Social Media Mining

Influence and Homophily

Dear instructors/users of these slides:

Please feel free to include these slides in your own material, or modify them as you see fit. If you decide to incorporate these slides into your presentations, please include the following note:

R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction*, Cambridge University Press, 2014.
Free book and slides at **http://socialmediamining.info/**

or include a link to the website:
http://socialmediamining.info/

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.

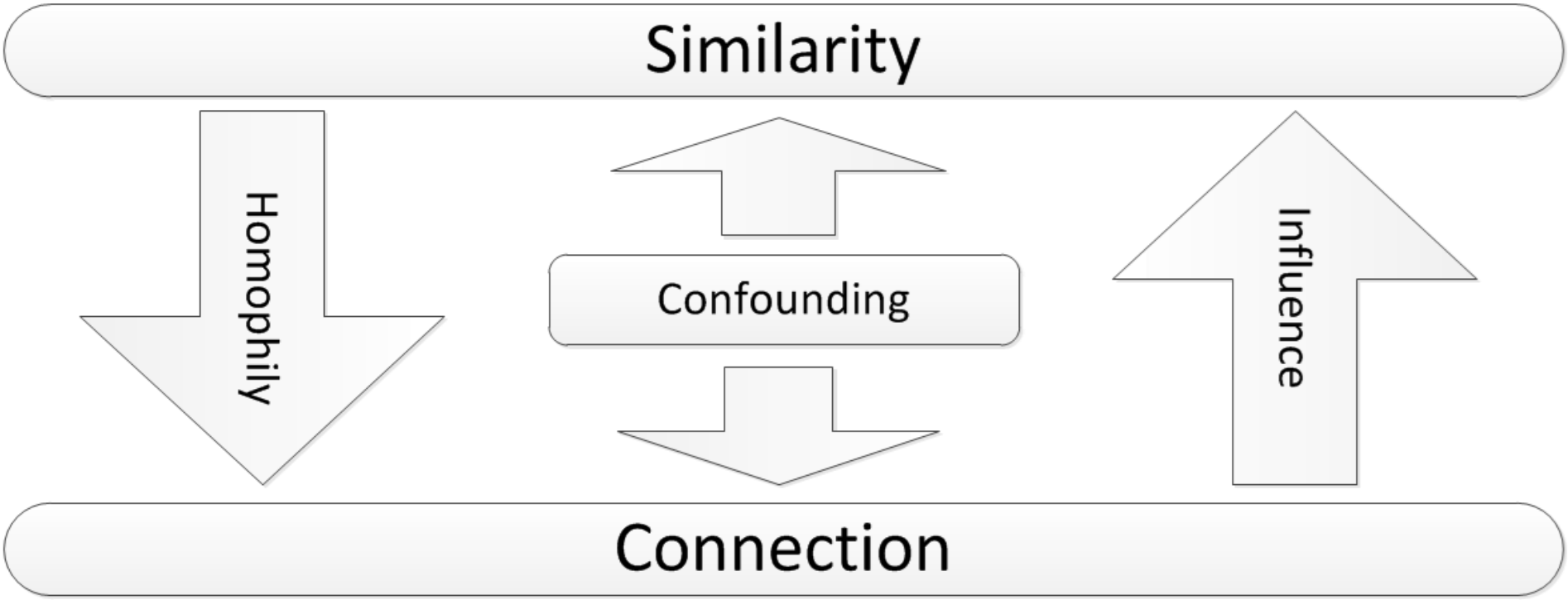
Homophily

- 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

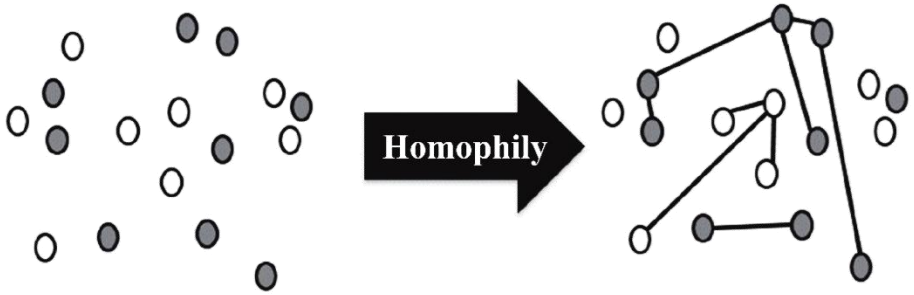
Both influence and Homophily generate similarity in social networks

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

BBC | News | Sport | Weather | Travel | TV | Radio | More... | Search

DEVON

BBC RADIO DEVON
Listen Live Listen Again

BBC Local
Devon
Things to do
People & Places
Nature & Outdoors
History
Religion & Ethics
Arts & Culture
BBC Introducing
TV & Radio

Local BBC Sites
News
Sport
Weather
Travel

Neighbouring Sites
Cornwall
Dorset
Somerset

Related BBC Sites
England

Page last updated at 14:58 GMT, Monday, 14 June 2010 15:58 UK

E-mail this to a friend Printable version

Patches for Plymouth's young smokers

By Jo Irving
BBC Devon website

More than 60% of Plymouth's under-16s smoke

MORE FROM DEVON
NEWS
SPORT
WEATHER
TRAVEL

ELSEWHERE ON THE WEB
Plymouth NHS Trust Stop Smoking Service

Why?

- Smoker friends influence their non-smoker friends

Influence

- Smokers become friends
  - Can this explain smoking behavior?

