

$$\sigma(t=0) = \vec{0}$$

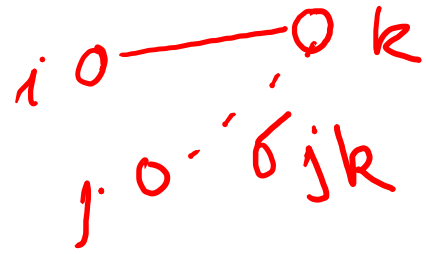
$$\sigma(t=1) = I$$

$$\sigma(t=2) = \alpha A^2 + I$$

$$\sigma(t=3) = \alpha^2 A^4 + \alpha A^2 + I$$

$$\vdots$$

$$\sigma = (I - \alpha A)^{-1}$$




$$\sigma_{ij} = \alpha \sum_{ik} A_{ik} \sigma_{kj} + \delta_{ij}$$

$$\boxed{\sigma = \alpha A \sigma + I}$$

$$\sigma(t=0) = \vec{0}$$

$$\sigma(t=1) = I$$

$$\sigma(t=2) = \alpha A + I$$

$$\sigma(t=3) = \alpha^2 A^2 + \alpha A + I$$

→ In real networks are the connections based on similarity in terms of attributes?
— Homophily / assortativity.

→ How do we quantify it? ←

→ Which ones in which groups?

Graph Partitioning →

Community Detection

$$Q = \frac{m_c - \langle m_c \rangle}{m}$$

c : class cluster group attribute

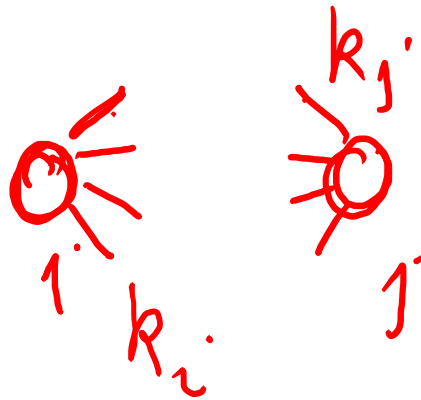
Modularity

Let $c_i$ represent class $i = 1, \ldots n$

$m_c$ = number of links between nodes with the same attribute $\frac{1}{2} \sum_{ij} A_{ij} \delta_{c_i, c_j}$

$\langle m_c \rangle$ : Expected number of links between nodes of the same type if connections are drawn at random.

$$\sum_{ij} \frac{k_i k_j}{2m} \delta_{c_i, c_j}$$




$$Q = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta_{c_i, c_j}$$

# Similarity

## Measures of Structural equivalence (extent)

- **Pearson correlation coefficient:**

$$r_{ij} = \frac{\sum_k (A_{ik} - \langle A_i \rangle) \sum_k (A_{jk} - \langle A_j \rangle)}{\sqrt{\sum_k (A_{ik} - \langle A_i \rangle)^2} \sqrt{\sum_k (A_{jk} - \langle A_j \rangle)^2}} = \frac{\mathrm{Cov}(\sigma_\mathrm{i}, \sigma_\mathrm{j})}{\sigma_i \sigma_j}$$

- **For unweighted, undirected graph**

$$r_{ij} = \frac{n_{ij} - \frac{k_i k_j}{n}}{\sqrt{k_i - \frac{k_i^2}{n}} \sqrt{k_j - \frac{k_j^2}{n}}}$$

$$-1 \leq r_{ij} \leq 1$$

# Similarity

## Measures of Structural equivalence (extent)

- **normalize common neighbors by the expected number of common neighbors when they are picked at random**

$$g_{ij} = \frac{n_{ij}}{\frac{k_i k_j}{n}} = \frac{n \sum_k A_{ik} A_{jk}}{\sum_k A_{ik} \sum_k A_{jk}}$$

- **Euclidean (Hamming) distance:**

$$d_{ij} = 1 - \frac{2n_{ij}}{k_i + k_j}$$

# Similarity

## Measures of Regular equivalence (extent)

- Two nodes $i$ and $j$ have high similarity score if they have neighbors $k$ and $l$ that themselves have high similarity. For an undirected network

$$\sigma_{ij} = \alpha \sum_{k,l} A_{ik} A_{jl} \sigma_{kl}$$

- Should have high self-similarity ($\sigma_{ii}$)

$$\sigma_{ij} = \alpha \sum_{k,l} A_{ik} A_{jl} \sigma_{kl} + \delta_{ij}$$

- problem: assuming a null matrix as the initial condition after $k$ iterations we have

$$\sigma^{(k)} = \sum_{r=0}^{k-1} \alpha^r A^{2r}$$

This measure of similarity is a weighted sum over the number of paths of even length between pairs of vertices.

