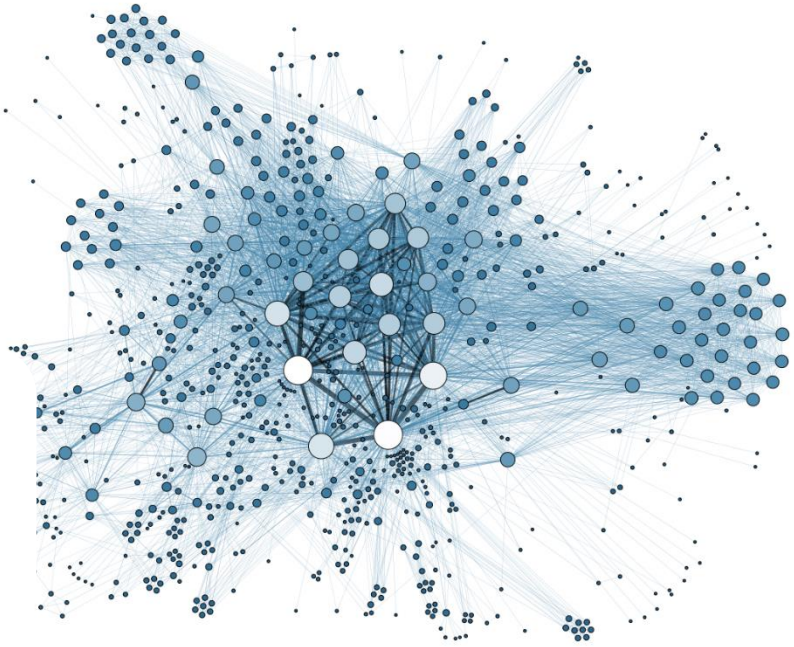
Social Media Analytics

INFS7450

Network Effects and Cascading Behavior

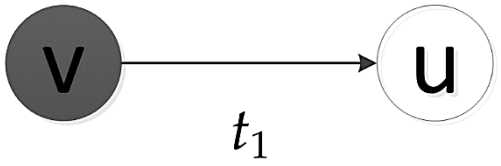


Dr. Hongzhi Yin
School of ITEE
The University of Queensland

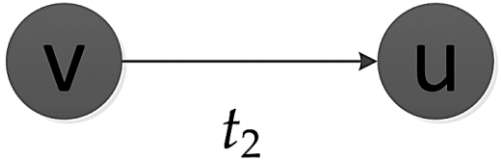
Influence Modeling

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

Linear Threshold Model (LTM)

- The influence process takes place in a network
- One widely studied model
  - Linear Threshold Model (LTM)
- Simple, yet effective methods for modeling influence in social networks
- Nodes make decision based on the influence coming from their already activated neighborhood

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

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

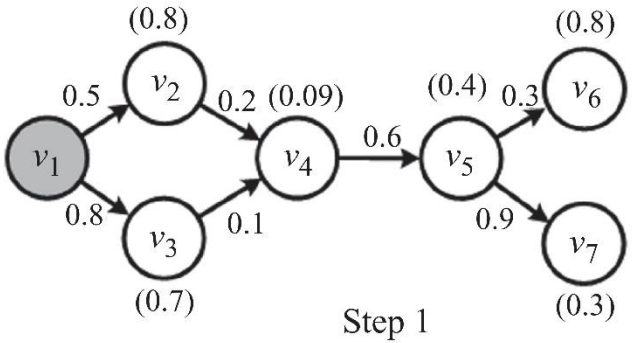
LTM Algorithm

Algorithm 1 Linear Threshold Model (LTM)

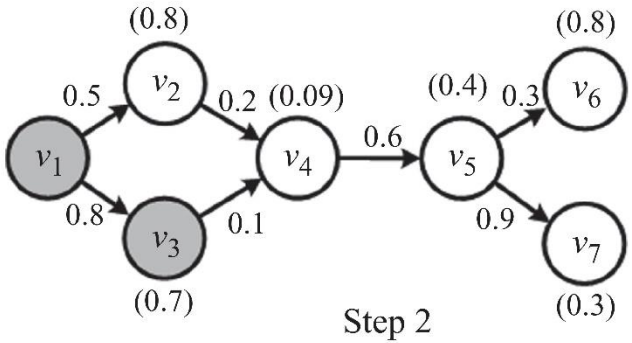
Require: Graph $G(V, E)$, set of initial activated nodes A_0

