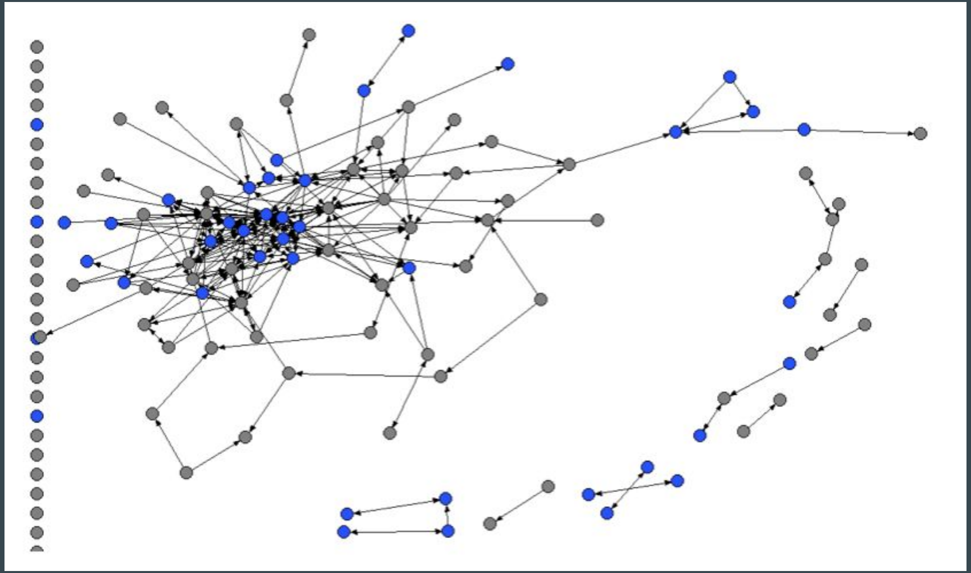
Dataism - Week 4



Social Networks and Influence Modeling

Goals for today...

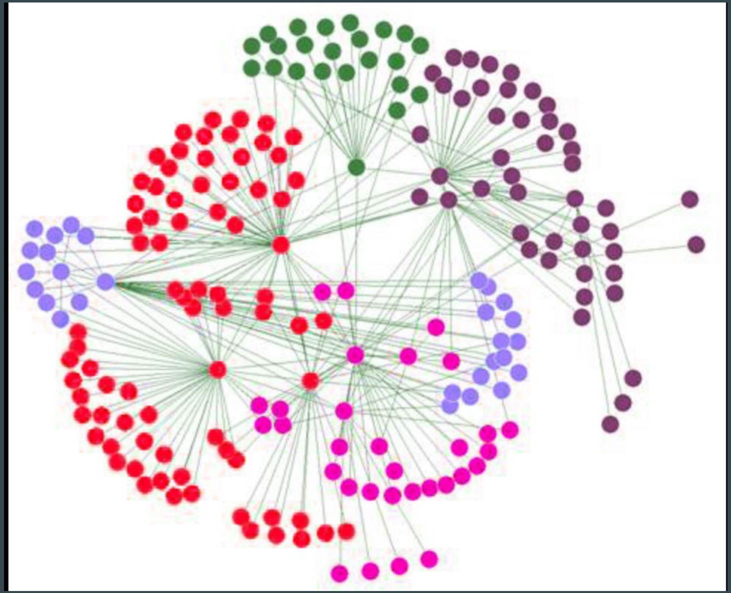
- Overview of case studies based on modeling social networks
- Examine social network data and understand how its visualized in a graph
- Learn how influence is modeled in social networks and understand the ‘influence maximization problem’
- Code and visualize the ‘linear threshold model’ with the Python networkx package

Case Studies

Today we concentrate on 'social networks:'

A set of social actors (individuals or organizations) and sets of pairwise social ties between them.

We will take a first look at 'social network analysis,' which attempts to identify local and global patterns, locate influential entities, examine change in the network over time...



Substance abuse prevention for homeless youth

Problem:

- In U.S., 4.2 million youth experience homelessness each year
- Substance abuse is very prevalent among them

SOCIAL NETWORKS AND SUBSTANCE ABUSE PREVENTION FOR HOMELESS YOUTH

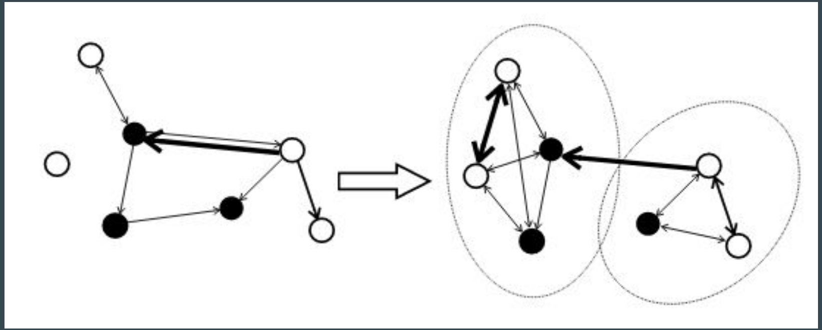
Substance abuse prevention for homeless youth

Problem:

- Researchers and community-based collaborators want to utilize peer-based prevention programs
- Low-cost, could engage hard to reach homeless youth, e.g., who are distrustful of adults
- But unclear how and with whom to implement these programs

SOCIAL NETWORKS AND SUBSTANCE ABUSE PREVENTION FOR HOMELESS YOUTH

Substance abuse prevention for homeless youth

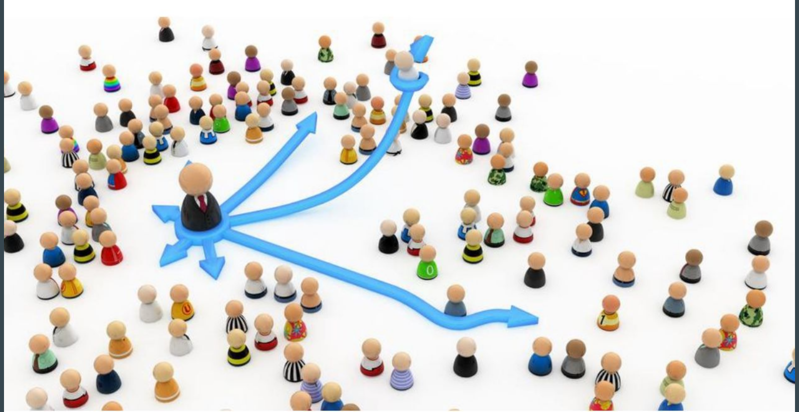


Response:

