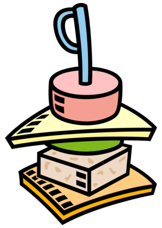
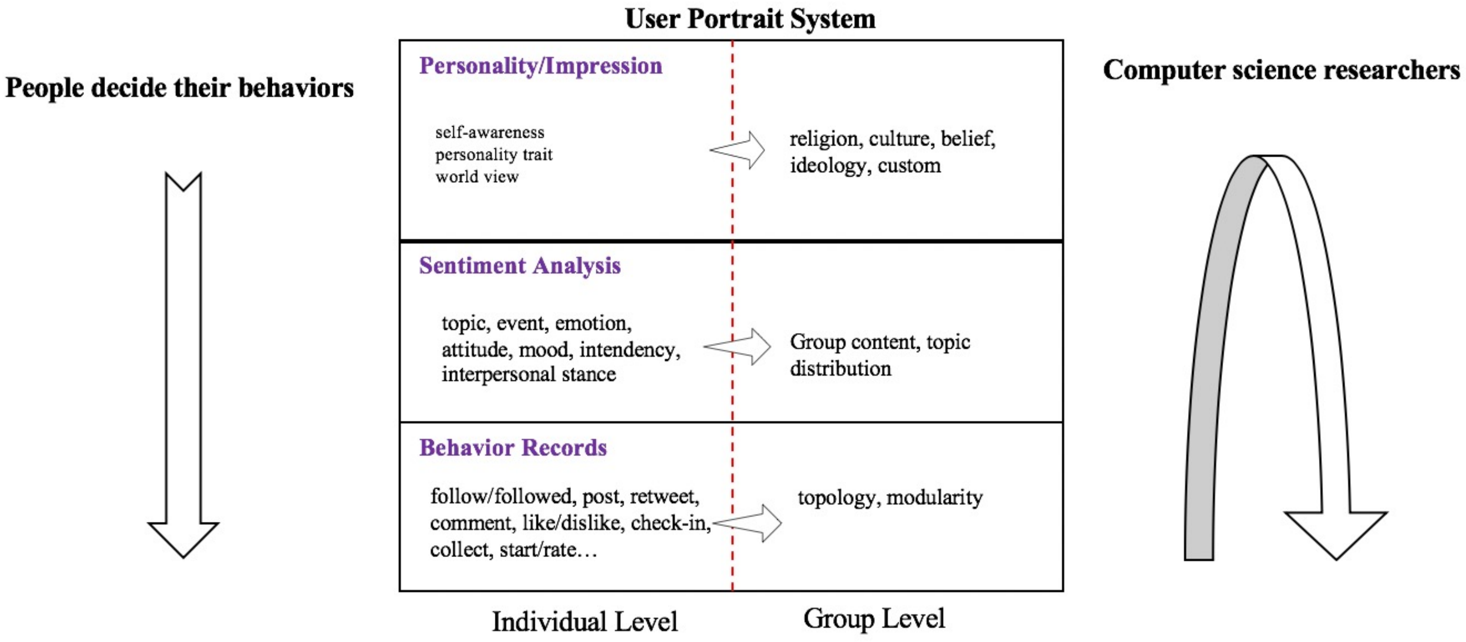


Appetizer/snacks before we start…

This is the INFS7450 online tutorial conducted by Xiangguo (Sheldon).
You see this slide because you come here earlier. Let's wait for the other students to come. ☺
(Appetizer will not be recorded in the video and won't be in our exam.)

Applications of the user portrait system

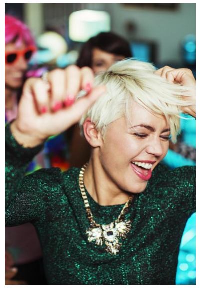


chatbot

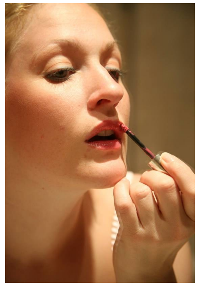
mental intervention

Socially Aware

individual RecSys

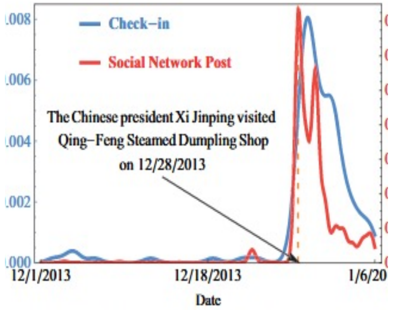


Extravert Ad



Introvert Ad

consumer behavior analysis



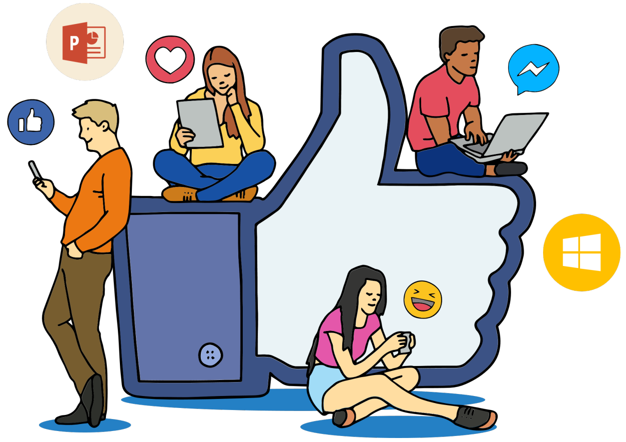
human resource management

SOCIAL MEDIA ANALYTICS INFS7450
Tutorial Week 7

Xiangguo Sun

School of ITEE

The University of Queensland



Online quiz 4

QUESTION 1

Which of the following statements is wrong?