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: whites
  - **Grey**: blacks
  - **Light Grey**: hispanics
  - **Black**: others
- 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 **hispanic**.
- 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)

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

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

This is modularity

Normalized Modularity [Finding the Maximum]

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

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

Modularity: Matrix Form

- 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

$$B = A - dd^T/2m$$

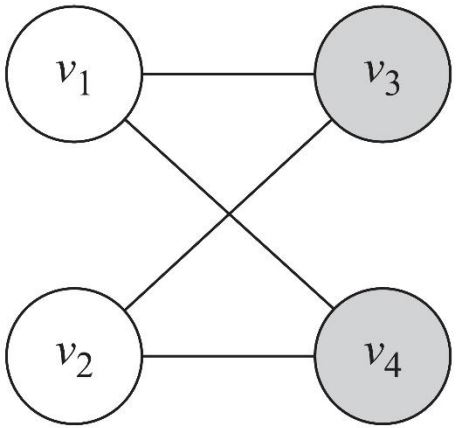
$\boldsymbol{d} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{1}}$ is the degree vector

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



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

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

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

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

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

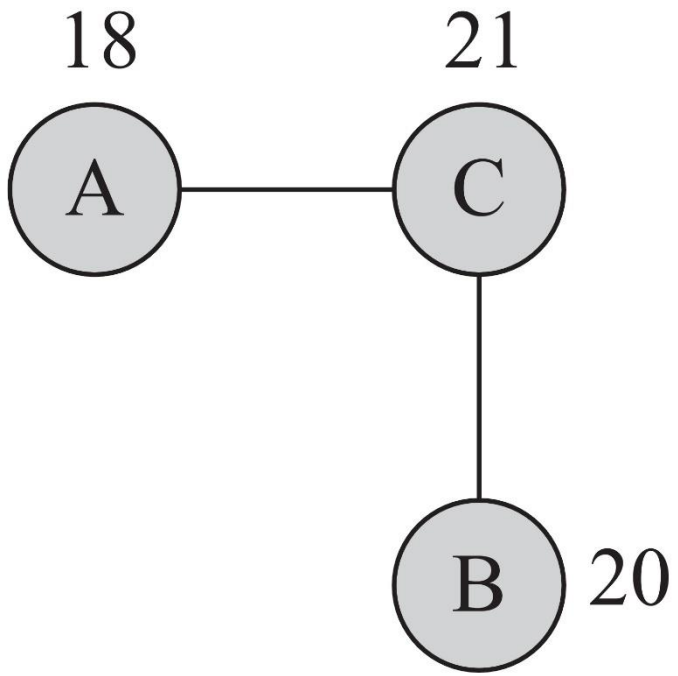
Covariance Variables

- 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

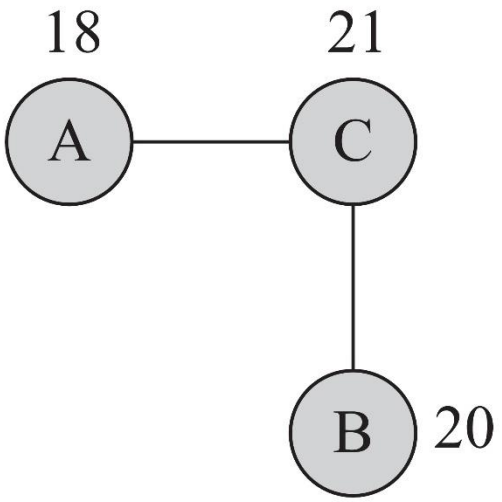
Pearson correlation $\rho(X,Y)$ is the normalized version of covariance

