Nodes, Ties and Influence

Chapter 2

IMPORTANCE OF NODES

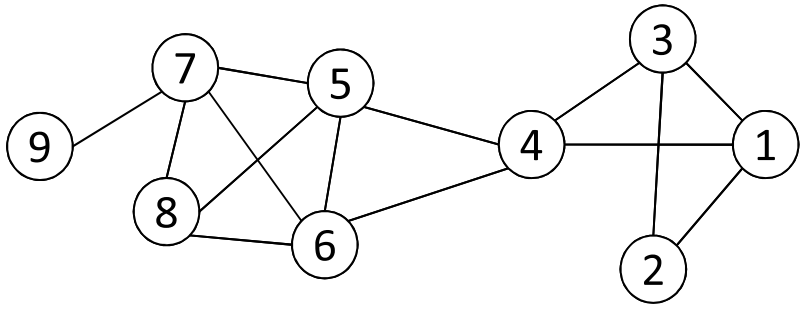
Importance of Nodes

- Not all nodes are equally important
- Centrality Analysis:
  - Find out the most important nodes in one network

- Commonly-used Measures
  - Degree Centrality
  - Closeness Centrality
  - Betweenness Centrality
  - Eigenvector Centrality

Degree Centrality

- The importance of a node is determined by the number of nodes adjacent to it
  - The larger the degree, the more import the node is
  - Only a small number of nodes have high degrees in many real-life networks
- Degree Centrality $C_D(v_i) = d_i = \sum_j A_{ij}$
- Normalized Degree Centrality: $C'_D(v_i) = d_i/(n-1)$



For node 1, degree centrality is 3; Normalized degree centrality is 3/(9-1)=3/8.

Closeness Centrality

- "Central" nodes are important, as they can reach the whole network more quickly than non-central nodes
- Importance measured by how close a node is to other nodes

- Average Distance: $$D_{avg}(v_i) = \frac{1}{n-1}\sum_{j \neq i}^{n} g(v_i, v_j)$$

- Closeness Centrality

$$C_C(v_i) = \left[\frac{1}{n-1}\sum_{j \neq i}^{n} g(v_i, v_j)\right]^{-1} = \frac{n-1}{\sum_{j \neq i}^{n} g(v_i, v_j)}$$

Closeness Centrality Example

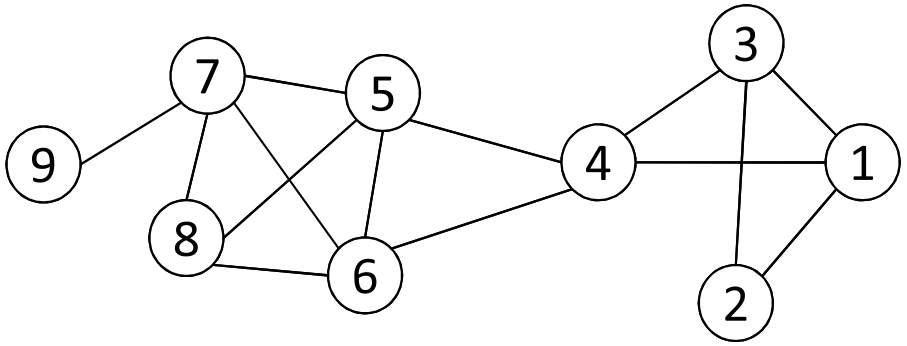


Table 2.1: Pairwise geodesic distance

Node	1	2	3	4	5	6	7	8	9
1	0	1	1	1	2	2	3	3	4
2	1	0	1	2	3	3	4	4	5
3	1	1	0	1	2	2	3	3	4
4	1	2	1	0	1	1	2	2	3
5	2	3	2	1	0	1	1	1	2
6	2	3	2	1	1	0	1	1	2
7	3	4	3	2	1	1	0	1	1
8	3	4	3	2	1	1	1	0	2
9	4	5	4	3	2	2	1	2	0