- ○ A. When individuals get connected, one can observe distinguishable assortativity in their connectivity networks.
- ◉ B. In networks with assortativity, dissimilar nodes are connected to one another more often than similar nodes.
- ○ C. Assortativity is the most commonly observed pattern among linked individuals.
- ○ D. Three common forces are influence, homophily, and confounding.

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.

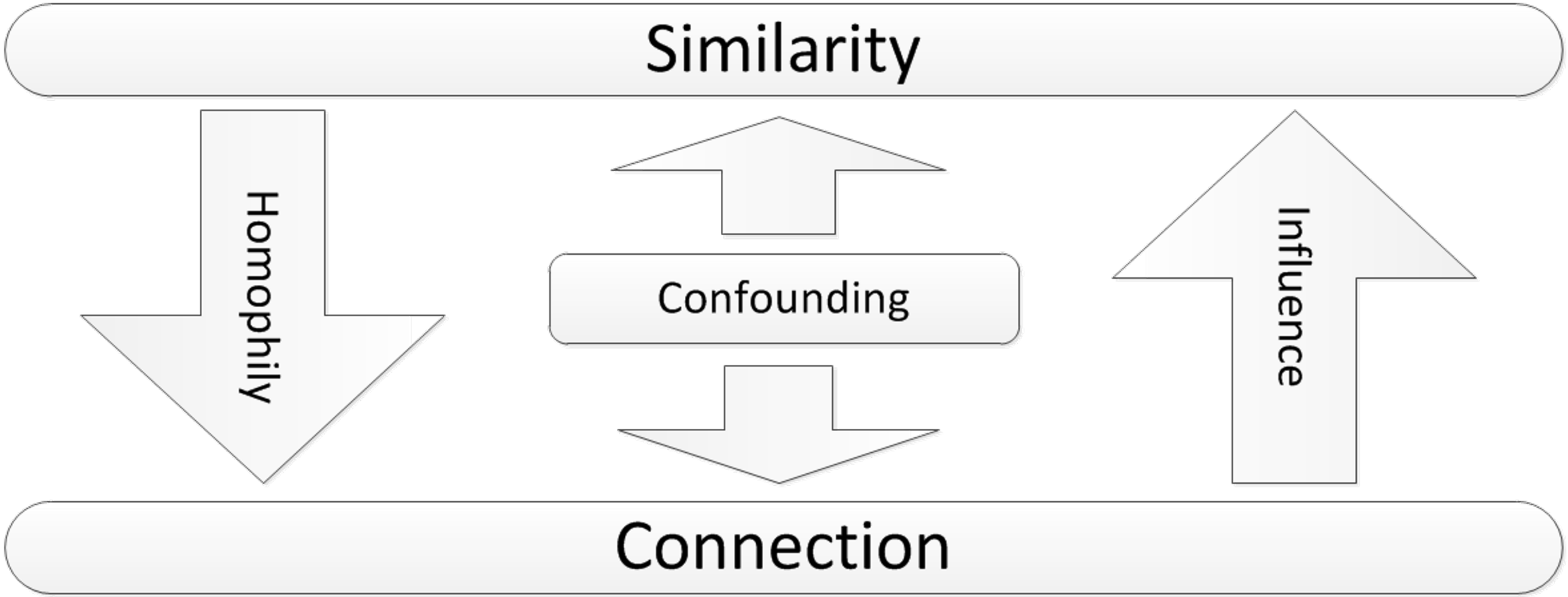
Homophily

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

Confounding

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

Influence, Homophily, and Confounding



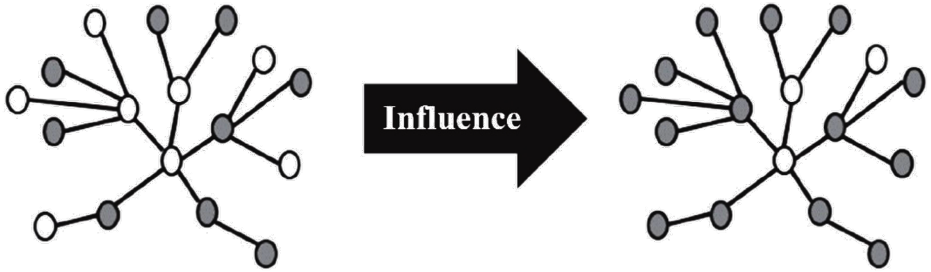
Cause and Effect

Source of Assortativity in Networks

Both influence and homophily generate social similarity in social networks

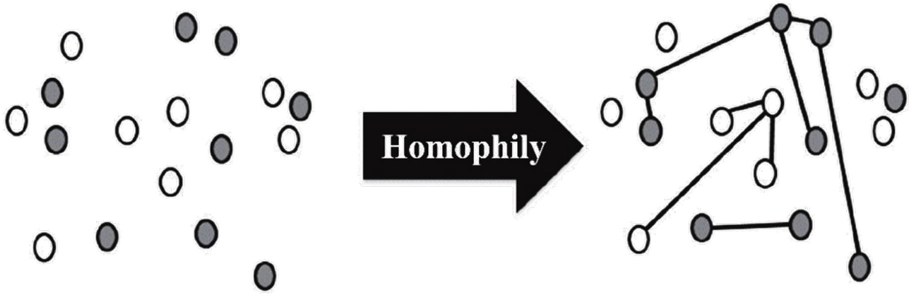
Influence

Makes connected nodes similar to each other



Homophily

Selects similar nodes to link together



Why?