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

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

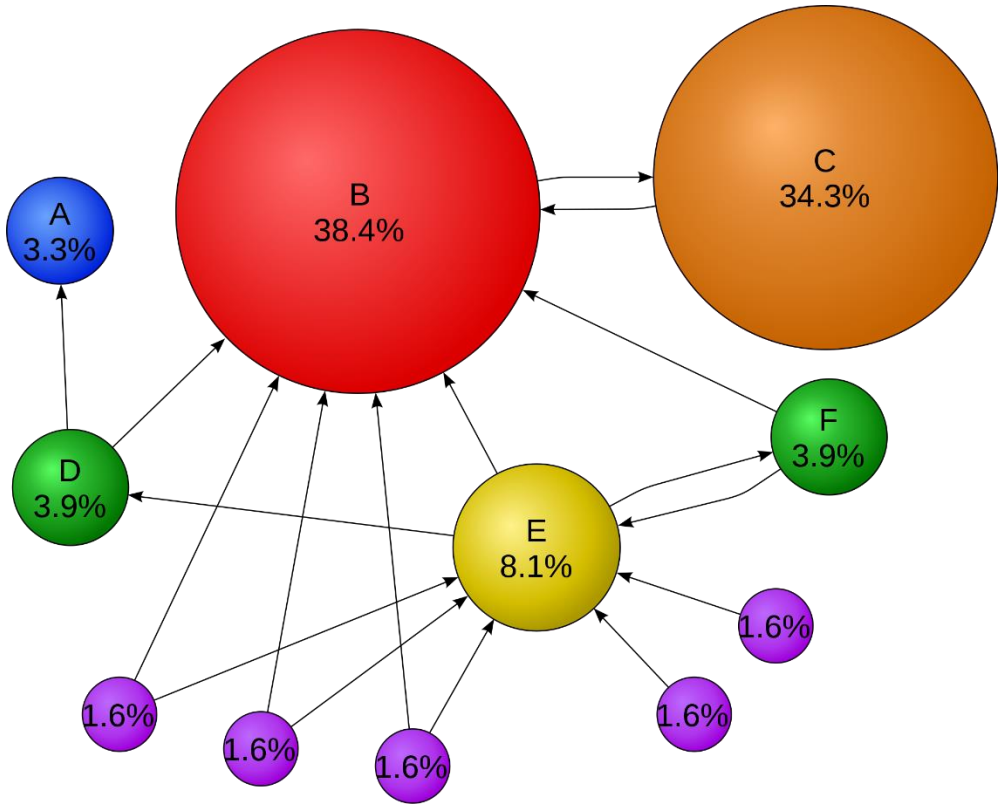
Influence

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

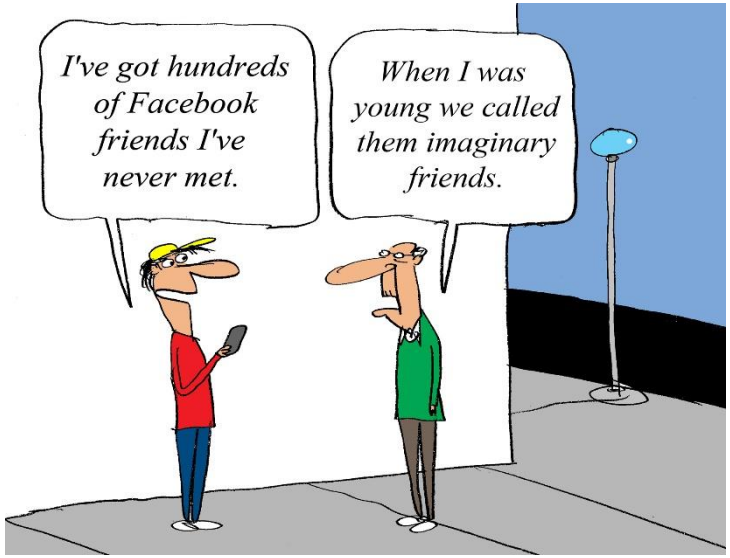
Prediction-based Measurement

We assume that

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

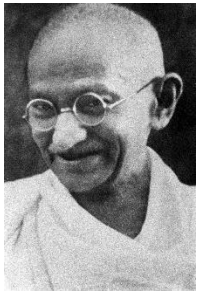
predicts how influential the user **will** be

- 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 (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 commonly used



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

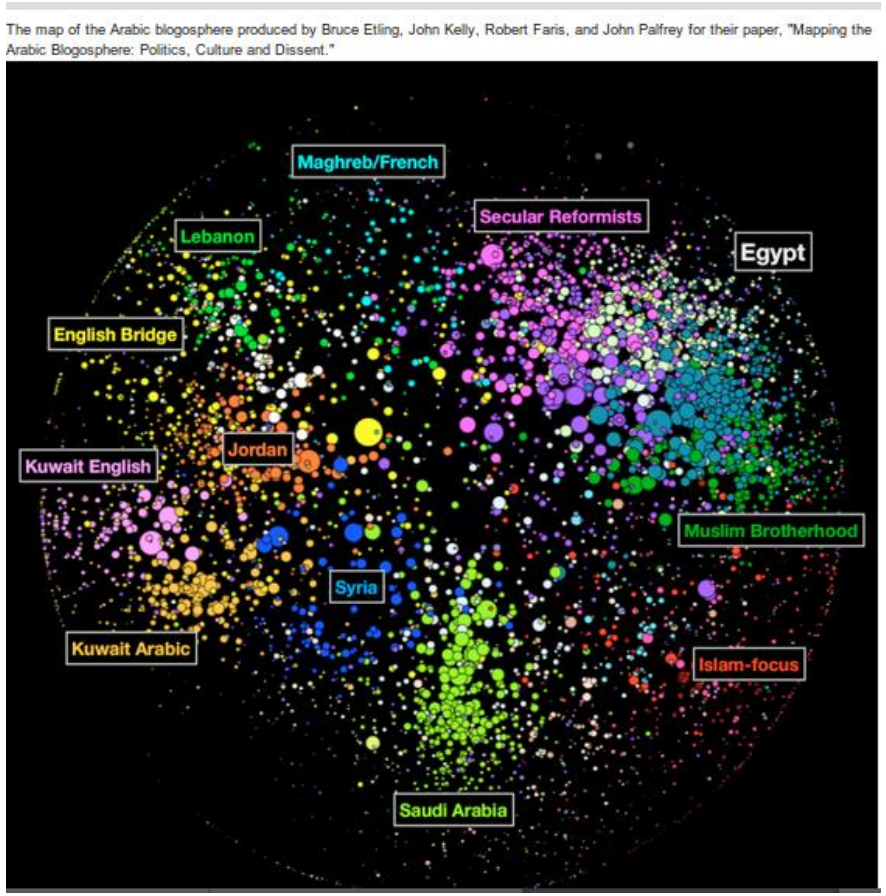
- Influence measure: the increase (or rate of increase) in the value of an item or action
  - The second person who bought the fax machine increased its value dramatically