$$C_C(3) = \frac{9-1}{1+1+1+2+2+3+3+4} = 8/17 = 0.47,$$

$$C_C(4) = \frac{9-1}{1+2+1+1+1+2+2+3} = 8/13 = 0.62.$$

Node 4 is more central than node 3

Betweenness Centrality

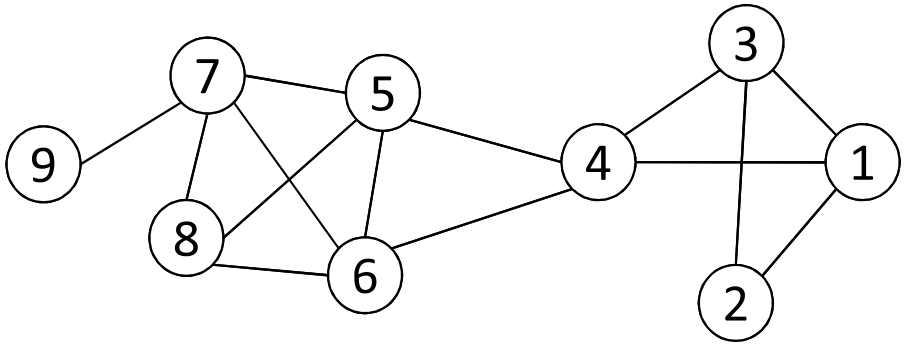
- Node betweenness counts the number of shortest paths that pass one node
- Nodes with high betweenness are important in communication and information diffusion

- Betweenness Centrality $$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

σ_{st} : The number of shortest paths between s and t

$\sigma_{st}(v_i)$: The number of shortest paths between s and t that pass v_i

Betweenness Centrality Example



$$C_B(4) = 15$$

Table 2.2: $\sigma_{st}(4)/\sigma_{st}$

	$s=1$	$s=2$	$s=3$
$t=5$	1/1	2/2	1/1
$t=6$	1/1	2/2	1/1
$t=7$	2/2	4/4	2/2
$t=8$	2/2	4/4	2/2
$t=9$	2/2	4/4	2/2

What's the betweenness centrality for node 5?

σ_{st} : The number of shortest paths between s and t

$\sigma_{st}(v_i)$: The number of shortest paths between s and t that pass v_i

$$C_B(v_i) = \sum_{v_s \neq v_i \neq v_t \in V, s<t} \frac{\sigma_{st}(v_i)}{\sigma_{st}}$$

Eigenvector Centrality

- One's importance is determined by his friends'
- If one has many important friends, he should be important as well.

$$C_E(v_i) \propto \sum_{v_j \in N_i} A_{ij} C_E(v_j)$$

$$\mathbf{x} \propto A\mathbf{x} \quad \Rightarrow \quad A\mathbf{x} = \lambda \mathbf{x}.$$

- The centrality corresponds to the top eigenvector of the adjacency matrix A.
- A variant of this eigenvector centrality is the PageRank score.

STRENGTHS OF TIES

Weak and Strong Ties

- In practice, connections are not of the same strength
- Interpersonal social networks are composed of strong ties (close friends) and weak ties (acquaintances).
- Strong ties and weak ties play different roles for community formation and information diffusion
- Strength of Weak Ties (*Granovetter, 1973*)
  - Occasional encounters with distant acquaintances can provide important information about new opportunities for job search

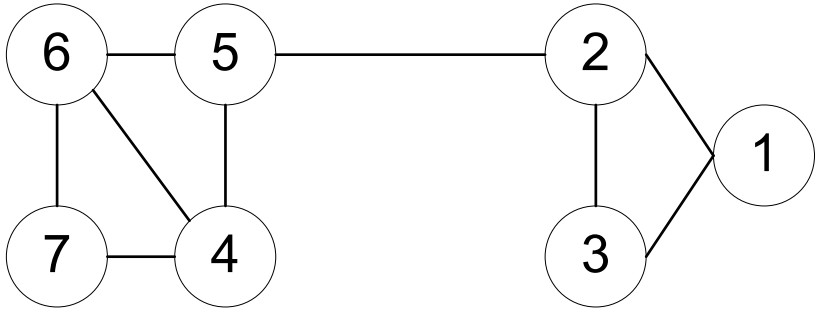
Connections in Social Media

- Social Media allows users to connect to each other more easily than ever
  - One user might have thousands of friends online
  - Who are the most important ones among your 300 Facebook friends?