- Smoker friends influence their non-smoker friends (social conformity)

Influence

- Smokers become friends

Homophily

- There are lots of places that people can smoke, and smoking is a type of local culture

Confounding

Online quiz 4

QUESTION 2

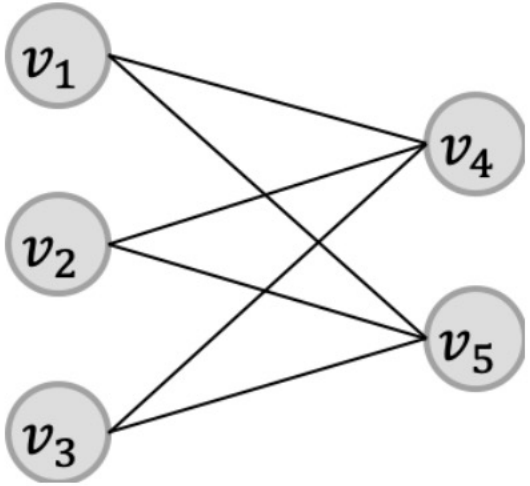
Which of the following statements is wrong?

- ○ A. 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.
- ○ B. Homophily is observed in already similar individuals.
- ○ C. Confounding is the environment's effect on making individuals similar.
- ◉ D. The confounding force is an internal factor that is dependent on inter-individual interactions.

Online quiz 4

QUESTION 3

What is the modularity value of the following bipartite graph? (let's say there are two types of nodes, the left side is one type, and the right side is the other type.)



- [x] A. -1/2
- [ ] B. 1/2
- [ ] C. -1/4
- [ ] D. 1/4

Modularity: Matrix Form

Let Modularity matrix be

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

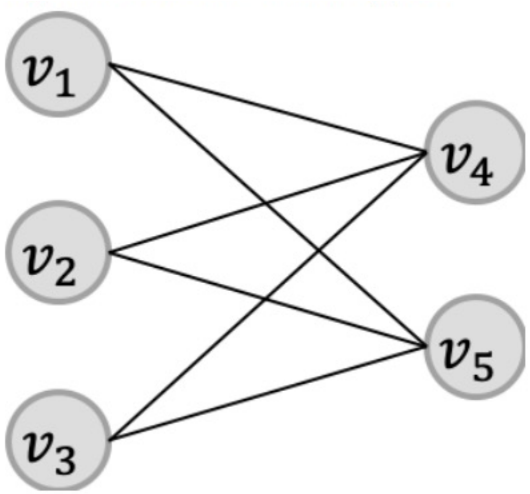
$\boldsymbol{d} \in \mathbb{R}^{n\times 1}$ is the degree vector

Modularity can be reformulated as

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

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

Online quiz 4



- A. -1/2
- B. 1/2
- C. -1/4
- D. 1/4

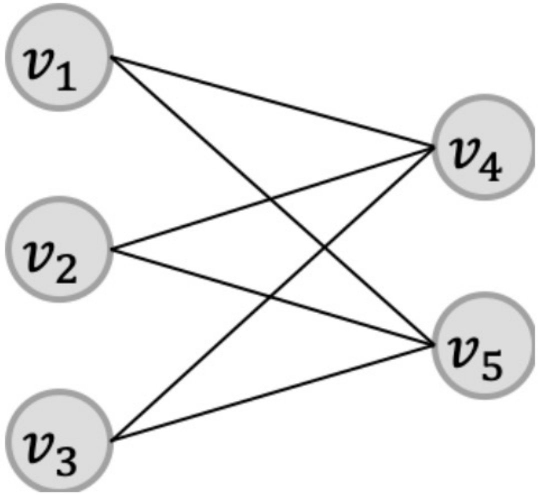
$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix} \quad \Delta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \quad d = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 3 \\ 3 \end{bmatrix} \quad m = 6$$

$$B = A - dd^T/2m = \begin{bmatrix} -1/3 & -1/3 & -1/3 & 1/2 & 1/2 \\ -1/3 & -1/3 & -1/3 & 1/2 & 1/2 \\ -1/3 & -1/3 & -1/3 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/2 & -3/4 & -3/4 \\ 1/2 & 1/2 & 1/2 & -3/4 & -3/4 \end{bmatrix}$$

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

QUESTION 4

Consider the following bipartite graph. For the modularity of this bipartite graph, which statement is wrong? (let's say there are two types of nodes, the left side is one type, and the right side is the other type.)



○ A.

$$A = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{bmatrix}$$

○ C.

$$d = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 3 \\ 3 \end{bmatrix} \text{ and } m=6$$

◉ B.

$$\Delta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

○ D.

$$B = A - \frac{dd^T}{2m} = \begin{bmatrix} -0.33 & -0.33 & -0.33 & 0.5 & 0.5 \\ -0.33 & -0.33 & -0.33 & 0.5 & 0.5 \\ -0.33 & -0.33 & -0.33 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & -0.75 & -0.75 \\ 0.5 & 0.5 & 0.5 & -0.75 & -0.75 \end{bmatrix}$$

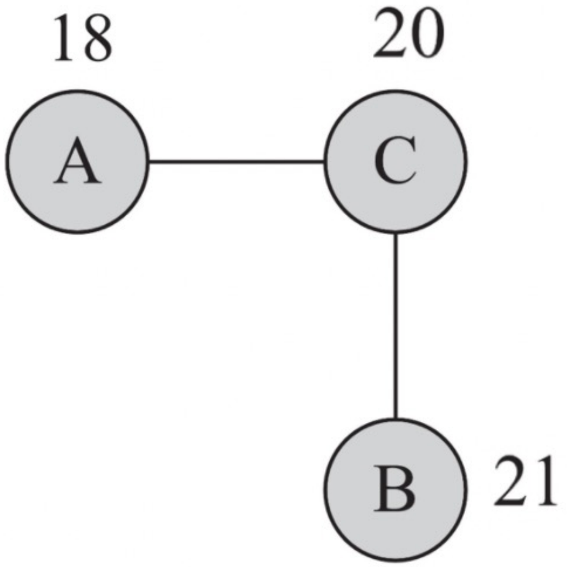
Online quiz 4

QUESTION 5

Consider following figure with values demonstrating the attributes associated with each node. Since this graph is undirected, we have the following edges:

$E = \{(a,c), (c,a), (c,b), (b,c)\}$

Let X_L as the value of the left end of an edge and X_R as the value of the right end of an edge:



Which of the following is correct?