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_∞;

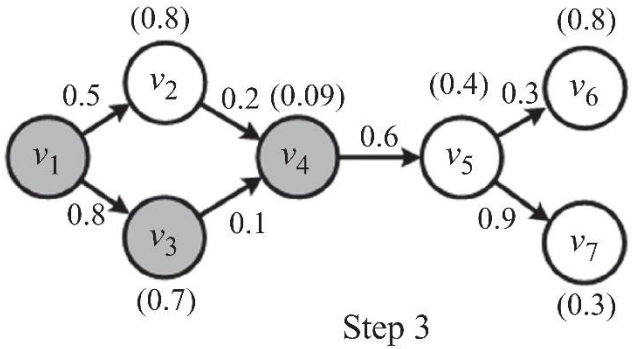
Linear Threshold Model (LTM) - An Example



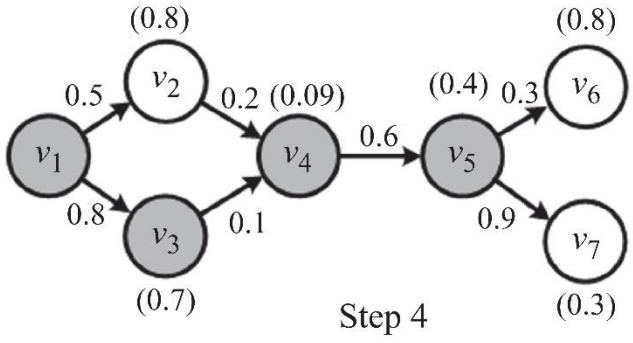
Step 1



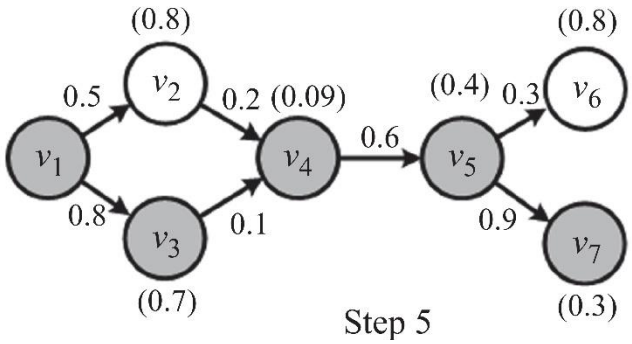
Step 2



Step 3



Step 4



Step 5

Thresholds are on top of nodes

Homophily

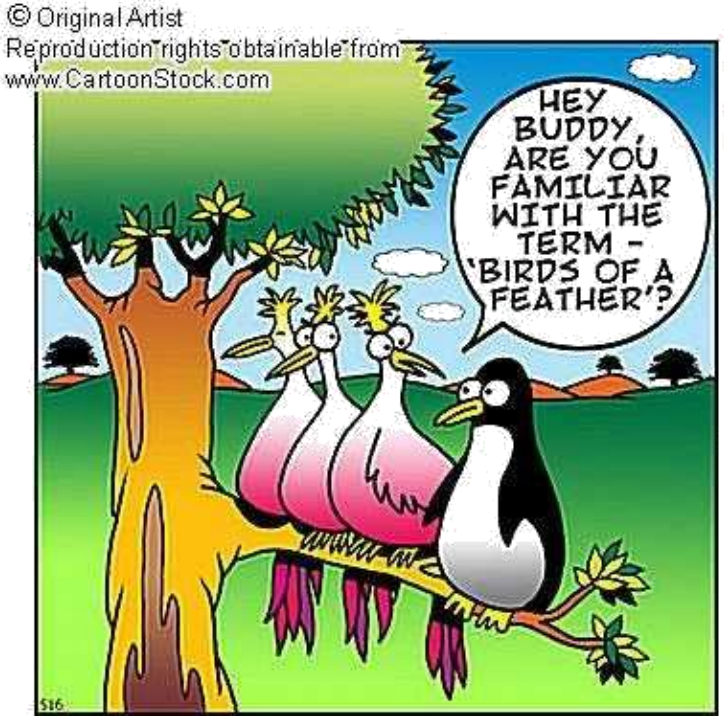
“Birds of a feather flock together”