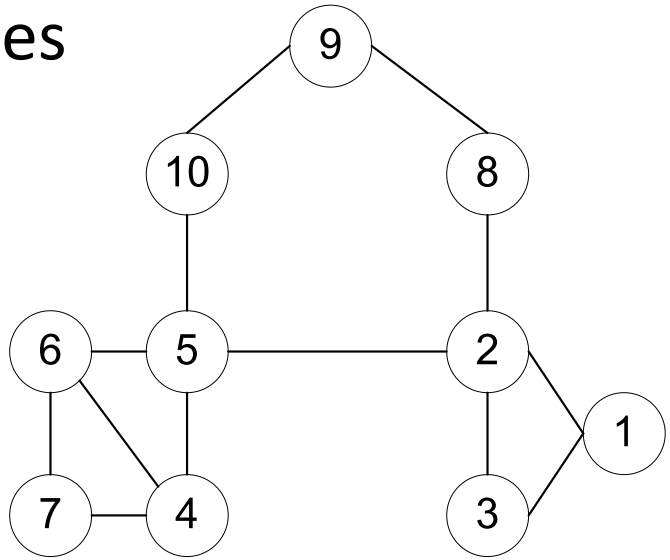
- Imperative to estimate the strengths of ties for advanced analysis
  - Analyze network topology
  - Learn from User Profiles and Attributes
  - Learn from User Activities

Learning from Network Topology

- Bridges connecting two different communities are weak ties
- An edge is a *bridge* if its removal results in disconnection of its terminal nodes



e(2,5) is a bridge

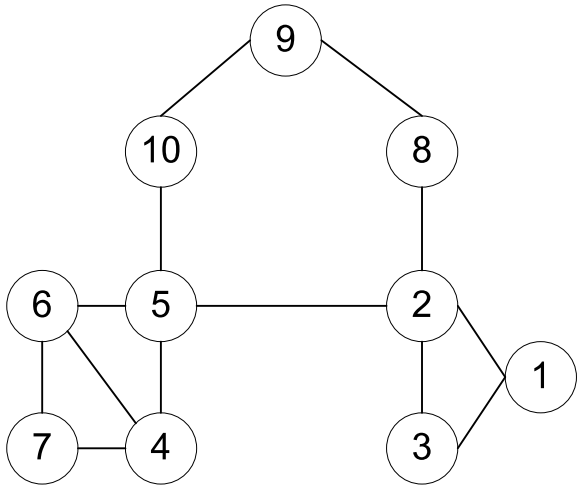


e(2,5) is NOT a bridge

“shortcut” Bridge

- Bridges are rare in real-life networks
- Alternatively, one can relax the definition by checking if the distance between two terminal nodes increases if the edge is removed
- The larger the distance, the weaker the tie is

- d(2,5) = 4 if e(2,5) is removed
- d(5,6) = 2 if e(5,6) is removed
- e(5,6) is a stronger tie than e(2,5)



Neighborhood Overlap

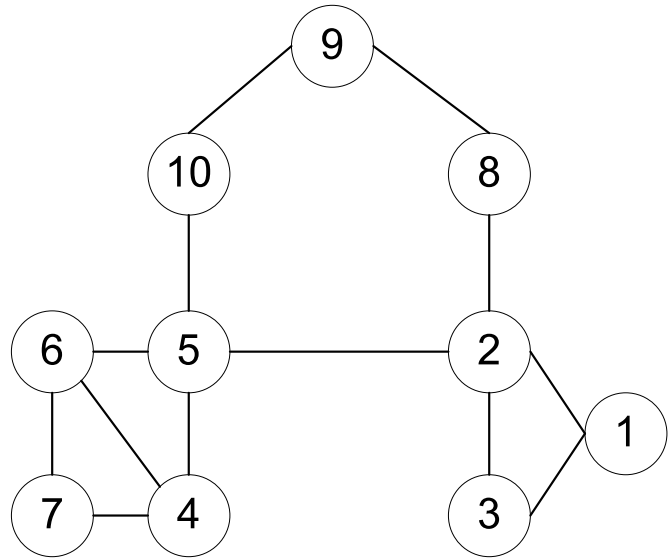
- Tie Strength can be measured based on neighborhood overlap; the larger the overlap, the stronger the tie is.

$$overlap(v_i, v_j) = \frac{\text{number of shared friends of both } v_i \text{ and } v_j}{\text{number of friends who are adjacent to at least } v_i \text{ or } v_j} = \frac{|N_i \cap N_j|}{|N_i \cup N_j| - 2}.$$

