Measuring Assortativity for Ordinal Attributes

- A common measure for analyzing the relationship between *ordinal* values is *covariance*

- It describes how two variables change together

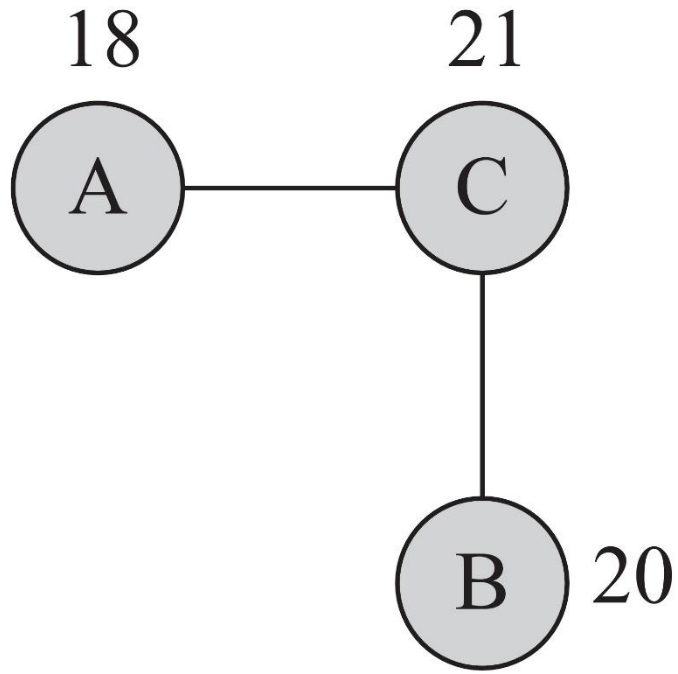
- In our case, we have a network
  - We are interested in how values assigned to nodes that are connected (via edges) are correlated

Covariance Variables

- The value assigned to node v_i is x_i
- We construct two variables X_L and X_R
- For any edge (v_i, v_j), we **assume** that x_i is observed from variable X_L and x_j is observed from variable X_R
- X_L represents the ordinal values associated with the left-node (the first node) of the edges
- X_R represents the values associated with the right-node (the second node) of the edges
- We need to compute the covariance between variables X_L and X_R

Covariance Variables: Example

List of edges:
(A, C)
(C, A)
(C, B)
(B, C)



- X_L : (18, 21, 21, 20)
  X_R : (21, 18, 20, 21)

$$\mathbf{E}(X_L) = \mathbf{E}(X_R)$$

$$\sigma(X_L) = \sigma(X_R)$$

Covariance

For two given column variables X_L and X_R, the covariance is

$$\begin{aligned}\sigma(X_L, X_R) &= \mathbf{E}[(X_L - \mathbf{E}[X_L])(X_R - \mathbf{E}[X_R])] \\ &= \mathbf{E}[X_L X_R - X_L \mathbf{E}[X_R] - \mathbf{E}[X_L] X_R + \mathbf{E}[X_L]\mathbf{E}[X_R]] \\ &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R] + \mathbf{E}[X_L]\mathbf{E}[X_R] \\ &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R]\end{aligned}$$

$E(X_L)$ is the mean of the variable and $E(X_L X_R)$ is the mean of the multiplication X_L and X_R

$$E(X_L) = E(X_R) = \frac{\sum_i (X_L)_i}{2m} = \frac{\sum_i d_i x_i}{2m}$$

$$E(X_L X_R) = \frac{1}{2m} \sum_i (X_L)_i (X_R)_i = \frac{\sum_{ij} A_{ij} x_i x_j}{2m}$$

Covariance

$$
\begin{aligned}
\sigma(X_L, X_R) &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R] \\
&= \frac{\sum_{ij} A_{ij} x_i x_j}{2m} - \frac{\sum_{ij} d_i d_j x_i x_j}{(2m)^2} \\
&= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2m} \right) x_i x_j
\end{aligned}
$$

Normalizing Covariance

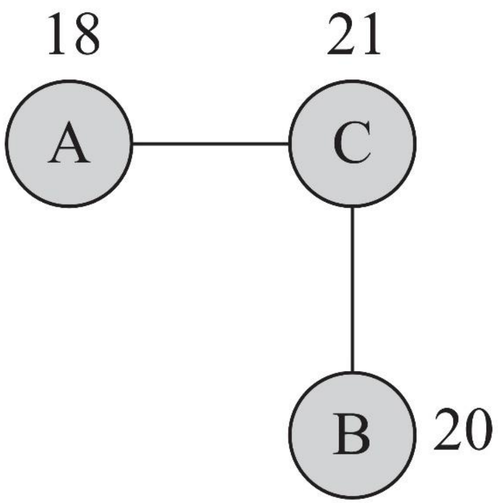
Pearson correlation $\rho(X, Y)$ is the normalized version of covariance

