中国科学院软件研究所
Institute of Software, Chinese Academy of Sciences

Big Data on Social Media Mining and Analytics
- Assortative Networks

May 8, 2015

Social Forces

- Social Forces connect individuals in different ways
- Among connected individuals, one often observes high social similarity or assortativity
  - In networks with assortativity, similar nodes are connected to one another more often than dissimilar nodes.
  - In social networks, a high similarity between friends is observed
  - This similarity is exhibited by similar behavior, similar interests, similar activities, and shared attributes such as language, among others.
- Friendship networks are examples of assortative networks

- Many social forces induce assortative networks. Three common forces are *influence*, *homophily*, and *confounding*.

Definition: **Assortativity**, or assortative mixing is a preference for a network's nodes to attach to others that are similar in some way.

Why connected people are similar?

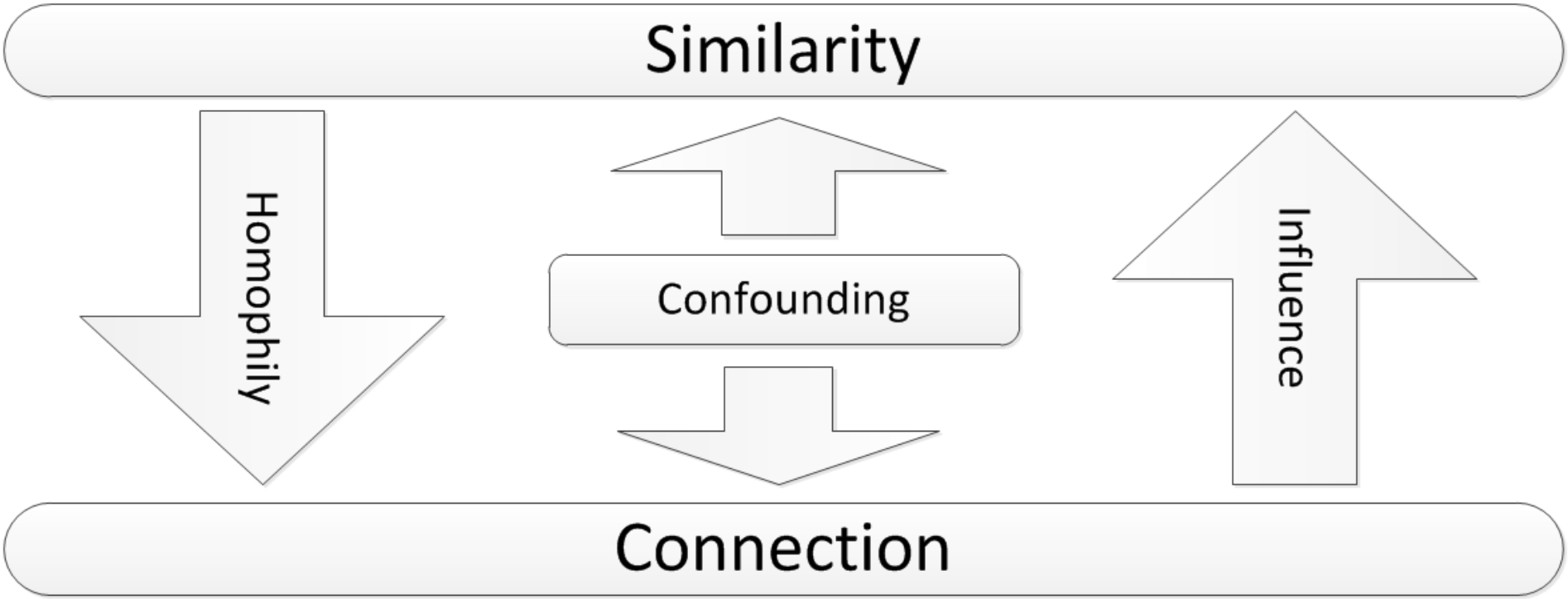
- **Influence**

Influence is the process by which an individual (the influential) affects another individual such that the influenced individual becomes more similar to the influential figure.

E.g. If most of one's friends switch to a mobile company, he might be influenced by his friends and switch to the company as well.

- **Homophily**
  - It is realized when similar individuals become friends due to their high similarity.
    - Two musicians are more likely to become friends.
- **Confounding**
  - Confounding is environment's effect on making individuals similar
    - Two individuals living in the same city are more likely to become friends than two random individuals

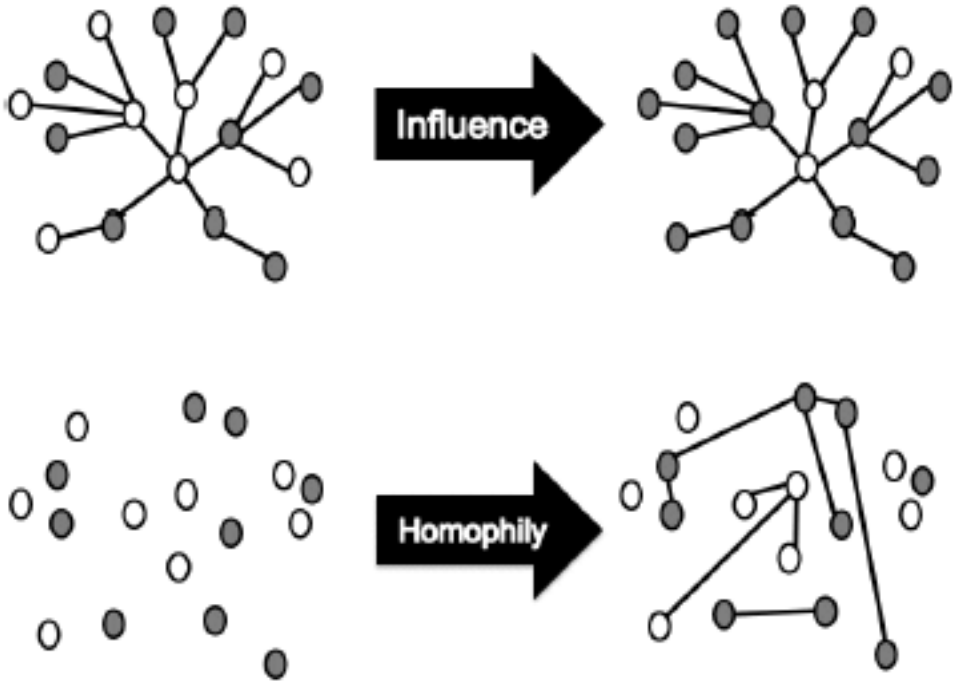
Influence, Homophily, and Confounding



Source of Assortativity in Networks

Both influence and Homophily generate similarity in social networks but in different ways

- **Influence** makes the connected nodes similar to each other
- **Homophily** selects similar nodes and links them together



Similarity of Connected Nodes in Social Networks

- Race
- Religion
- Education
- Income level
- Job and skills
- Language
- Interests and preferences

Assortativity: An Example

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

Smoking Behavior In a Group of Friends: why is happening?