- -2 in the denominator is to exclude v_i and v_j

$$overlap(2, 5) = 0,$$

$$overlap(5, 6) = \frac{|\{4\}|}{|\{2, 4, 5, 6, 7, 10\}| - 2} = 1/4$$



Learning from Profiles and Interactions

- Twitter: one can follow others without followee's confirmation
  - The real friendship network is determined by the frequency two users talk to each other, rather than the follower-followee network
  - The real friendship network is more influential in driving Twitter usage
- Strengths of ties can be predicted accurately based on various information from Facebook
  - Friend-initiated posts, message exchanged in wall post, number of mutual friends, etc.
- Learning numeric link strength by maximum likelihood estimation
  - User profile similarity determines the strength
  - Link strength in turn determines user interaction
  - Maximize the likelihood based on observed profiles and interactions

Learning from User Activities

- One might learn how one influences his friends if the user activity log is accessible
- Depending on the adopted influence model
  - Independent cascading model
  - Linear threshold model
- Maximizing the likelihood of user activity given an influence model

INFLUENCE MODELING

Influence modeling

Influence modeling is one of the fundamental questions in order to understand the information diffusion, spread of new ideas, and word-of-mouth (viral) marketing

Well known Influence modeling methods

1. Linear threshold model (LTM)
2. Independent cascade model (ICM)

Common properties of Influence modeling methods

- A social network is represented a *directed graph, with each actor being one node;*
- Each node is started as active or inactive;
- A node, once activated, will activate his neighboring nodes;
- Once a node is activated, this node cannot be deactivated.

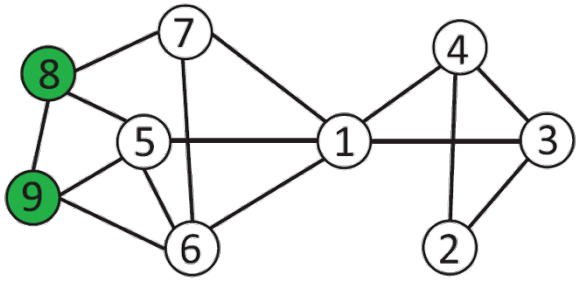
Linear Threshold Model

An actor would take an action if the number of his friends who have taken the action exceeds (reaches) a certain threshold

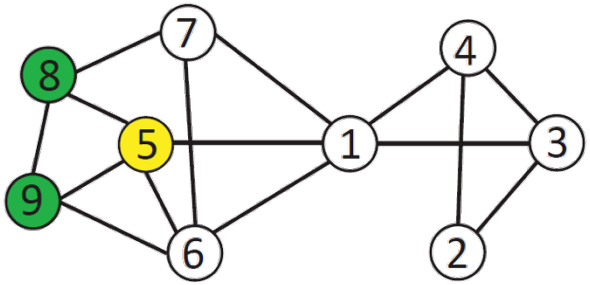
- Each node *v chooses a threshold* Θ_v *randomly from a uniform* distribution in an interval between 0 and 1.
- In each discrete step, all nodes that were active in the previous step remain active
- The nodes satisfying the following condition will be activated

$$\sum_{w \in N_v, w \text{ is active}} b_{w,v} \geq \theta_v$$

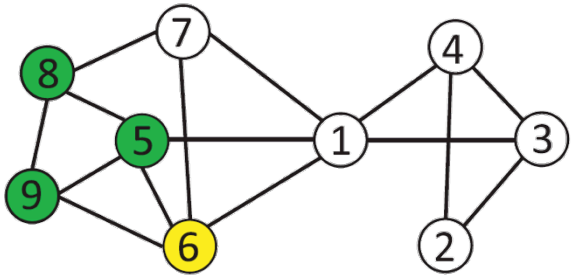
Linear Threshold Model- Diffusion Process (Threshold = 50%)