- A. $X_L \begin{bmatrix} 18 \\ 20 \\ 20 \\ 21 \end{bmatrix}, X_R \begin{bmatrix} 20 \\ 18 \\ 21 \\ 20 \end{bmatrix}$
- B. $X_L \begin{bmatrix} 20 \\ 20 \\ 18 \\ 21 \end{bmatrix}, X_R \begin{bmatrix} 18 \\ 11 \\ 21 \\ 20 \end{bmatrix}$
- C. $X_L \begin{bmatrix} 20 \\ 18 \\ 21 \\ 20 \end{bmatrix}, X_R \begin{bmatrix} 18 \\ 20 \\ 20 \\ 21 \end{bmatrix}$
- D. $X_L \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix}, X_R \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}$

List of edges:

(A, C)
(C, A)
(C, B)
(B, C)

$X_L : (18, 21, 21, 20)$

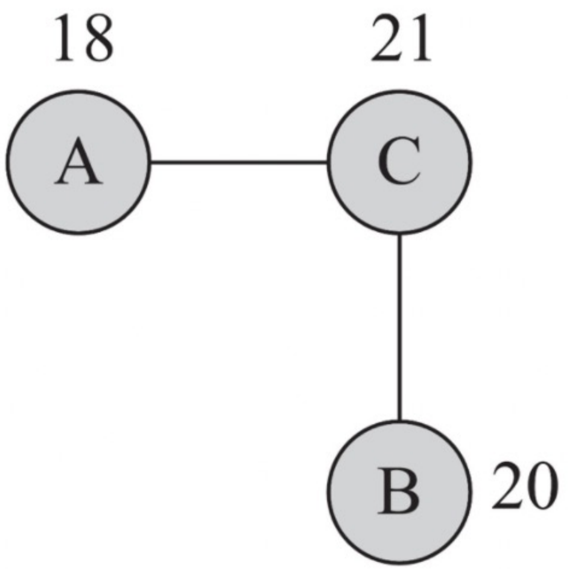
$X_R : (21, 18, 20, 21)$

QUESTION 6

Consider the following figure with values demonstrating the attributes associated with each node. Since this graph is undirected, we have the following edges:

$E = \{(a,c),\ (c,a),\ (c,b),\ (b,c)\}$

Let X_L as the value of the left end of an edge and X_R as the value of the right end of an edge:



$$E(X_L) = E(X_R) = \frac{\sum_i (X_L)_i}{2m} = \frac{\sum_i d_i x_i}{2m}$$

$$E(X_L X_R) = \frac{1}{2m} \sum_i (X_L)_i (X_R)_i = \frac{\sum_{ij} A_{ij} x_i x_j}{2m}$$

$(18 + 21 + 21 + 20)/2*2 = 20$

$(1/(2*2)) * (2*A_{ac}X_aX_c + A_{cb}X_cX_b) = 399$

$399 - 20 * 20 = -1$

What is the covariance $\sigma(X_l, X_R)$ between X_L and X_R?

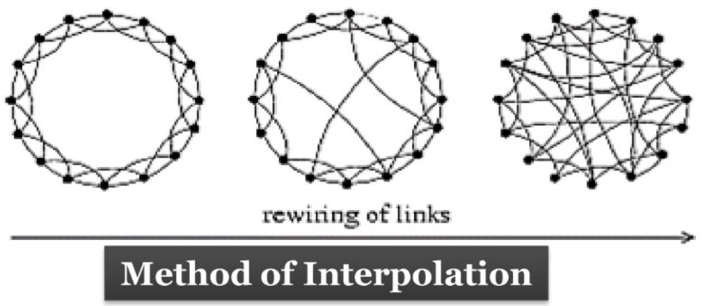
- ○ A. -0.67
- ● B. -1
- ○ C. 55.875
- ○ D. 0.67

QUESTION 7

Which of the following is WRONG according to the Small-World Graph Model?

- [x] A. A network with high clustering can not be at the same time a small world.
- [ ] B. The Watts Strogatz Model provides insight on the interplay between clustering and the regular lattice
- [ ] C. In the small-world model, a parameter $0 \leq \beta \leq 1$ controls randomness in the model, when $\beta = 0$, the model is basically a regular lattice.
- [ ] D. In the small-world model, a parameter $0 \leq \beta \leq 1$ controls randomness in the model, when $\beta = 1$, the model becomes a random graph.

- The lattice has a **high**, but **fixed**, clustering coefficient
- The lattice has a **high** average path length



- In the small-world model, a parameter $0 \leq \beta \leq 1$ controls randomness in the model
  - When β is 0, the model is basically a regular lattice
  - When $\beta = 1$, the model becomes a random graph

QUESTION 8

Which of the following statements regarding Influence Modeling is wrong?