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, 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 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 $sim(v, u)$

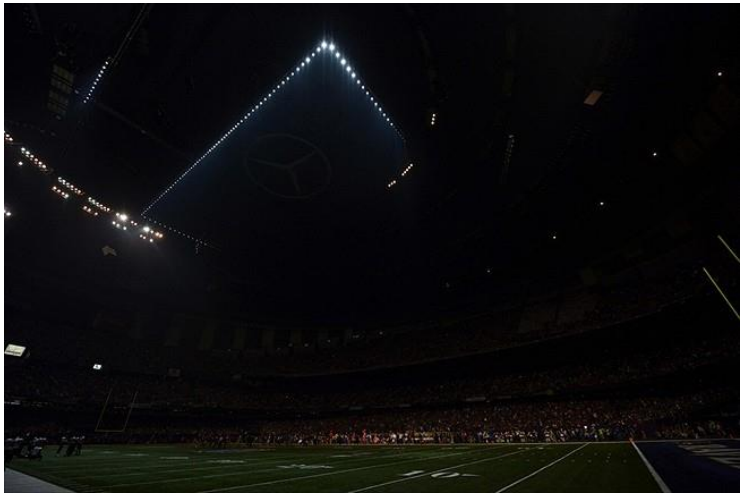
return Set of edges E
for all $v \in V$ **do**
  θ_v = generate a random number in [0,1];
  for all $(v, u) \notin E$ **do**
    if $\theta_v < sim(v, u)$ **then**
      $E = E \cup (v, u)$;
    end if
  end for
end for
Return E;

Information Diffusion

Example of Information Diffusion

In February 2013, during the third quarter of Super Bowl XLVII, a power outage stopped the game for 34 minutes.

- Oreo, a sandwich cookie company, tweeted during the outage:
  - "Power out? No Problem. You can still dunk it in the dark".
- The tweet became popular almost immediately, reaching
  - 15,000 retweets on Twitter and 20,000 likes on Facebook in less than 2 days.
- Cheap advertisement reaching a large population of individuals.
  - companies spent as much as 4 million dollars to run a 30 second ad during the super bowl on TV.



Example of **Information Diffusion**

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.

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

- Only public information is available
- Global information is available

Herd Behavior: Popular Restaurant Example

- Assume you are on a trip in a metropolitan area that you are less familiar with.
- Planning for dinner, you find restaurant **A** with excellent reviews online and decide to go there.
- When arriving at **A**, you see **A** is almost empty and restaurant **B**, which is next door and serves the same cuisine, almost full.
- Deciding to go to **B**, based on the belief that other people have also had the chance of going to **A** but finally chose **B**

Herd Behavior: Milgram's Experiment

Stanley Milgram asked one person to stand still on a busy street corner in New York City and stare straight up at the sky

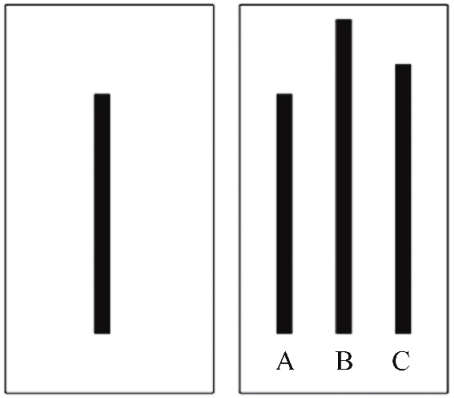
- About 4% of all passersby stopped to look up.

When 5 people stand on the sidewalk and look straight up at the sky, 20% of all passersby stopped to look up.