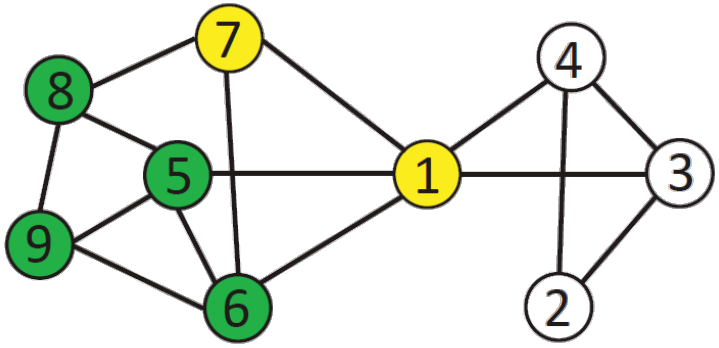
Step 0



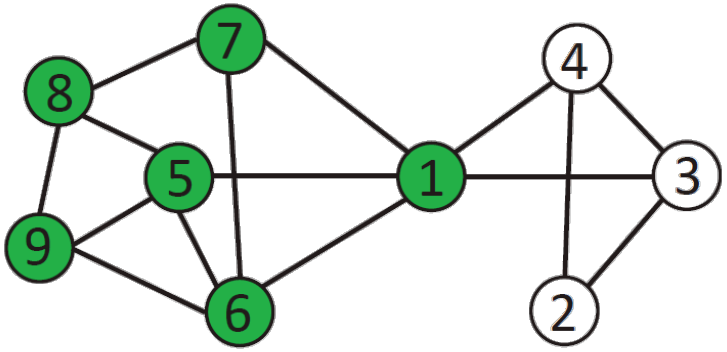
Step 1



Step 2



Step 3



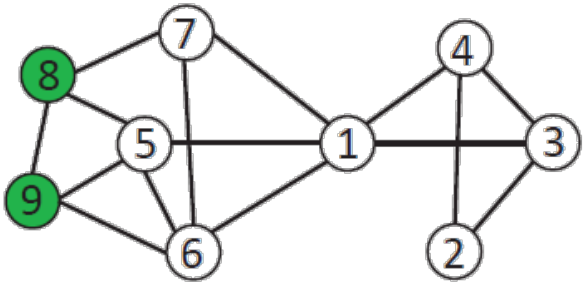
Final Stage

Independent Cascade Model (ICM)

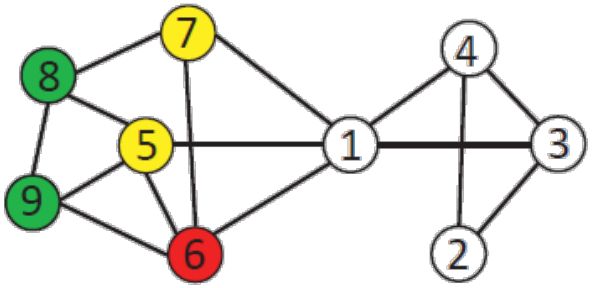
The independent cascade model focuses on the sender's rather than the receiver's view

- A node *w, once activated* at step *t , has one chance to activate each of its neighbors randomly*
  - For a neighboring node (say, *v), the activation* succeeds with probability $p_{w,v}$ (e.g. p = 0.5*)*
- If the activation succeeds, then *v will become active at step t + 1*
- In the subsequent rounds, *w will not attempt to activate v anymore.*
- *The diffusion process,* starts with an initial activated set of nodes, then continues until no further activation is possible

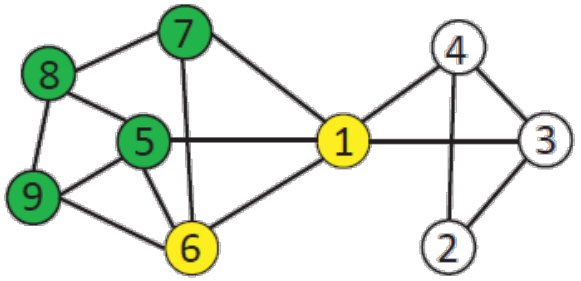
Independent Cascade Model- Diffusion Process



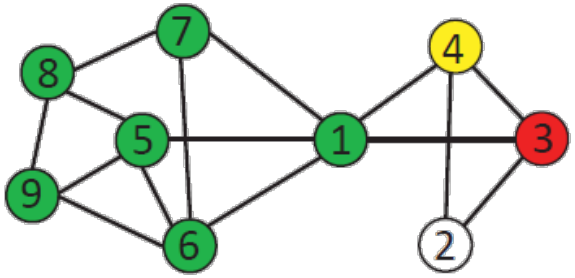
Step 0



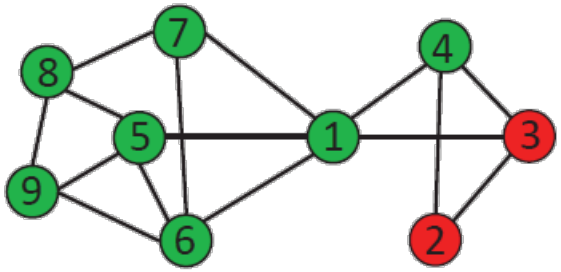
Step 1



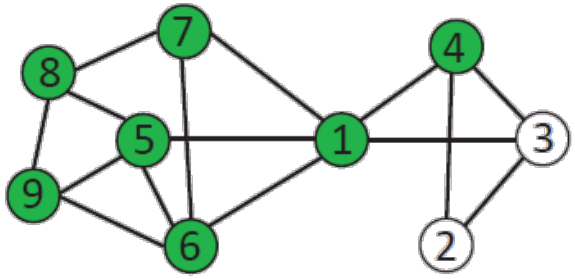
Step 2



Step 3



Step 4



Final Stage

Influence Maximization

Given a network and a parameter *k, which k nodes should be selected to be in the* activation set *B in order to maximize the influence in terms of active nodes at the* end?