- Smoker friends influence their non-smoker friends

Influence

- Smokers become friends

Homophily

- There are lots of places that people can smoke

Confounding

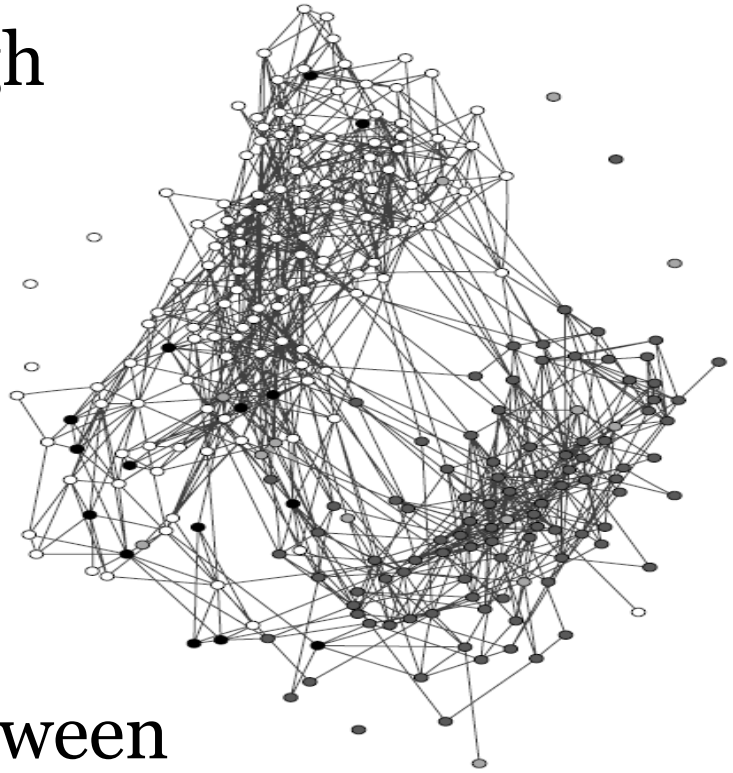
Our goal

- How can we measure assortativity?
- How can we measure influence or homophily?
- How can we model influence or homophily?
- How can we distinguish the two?

Measuring Assortativity

Assortativity: An Example

- The friendship network in a high school in the US in 1994
- Colors in represent races,
  - Whites are white,
  - Blacks are grey
  - Hispanics are light grey
  - Others are black
- There is a high assortativity between individuals of the same race



This technique works for nominal attributes, such as race, but does not work for ordinal ones such as age. Consider a network where individuals are friends with people of different ages.

Measuring Assortativity for Nominal Attributes

- Where nominal attributes are assigned to nodes (race), we can use edges that are between nodes of the same type (i.e., attribute value) to measure assortativity of the network
  - 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 measures the difference between the measured assortativity and its expected assortativity
  - The higher this value, the more significant the assortativity observed
- **Example**
  - Consider a school where half the population is white and half the population is Hispanic. It is expected for 50% of the connections to be between members of different races. If all connections in this school were between members of different races, then we have a significant finding

Assortativity Significance: Measuring

The expected number of edges between v_i and v_j is $d_i d_j / 2m$

The expected number of edges between v_i and v_j that are of the same type is

Assortativity

The expected assortativity in the whole graph

$$Q = \frac{1}{2m}\sum_{ij} A_{ij}\delta(t(v_i), t(v_j)) - \frac{1}{2m}\sum_{ij} \frac{d_i d_j}{2m}\delta(t(v_i), t(v_j))$$

$$= \frac{1}{2m}\sum_{ij} (A_{ij} - \frac{d_i d_j}{2m})\delta(t(v_i), t(v_j)).$$

This measure is called modularity.

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

(i.e., when $A_{ij} = 1$, $\delta(t(v_i), t(v_j)) = 1$)

Normalized Modularity

We can normalize modularity by dividing it by the maximum it can take:

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

Modularity: Matrix Form

Modularity can be simplified using a matrix format.

- Let $\Delta \in \mathbb{R}^{n \times k}$ denote the indicator matrix and let k denote the number of types

$$\Delta_{x,k} = \begin{cases} 1, & \text{if } t(x) = k; \\ 0, & \text{if } t(x) \neq k \end{cases}$$

- The Kronecker delta function can be reformulated using the indicator matrix

$$\delta(\, t(v_i), t(v_j)\,) = \sum_k \Delta_{v_i,k} \Delta_{v_j,k}.$$

- Therefore, $(\Delta\Delta^T)_{i,j} = \delta(t(v_i), t(v_j))$.

Normalized Modularity: Matrix Form

Let Modularity matrix be:

$$B = A - dd^T/2m$$ where $\mathbf{d} \in \mathbb{R}^{n\times 1}$ is the degree vector for all nodes.

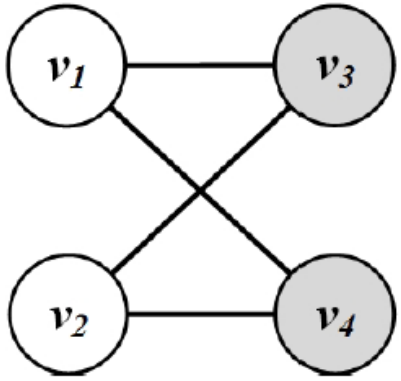
Then, modularity can be reformulated as

Given that the trace of multiplication of two matrices X and Y^T is $Tr(XY^T) = \sum_{i,j} X_{i,j} Y_{i,j}$ and $Tr(XY) = Tr(YX)$, modularity can be reformulated as

$$Q = \frac{1}{2m} \sum_{ij} \underbrace{\left(A_{ij} - \frac{d_i d_j}{2m} \right)}_{B_{ij}} \underbrace{\delta\left(t(v_i), t(v_j) \right)}_{(\Delta\Delta^T)_{i,j}} = \frac{1}{2m} \mathrm{Tr}(B\Delta\Delta^T)$$

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

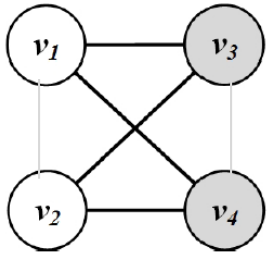
Modularity Example



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

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

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



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

Measuring Assortativity for Ordinal Attributes

- A common measure for analyzing the relationship between two variables with ordinal values is *covariance*.

