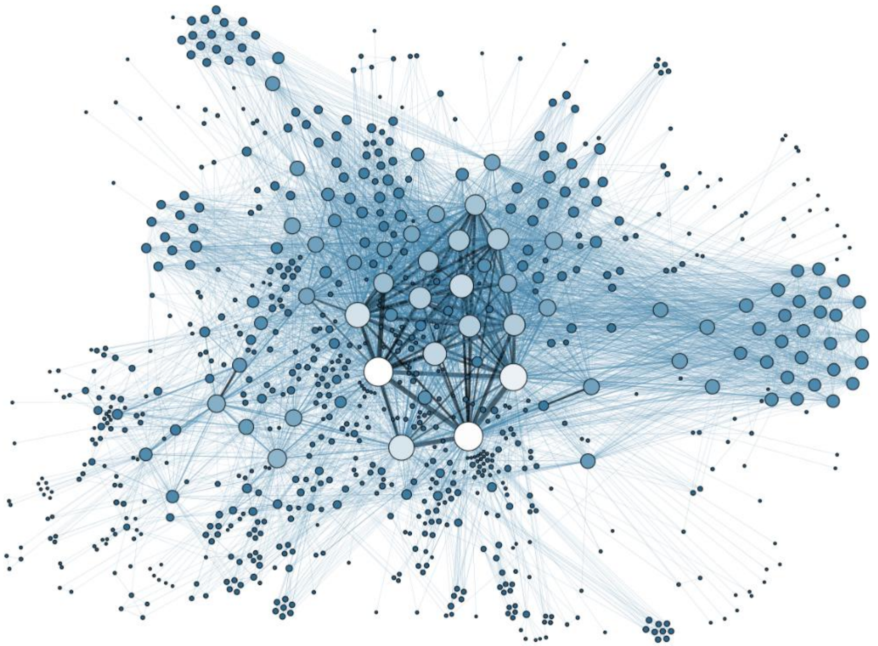
Social Media & Network Analytics

Influence and Homophily



Social Forces

- **Social Forces** connect individuals in different ways

- When individuals get connected, we observe distinguishable patterns in their connectivity networks.
  - **Assortativity**, also known as *social similarity*

- In networks with assortativity:
  - Similar nodes are connected to one another more often than dissimilar nodes.

- Social networks are assortative
  - A high similarity between friends is observed
  - We observe similar behavior, interests, activities, or shared attributes such as language among friends

Why are connected people similar?

Influence

- The process by which a user (i.e., influential) affects another user
- The influenced user becomes more similar to the influential figure.
  - **Example:** If most of our friends/family members switch to a cellphone company, we might switch [i.e., become influenced] too.