- Let *σ(B) denote the expected number of nodes that can be influenced by B, the* optimization problem can be formulated as follows:

$$\max_{B \subseteq V} \sigma(B) \; s.t. \; |B| \leq k$$

Influence Maximization- A greedy approach

Maximizing the influence, is a NP-hard problem but it is proved that the greedy approaches gives a solution that is 63 % of the optimal.

A greedy approach:

- Start with $B = \emptyset$
- *Evaluate σ(v) for each node, and pick the node with maximum σ as the first node v1 to form B = {v1}*
- Select a node which will increase *σ(B) most if the node is included in B.*

- *Essentially, we greedily* find a node $v \in V \setminus B$ *such that*

$$v = \arg \max_{v \in V \setminus B} \sigma(B \cup \{v\})$$

DISTINGUISH BETWEEN INFLUENCE AND CORRELATION

Correlation

It has been widely observed that user attributes and behaviors tend to correlate with their social networks

- Suppose we have a binary attribute with each node (say, whether or not being smoker)
- If the attribute is correlated with the network, we expect actors sharing the same attribute value to be positively correlated with social connections
- That is, smokers are more likely to interact with other smokers, and non-smokers with non-smokers

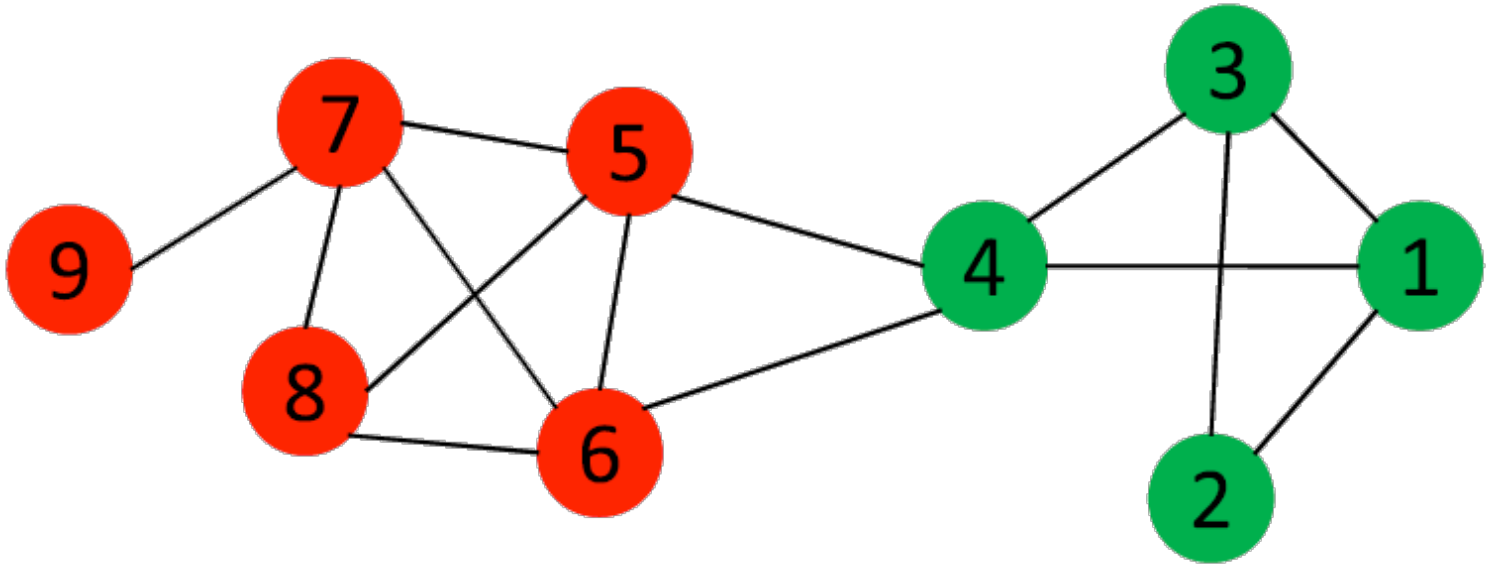
Test For Correlation

If the fraction of edges linking nodes with different attribute values are significantly less than the expected probability, then there is evidence of correlation