Case Studies for Measuring Influence in Social Media

- **Measuring Influence on Blogosphere**
- **Measuring Influence on Twitter**

Measuring Social Influence on Blogosphere

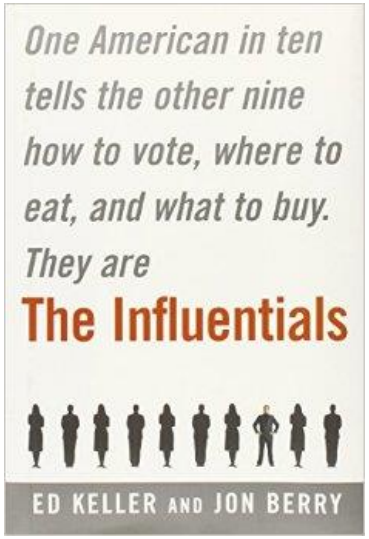
- **Goal:** figure out most influential bloggers on the blogosphere
- **Why?** We have limited time
  - Following the influentials is often a good heuristic of filtering what's uninteresting
- Common measure for quantifying influence of bloggers is to use in-degree centrality
- In-links are sparse
  - More detailed analysis is required to measure influence

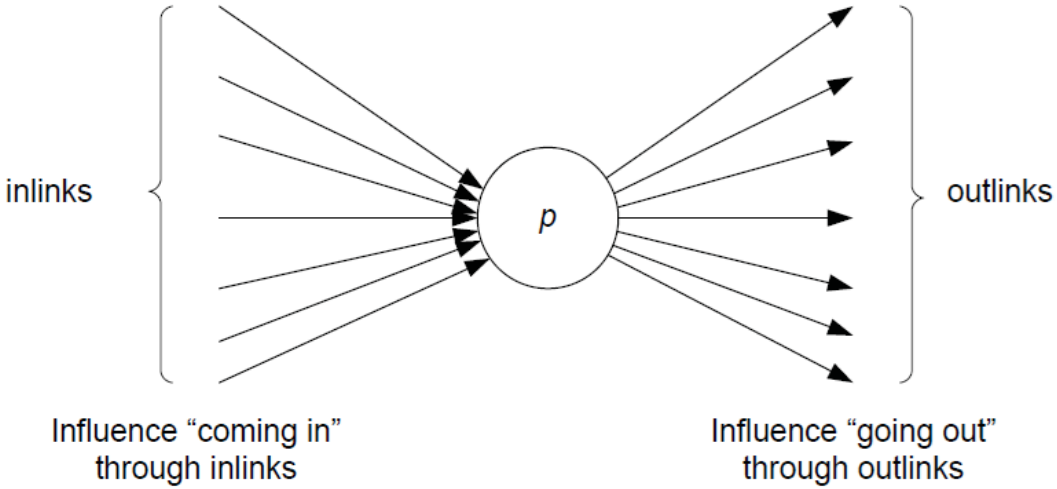


iFinder: Characterizing Influence in Blogs

Keller and Berry argue that the **influentials** are

1. Recognized by others [***Recognition***]
2. Their activities result in follow-up activities [***Activity Generation***]
3. Have novel perspectives [***Novelty***]
4. Are eloquent [***Eloquence***]





We can model each one of these properties using a graph

- p is a blogpost referred to by other links

Social Gestures [Features for a Blogpost]

Recognition

- Feature: 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

- Feature: the number of comments that p receives.
- c_p denotes the number of comments that blogpost p receives.

Novelty

- Feature: inversely correlated with the number of references a blogpost employs. i.e., the more citations a blogpost has it is considered less novel.
- O_p denotes the set of out-links for blogpost p.

Eloquence

- Feature: estimated by the length of the blogpost.
- Bloggers tend to write short blogposts. Longer blogposts are believed to be more eloquent.
- 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).

$$InfluenceFlow(p) = w_{\text{in}} \sum_{m=1}^{|\mathcal{I}_p|} I(P_m) - w_{\text{out}} \sum_{n=1}^{|\mathcal{O}_p|} I(P_n)$$

- $I(.)$ denotes the influence a blogpost
- p_m is the number of blogposts that point to blog post p
- p_n is the number of blog posts referred to in p
- w_{in} and w_{out} are the weights that adjust the contribution of in- and out-links, respectively

Blogpost Influence

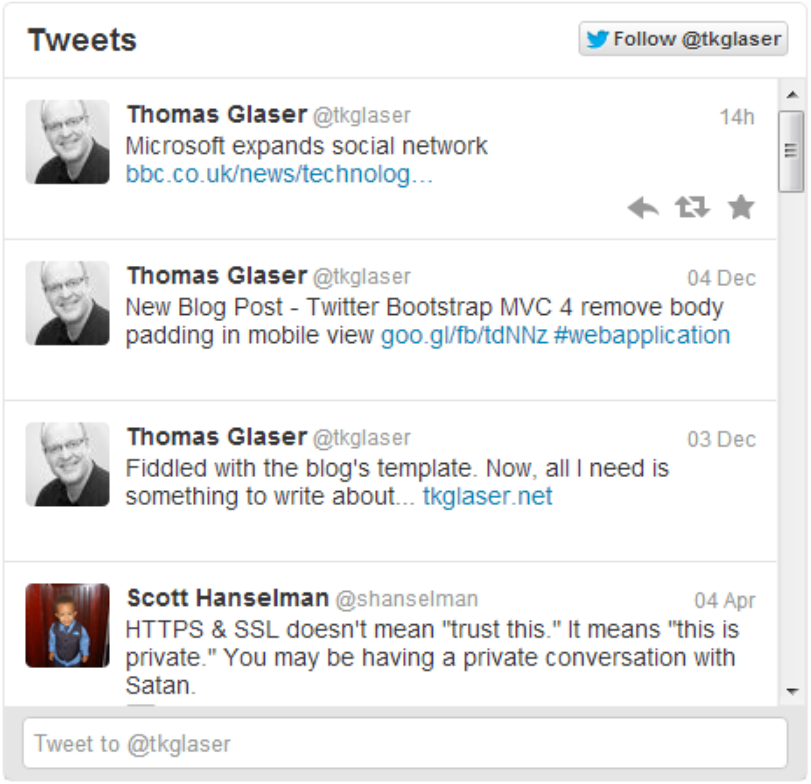
$$I(p) = w_{\text{length}} l_p (w_{\text{comment}} c_p + InfluenceFlow(p))$$