- It describes how two variables change with respect to each other.

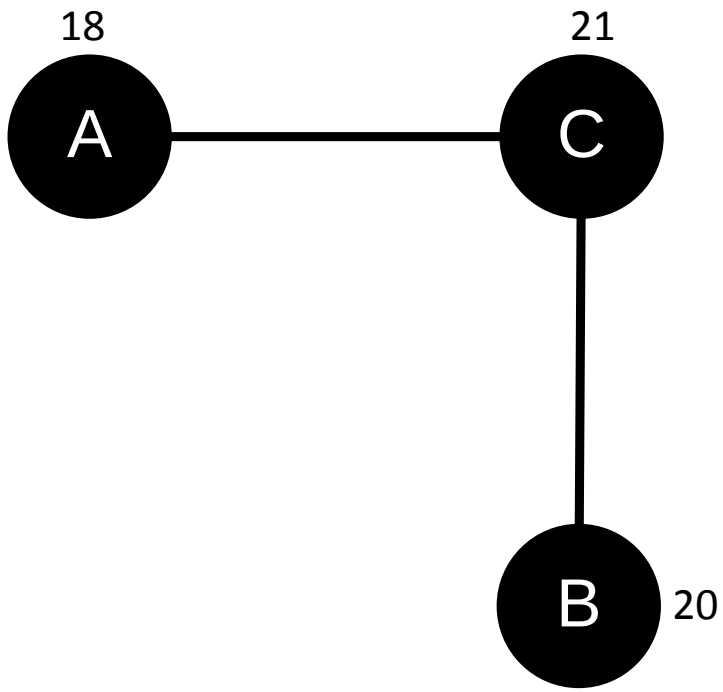
- In our case we are interested in how attribute values of nodes that are connected via edges are correlated.

Covariance Variables

- We construct two variables X_L and X_R, where 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.
- In other words, X_L represents the ordinal values associated with the left node of the edges and X_R represents the values associated with the right node of the edges
- Our problem is therefore reduced to computing 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)

→

$$E(X_L) = E(X_R),$$
$$\sigma(X_L) = \sigma(X_R).$$

Covariance

For two given column variables X_L and X_R the covariance is

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

$\mathrm{E}(X_L)$ is the mean of the variable X_L and $\mathrm{E}(X_L X_R)$ is the mean of the multiplication of 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

$$\sigma(X_L, X_R) = \mathrm{E}[X_L X_R] - \mathrm{E}[X_L]\mathrm{E}[X_R]$$

$$= \frac{\sum_{ij} A_{ij} x_i x_j}{2m} - \frac{\sum_{ij} d_i d_j x_i x_j}{(2m)^2}$$

$$= \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{d_i d_j}{2m} \right) x_i x_j.$$

Pearson correlation (X_L, X_R) is the ***normalized*** version of covariance:

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

Normalizing Covariance

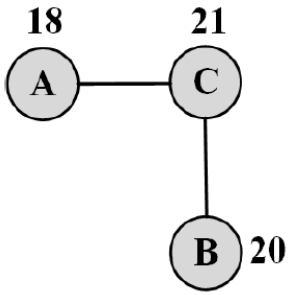
$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)^2},$$

$$= \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)x_i x_j}{\mathrm{E}[(X_L)^2] - (\mathrm{E}[X_L])^2}$$

$$= \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)x_i x_j}{\frac{1}{2m}\sum_{ij} A_{ij}x_i^2 - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}x_i x_j}.$$

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

Example



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

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

Social Influence

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

- Measuring influence in social media
- Modeling Influence in social media

Measuring Influence

- Measuring influence is assigning a number to each node that represents the influential power of that node.
- The influence can be measured either based on *prediction* or *observation*.

Prediction-based Measurement

- In prediction-based measurement, we assume that an individual's attribute or the way she is situated in the network predicts *how influential she will be*.

- For instance, we can assume that the *gregariousness* (e.g., number of friends) of an individual is correlated with how influential she will be. Therefore, it is natural to use any of the centrality measures discussed in *Network Measures* for prediction-based influence measurements.
- An example:
  - On Twitter, in-degree (number of followers) is a benchmark for measuring influence commonly used