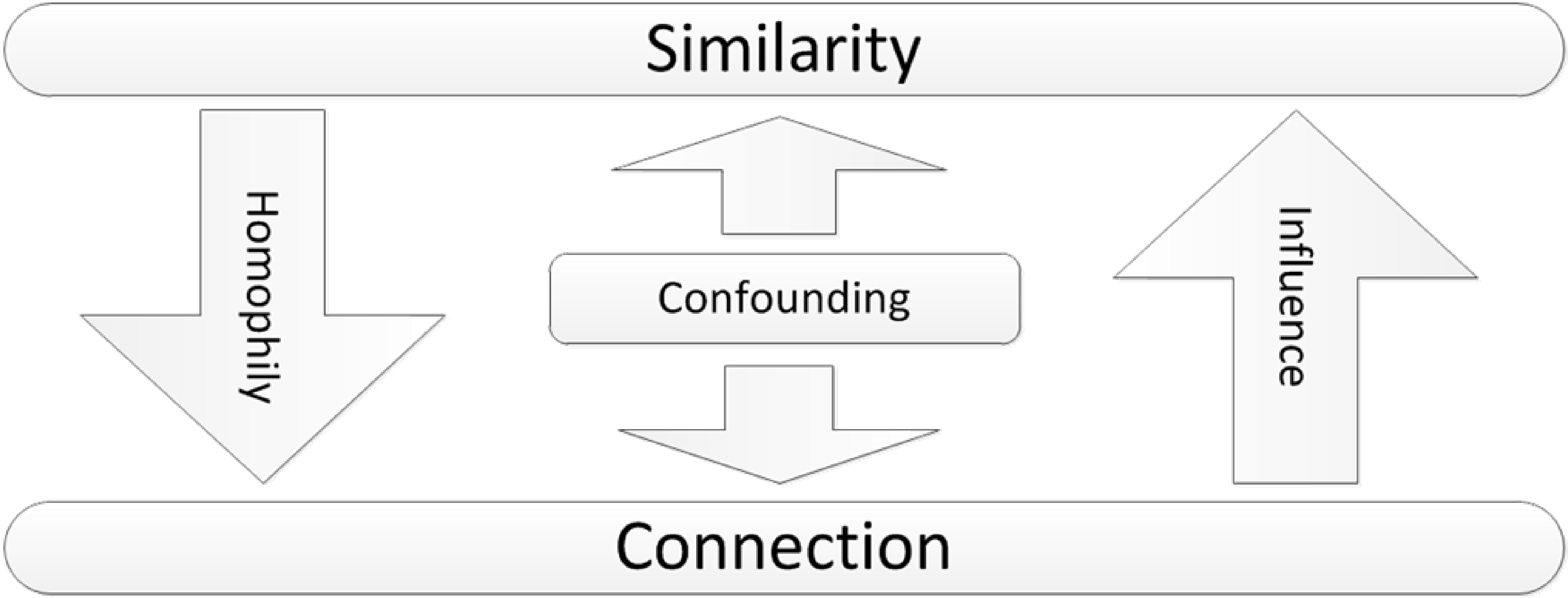
Homophily

- Similar individuals becoming friends due to their high similarity
  - **Example:** Two musicians are more likely to become friends.

Confounding

- The environment's effect on making individuals similar
  - **Example:** Two individuals living in the same city are more likely to become friends than two random individuals

Influence, Homophily, and Confounding



Source of Assortativity in Networks

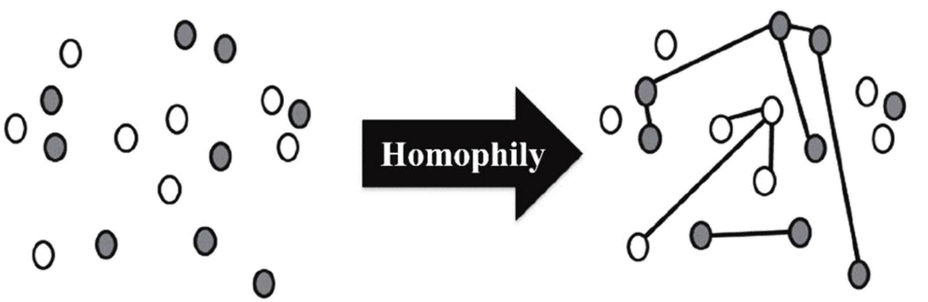
Both influence and Homophily generate similarity in social networks

Influence

Makes connected nodes similar to each other

Homophily

Selects similar nodes and links them together



Assortativity Example

The city's draft tobacco control strategy says more than 60% of under-16s in Plymouth smoke regularly

Why?

- Smoker friends influence their non-smoker friends — **Influence**

- Smokers become friends — **Homophily**

- There are lots of places that people can smoke — **Confounding**

Our goal?

1. How can we **measure assortativity**?
2. How can we **measure influence** or **homophily**?
3. How can we **model influence** or **homophily**?
4. How can we **distinguish between the two**?

Measuring Assortativity

Assortativity: An Example

- The friendship network in a US high school in 1994
- Colors represent races,
  - White
  - **Grey**
  - **Light Grey**
  - **Black**
- High assortativity between individuals of the same race

Measuring Assortativity for Nominal Attributes

- Assume **nominal** attributes are assigned to nodes
  - Example: race

- Edges between nodes of the same type can be used to measure assortativity of the network
  - Same type = nodes that share an attribute value
  - Node attributes could be nationality, race, sex, etc.

$$\frac{1}{m}\sum_{(v_i,v_j)\in E}\delta(t(v_i),t(v_j)) = \frac{1}{2m}\sum_{ij}A_{ij}\delta(t(v_i),t(v_j))$$

$t(v_i)$ denotes type of vertex v_i

$$\delta(x,y) = \begin{cases} 0, & \text{if } x \neq y \\ 1, & \text{if } x = y \end{cases}$$

Kronecker delta function

Assortativity Significance

- **Assortativity significance**
  - The difference between **measured assortativity** and **expected assortativity**
  - The higher this difference, the more significant the assortativity observed

Example

- In a school, 50% of the population is **white** and the other 50% is **Light Grey**.
- We expect 50% of the connections to be between members of different races.
- If all connections are between members of different races, then we have a significant finding