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

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 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]

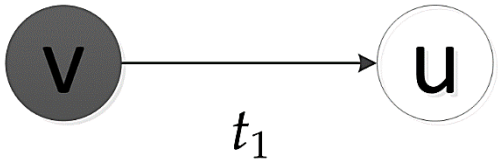
- Popular users (users with high in-degree) do not necessarily have high ranks in terms of number of retweets or mentions.

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

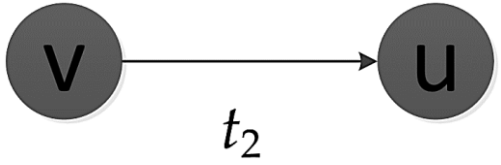
Influence Modeling

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

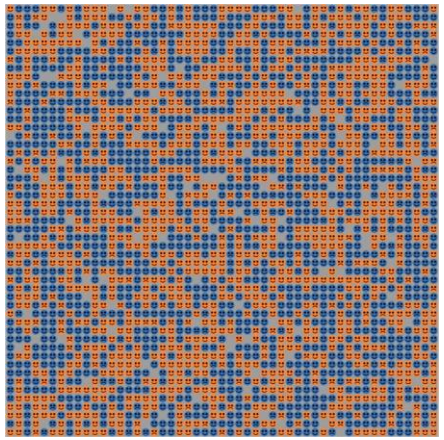
Influence Modeling: Assumptions

- 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



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

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

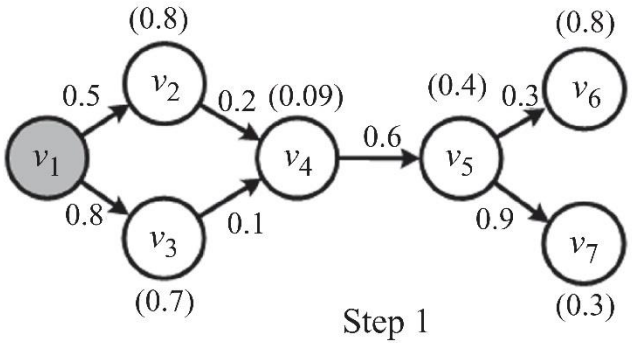
LTM Algorithm

Algorithm 1 Linear Threshold Model (LTM)

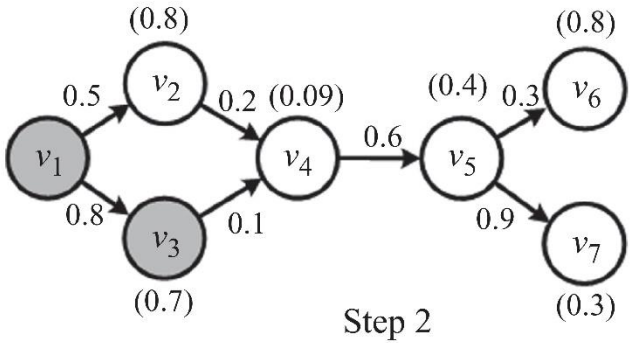
Require: Graph $G(V, E)$, set of initial activated nodes A_0

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_∞;

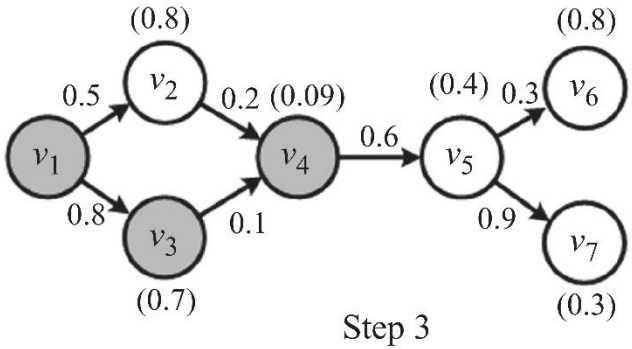
Linear Threshold Model (LTM) - An Example