$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)\sigma(X_R)}.$$

In our case: $\sigma(X_L) = \sigma(X_R)$

$$\begin{aligned}
\rho(X_L, X_R) &= \frac{\sigma(X_L, X_R)}{\sigma(X_L)^2}, \\
&= \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m} \right) x_i x_j}{\mathbf{E}[(X_L)^2] - (\mathbf{E}[X_L])^2} \\
&= \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m} \right) x_i x_j}{\frac{1}{2m}\sum_{ij} A_{ij} x_i^2 - \frac{1}{2m}\sum_{ij} \frac{d_i d_j}{2m} x_i x_j}
\end{aligned}$$

Correlation Example



$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix} \qquad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}$$

$$\rho(X_L, X_R) = -0.67$$

Influence

- **Measuring Influence**
- **Modeling Influence**

Influence: Definition

Influence

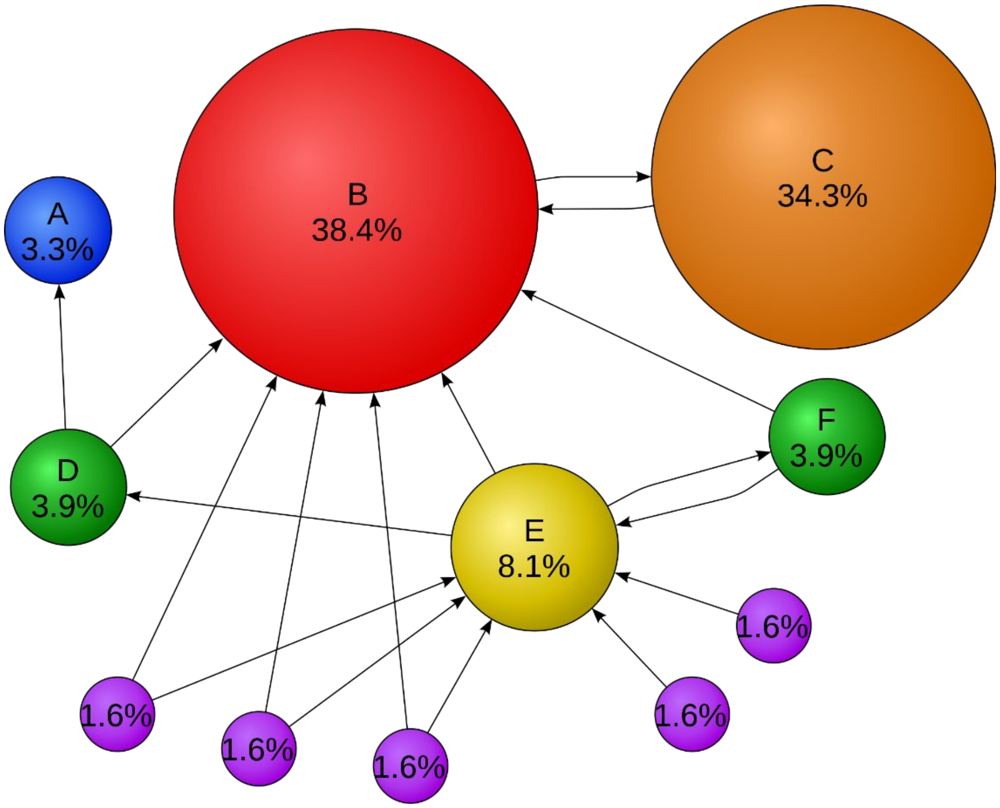
The act or power of producing an effect without apparent exertion of force or direct exercise of command

Measuring Influence

Measuring Influence

- Measuring influence
  - Assigning a **number** (or a set of numbers) to each node that represents the influential power of that node