- Use A.I. to construct optimal peer-intervention groups
  - Maximize positive/minimize negative influence
  - Goal: Minimize total amount of substance abuse in the network
- Led by researchers from the USC Center for AI for Society
- In collaboration with Urban Peak (overnight shelter, drop-in center, outreach programs, education & employment training for Denver, CO area)

Other Projects

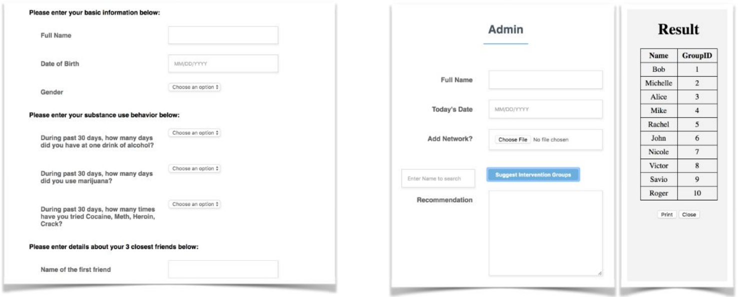
- **HIV prevention amongst homeless youth by influence maximization**
  - Spread information driving h.y. toward safer practices as efficiently as possible
- **Violence minimization amongst homeless youth**
- **Suicide prevention amongst college students**
  - ‘Gatekeeper training’ - train individuals to monitor their social network for warning signs of suicide
  - Strategically select gatekeepers for ‘maximal coverage’ of entire student network
- **Further applications:** simulating riot breakouts, viral marketing, spread of (dis)information (e.g. on twitter)



The Data

The USC researchers tested their model on Social Network Data gathered from the Youthnet Study.

- For each homeless youth: hard-drug using behavior, who they know, strength of that relationship
- The study of social networks is tied to an area of mathematics called 'graph theory'



Please enter your basic information below:

Full Name

Date of Birth

Gender

Please enter your substance use behavior below:

During past 30 days, how many days did you have at one drink of alcohol?

During past 30 days, how many days did you use marijuana?

During past 30 days, how many times have you tried Cocaine, Meth, Heroin, Crack?

Please enter details about your 3 closest friends below:

Name of the first friend

Admin

Full Name

Today's Date

Add Network?

Recommendation

Result

Name	GroupID
Bob	1
Michelle	2
Alice	3
Mike	4
Rachel	5
John	6
Nicole	7
Victor	8
Savio	9
Roger	10

(a) A snapshot of the data collection form

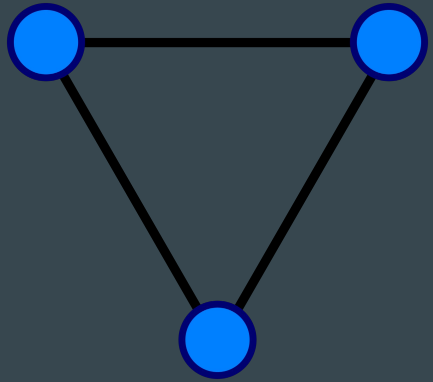
(b) A snapshot of the GUI for the group recommendation

Figure 9: Two snapshots of the GUIDE application. (a) First, detailed information about the substance-use behavior of each individual is collected and saved in a data base. (b) Second, the practitioners will query the application for a group recommendation.

Social Network Data - Graphs

Social Network Data is typically encoded in a *graph:*

- A graph G = (V,E) is a set of vertices V connected by a set of edges E.
- The vertices are also called 'nodes'
- In a social network, the nodes are people, connected by edges representing some defined relation (e.g. 'are facebook friends')

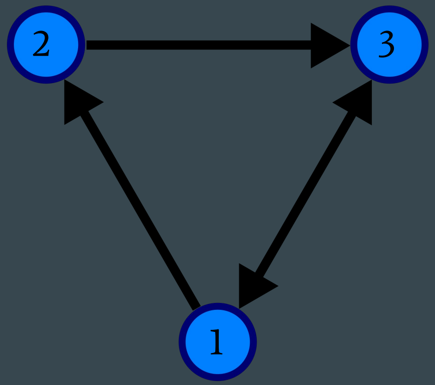


A graph G with three vertices
V = {1, 2, 3}
and three edges
E={ (1,2), (2,3), (3,1)}

Social Network Data - Directed Graphs

A graph in which the edges have a specified direction is called a *directed graph*.

- Usually written as a pair (a,b), meaning arrow **from** vertex a **to** vertex b.
- Ex. twitter is a directed graph, where (a,b) means ‘a follows b’.



A directed graph G with three vertices

V = {1, 2, 3}

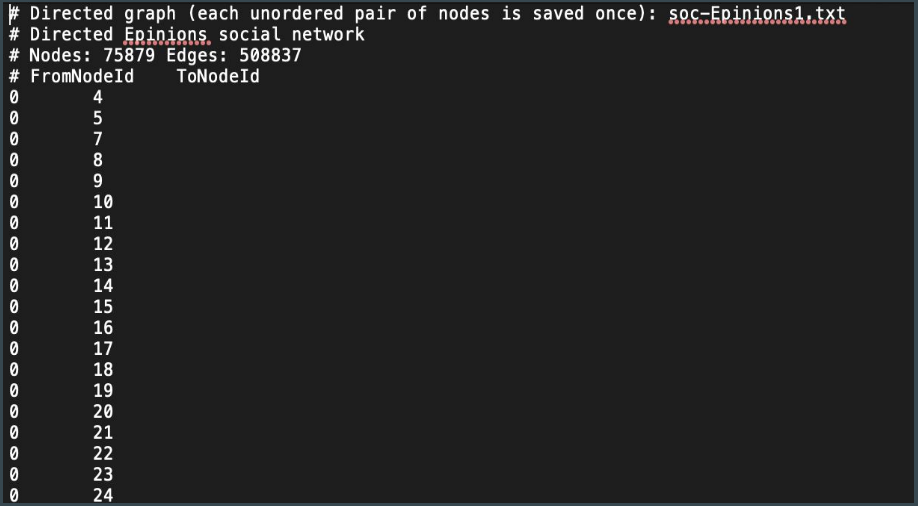
and 4 edges

E={ (1,2), (2,3), (3,1), (1,3)}

Example

This is a dataset from the consumer review site eopinions.com:

- Members can decide whether to ‘trust’ each other.
- Edges (a,b), ‘a’ in the first column, ‘b’ in the second, means ‘a trusts b.’