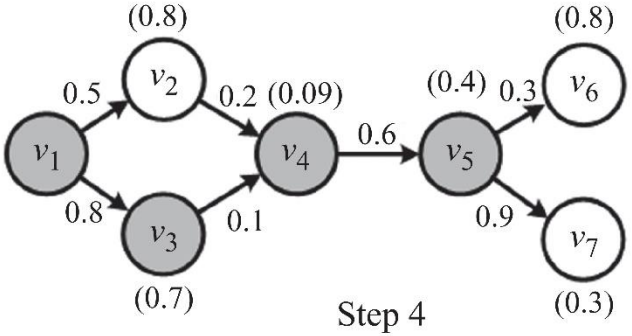
Step 1



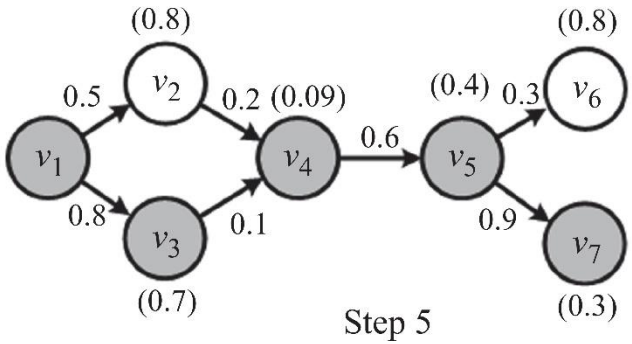
Step 2



Step 3



Step 4



Step 5

Thresholds are on top of nodes

Influence in Implicit Networks

- An implicit network is one where the influence spreads over nodes in the network
- Unlike the threshold model, we cannot observe users who are responsible for influencing others (the influentials), but only those who get influenced
- The information available:
  - 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 user u can influence $I(u,t)$ non-influenced users after t steps

- Assuming discrete time, we can formulate the size of influence population as

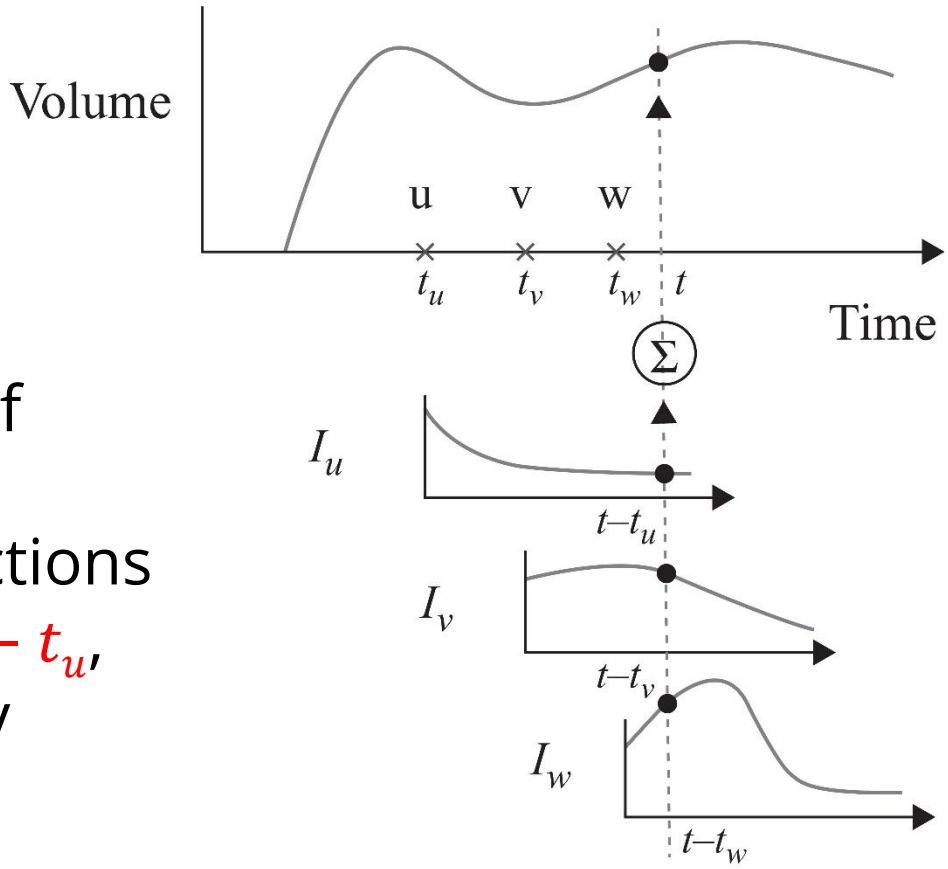
$$|P(t)| = \sum_{u \in P(t)} I(u, t - t_u)$$

GOAL: estimate $I(.,.)$ given activation time (t_u) and the number of influenced users at any time ($|P(t)|$)

The Size of the Influenced Population

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

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 is the summation of number of users influenced by activated individuals

Estimating Influence Function

Estimating $I(.,.)$

- **Parametric estimation**
  - Use some distribution to estimate I function.
  - Assume **all users** influence others in the same parametric form
    - For instance, one can use the power-law distribution to estimate influence:

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

    - Here we need to estimate the coefficients
- **Non-Parametric estimation**

Non-Parametric Estimation

Assume that nodes can get deactivated over time and can no longer influence others.

- $A(u,t) = 1$ denotes that u is active at time t
- $A(u,t) = 0$ denotes that u is either deactivated or still not influenced,
- $|V|$ is the population size and T is the last time step

$$|P(t)| = \sum_{u=1}^{|V|} \sum_{t=1}^{T} A(u,t) I(u,t),$$

$$P = AI$$

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

Can be solved using non-negative least-square methods.