Observation-based Measurement

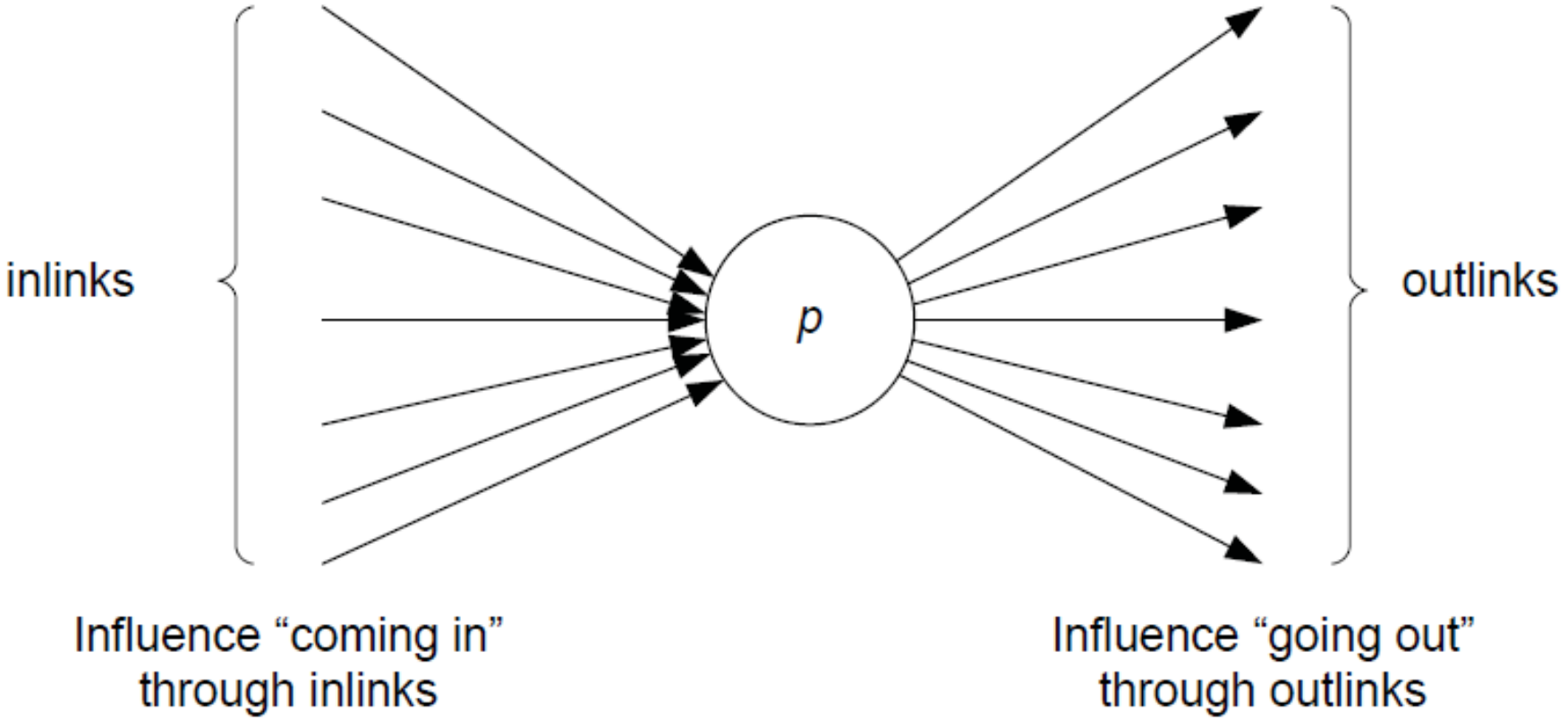
- In observation-based we quantify influence of an individual by measuring the amount of influence attributed to the individual
  - When an individual is the role model
    - Influence measure: size of the audience that has been influenced
  - When an individual spreads information:
    - Influence measure: the size of the cascade, the population affected, the rate at which the population gets influenced
  - When an individual's participation increases values:
    - Influence measure: the increase (or rate of increase) in the value of an item or action

Case Studies for Measuring Influence in Social Media

Measuring Social Influence in the Blogosphere

- The goal of measuring influence in blogosphere is to figure out most influential bloggers on the blogosphere
- Due to limited time an individual has, following the influentials is often a good heuristic of filtering what's uninteresting
- One common measure for quantifying influence of bloggers is to use ***indegree centrality***
- Due to the sparsity of in-links, more detailed analysis is required to measure influence in blogosphere

iFinder: A System to measure influence on blogsphore



Social Gestures

- **Recognition**
  - Recognition for a blogpost is the number of the links that point to the blogpost (in-links).
    - Let I_p denotes the set of in-links that point to blogpost p.
- **Activity Generation**
  - Activity generated by a blogpost is the number of comments that p receives.
    - c_p denotes the number of comments that blogpost p receives.
- **Novelty**
  - The blogpost's novelty is inversely correlated with the number of references a blogpost employs. In particular the more citations a blogpost has it is considered less novel.
    - O_p denotes the set of out-links for blogpost p.
- **Eloquence**
  - Eloquence is estimated by the length of the blogpost. Given the unformal nature of blogs and the bloggers tendency to write short blogposts, longer blogposts are believed to be more eloquent. So the length of a blogpost l_p can be employed as a measure of eloquence

Influence Flow

Influence flow describes a measure that accounts for in-links (recognition) and out-links (novelty).

$$InfluenceFlow(p) = w_{\text{in}} \sum_{m=1}^{|I_p|} I(P_m) - w_{\text{out}} \sum_{n=1}^{|O_p|} I(P_n),$$

I(.) denotes the influence of a blogpost and w_{in} and w_{out} are the weights that adjust the contribution of in- and out-links, respectively.

p_m is the number of blogposts that point to blog post p and p_n is the number of blog posts referred to in p.

Blogpost Influence

$$I(p) = w_{\text{length}} l_p (w_{\text{comment}} c_p + InfluenceFlow(p)).$$

- w_{length} is the weight for the length of the blogpost.

- $w_{comment}$ describes how the number of comments is weighted.
- Weights w_{in}, w_{out}, $w_{comments}$, and w_{length} can be tuned to make the model suitable for different domains.

Finally, a blogger's influence index (iIndex) can be defined as the maximum influence value among all his or her N blogposts,

$$iIndex = \max_{p_n \in N} I(p_n).$$

Ref. N. Agarwal, et al. Identifying the influential bloggers in a community, WSDM, ACM, 2008, pp. 207–218.

Measuring Social Influence on Twitter: Measures

- **Indegree**
  - 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**:
  - Tweeter 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.

Each one of these measures by itself can be used to identify influential users in Twitter. This can be performed by utilizing the measure for each individual and then ranking individuals based on their measured influence value.

Measuring Social Influence on Twitter: Measures

- Contrary to public belief, number of followers is considered an inaccurate measure compared to the other two.

- We can rank individuals on twitter ndependently 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

In order to compare ranks across more than one measure (say, indegree and mentions), we can use Spearman's Rank Correlation Coefficient

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

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.

Comparing Ranks Across Three Measures

- Spearman's rank correlation is the Pearsons correlation coefficient for ordinal variables that represent ranks (i.e., takes values between 1. . . n); hence, the 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
Indegree vs Retweets	0.122
Indegree vs Mentions	0.286
Retweets vs Mentions	0.638

M. Cha,et al, Measuring user influence in twitter: The million follower fallacy, AAAI Conference on Weblogs and Social Media, vol. 14, 2010.

Influence Modeling

In influence modeling, the goal is to design models that can explain *how individuals influence one another* in explicit and implicit networks.