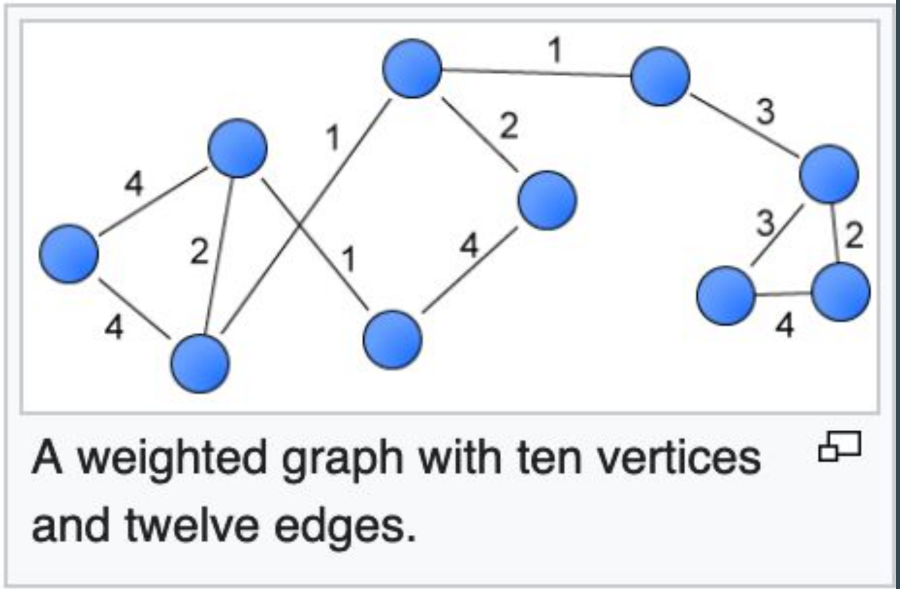


```
# Directed graph (each unordered pair of nodes is saved once): soc-Epinions1.txt
# Directed Epinions social network
# Nodes: 75879 Edges: 508837
# FromNodeId    ToNodeId
0       4
0       5
0       7
0       8
0       9
0       10
0       11
0       12
0       13
0       14
0       15
0       16
0       17
0       18
0       19
0       20
0       21
0       22
0       23
0       24
```

Social Network Data - Weighted Graphs

The edges can be labeled with a number, resulting in a *weighted graph*

- In USC study, edges are labeled ‘0’ or ‘1,’ where 0 indicates a weak relationship and 1 indicates a strong one.

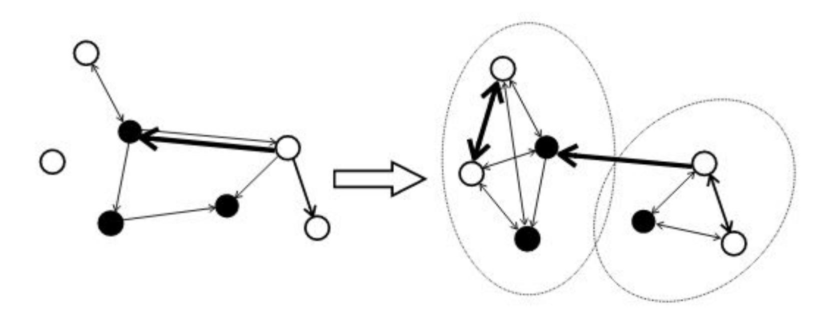


A weighted graph with ten vertices and twelve edges.

Social Network Data - Colored Graphs

Lastly, the people/nodes might be classified into different classes (represented by colors), resulting in a *colored graph.*

- E.g. hard drug ‘users’ are represented with black nodes and ‘non-users’ with white nodes.



Effect of intervention groups

Idea:

- People in the same group will make new/strengthen existing connections.
- The connections will be stronger between similar people
- Connections to users in separate groups will be weakened/broken
- This changes the ‘topology’ of the social network

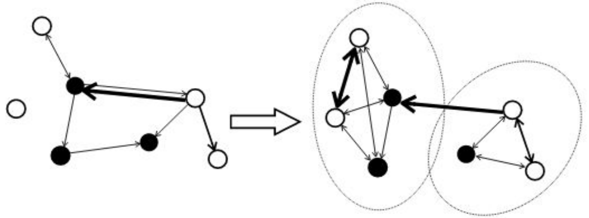
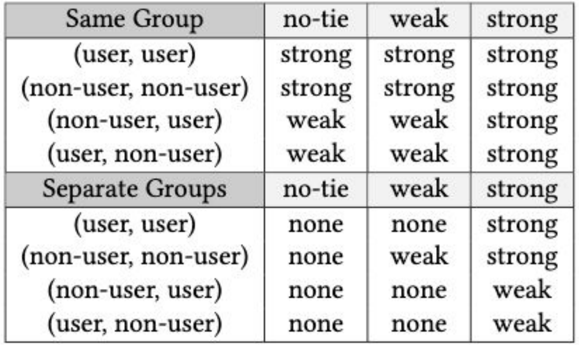


Figure 2: An example network pre- (left) and post- (right) intervention. The black (resp. white) circles indicate “user” (resp. “non-user”) nodes. Weak (resp. strong) links are denoted by thin (resp. thick) arrows. The ellipsoids represent the two groups that are formed for the intervention. As seen, new edges are created within the groups, while some edges are cut across the groups. partitions



Same Group	no-tie	weak	strong
(user, user)	strong	strong	strong
(non-user, non-user)	strong	strong	strong
(non-user, user)	weak	weak	strong
(user, non-user)	weak	weak	strong
Separate Groups	**no-tie**	**weak**	**strong**
(user, user)	none	none	strong
(non-user, non-user)	none	weak	strong
(non-user, user)	none	none	weak
(user, non-user)	none	none	weak

Table 1: Changes in tie strength post-intervention. The existing relationships, and the behavior of the individuals as well as their assignment to groups impacts the changes.