- The influence can be measured based on
  1. Prediction or
  2. Observation

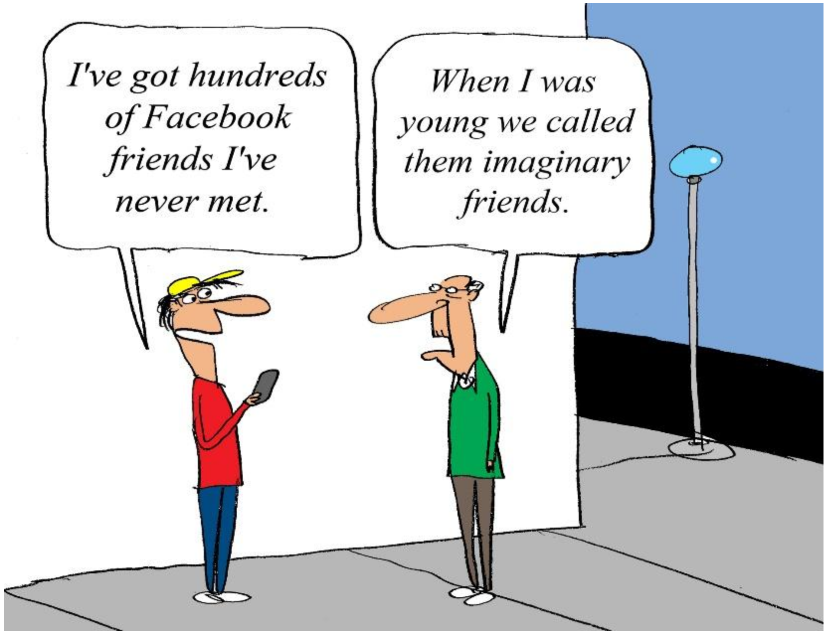


Prediction-based Measurement

We assume that
- an individual's attribute, or
- the way the user is situated in the network

can **predict** how influential the user **will** be

- Example 1
  - The number of *friends* of an individual is correlated with how influential she is
    - It is natural to use any of the ***centrality measures*** discussed (Chapter 3) for prediction-based influence measurements
    - How strong are these friendships?
- Example 2
  - On Twitter, in-degree (number of ***followers***) is a benchmark for measuring influence



Observation-based Measurement

We quantify influence of an individual by measuring the amount of influence ***attributed*** to the individual

I. When an individual is the role model

- Influence measure: size of the audience that has been influenced

II. When an individual spreads information

- Influence measure: the size of the cascade, the population affected, the rate at which the population gets influenced

III. When an individual increases values

- Influence measure: the increase (or rate of increase) in the value of an item or action
  - The second person who bought the fax machine increased its value dramatically

Case Studies for Measuring Influence in Social Media

- **Measuring Influence on Twitter**

Measuring Social Influence on Twitter

- In **Twitter**, users have an option of following individuals, which allows users to receive tweets from the person being followed

- Intuitively, one can think of the number of followers as a measure of **influence** (in-degree centrality)

Measuring Social Influence on Twitter: Measures

- **In-degree**
  - The number of users following a person on **Twitter**
  - Indegree denotes the “audience size” of an individual.
- **Number of Mentions**
  - The number of times an individual is mentioned in a tweet, by including @username in a tweet.
  - The number of mentions suggests the “ability in engaging others in conversation”
- **Number of Retweets**
  - **Twitter** users have the opportunity to forward tweets to a broader audience via the retweet capability.
  - The number of retweets indicates individual’s ability in generating content that is worth being passed on.

Measuring Social Influence on Twitter: Measures

- Each one of these measures by itself can be used to identify influential users in Twitter.
  - We utilizing the measure for each individual and then rank users based on their measured influence value.
- **Observation:** contrary to public belief, number of followers is considered an *inaccurate* measure compared to the other two.
- We can rank individuals on Twitter independently based on these three measures.
- To see if they are correlated or redundant, we can compare ranks of individuals across three measures using **rank correlation** measures.

Comparing Ranks across Three Measures

To compare ranks across more than one measure (say, in-degree and mentions), we can use **Spearman's Rank Correlation** Coefficient

$$\rho = 1 - \frac{6\sum(m_1^i - m_2^i)^2}{n^3 - n}$$

m_1^i and m_2^i are ranks of individual i based on measures m_1 and m_2, and n is the total number of usernames.

In-degrees do not carry much information

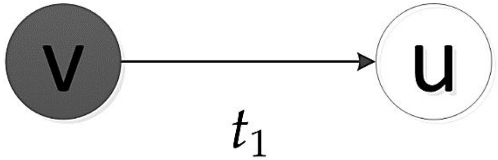
- **Spearman's rank correlation** is the **Pearson correlation coefficient** for **ordinal variables that represent ranks**
  - i.e., input range [1. . . n]
  - Output value is in range [-1,1]
- Popular users (users with high in-degree) do not necessarily have high ranks in terms of number of retweets or mentions.