Finally, when a group of 18 people look up simultaneously, almost 50% of all passersby stopped to look up.

Herd Behavior: Asch's Vision Test Experiment

- Groups of students participated in a vision test
- They were shown two cards, one with a single line segment and one with 3 lines
- The participants were required to match line segments with the same length
- Each participant was put into a group where all other group members were collaborators (cappers) with Asch.



- These collaborators were introduced as participants to the subject.
- Asch found that in control groups with no pressure to conform, only 3% of the subjects provided an incorrect answer.
- However, when participants were surrounded by individuals providing an incorrect answer on purpose (i.e., experimental groups), up to 32% of the responses were incorrect.

Herding: Asch Elevator Experiment

https://www.youtube.com/watch?v=BgRoiTWkBHU

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

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

Designing a Herd Behavior Experiment

1. There needs to be a decision made.
   - In our example, it is going to a restaurant

2. Decisions need to be in a sequential order

3. Decisions are not mindless and people have private information that helps them decide

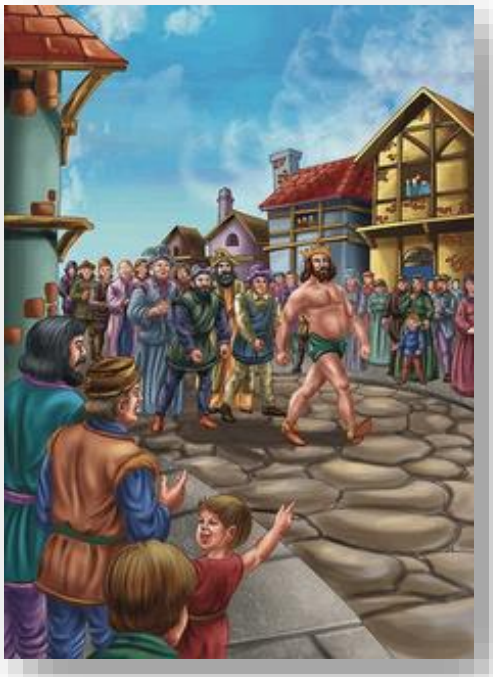
4. Message passing is not allowed. Individuals don't know the private information of others, but can infer what others know from what is observed from their behavior.

Herding Intervention

Herding: we only have access to public information

- Herding may be intervened by releasing private information which was not accessible before

The little boy in
"**The Emperor's New Clothes**"
story *intervenes* the herd by shouting
"**There is no clothe**"



Herding Intervention

To intervene the herding effect, we need one person to tell the herd that there is nothing in the sky, and the first person did this just to stop his nose bleeding

How Does Intervention Work?

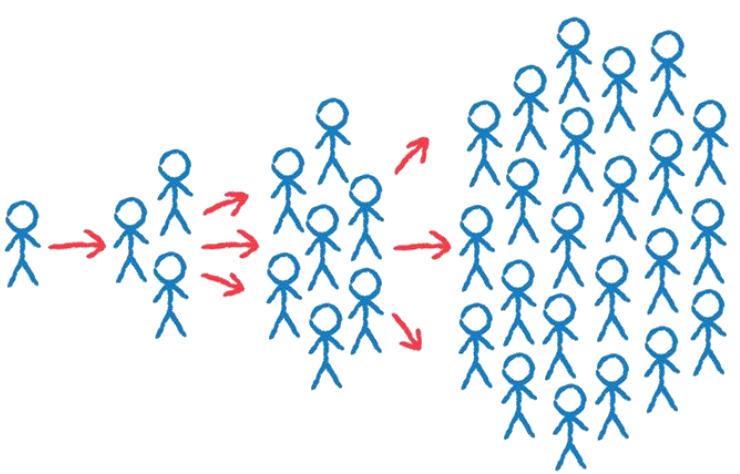
- When a new piece of private information releases,
  - The herd reevaluate their guesses and this may create completely new results
- The Emperor's New Clothes
  - When the boy gives his private observation, other people compare it with their observation and confirm it
  - This piece of information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

Information Cascade

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

Information Cascade

- Users often repost content posted by others in social networks
  - Content is often received via immediate neighbors (friends).