Example; if connections are independent of the smoking behavior:

- p fraction are smokers (1-p non-smoker)
  - one edge is expected to connect two smokers with probability $p \times p$,
  - *two non-smokers with* probability: $(1 - p) \times (1 - p)$
  - *A smoker and a non-smoker:* $2\ p\ (1\text{-}p)$

Test For Correlation- An example



Red nodes denote non-smokers, and green ones are smokers. If there is no correlation, then the probability of one edge connecting a smoker and a non-smoker is $2 \times 4/9 \times 5/9 = 49\%$.

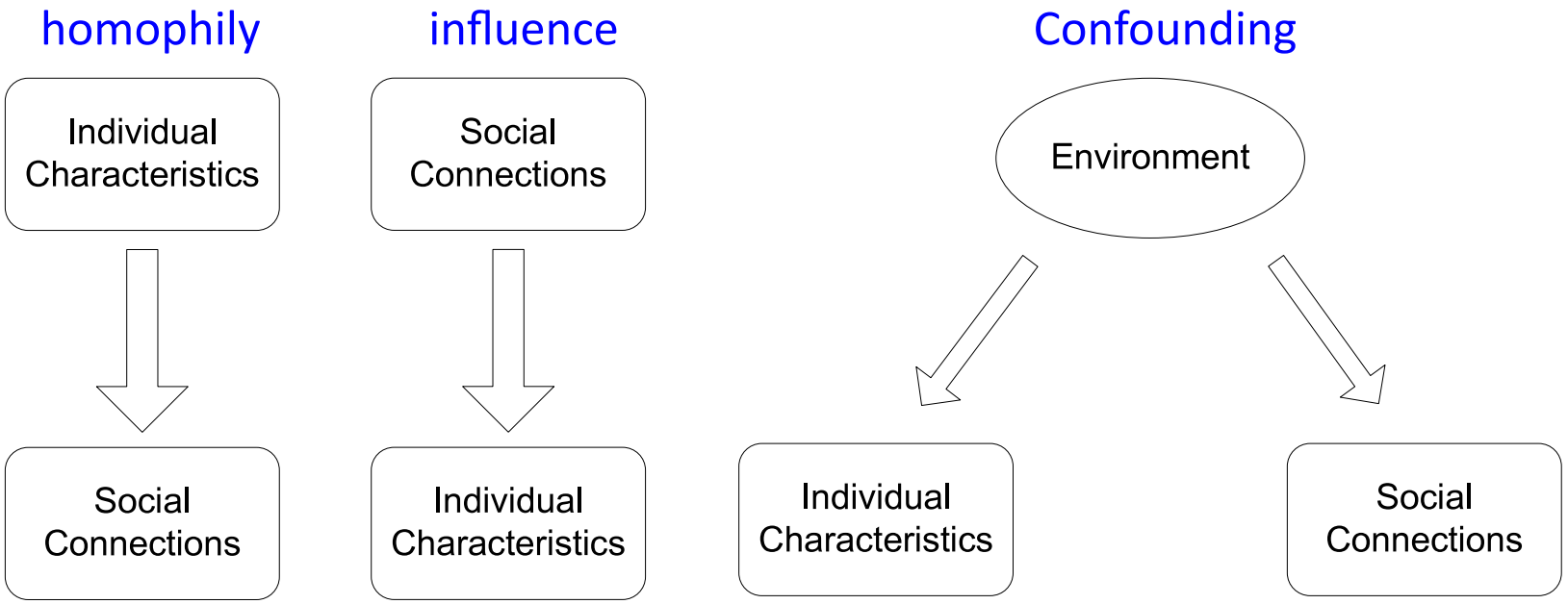
In this example the fraction is $2/14 = 14\% < 49\%$ *, so* this network demonstrates some degree of correlation with respect to the smoking behavior.

A more formal way is to conduct a $\chi 2$ *test for independence of social connections and attributes (La Fond and Neville,* 2010)

Correlation in social networks

It is well known that there exist correlations between behaviors or attributes of adjacent actors in a social network.