Assortativity Significance

Assortativity

Expected assortativity
(according to configuration model)

$$
\begin{aligned}
Q &= \boxed{\frac{1}{2m}\sum_{ij} A_{ij}\delta(t(v_i), t(v_j))} - \boxed{\frac{1}{2m}\sum_{ij} \frac{d_i d_j}{2m}\delta(t(v_i), t(v_j))} \\
&= \frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\delta(t(v_i), t(v_j)).
\end{aligned}
$$

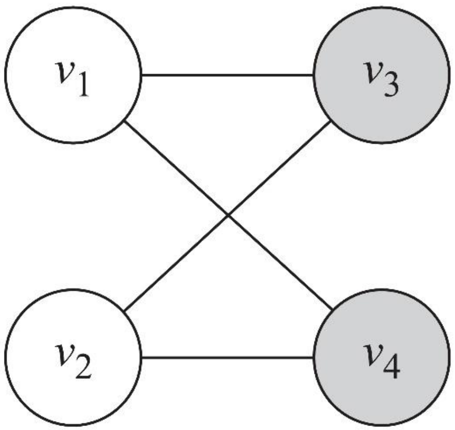
This is modularity

Normalized Modularity [Finding the Maximum]

The maximum happens when all vertices of the same type are connected to one another

$$
\begin{aligned}
Q_{\text{normalized}} &= \frac{Q}{Q_{\max}} \\
&= \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m} \right)\delta(\, t(v_i), t(v_j)\,)}{\max\left[\frac{1}{2m}\sum_{ij} A_{ij}\delta(\, t(v_i), t(v_j)\,) - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}\delta(\, t(v_i), t(v_j)\,)\right]} \\
&= \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m} \right)\delta(\, t(v_i), t(v_j)\,)}{\frac{1}{2m}2m - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}\delta(\, t(v_i), t(v_j)\,)} \\
&= \frac{\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m} \right)\delta(\, t(v_i), t(v_j)\,)}{2m - \sum_{ij}\frac{d_i d_j}{2m}\delta(\, t(v_i), t(v_j)\,)}
\end{aligned}
$$

Modularity Example



$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \quad \Delta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}, m = 4$$

$$B = A - \mathbf{d}\mathbf{d}^T/2m = \begin{bmatrix} -0.5 & -0.5 & 0.5 & 0.5 \\ -0.5 & -0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \end{bmatrix}$$

$$Q = \frac{1}{2m}\mathrm{Tr}(\Delta^T B \Delta) = -0.5$$

The number of edges between nodes of the **same color** is less than the **expected** number of edges between them

Influence

- Measuring Influence
- Modeling Influence

Influence: Definition

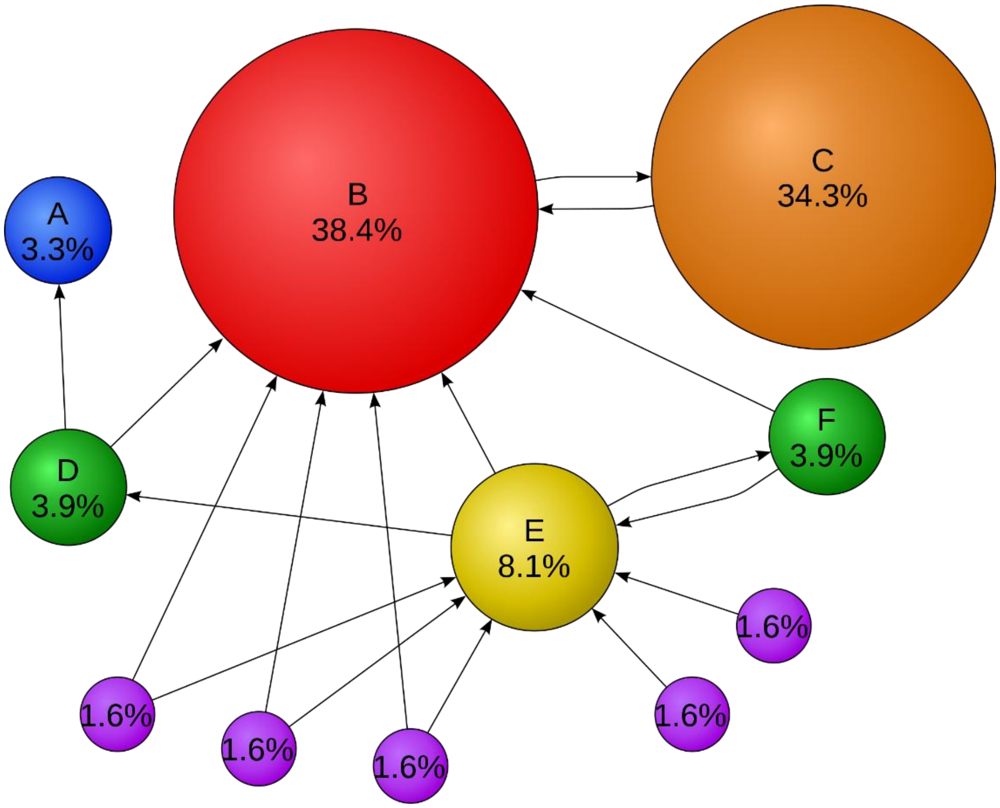
Influence

The act or power of producing an effect, how someone behaves or thinks, possibly 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