Influence Modeling

- At time stamp t_1, node v is activated and node u is not activated
- Node u becomes activated at time stamp t_2, as the effect of the influence
- Each node is started as active or inactive;
- A node, once activated, will activate its neighboring nodes
- Once a node is activated, this node cannot be deactivated

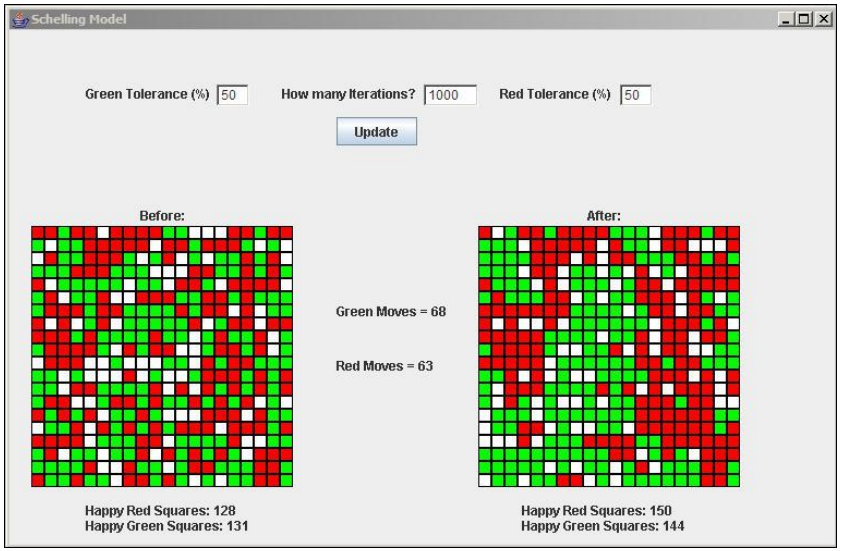
Influence Modeling: Assumptions

- In general we can assume that the influence process takes place in a network of connected individuals.
- Sometimes this network is observable (an explicit network) and sometimes not (an implicit network).
  - In the observable case, we can resort to threshold models such as the linear threshold model (LTM)
  - In the case of implicit networks, we can employ methods such as the Linear Influence Model (LIM) that take the number of individuals who get influenced at different times as input, e.g., the number of buyers per week

Threshold Models for Explicit Networks

- Threshold models are simple, yet effective methods for modeling influence in explicit networks
- In threshold model actors make decision based on the number or the fraction (the threshold) of their neighborhood that have already decided to make the same decision

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



Linear Threshold Model (LTM)

Assume a weighted directed graph $G(V, E)$, where nodes v_j and v_i are connected with weight $w_{j,i} \geq 0$. This weight denotes how much node v_j can affect node v_i's decision.

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

Where $N_{in}(v_i)$ denotes the incoming neighbors of node v_i.

In a linear threshold model, each node $\boldsymbol{v_i}$ is assigned a threshold θ_i such that when the amount of influence exerted toward $\boldsymbol{v_i}$ by its active incoming neighbors is more than θi , then vi becomes active, if still inactive.

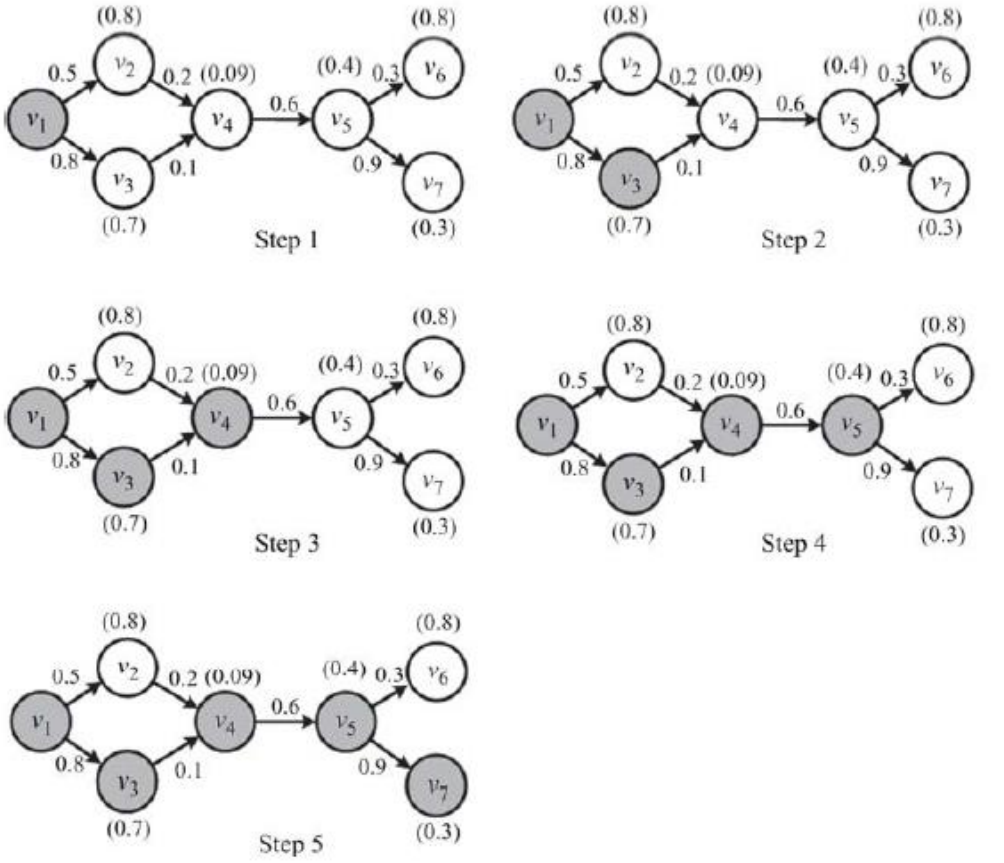
Linear Threshold Model (LTM)

Thus, for v_i to become active at time t, we should have

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

where A_{t-1} denotes the set of active nodes at the end of time *t-1*. The threshold values are generally assigned uniformly at random to nodes from the interval [0,1]. Note that the threshold θ_i defines how resistant to change node v_i is: a very small θ_i value might indicate that a small change in the activity of v_i's neighborhood results in v_i becoming active and a large θ_i shows that v_i resists changes.