Information propagates through friends

An information cascade: a piece of information/decision cascaded among some users, where

individuals are connected by a network and
individuals are only observing decisions of their immediate neighbors (friends).

Cascade users have less information available

- Herding users can observe almost all others' decisions/actions

Notable example

- Between 1996/1997,
  - Hotmail was one of the first internet business to become extremely successful utilizing viral marketing
  - By inserting the tagline "*Get your free e-mail at Hotmail*" at the bottom of every e-mail sent out by its users.

- Hotmail was able to sign up **12 million users** in 18 months.

- At that time, this was the fastest growth of any user- based company.
  - By the time Hotmail reached **66 million** users, the company was establishing **270,000** new accounts each day.

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

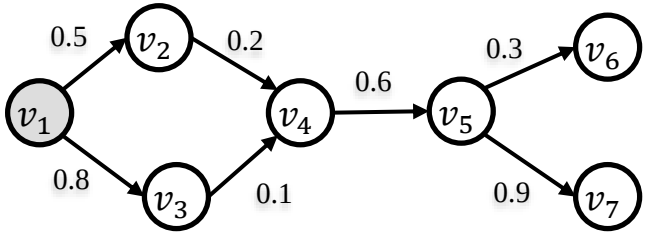
Algorithm 1 Independent Cascade Model (ICM)

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

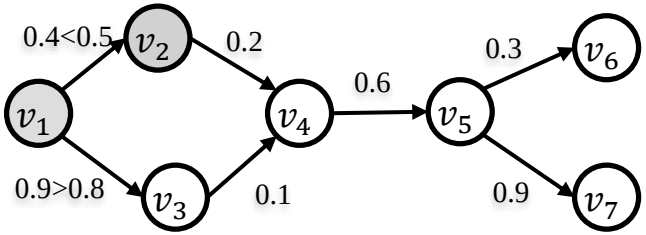
return Final set of activated nodes A_∞
i = 0;
while $A_i \neq \{\}$ **do**

  $i = i + 1$;
  $A_i = \{\}$; \\ 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_∞;

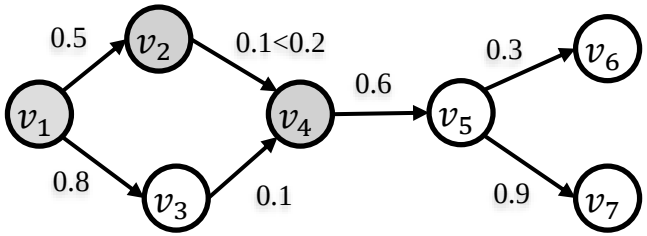
Independent Cascade Model: An Example



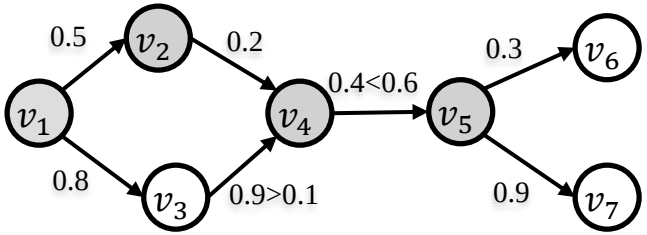
Step 1



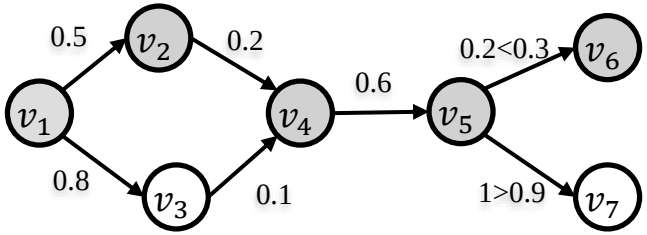
Step 2



Step 3



Step 4



Step 5

The edge weight represents the influence probability.