Choosing ideal groups

- The object is to create an algorithm that chooses the best possible peer intervention groups (i.e. leading to the fewest number of users)
- (This is NP-hard, so don’t have an efficient algorithm for finding “best” solution. Have algorithms for making good guesses.)

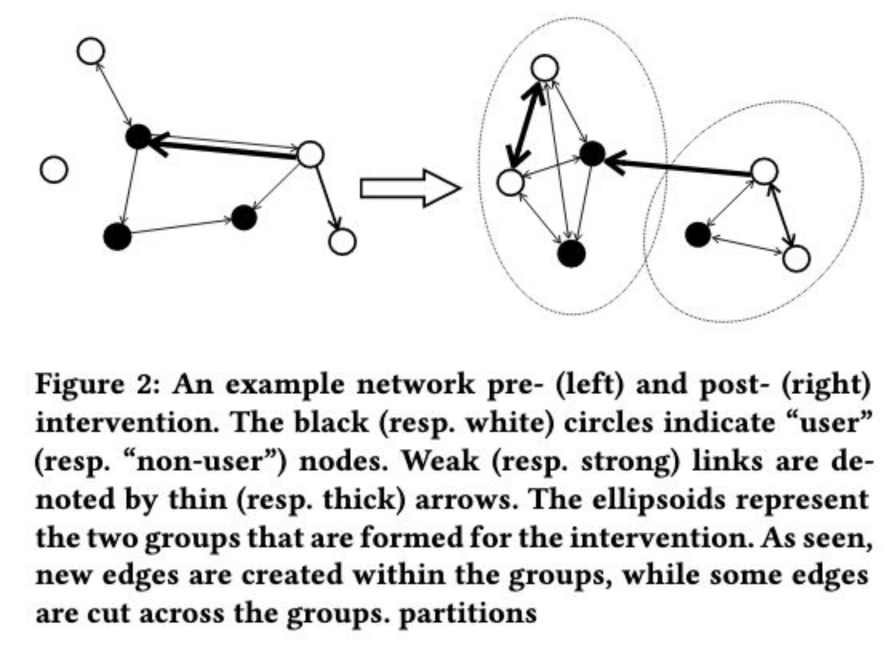
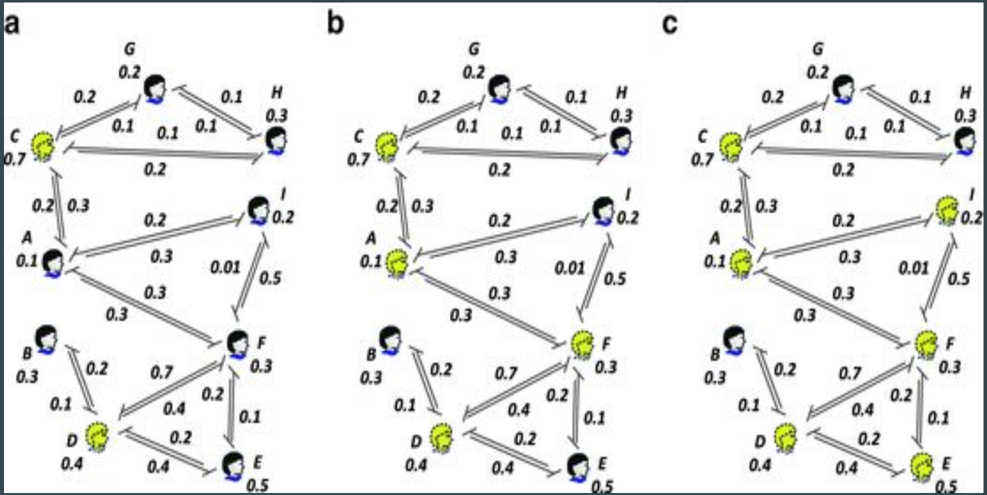


Figure 2: An example network pre- (left) and post- (right) intervention. The black (resp. white) circles indicate “user” (resp. “non-user”) nodes. Weak (resp. strong) links are denoted by thin (resp. thick) arrows. The ellipsoids represent the two groups that are formed for the intervention. As seen, new edges are created within the groups, while some edges are cut across the groups. partitions

Modeling Influence in Social Networks

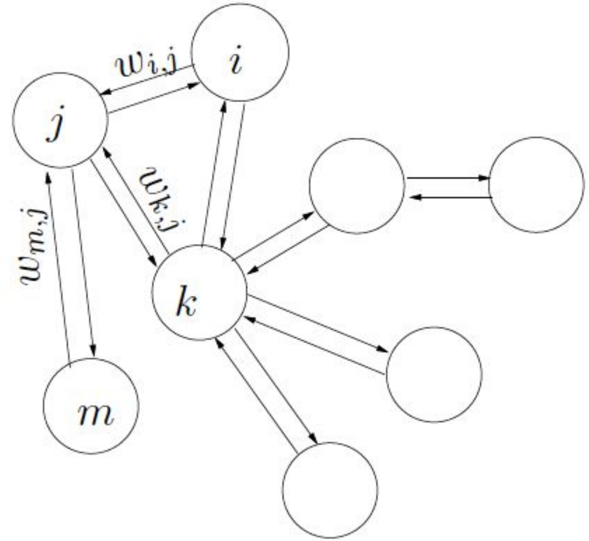
- Graph changes after intervention
- In order to model reduction in substance abuse, must model how connections to users/non-users influences an individual's drug use
- One mainstream model for influence is the 'Linear threshold model'
  - Used in USC study



Information or behavior diffuses through the social network as people influence each other. This is depicted above by the increasing number of yellow heads.