- Three major social processes to explain correlation are:
  - *Homophily, confounding, and influence*



Correlation in social networks

Homophily; is a term to explain our tendency to link to others that share certain similarity with us

Confounding; correlation between actors can also be forged due to external influences from environment. "two individuals living in the same city are more likely to become friends than two random individuals"

Influence; *a* process that causes behavioral correlations between adjacent actors. "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."

Influence or Correlation

In many studies about influence modeling, influence is determined by timestamps

- **Shuffle test** is an approach to identify whether influence is a factor associated with a social system
- The probability of one node being active is a logistic function of the number of his active friends as follows

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

  - *a is the number of active friends,*
  - *α the social correlation coefficient and β a constant to* explain the innate bias for activation

Activation likelihood

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 at time *t is*

$$\prod_t \prod_a p(a)^{Y_{a,t}} (1 - p(a))^{N_{a,t}}$$

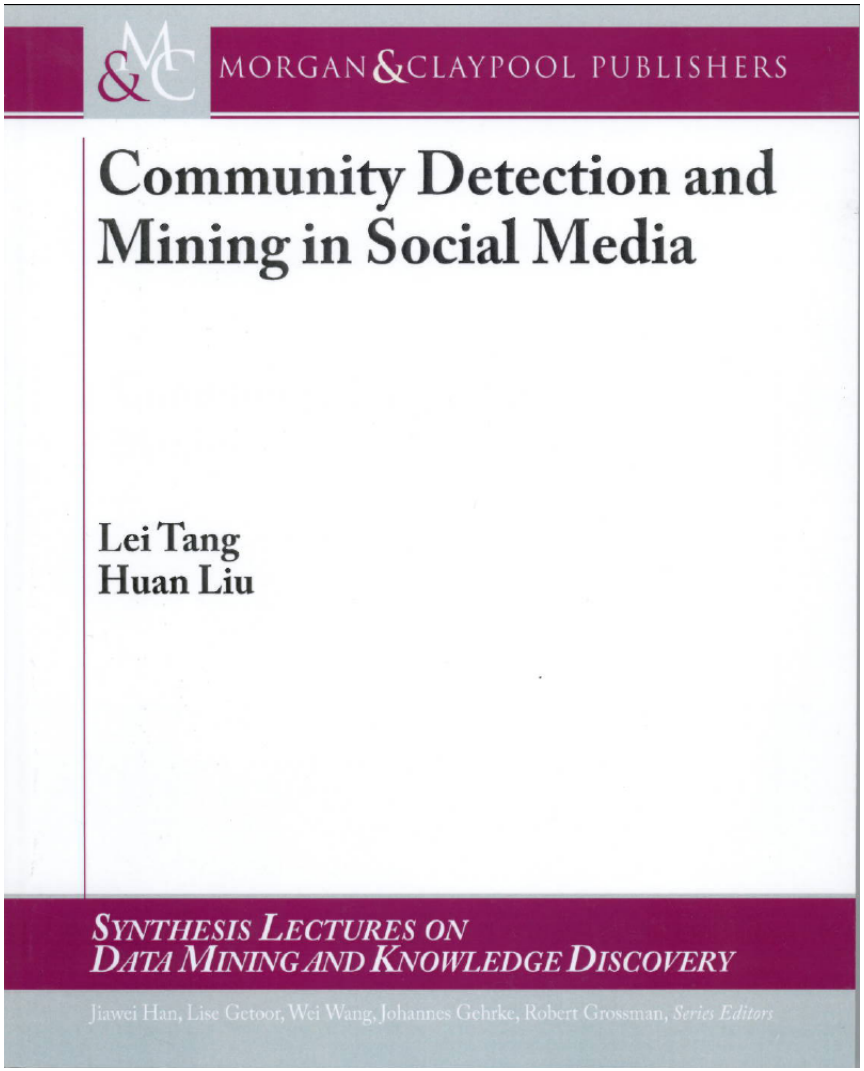
Given the user activity log, we can compute a correlation coefficient *α to maximize the above* likelihood.

Shuffle test

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

Test for 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 activity log, then **there is evidence of influence**.



Book Available at

- Morgan & claypool Publishers
- Amazon

If you have any comments, please feel free to contact:

- **Lei Tang**, Yahoo! Labs, ltang@yahoo-inc.com
- **Huan Liu**, ASU huanliu@asu.edu