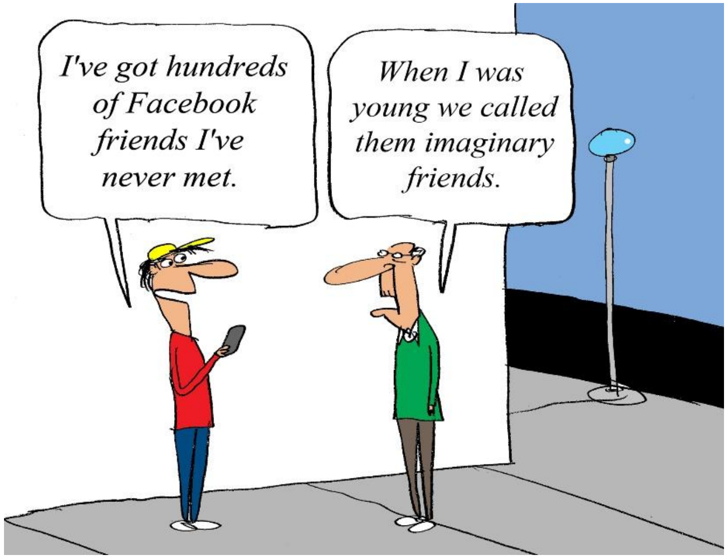
Prediction-based Measurement

We assume that

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

predicts how influential the user **will** be

- Example 1:
  - We can assume that the number of friends of an individual is correlated with how influential she will be
    - It is natural to use any of the centrality measures discussed for prediction-based influence measurements
    - How strong are these friendships?

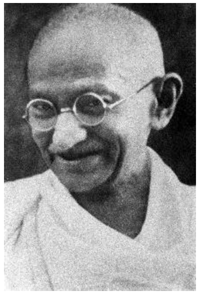


- Example 2:
  - On Twitter, in-degree (number of followers) is a benchmark for measuring influence commonly used

TWEETS	FOLLOWING	FOLLOWERS
42.7K	117K	214K

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 Influence on Blogosphere (see book)

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 may not be 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 an 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]

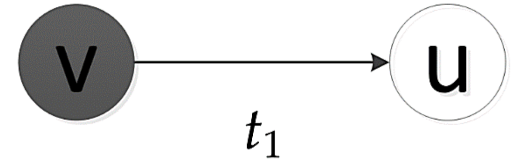
- Popular users (users with high in-degree) do not necessarily have high ranks in terms of number of retweets or mentions.

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

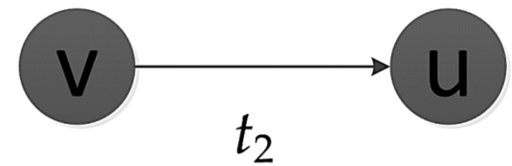
Influence Modeling

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

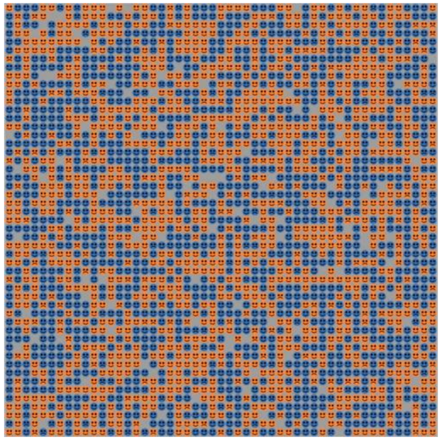


- Each node is started as active or inactive
- A node, once activated, will activate its neighbors
- An activated node cannot be deactivated
- The influence process takes place in a network

Threshold Models

- Simple, yet effective methods for modeling influence in explicit networks
- Nodes make decision based on the influence coming from 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



From: 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 $[0..t-1]$ remain active, but only nodes activated at time $t-1$ get the chance to activate
- Nodes satisfying the following condition will be activated