Linear Threshold Model

- Model influence and diffusion of information/behavior in a network
- Need to know: Basic understanding of graphs, high school algebra.
- To code: Some familiarity with the 'networkx' Python library
    - "NetworkX is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks."
- Widely used in combination with the 'influence maximization problem'
    - Find the most influential nodes in a social network
    - These nodes can spread information most efficiently

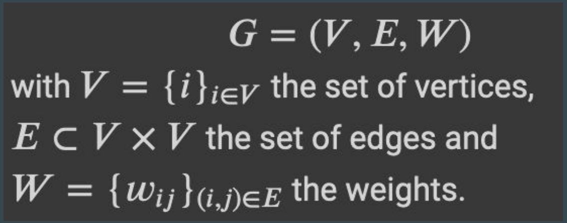


$$\Sigma_i \, w_{i,j} \leq 1, \forall j$$

Fig. 1. Linear Threshold model

Linear Threshold Model - Steps

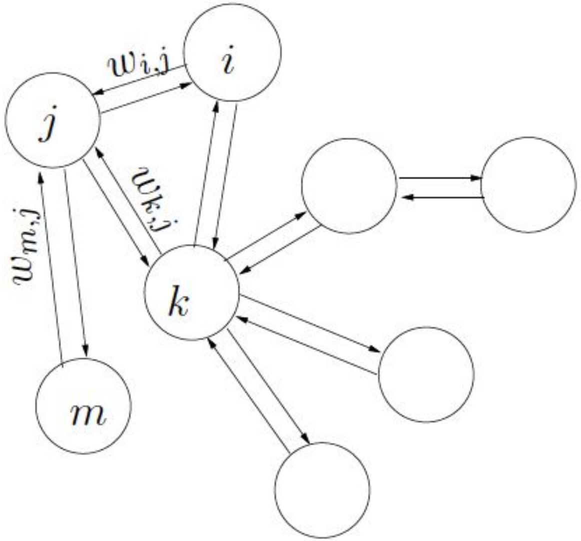
1) **Start with social network data as a weighted directed graph**



$$G = (V, E, W)$$

with $V = \{i\}_{i \in V}$ the set of vertices,
$E \subset V \times V$ the set of edges and
$W = \{w_{ij}\}_{(i,j) \in E}$ the weights.

- $V \times V$ is notation for pairs of vertices. So (i,j) in VxV means there is an edge from vertex i to vertex j.
- The weight w_ij is interpreted as the 'amount of influence i has on j'
- The weights going into a single node **sum to 1.**



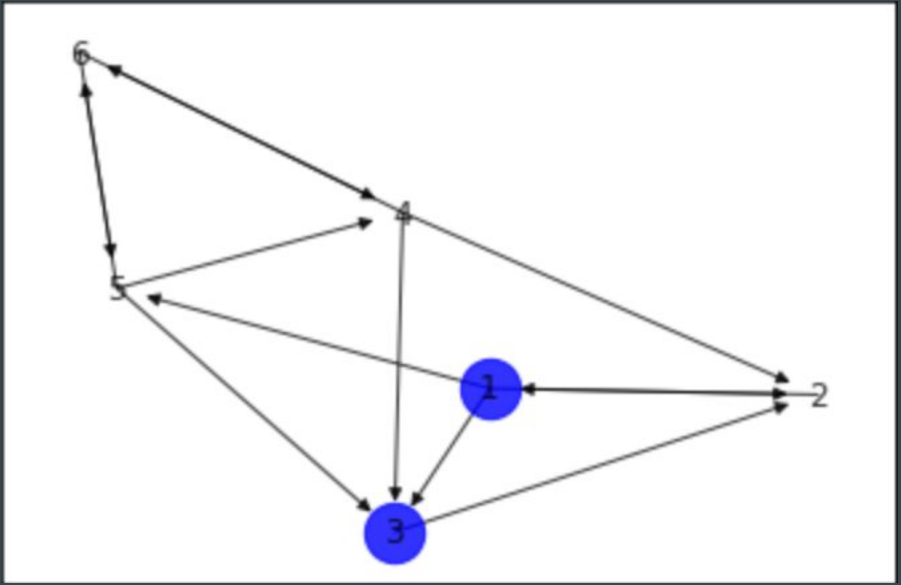
$$\Sigma_i \, w_{i,j} \leq 1, \, \forall j$$

Fig. 1. Linear Threshold model

Linear Threshold Model - Steps

2) **Color the 'seeds' in graph**, i.e., the nodes with a certain behavior or information (e.g. represent users with blue nodes)

3) **For each node, set a 'threshold value.'** This is a number between 0 and 1 that determines how much an individual needs to be influenced to change their behavior. 0 -> sheep, 1 -> very stubborn



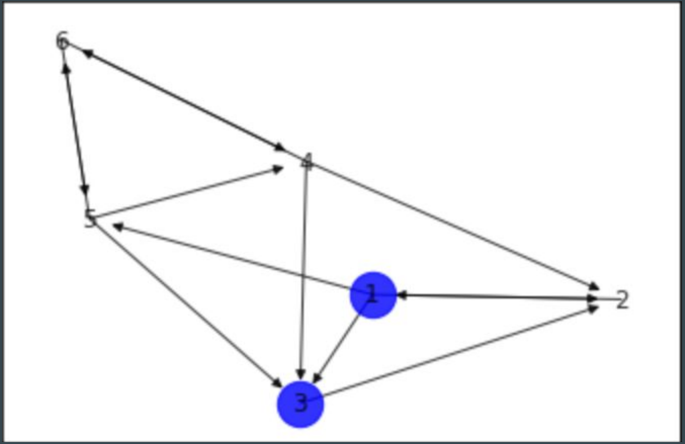
Linear Threshold Model - Steps

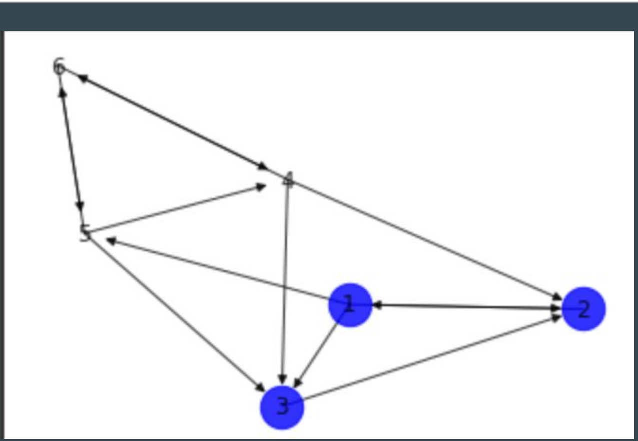
4) **‘Spread influence’** - In each round, do the following