Linear Threshold Model- An Example



Algorithm Linear Threshold Model (LTM)

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

1: **return** Final set of activated nodes A_∞
2: i=0;
3: Uniformly assign random thresholds θ_v from the interval [0, 1];
4: **while** $i = 0$ or $(A_{i-1} \neq A_i, i \geq 1)$ **do**
5: $A_{i+1} = A_i$
6: inactive $= V - A_i$;
7: **for all** $v \in$ inactive **do**
8: **if** $\sum_{j \text{ connected to } v, j \in A_i} w_{j,v} \geq \theta_v$. **then**
9: activate v;
10: $A_{i+1} = A_{i+1} \cup \{v\}$;
11: **end if**
12: **end for**
13: $i = i + 1$;
14: **end while**
15: $A_\infty = A_i$;
16: Return A_∞;

Linear Threshold Model (LTM) Simulation. The values attached to nodes denote thresholds θ_i, and the values on the edges represent weights $w_{i,j}$.

Influence in Implicit Networks

- An implicit network is one where the influence spreads over nodes in the network
- Unlike the threshold model, one cannot observe individuals who are responsible for influencing others (the influentials), but only those who get influenced
- The information available is:
  - The set of influenced individuals at any time, P(t)
  - Time t_u, where each individual u gets initially influenced (activated)

Influence in Implicit Networks

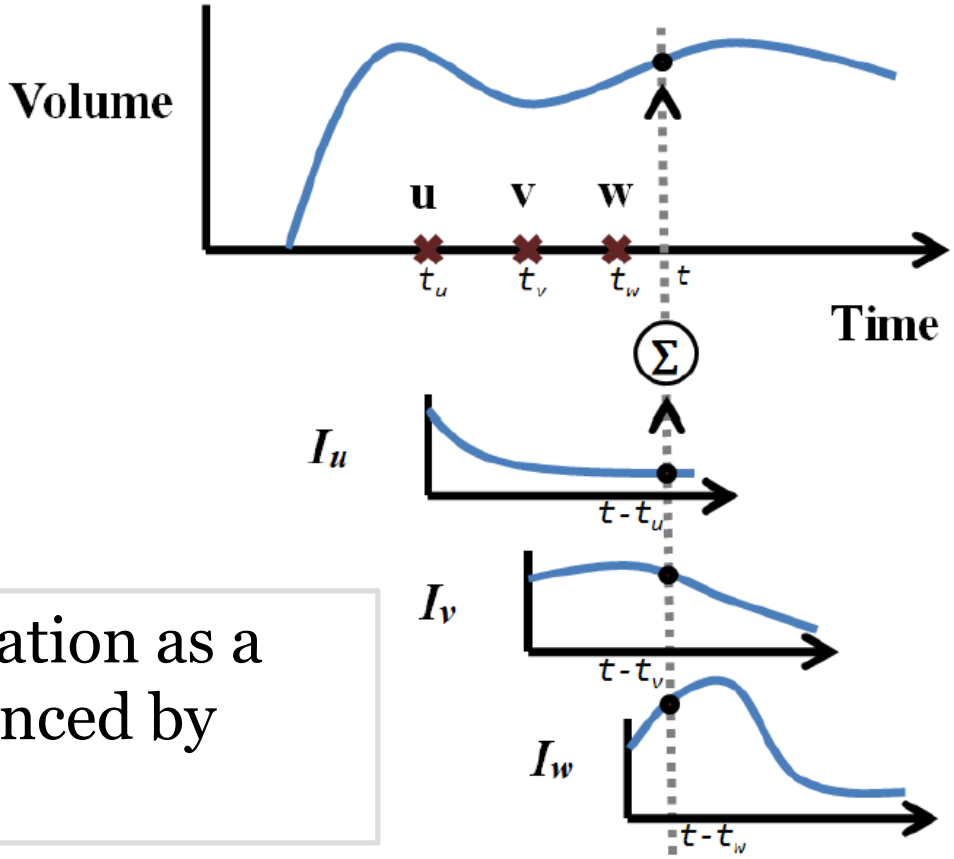
Assume that any influenced individual u can influence I(u, t) non-influenced individuals at time t.

- Assuming discrete timesteps, we can formulate the size of influenced population as

$$|P(t)| = \sum_{u \in P(t)} I(u, t - t_u).$$

The Size of the Influenced Population

At time t, the total number of influenced individuals is the summation of influence functions I_u, I_v, and I_w at time steps $t - t_u$, $t - t_v$, and $t - t_w$, respectively



The size of the influenced population as a summation of individuals influenced by activated individuals

Individuals u, v, and w are activated at time steps t_u, t_v, and t_w, respectively

Parametric estimation

The **goal** is to estimate I(., .) given activation time and the number of influenced individuals at any time.

A simple way to use a probability distribution to estimate *I* function. Assume that all users influence others in the same parametric form. For instance, , one can use the *power-law distribution* to estimate influence:

$$I(u,t) = c_u(t - t_u)^{-\alpha_u}$$

where we estimate coefficients c_u and α_u for any *u* by methods such as *maximum likelihood estimation.*

Non-Parametric Estimation: linear influence model (LIM)

A more flexible approach is to assume a nonparametric function and estimate the influence function's form.

In LIM, assume that nodes can get deactivated over time and can no longer influence others.

- Let A(u, t) = 1 denote node u is active at time t
- A(u, t) = 0 denotes that u is either deactivated or still not influenced,
- |V| is the total size of population and T is the last time stamp

$$|P(t)| = \sum_{u \in P(t)} I(u, t - t_u). \quad \Rightarrow \quad |P(t)| = \sum_{u=1}^{|V|} \sum_{t=1}^{T} A(u,t) I(u,t),$$

Equivalently in matrix form

$$P = AI.$$

$$\begin{aligned} \text{minimize} \quad & \|P - AI\|_2^2 \\ \text{subject to} \quad & I \geq 0. \end{aligned}$$

This can be solved using ***non-negative least square*** methods.

Homophily

Homophily is the tendency of similar individuals to become friends.

Homophily- Definition

- Homophily (i.e., "love of the same") is the tendency of individuals to associate and bond with similar others
- 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, Believes, and Aspirations

Measuring Homophily

- To measure homophily, one can measure how the assortativity of the network changes over time
  - Consider two snapshots of a network Gt(V, E) and Gt' (V, E') at times t and t', respectively, where t' > t
  - Without loss of generality, assume that the number of nodes stay fixed and edges connecting them are added or removed over time.