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

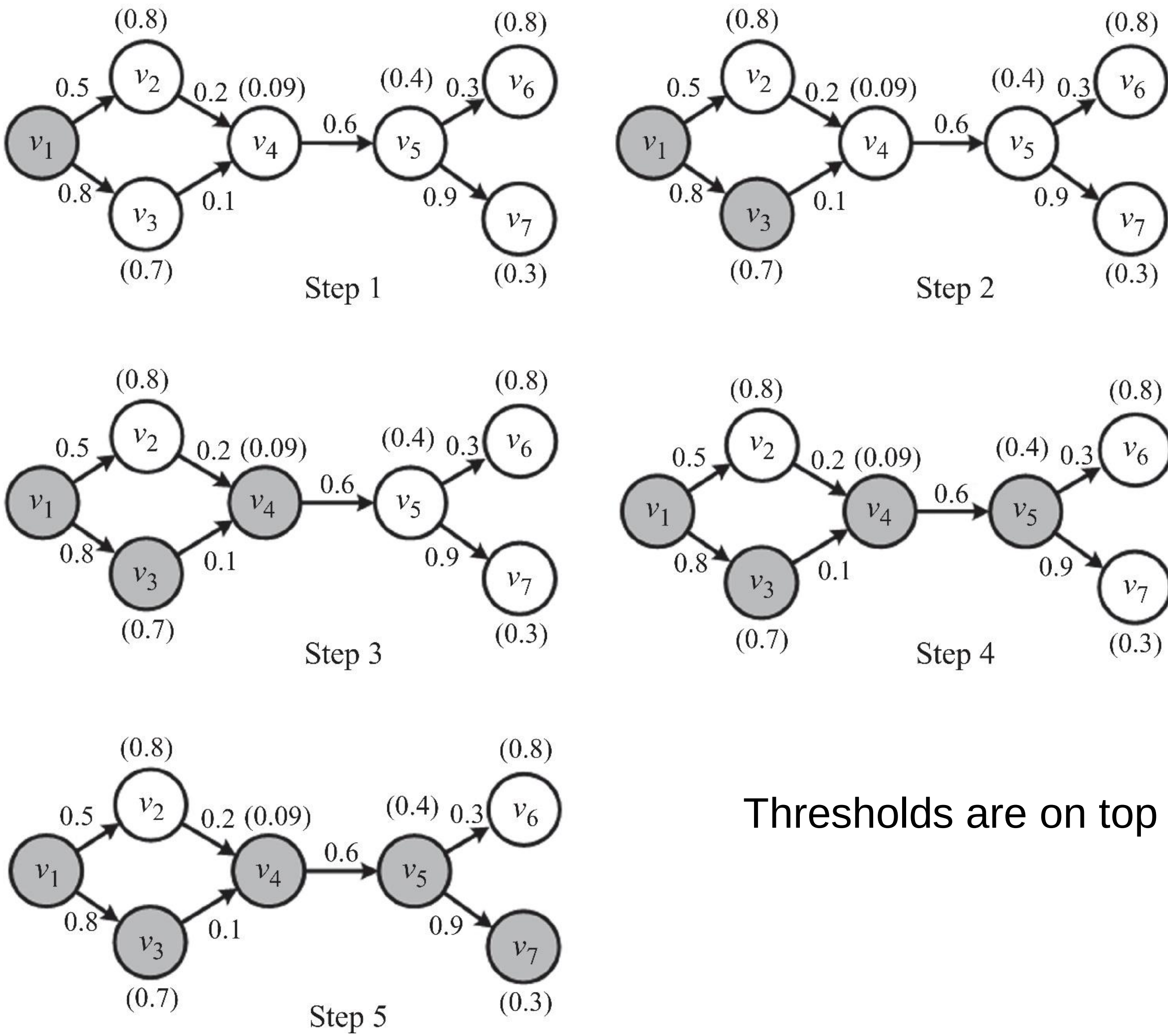
LTM Algorithm

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

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

Modeling Homophily

Homophily can be modeled using a variation of ICM

- At each time step, a single node gets activated.

  This is used for evaluating.

- $P_{v,w}$ 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, w) 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 activated.

Homophily Model

Algorithm 1 Homophily Model

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

return Set of edges E
for all $v \in V$ **do**
    θ_v = generate a random number in [0,1];
    for all $(v, u) \notin E$ **do**
        if $\theta_v < sim(v, u)$ **then**
            $E = E \cup (v, u)$;
        end if
    end for
end for
Return E;

Summary

- Assortivitiy in networks
- Influence and Homophily
- Tests for above