- For each node *i* that isn’t a seed, define the influence on *i* as the sum of the weights corresponding to neighboring seeds:

$$\text{Influence}(i) = \sum_{\text{neighbor seeds}} w_{ij}$$

- If Influence(*i*) > Threshold(i), then vertex *i* becomes a seed



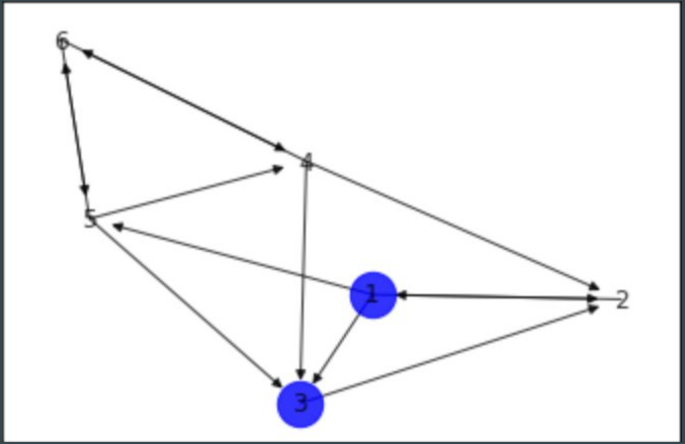


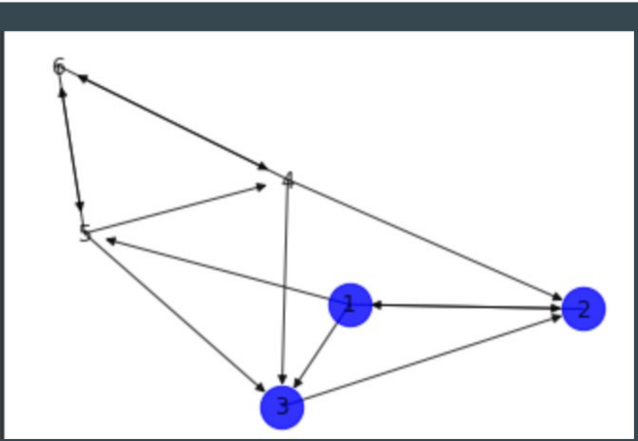
Linear Threshold Model - Steps

5) **Repeat as many times as there are opportunities for nodes to influence each other.**

- If influence is continuous, then run until no more seeds can be added

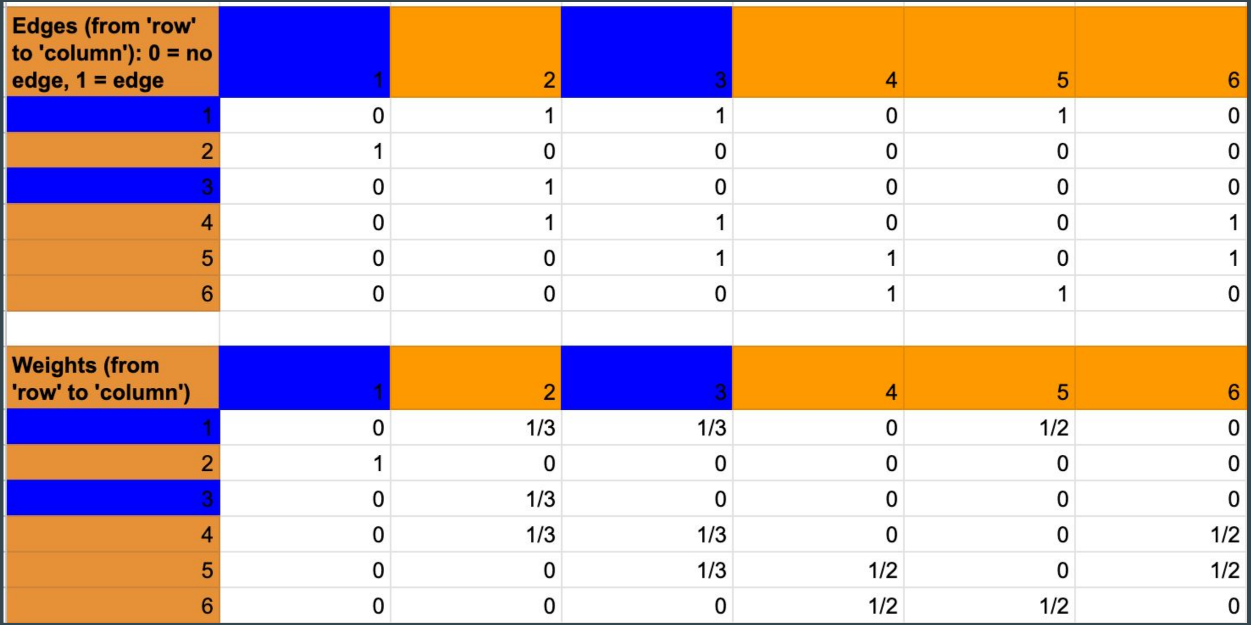
Note: The substance abuse model does LTM twice, once for non-users influencing users and once for users influencing non-users





Linear Threshold - Example

Suppose we are looking at a small network of 6 people. Two of them are convinced that Trump should be re-elected and are trying to convince their friends of the same. We assume that a given individual's connections in the network all influence them equally.