# Similarity

## Measures of Regular equivalence (extent)

- **Modified definition of regular equivalence:** Nodes $i$ and $j$ are similar if $i$ has a neighbor of $k$ that is itself similar to $j$.

$$\sigma_{ij} = \alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}$$

- Iterating again starting with $\sigma = 0$, we get:

$$\sigma = \sum_m (\alpha A)^m = (1 - \alpha A)^{-1}$$

  Similarity of two nodes is the weighted sum of the number of paths of different length that connect them.

- Bias in favor of high degree node can be removed by dividing with node degree

$$\sigma_{ij} = \frac{1}{k_i}\left[\alpha \sum_k A_{ik}\sigma_{kj} + \delta_{ij}\right]$$

# Similarity

## Homophily and Assortative Mixing

- People have a strong tendency to associate with others whom they perceive as being similar to themselves. This property is called *homophily* or *assortative* mixing.
- Disassortative mixing: tendency to people to associate with others who are unlike them.
- Political polarization, mixing on the basis of race, obesity etc. (Assortative)
- Dating networks, food web (predator-prey), economic networks (producers/consumers) (Disassortative)

# Similarity

## Mixing by Categorical attributes

- Characterize by some numbers the value of the mixing in the network. One of the ways to do this is called (modularity)
- Every vertex has a label ($c_i$)
- How much more often do attributes match across edges than what is expected at random? (Pearson correlation coeff.)
- Modularity

$$Q = \frac{m_c - \langle m_c \rangle}{m} = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

- The total number of edges that run between nodes of the same class

$$\sum_{ij} A_{ij} \delta(c_i, c_j)$$

- expected number of edges between all pairs of vertices of the same type if edges are placed at random

$$\frac{k_i k_j}{2m}$$

# Similarity

## Mixing by Categorical attributes

- Modularity

$$Q = \frac{m_c - \langle m_c \rangle}{m} = \frac{1}{2m} \sum_{ij} \left( A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

$$\begin{aligned} Q &= 0 \quad \textbf{(single class or completely random)} \\ Q &> 0 \quad \textbf{(Assortative mixing)} \\ Q &< 0 \quad \textbf{(Disassortative)} \end{aligned}$$

# Similarity

## Mixing by Categorical attributes

- Modularity matrix $B$

$$[B]_{ij} = A_{ij} - \frac{k_i k_j}{2m}$$

- $Q$ is never equal to 1

$$Q_{max} = \frac{1}{2m}\left(2m - \sum_{ij} \frac{k_i k_j}{2m}\right)$$

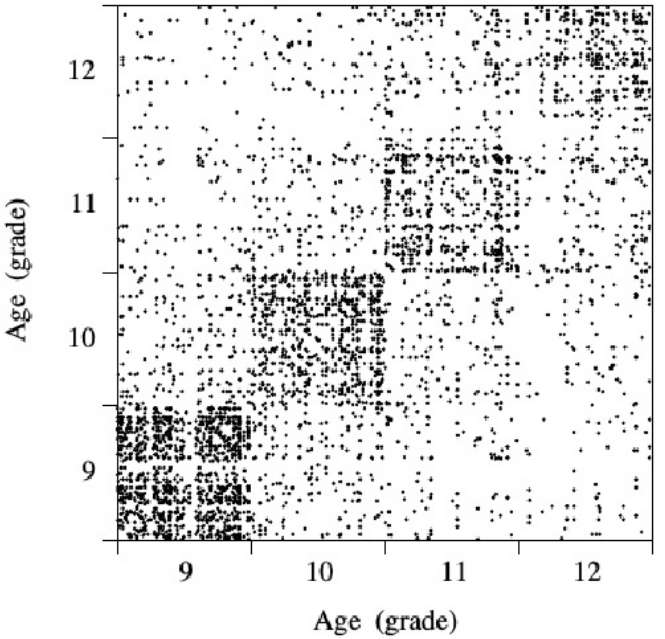
- Value of $Q$ can be significantly smaller than 1 even for perfectly mixed networks. So we normalize it by the maximum value (Assortativity coefficient $Q/Q_{max}$)

# Similarity

## Mixing by ordered characteristics

- We can also have mixing if characteristics are approximately the same.

$$r = \frac{\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) x_i x_j}{\sum_{ij}\left(k_i \delta_{ij} - \frac{k_i k_j}{2m}\right) x_i x_j}$$



# Similarity

## Assortative mixing by degree

- Assortative: core-periphery
- Disassortative: star-like structure

$$r = \frac{\sum_{ij}\left(A_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}{\sum_{ij}\left(k_i \delta_{ij} - \frac{k_i k_j}{2m}\right) k_i k_j}$$