- A. Linear Threshold Model (LTM) can be used to model the network influence.
- B. The LTM model is a simple but effective model for modeling influence in social networks, in which nodes make decision based on the influence coming from their already activated neighborhood.
- C. In LTM, each node chooses a threshold θ_i randomly from a uniform distribution in an interval between 0 and 1.
- D. At time t, all nodes that were active in the previous steps [0.. −1] become inactive, and only nodes already activated before time t get the chance to activate others at time

QUESTION 9

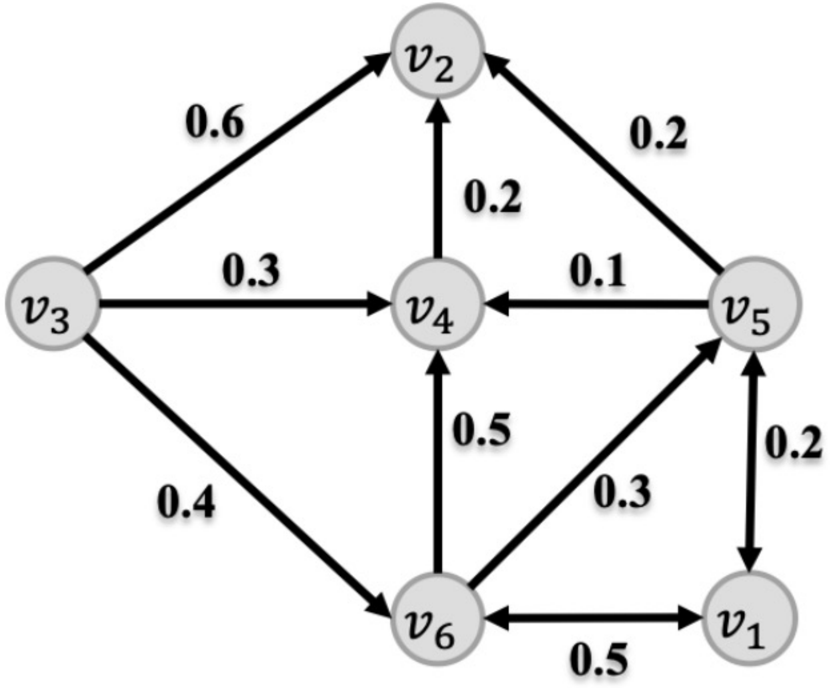
Here is a Regular Lattice with degree $c = 4$ and 10 nodes, which of the following is correct?

- ○ A. A regular (ring) lattice is a special case of regular networks where there exist multiple patterns on how ordered nodes are connected to one another.
- ○ B. there is an edge connecting node v_1 and node v_4
- ◉ C. there is an edge connecting node v_1 and node v_9
- ○ D. In this regular lattice, nodes are connected to their previous 4 and following 4 neighbors

- In a regular lattice of degree c, nodes are connected to their previous $c/2$ and following $c/2$ neighbors

QUESTION 10

Follow the LTM procedure for the following graph. Assume all the thresholds are 0.5 and node v3 is activated at time 0.



Which of the following nodes will be activated at time 1 ?

- A. v2
- B. v4
- C. v6
- D. v1

Suggestions from Xiangguo

S1:Which of (the following) nodes are/is activated after time 1: (v3,v2)

S2: you can continue the process until the model stops, and draw the sanpshots in each step.

Linear Threshold Model (LTM)

Assume that for any given node v_i , the sum of incoming influence ($w_{j,i}$) from its incoming neighbors is

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

- 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. Only nodes already activated before time t (i.e., A_{t-1}) can activate others at time t