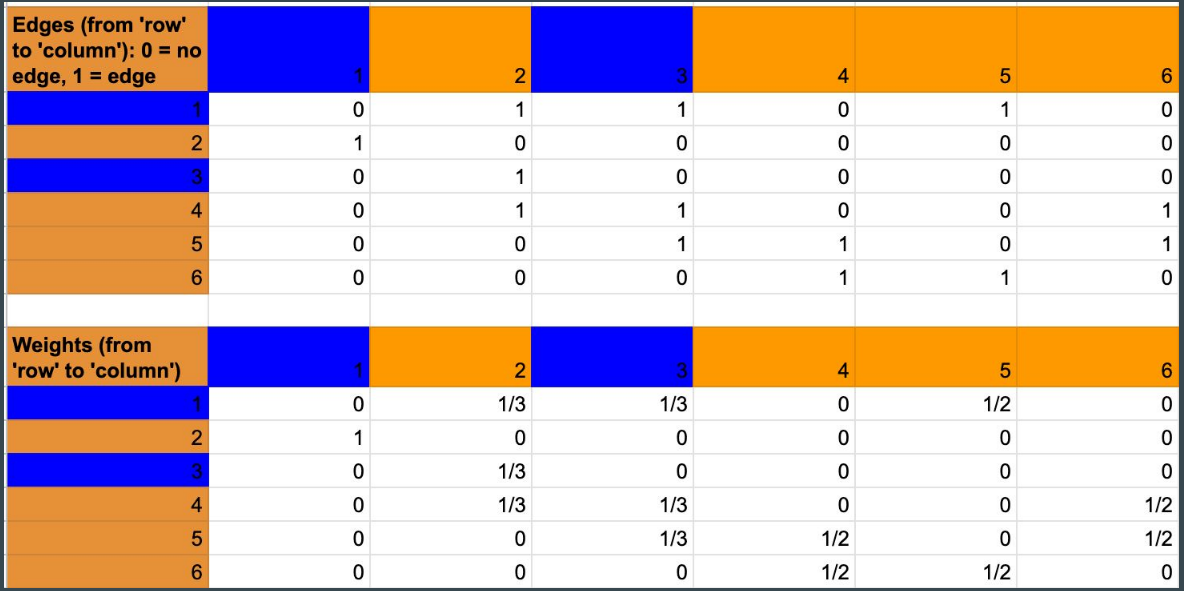


Edges (from 'row' to 'column'): 0 = no edge, 1 = edge	1	2	3	4	5	6
1	0	1	1	0	1	0
2	1	0	0	0	0	0
3	0	1	0	0	0	0
4	0	1	1	0	0	1
5	0	0	1	1	0	1
6	0	0	0	1	1	0

Weights (from 'row' to 'column')	1	2	3	4	5	6
1	0	1/3	1/3	0	1/2	0
2	1	0	0	0	0	0
3	0	1/3	0	0	0	0
4	0	1/3	1/3	0	0	1/2
5	0	0	1/3	1/2	0	1/2
6	0	0	0	1/2	1/2	0

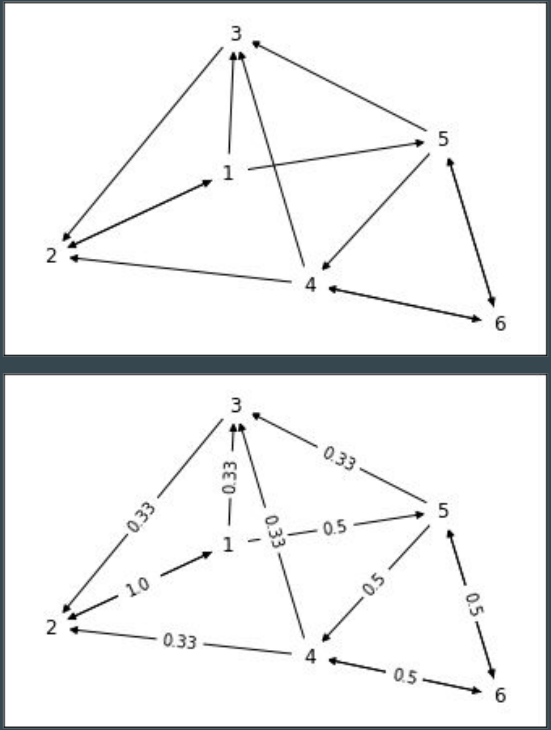
The weights going into a single node should sum to 1. We can see that the values in each column indeed sum to 1.

1) Start with social network data as a weighted directed graph



Edges (from 'row' to 'column'): 0 = no edge, 1 = edge	1	2	3	4	5	6
1	0	1	1	0	1	0
2	1	0	0	0	0	0
3	0	1	0	0	0	0
4	0	1	1	0	0	1
5	0	0	1	1	0	1
6	0	0	0	1	1	0

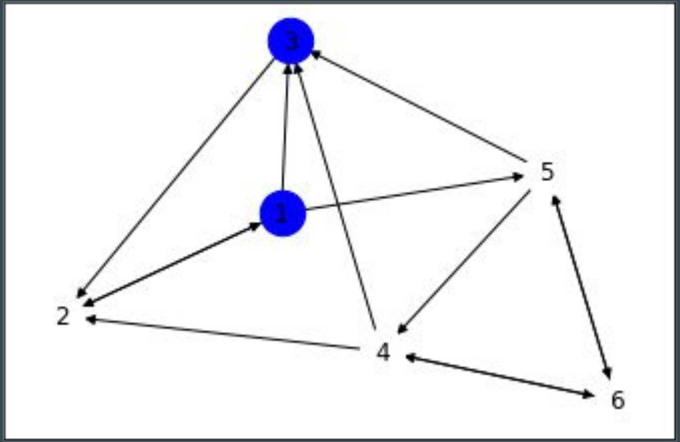
Weights (from 'row' to 'column')	1	2	3	4	5	6
1	0	1/3	1/3	0	1/2	0
2	1	0	0	0	0	0
3	0	1/3	0	0	0	0
4	0	1/3	1/3	0	0	1/2
5	0	0	1/3	1/2	0	1/2
6	0	0	0	1/2	1/2	0



Graphical representation of the data.