Measures	Correlation Value
In-degree vs. retweets	0.122
In-degree vs. mentions	0.286
Retweets vs. mentions	0.638

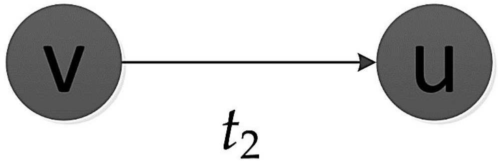
Influence Modeling

Influence Modeling

- At time t_1, node v is activated and node u is not



- Node u becomes activated at time t_2 due to influence



- Each node is started as active or inactive
- A node, once activated, will activate its neighbors
- An activated node cannot be deactivated

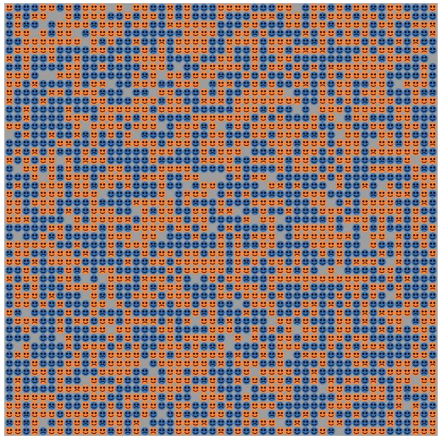
Influence Modeling: Assumptions

- The influence process takes place in a network
- Sometimes this network is observable (an explicit network) and sometimes not (an implicit network).
- **Observable network:** we can use threshold models, e.g., linear threshold model
- **Implicit Network:** we can use methods that take the number of individuals who get influenced at different times as input, e.g., the number of buyers per week
  - *Linear Influence Model (**LIM**)*

Threshold Models

- Simple, yet effective methods for modeling influence in **explicit** networks
- Nodes make decision based on the influence coming from of their **already activated neighborhood**

- Using a threshold model, *Schelling* demonstrated that minor preferences in having neighbors of the same color leads to complete racial segregation



http://www.youtube.com/watch?v=dnffIS2EJ30

Linear Threshold Model (LTM)

A node i would become active if incoming influence ($w_{j,i}$) from friends exceeds a certain threshold

$$\sum_{v_j \in N_{\text{in}}(v_i)} w_{j,i} \leq 1$$

- Each node i chooses a threshold θ_i randomly from a uniform distribution in an interval between 0 and 1

- At time t, all nodes that were active in the previous steps $[0 \ldots t-1]$ remain active, but only nodes activated at time $t-1$ get the chance to activate