- Nodes satisfying the following condition will be activated

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

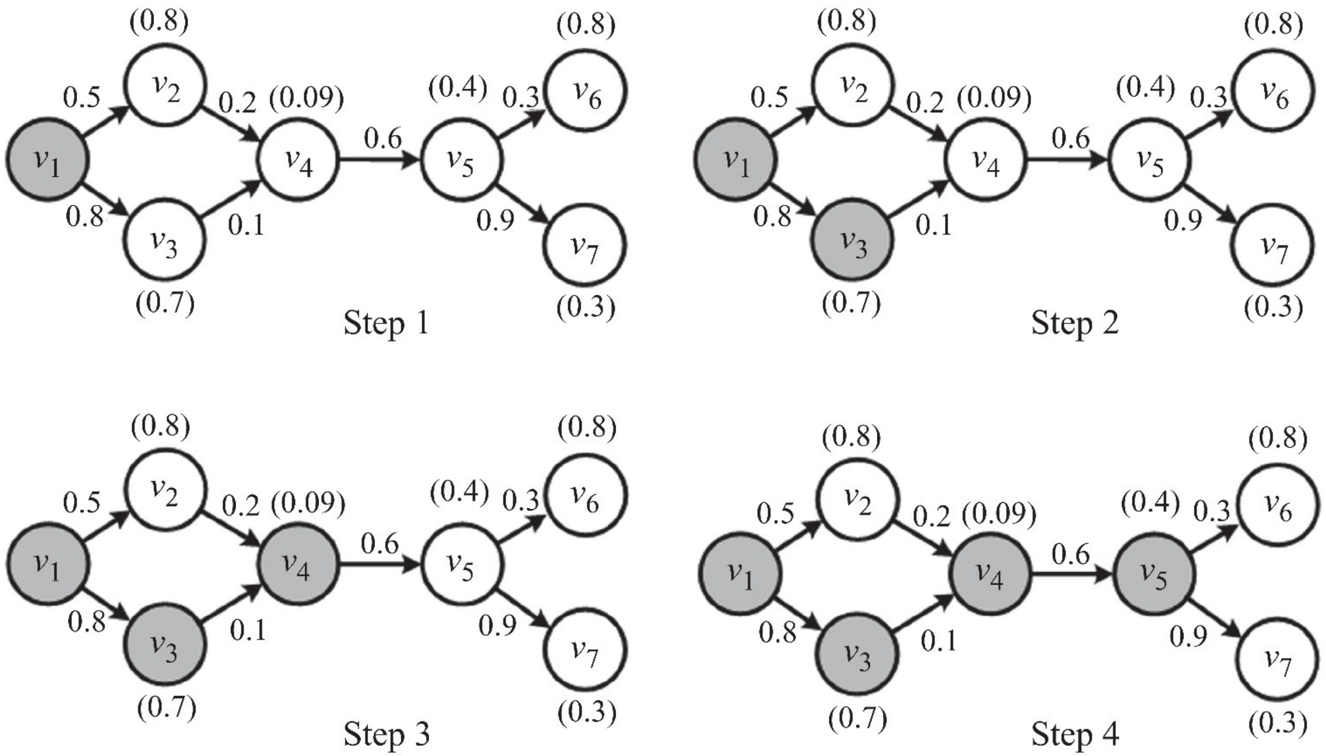
LTM Algorithm

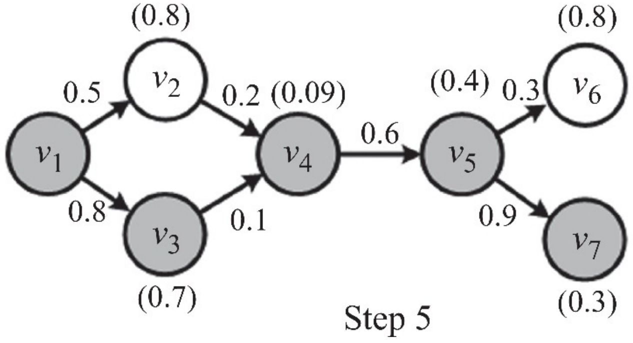
Algorithm 1 Linear Threshold Model (LTM)

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

return Final set of activated nodes A_∞
i=0;
Uniformly assign random thresholds θ_v from the interval [0, 1];
while $i = 0$ or $(A_{i-1} \neq A_i, i \geq 1)$ **do**
    $A_{i+1} = A_i$
    inactive $= V - A_i$;
    for all $v \in$ inactive **do**
        if $\sum_{j \text{ connected to } v, j \in A_i} w_{j,v} \geq \theta_v$. **then**
            activate v;
            $A_{i+1} = A_{i+1} \cup \{v\}$;
        end if
    end for
    $i = i + 1$;
end while
$A_\infty = A_i$;
Return A_∞;

Linear Threshold Model (LTM) - An Example





Thresholds are on top of nodes

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 the change in Modularity:

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

Ordinal attributes. The Homophily index is defined as the change in Pearson correlation:

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

Modeling Homophily

- At each time step, a single node gets activated.
  - A node once activated will remain activated.

- 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 that node v requires for being connected to other nodes.

- For any edge (v, w) that is still not examined, if the similarity $sim(v, w) \ > \theta_v$, then edge (v, w) is added.
  - The formation of edges is the result of similarity.
- This continues until all vertices are activated.

Homophily Model

Algorithm 1 Homophily Model

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

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

Information Diffusion

- Information diffusion: process by which a piece of information (knowledge) is spread and reaches individuals through interactions.

- We focus on techniques that can model and predict information diffusion in social networks.

Information Diffusion

- **Sender(s).** A sender or a small set of senders that initiate the information diffusion process;

- **Receiver(s).** A receiver or a set of receivers that receive diffused information. Commonly, the set of receivers is much larger than the set of senders; and

- **Medium.** It is the place through which the diffusion happens. For example, when a rumor is spreading, the medium can be social media platforms such as Twitter and Facebook.

Herd Behavior

- Network is observable
- Only public information is available
- Global information is available

Information Diffusion Types

- **Herd Behavior**
- **Information Cascade**

We define the process of interfering with information diffusion by expediting, delaying, or even stopping diffusion as Intervention

Herd Behavior

Herd behavior describes when a group of individuals performs the same or similar actions without any plan

Main Components of Herd Behavior

- Connections between individuals
- A method to transfer behaviors among individuals or to observe the others' behaviors (global information)

Examples of Herd Behavior

- Flocks, herds of animals, and audiences during sporting events

Network Observability in Herb Behavior

In herd behavior, individuals make decisions by observing all other individuals' decisions

- In general, herd behavior's network is close to a complete graph where nodes can observe at least most other nodes and they can observe public information
  - For example, they can see the crowd

How Does Intervention Work?

- When a new piece of private information releases,
  - The herd reevaluate their guesses and this may create completely new results
- The Emperor’s New Clothes
  - When the boy gives his private observation, other people compare it with their observation and confirm it
  - This piece of information may change others’ guess and ends the herding effect

Information Cascade

- In the presence of a network
- Only local information is available

Independent Cascade Model (ICM)

- A node activated at time t, has one chance, at time step $t+1$, to activate its neighbors

- Assume v is activated at time t
  - For any neighbor w of v, there's a probability p_{vw} that node w gets activated at time $t+1$.

- Node v activated at time t has a single chance of activating its neighbors
  - The activation can only happen at $t+1$

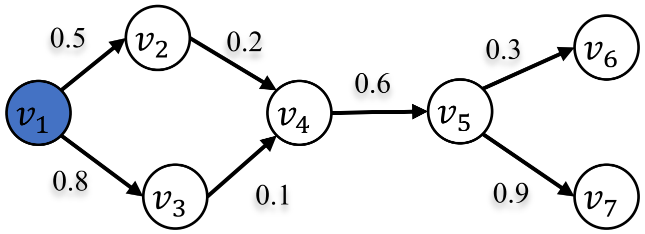
ICM Algorithm

Algorithm 1 Independent Cascade Model (ICM)