2) Color the seeds in the graph

- Begin with a list of ‘seeds’ exhibiting a certain behavior.



Persons 1 and 3 believe Trump should be re-elected. The others are undecided.

3) Set the threshold values

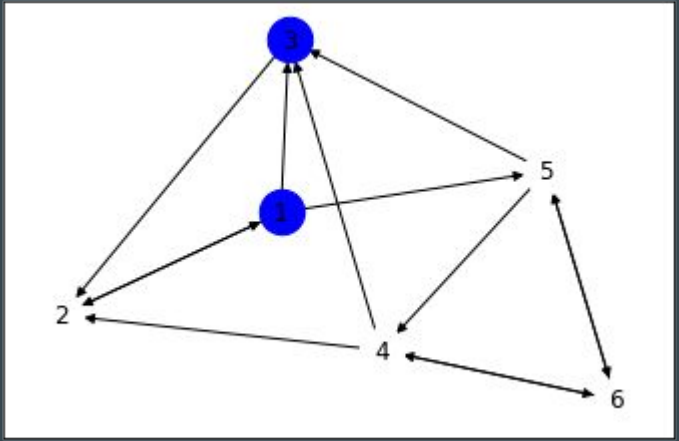
- The threshold value is a number between 0 and 1
- Each individual has its own threshold value
- Higher threshold values mean the person is harder to influence

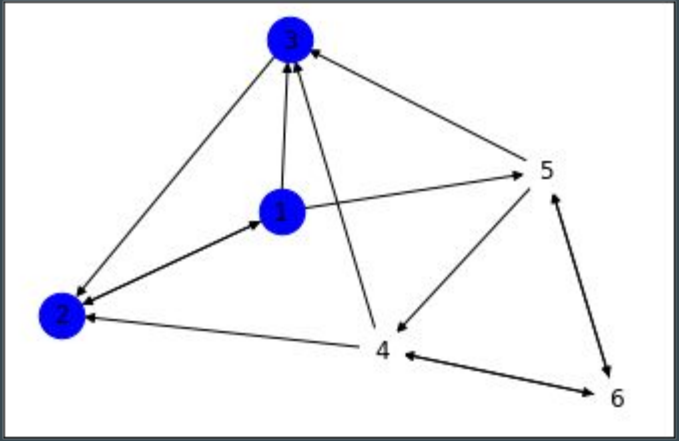
For simplicity, we assume every individual has a threshold of 0.5

- There are many options for the choice of threshold. The USC study chooses the thresholds randomly.

4) Spread Influence

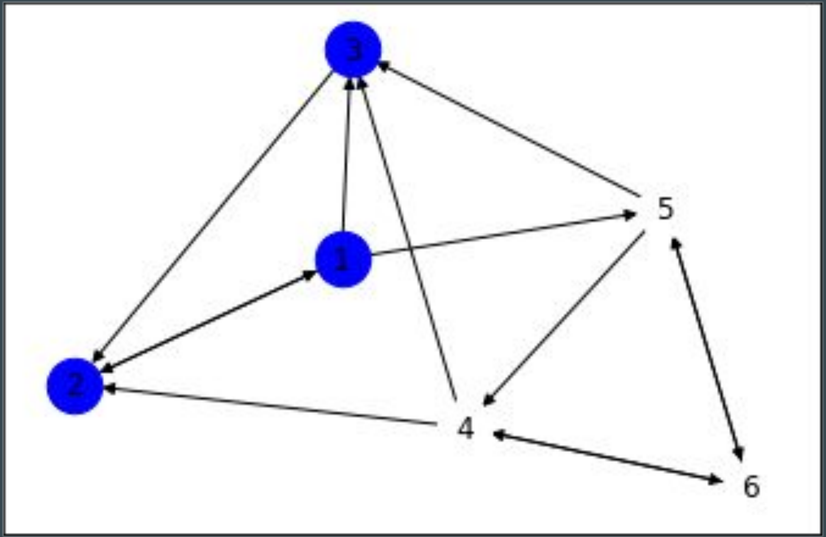
- For each node that isn't a seed, i.e., each undecided voter, take the sum of the influence weights from neighboring seeds
  - Person 2 is connected to Persons 1, 3 and 4.
  - 1 and 3 are seeds, each with influence weight ⅓ on 2
  - The sum is ⅓ + ⅓ = ⅔
- If the sum is greater than the threshold (½), the node becomes a seed.





5) Repeat for each influence opportunity

- Iterate for each influence opportunity or until the behavior cannot spread any further.
- In our example, the idea that Trump should be re-elected can’t spread any further



Simplifying assumptions

We made several assumptions for simplicity that might be altered in practice:

- All of a person’s connections influence them equally
  - In practice some connections influence a person more strongly
- If the sum of influence weights from neighboring seeds is greater than the threshold, the individual becomes a seed.
  - Many models instead assume that the individual becomes a seed *with a certain probability*
- All threshold values are 0.5
  - Thresholds can vary from person to person. The USC model uses a random number between 0 and 1.
- There’s only one kind of seed
  - Can have competing influences. In the USC model we do the process once for users influencing non-users and once vice versa

Resources

- Applied Social Network Analysis in Python, Coursera
- https://github.com/melaniewalsh/sample-social-network-datasets - practice datasets for social network analysis
- https://snap.stanford.edu/data/ - Stanford Large Network Dataset Collection
  - Collection of social network data from large online networks (including FB, twitter, Google+)
- https://networkx.github.io/documentation/stable/ - Documentation for networkx python library
- https://www.kirenz.com/post/2019-08-13-network_analysis/ - A short intro to social network analysis with Python

Discussion Questions

How do you measure the effectiveness of an AI-based intervention?

How is the data for social influence algorithms gathered?

How can such algorithms be abused?

Which assumptions seem unrealistic? How could they be modified?

What problems could such models cause?