lsqnonneg in MATLAB

Homophily

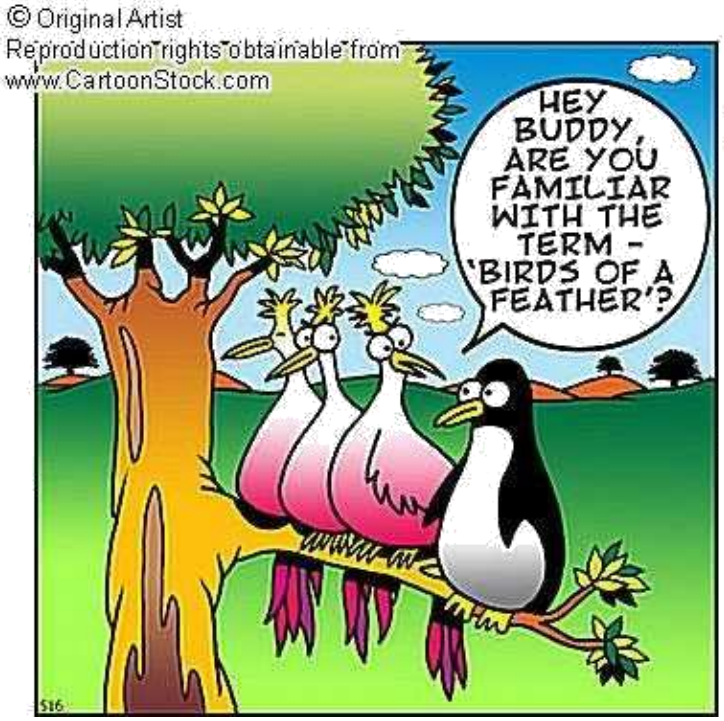
“Birds of a feather flock together”

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 activated.

Homophily Model

Algorithm 1 Homophily Model

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

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

Distinguishing Influence and Homophily

- **Shuffle Test**
- **Edge-Reversal Test**
- **Randomization Test**

Distinguishing Influence and Homophily

- 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
- 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

I. Shuffle Test (Influence)

IDEA:

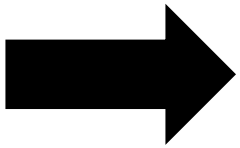
- Influence is temporal.
- When u influences v, then u should have been activated before v.
- Define a temporal assortativity measure.
- If there is no influence, then a shuffling of the activation timestamps should not affect the temporal assortativity measurement.

Shuffle Test

If influence does not play a role, the timing of activations should be independent of users.

Even if we randomly shuffle the timestamps of user activities, we should obtain a similar temporal assortativity value

User	A	B	C
Time	1	2	3



User	A	B	C
Time	2	3	1

Test of Influence

After we shuffle the timestamps of user activities, if the new estimate of temporal assortativity is significantly different from the original estimate based on the user's activity log,

there is evidence of influence.

Measuring Temporal Assortativity

- Assume node activation probability depends on a, the number of already-active friends of the node.
  - Denote the probability as $\mathrm{p}(a)$

- Assume $p(a)$ can be estimated using a logistic function

$$p(a) = \frac{e^{\alpha a+\beta}}{1+e^{\alpha a+\beta}} \quad \Rightarrow \quad \ln \frac{p(a)}{1-p(a)} = \alpha a + \beta$$

- a is the number of active friends,
- α is the temporal assortativity (**social correlation**) : **variable**
- β is a constant to explain the innate bias for activation : **variable**

Activation Likelihood

Suppose at time t

- $y_{a,t}$ users with a active friends become active
- $n_{a,t}$ users with a active friends, stay inactive
- Number of users with a friends activated/not-activated at any time

$$y_a = \sum_t y_{a,t} \qquad n_a = \sum_t n_{a,t}$$

The probability of observing your data (**likelihood function**) is

$$\prod_a p(a)^{y_a} (1 - p(a))^{n_a}$$

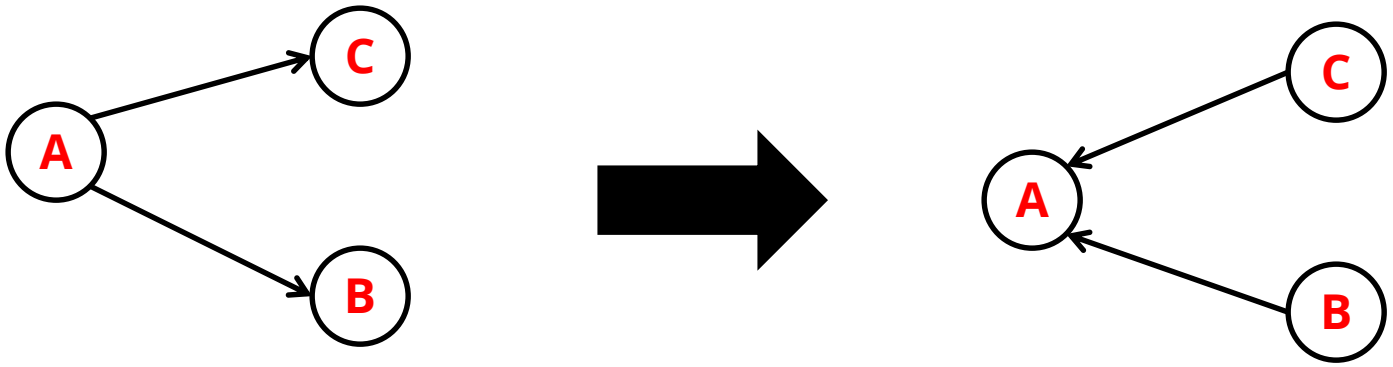
Given the user's activity log, we can compute a correlation coefficient α and bias β to maximize the above likelihood