For nominal attributes, the homophily index is defined as

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

where $Q_{normalized}$ is defined

$$Q_{normalized} = \frac{\sum_{ij}\left(A_{ij} - \frac{d_i d_j}{2m}\right)\delta\left(t(v_i), t(v_j)\right)}{2m - \sum_{ij}\frac{d_i d_j}{2m}\delta\left(t(v_i), t(v_j)\right)}.$$

Similarly, for ordinal attributes, the homophily index can be defined as the change in the Pearson correlation

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

$$\rho(X_L, X_R) = \frac{\frac{1}{2m}\sum_{ij}\left(A_{ij} - \frac{a_i a_j}{2m}\right)x_i x_j}{\frac{1}{2m}\sum_{ij}A_{ij}x_i^2 - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}x_i x_j}.$$

Modeling Homophily

We model homophily using a variation of Independent Cascade Model

- At each time step, a single node gets activated. and the activated node gets a chance of getting connected to other nodes due to homophily
  - A node once activated will remain activated.
- $\boldsymbol{P_{v,w}}$ in the ICM model is replaced with the similarity between nodes v and w, $\boldsymbol{sim(v, w)}$. (Let *v* be an active node at time *t*. Then, for any neighbor *w*, there is a probability $\boldsymbol{P_{v,w}}$ that node w gets activated at *t+1*.)

When a node v is activated, we generate a random tolerance value θ_v for the node, between 0 and 1.

The tolerance value defines the minimum similarity, node ***v*** tolerates or requires for being connected to other nodes.

Then for any edge (*v*, *u*) that is still not in the edge set, if the similarity $\boldsymbol{sim(v, w) > \theta_v}$, the edge (*v*, *w*) is added.

This continues until all vertices are visited.

Algorithm Homophily Model

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

```
1: return Set of edges E
2: for all v ∈ V do
3:    θ_v = generate a random number in [0,1];
4:    for all (v, u) ∉ E do
5:       if θ_v < sim(v, u) then
6:          E = E ∪ (v, u);
7:       end if
8:    end for
9: end for
10: Return E;
```

Distinguishing Influence and Homophily

Distinguishing Influence and Homophily

- We are often interested in understanding which social force (influence or homophily) resulted in an assortative network.

- To distinguish between an influence-based assortativity or homophily-based one, *statistical tests* can be used.

(The *shuffle test*, the *edge-reversal test*, and the *randomization test*)

- Note that in all these tests, we assume that several temporal snapshots of the dataset are available (like the LIM model) where we know exactly, when each node is activated, when edges are formed, or when attributes are changed.

Shuffle Test

The basic idea behind the shuffle test comes from the fact that *influence is temporal*. In other words, when u influences v, then v should have been activated after u.

So, in shuffle test, we define a temporal assortativity measure (called *social correlation*). We assume that if there is no influence, then a shuffling of the activation timestamps should not affect the temporal assortativity measurement.

- *a is the number of active friends,*
- *α measures the social correlation and β* denotes activation bias.

Assume the probability of activation of node v depends on a, the number of already-active friends of v. This activation probability is calculated using a *logistic function*

$$p(a) = \frac{e^{\alpha a+\beta}}{1+e^{\alpha a+\beta}}, \quad \Longleftrightarrow \quad \ln\left(\frac{p(a)}{1-p(a)}\right) = \alpha a + \beta,$$

Activation Likelihood

For computing the number of already active nodes of an individual, we need to know the activation time stamps of the nodes.

Suppose at one time point t , $y_{a,t}$ *users with a active friends* become active, and $n_{a,t}$ *users who also have a active friends yet stay inactive at time* t .

- *The* likelihood function *is*

$$\prod_a p(a)^{y_a}(1-p(a))^{n_a} \qquad y_a = \sum_t y_{a,t} \qquad n_a = \sum_t n_{a,t}$$

Given the user's activity log, we can compute a correlation coefficient α *and bias* β *to maximize the above* likelihood (optional: using a maximum likelihood iterative method).

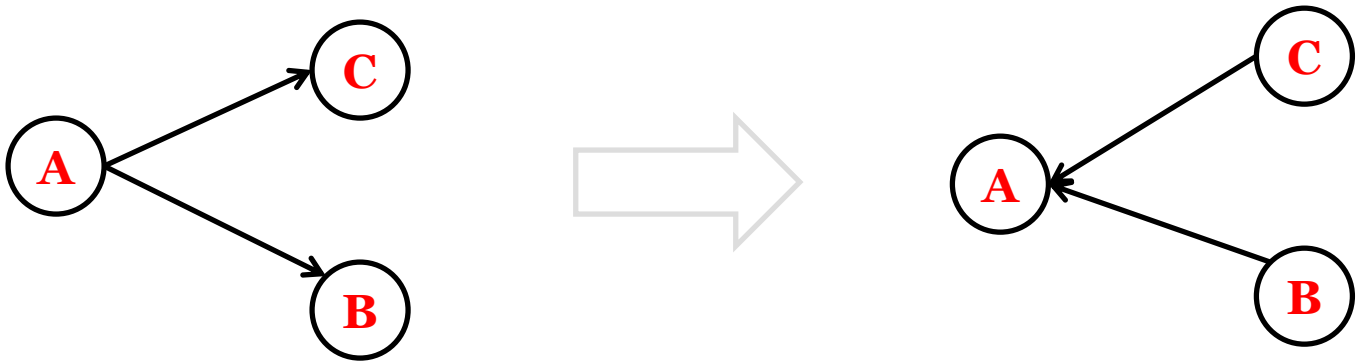
The key idea of the shuffle test is that if influence does not play a role, the timing of activations should be independent of users. Thus, even if we randomly shuffle the timestamps of user activities, we should obtain a similar α *value.*

Test of Influence:
After we shuffle the timestamps of user activities, if the new estimate of social correlation is significantly different from the estimate based on the user's activity log, **there is evidence of influence**.