- Nodes satisfying the following condition will be activated

$$\sum_{v_j \in N_{\text{in}}(v_i), v_j \in A_{t-1}} w_{j,i} \geq \theta_i$$

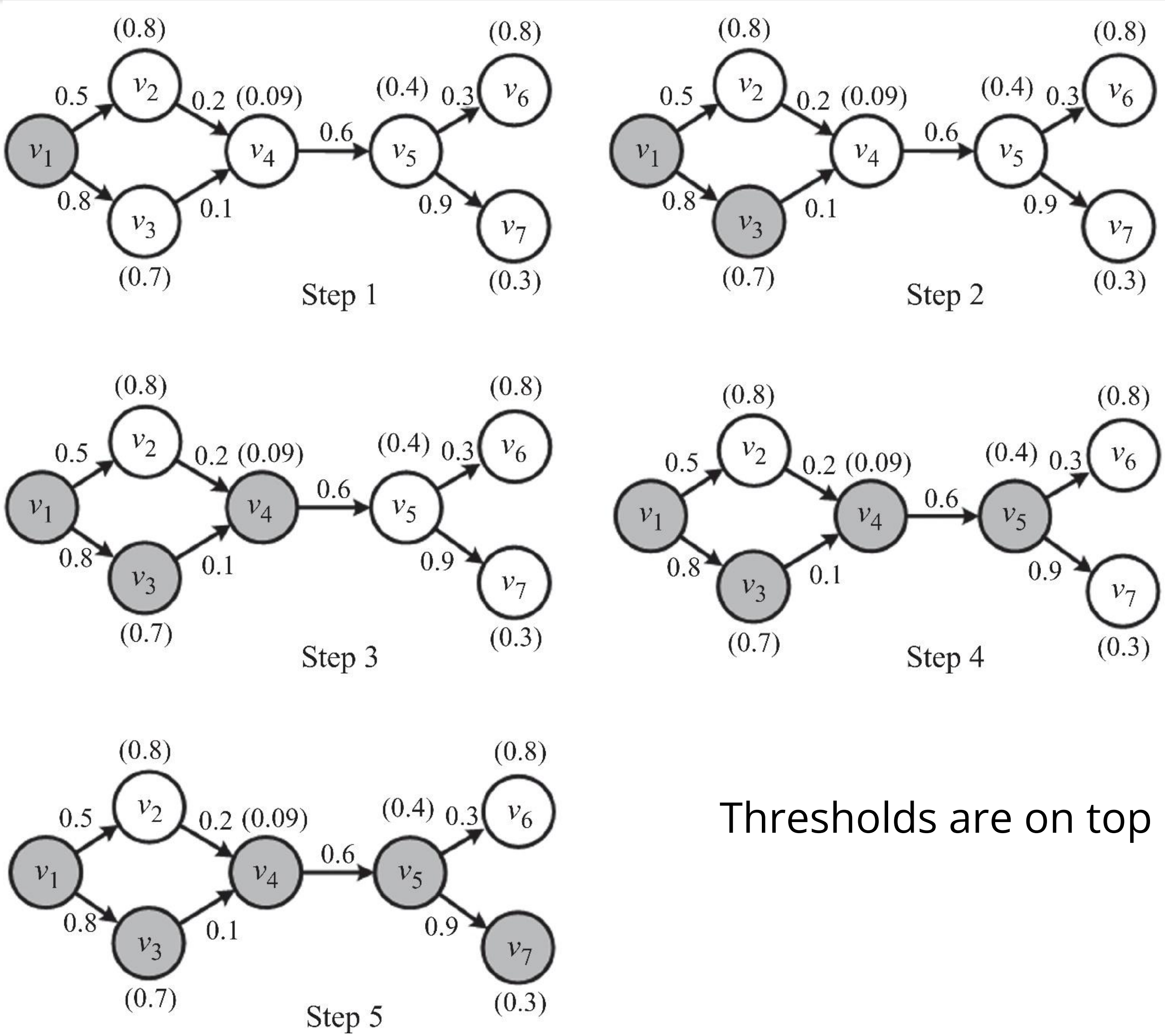
LTM Algorithm

Algorithm 1 Linear Threshold Model (LTM)

Require: Graph $G(V, E)$, set of initial activated nodes A_0

return Final set of activated nodes A_∞
i=0;
Uniformly assign random thresholds θ_v from the interval [0, 1];
while $i = 0$ or $(A_{i-1} \neq A_i, i \geq 1)$ **do**
    $A_{i+1} = A_i$
    inactive $= V - A_i$;
    for all $v \in$ inactive **do**
        if $\sum_{j \text{ connected to } v, j \in A_i} w_{j,v} \geq \theta_v$. **then**
            activate v;
            $A_{i+1} = A_{i+1} \cup \{v\}$;
        end if
    end for
    $i = i + 1$;
end while
$A_\infty = A_i$;
Return A_∞;

Linear Threshold Model (LTM) - An Example



Thresholds are on top of nodes

Homophily

“Birds of a feather flock together”

Definition

Homophily: the tendency of individuals to associate and bond with similar others

- i.e., love of the same

- People interact more often with people who are "*like them*" than with people who are dissimilar

What leads to Homophily?

- Race and ethnicity, Sex and Gender, Age, Religion, Education, Occupation and social class, Network positions, Behavior, Attitudes, Abilities, Beliefs, and Aspirations

Measuring Homophily

- We can measure how the assortativity of the network changes over time
  - Consider two snapshots of a network $G_t(V,E)$ and $G_{t'}(V,E')$ at times t and t', respectively, where $t' > t$
  - $\boldsymbol{V}$: fixed, $\boldsymbol{E}$: edges are added/removed over time.

Nominal attributes. The Homophily index is defined as

$$H = Q^{t'}_{normalized} - Q^{t}_{normalized}$$

Ordinal attributes. The Homophily index is defined as the change in Pearson correlation

$$H = \rho^{t'} - \rho^{t}$$

Modeling Homophily

Homophily can be modeled using a variation of ICM

- At each time step, a single node gets activated
  - A node once activated will remain activated
- $P_{v,w}$ in the ICM model is replaced with the ***similarity*** between nodes v and w, $sim(v,w)$
- When a node v is activated, we generate a random tolerance value θ_v for the node, between 0 and 1
  - The tolerance value is the minimum similarity, node v requires for being connected to other nodes
- For any edge (v,u) that is still not in the edge set, if the similarity $sim(v,w) > \theta_v$, then edge (v,w) is added
- This continues until all vertices are visited

Homophily Model

Algorithm 1 Homophily Model

Require: Graph $G(V, E)$, $E = \emptyset$, similarities $sim(v, u)$

```
return  Set of edges E
for all v ∈ V do
    θ_v = generate a random number in [0,1];
    for all (v, u) ∉ E do
        if θ_v < sim(v, u) then
            E = E ∪ (v, u);
        end if
    end for
end for
Return E;
```