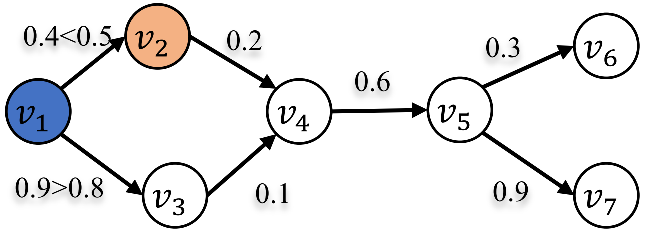
Require: Diffusion graph $G(V, E)$, set of initial activated nodes A_0, activation probabilities $p_{v,w}$

return Final set of activated nodes A_∞
i = 0;
while $A_i \neq \{\}$ **do**
    $i = i + 1$;
    $A_i = \{\}$; \\ the set of nodes being activated at time i
    for all $v \in A_{i-1}$ **do**
        for all w neighbor of $v, w \notin \cup_{j=0}^{i} A_j$ **do**
            rand = generate a random number in [0,1];
            if rand $< p_{v,w}$ **then**
                activate w;
                $A_i = A_i \cup \{w\}$;
            end if
        end for
    end for
end while
$A_\infty = \cup_{j=0}^{i} A_j$;
Return A_∞;

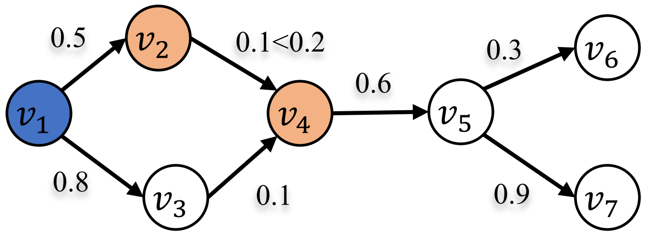
Independent Cascade Model: An Example



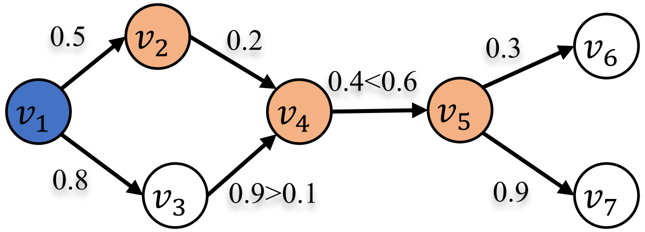
Step 1



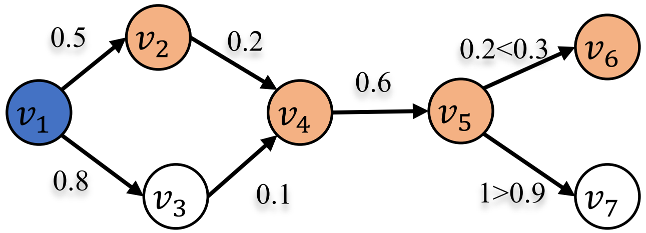
Step 2



Step 3



Step 4



Step 5

The edge weight represents the influence probability.

Chatting box

Zhi Zeng to Everyone 1:03 pr

In the independent model, if v1 is able to activate both v2 and v3, will v2 and v3 be both activated? ...

Shobhan Mitra to Everyone 1:04 p

here they don't have threshold or they have randomly generated threshold? ..

Zhaoyu Zhang to Everyone 1:06 p

And if v2 and v3 are both activated by v1, v4's threshold
is 0.25, v4 will not be activated right?

Xue Li to Everyone 1:09

According to the lecture, the biggest difference between LTM and ICM is the activate time, a node in LTM can active other nodes at time 1, or time 2, or time 3, but ICM just can active exact node at time n (a specific time). In terms of lecture , a node has been already activated, it should remain the activate status. So what the meaning for LTM one node can active other node at time 1, or time 2 or time 3?

Vy Hoa Phun to Everyone 1:18 pm

I'm curious why that algorithm is called "Linear" Threshold model, is it simply because the criterion to activate a node has linear form?

Fan Xie to Everyone 1:20 pm

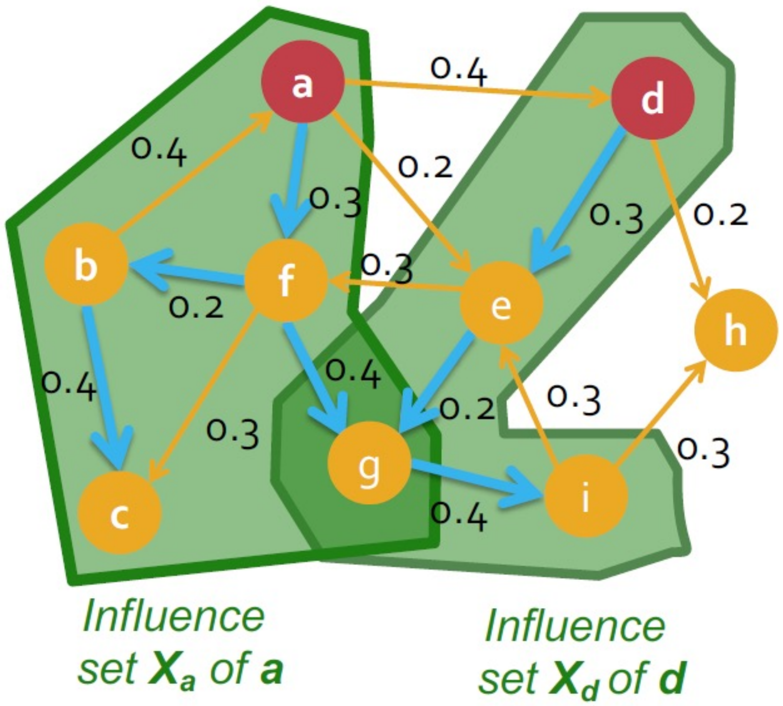
Summary of ICM

- Independent Cascade Model
  - Directed finite $G = (V, E)$
  - Set S starts out with new behavior
    - Say nodes with this behavior are "**active**"
  - Each edge (v, w) has a probability p_{vw}
  - If node v is active, it gets **one** chance to make w active, with probability p_{vw}
    - Each edge fires at most once
- **Does scheduling matter? No**
    - If u, v are both active at the same time, it doesn't matter which tries to activate w first
  - **But the time moves in discrete steps**