The Edge-reversal Test

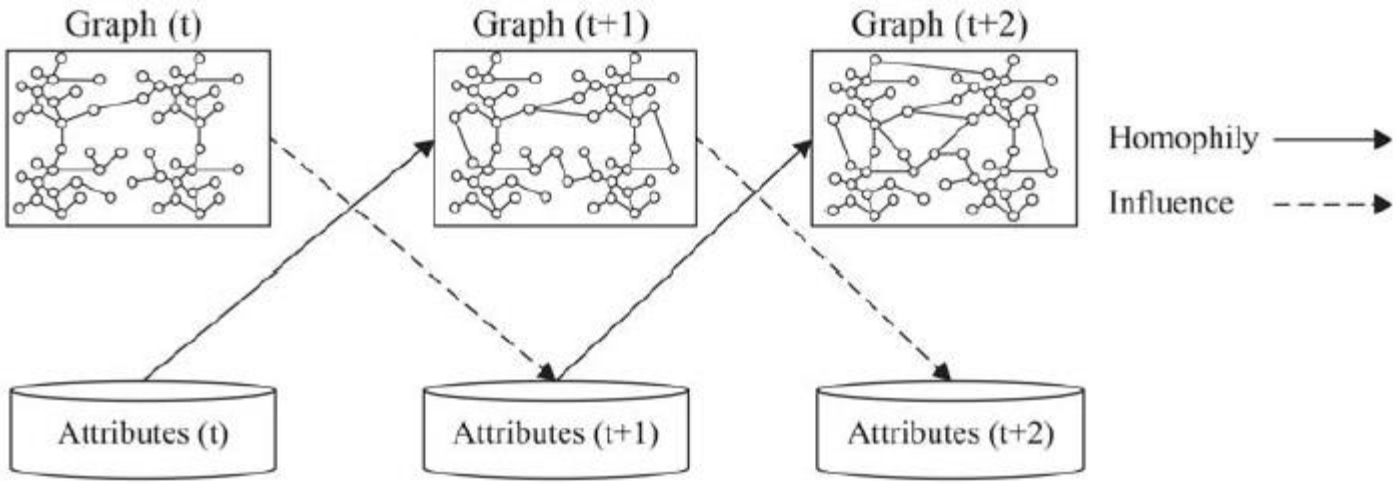
If influence resulted in activation, then the *direction* of edges should be important (who influenced whom).

- Reverse directions of all the edges
- Run the same logistic regression on the data using the new graph $\ln(\frac{p(a)}{1-p(a)}) = \alpha a + \beta,$
- If correlation is not due to influence, then α should not change dramatically.



Randomization Test

Unlike the other two tests, the randomization test is capable of detecting *both influence and homophily in networks*.



What are the differences between influence and homophily performing in networks?

Ans: individuals already linked to one another change their attributes (e.g., a user changes habits), whereas in homophily, attributes do not change but connections are formed due to similarity.

Randomization Test

The assumption is that, if influence or homophily happens in a network, then networks become more assortative.

Let $A(G_t,X_t)$ denote the assortativity of network G and attributes X at time t.

where X denote the attributes associated with nodes (age, gender, location, etc.) and X_t denote the attributes at time t.

Then, the network becomes more assortative at time t+1 if $A(G_{t+1}, X_{t+1}) - A(G_t, X_t) > 0.$

Now, we can assume that part of this assortativity is due to influence if the *influence gain* $G_{Influence}$ is positive,

$$G_{\text{Influence}}(t) = A(G_t, X_{t+1}) - A(G_t, X_t) > 0,$$

and part is due to homophily if we have positive *homophily gain* $G_{Homophily}$:

$$G_{\text{Homophily}}(t) = A(G_{t+1}, X_t) - A(G_t, X_t) > 0.$$

In randomization tests, one determines whether changes in them are significant or not!

Significance Test on influence

1. The influence significance algorithm starts with computing influence gain, which is the assortativity difference observed due to influence (g_0).
2. It then forms a random attribute set at time t+1 (null-hypotheses), assuming that attributes changed randomly at t+1 and not due to influence. This random attribute set XR^i_{t+1} is formed from X_{t+1} by making sure that effects of influence in changing attributes are removed. This randomized set is constructed n times. This set is then used to compute influence gains $\{g_i\}_{i=1}^n$

3. We can assume that whenever g_0 is smaller than $\alpha/2$ % or larger than $1-\alpha/2\%$ of $\{g_i\}_{i=1}^n$ values, it is significant. The value of α is set empirically.

Algorithm Influence Significance Test

Require: G_t, G_{t+1}, X_t, X_{t+1}, number of randomized runs n, α

1: **return** Significance
2: $g_0 = G_{Influence}(t)$;
3: **for all** $1 \le i \le n$ **do**
4: $XR^i_{t+1} = randomize_1(X_t, X_{t+1})$;
5: $g_i = A(G_t, XR^i_{t+1}) - A(G_t, X_t)$;
6: **end for**
7: **if** g_0 larger than $(1-\alpha/2)\%$ of values in $\{g_i\}_{i=1}^n$ **then**
8: return significant;
9: **else if** g_0 smaller than $\alpha/2\%$ of values in $\{g_i\}_{i=1}^n$ **then**
10: return significant;
11: **else**
12: return insignificant;
13: **end if**

Significance Test on influence

Similarly, in the homophily significance test, we compute the original homophily gain and construct random graph links GR^i_{t+1} at time t+1, such that no homophily effect is exhibited in how links are formed. To perform this for any two (randomly selected) links e_{ij} and e_{kl} formed in the original G_{t+1} graph, we form edges e_{il} and e_{kj} in GR^i_{t+1}. This is to make sure that the homophily effect is removed and that the degrees in GR^i_{t+1} are equal to that of G_{t+1}.

Algorithm Homophily Significance Test

Require: G_t, G_{t+1}, X_t, X_{t+1}, number of randomized runs n, α

1: **return** Significance
2: $g_0 = G_{Homophily}(t)$;
3: **for all** $1 \leq i \leq n$ **do**
4: $GR^i_{t+1} = randomize_H(G_t, G_{t+1})$;
5: $g_i = A(GR^i_{t+1}, X_t) - A(G_t, X_t)$;
6: **end for**
7: **if** g_0 larger than $(1 - \alpha/2)\%$ of values in $\{g_i\}^n_{i=1}$ **then**
8: return significant;
9: **else if** g_0 smaller than $\alpha/2\%$ of values in $\{g_i\}^n_{i=1}$ **then**
10: return significant;
11: **else**
12: return insignificant;
13: **end if**

Summary

1. Social forces across social media: *Influence* and *Homophily*.
2. Both influence and homophily result in networks where similar individuals are connected to each other. These are assortative networks.
3. To estimate the assortativity of networks:
   - modularity for nominal attributes
   - correlation for ordinal ones.

4. Measuring and Modeling *Influence*
   prediction-based or observation-based methods
   linear threshold model (LTM) (network information is available)
   linear influence model (LIM). (network information is not available)
5. Measuring and Modeling *Homophily.*
   measured by computing the assortativity dierence in time
   modeled using a variant of independent cascade models.
6. To determine the source of assortativity in social networks: three tests are introduced