- Using a maximum likelihood iterative method

2. The Edge-reversal Test (Influence)

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

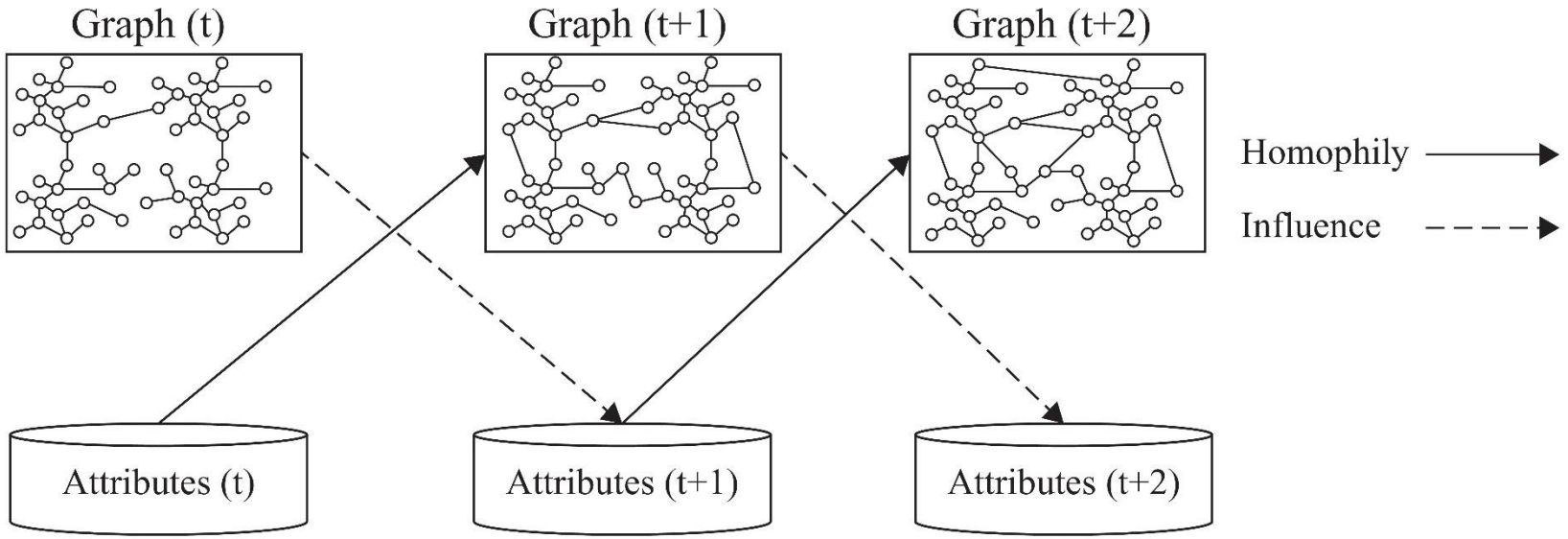
- Reverse directions of all the edges
- Run the same logistic regression on the data using the new graph
- If correlation is not due to influence, then α should not change



3. Randomization Test (Influence/Homophily)

- Capable of detecting both Influence and Homophily in networks

- Influence changes attributes and Homophily changes connections



Notation and Preliminaries

- X denotes node attributes
  - X^i denotes the attributes of node v_i
  - X_t denotes the attributes of nodes at time t

- $A(G_t, X_t)$ denotes the assortativity of network G and attributes X at time t

- The network becomes more assortative at time t if

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

Influence Gain and Homophily Gain

- If the assortativity is due to influence, **Influence gain** is positive

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

- If the assortativity is due to homophily, **Homophily gain** is positive

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

- In randomization test, we check if these gains are **significant**

Influence Significance Test

- Compute influence gain at time t
  - Denote as g_0
- Compute n random attributes sets for time t+1
  - Denote as XR_{t+1}^i , $1 \leq i \leq n$
  - **Example**.
    - u has influence over v.
    - movies is in hobbies of u at time t, but not in hobbies of v at time t.
    - At time $t+1$ movies is added to hobbies of v.
    - To remove influence effect, we can remove movies from hobbies of v at time $t+1$ and replace it with some random hobby (e.g., reading)
- Compute the [random] influence gain for all XR_{t+1}^i sets
  - Call them g_i
- If g_0 is greater than $\left(1-\frac{\alpha}{2}\right)\%$ of all g_i's (or smaller than $\left(\frac{\alpha}{2}\right)\%$ of them)
  - The influence gain is **significant**

Influence Significance Test

Algorithm 1 Influence Significance Test

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

return Significance
$g_0 = G_{Influence}(t)$;
for all $1 \leq i \leq n$ **do**
    $XR^i_{t+1} = randomize_I(X_t, X_{t+1})$;
    $g_i = A(G_t, XR^i_{t+1}) - A(G_t, X_t)$;
end for
if g_0 larger than $(1 - \alpha/2)\%$ of values in $\{g_i\}_{i=1}^n$ **then**
    return significant;
else if g_0 smaller than $\alpha/2\%$ of values in $\{g_i\}_{i=1}^n$ **then**
    return significant;
else
    return insignificant;
end if

Homophily Significance Test

- We construct random graphs, with fixed attribute sets

- We remove the effect of homophily by generating n random graphs GR_{t+1}^{i} at time $t+1$
  - For any two (randomly selected) edges e_{ij} and e_{kl} formed in the original graph G_{t+1}
    - We form edges e_{il} and e_{kj}
    - Homophily effect removed / degrees stay the same

Homophily Significance Test

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