Most Influential Set

Problem: (*k* is a user-specified parameter)

- **Most influential set of size *k*:** set ***S*** of ***k*** nodes producing largest **expected** cascade size ***f(S)*** if activated [Domingos-Richardson '01]
- **Optimization problem:** $\max_{\text{S of size k}} f(S)$

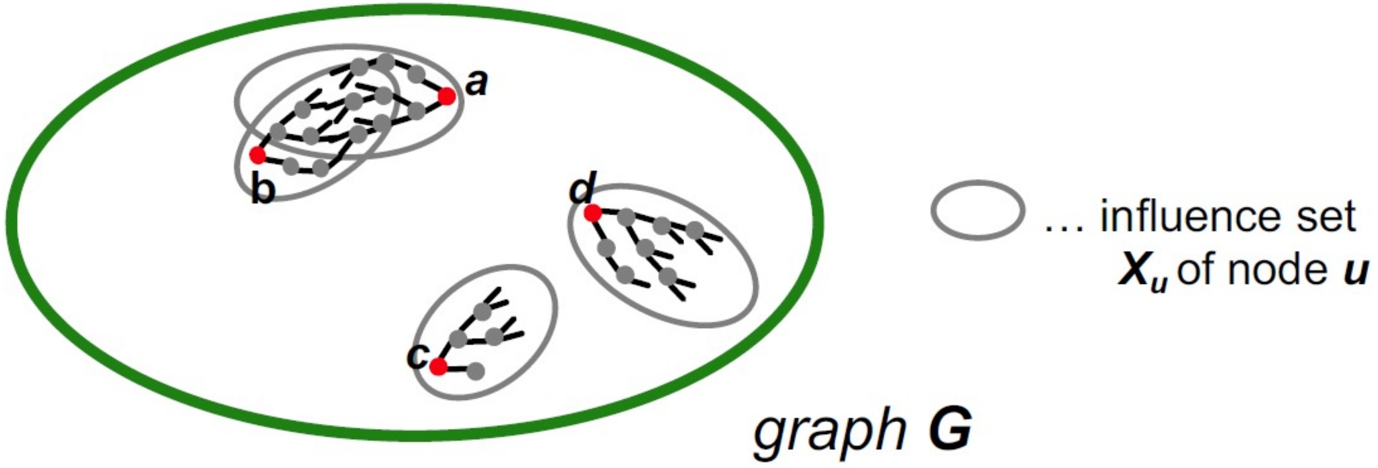


Why "expected cascade size"? X_a is a result of a random process. So in practice we would want to compute X_a for many random realizations and then maximize the "average" value *f(S)*. For now let's ignore this nuisance and simply assume that each node *u* influences *a* set of nodes X_u

$$f(S) = \frac{1}{|I|} \sum_{\text{Random realizations } i} f_i(S)$$

Most Influential Set of Nodes

- ***S***: is initial active set
- ***f(S)***: **The expected size of final active set**
  - ***f(S)*** is the size of the union of X_u: $\boldsymbol{f(S)} = |\cup_{\boldsymbol{u \in S}} \boldsymbol{X_u}|$

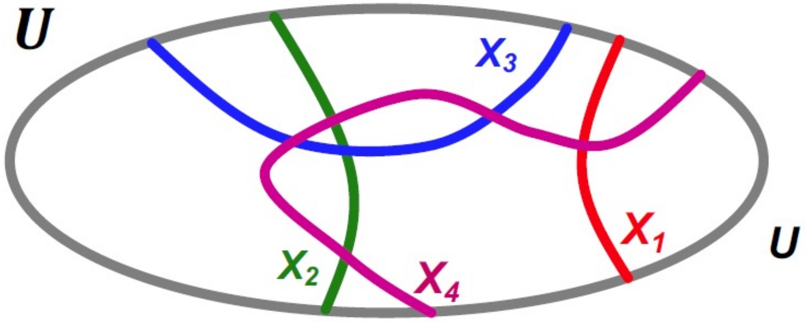


- **Set *S* is more influential if *f(S)* is larger**

  $$\boldsymbol{f(\{a, b\}) < f(\{a, c\}) < f(\{a, d\})}$$

Most Influential Set of Nodes

- How hard is this problem?
  - NP-Hard (How to prove it)
  - Finding most influential set of nodes is at least as hard as set cover problem (a known NP-hard problem).
- Set cover problem
  - Given universe of elements $U = \{u_1, \dots, u_n\}$ and sets $X_1, \dots, X_m \subseteq U$



  - Q: Are there *k* sets among $X_1, \dots, X_m$ such that their union is *U*?
- Goal:

  Encode set cover as an instance of $\max_{S \text{ of size } k} f(S)$

Summary So far

- **Extremely bad news**
  - **Influence Maximization is NP-hard**

- **Next, good news:**
  - **There exists an <u>approximation</u> algorithm!**
    - For some inputs the algorithm won't find globally optimal solution/set ***OPT***
    - But we will also prove that the algorithm will never do too badly either. More precisely, the algorithm will find a set ***S*** that where $\boldsymbol{f(S) \geq 0.63 * f(OPT)}$, where *OPT* is the globally optimal set.

Greedy Algorithm

Algorithm 1 Maximizing the spread of cascades – Greedy algorithm

Require: Diffusion graph $G(V, E)$, budget k

return Seed set S (set of initially activated nodes)
$i = 0$;
$S = \{\}$;
while $i \neq k$ **do**
    $v = \arg\max_{v \in V \setminus S} f(S \cup \{v\})$;
    `or equivalently` $\arg\max_{v \in V \setminus S} f(S \cup \{v\}) - f(S)$
    $S = S \cup \{v\}$;
    $i = i + 1$;
end while
Return S;