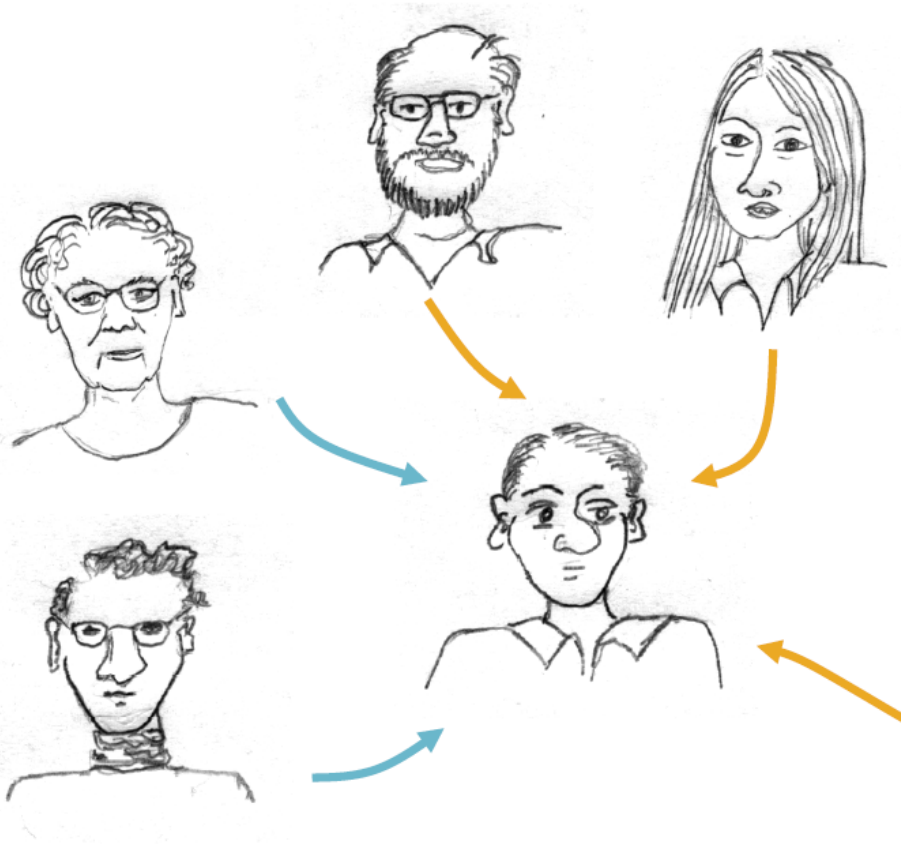
Summary of ICM

- **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**

Influence/Cascade Maximization

Viral Marketing

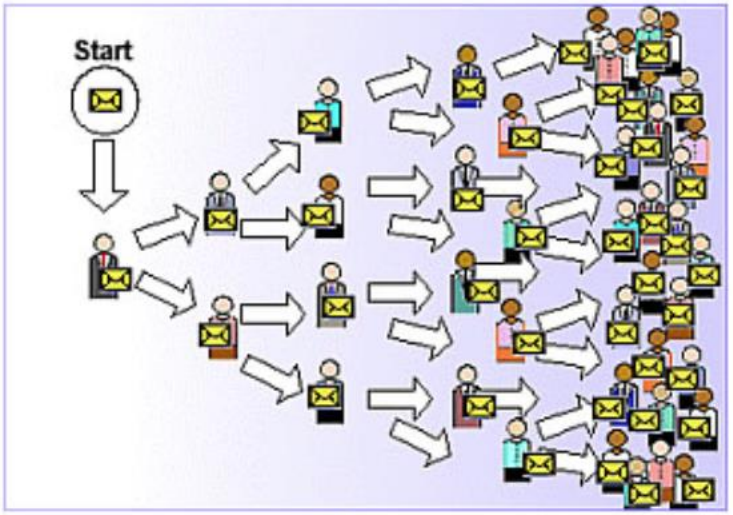
- **We are more influenced by our friends than strangers**



- 68% of consumers consult friends and family before purchasing home electronics
- 50% do research online before purchasing electronics

Viral Marketing

Identify influential customers



Convince them to adopt the product – Offer discount or free samples

These customers endorse the product among their friends

Kate Middleton Effect

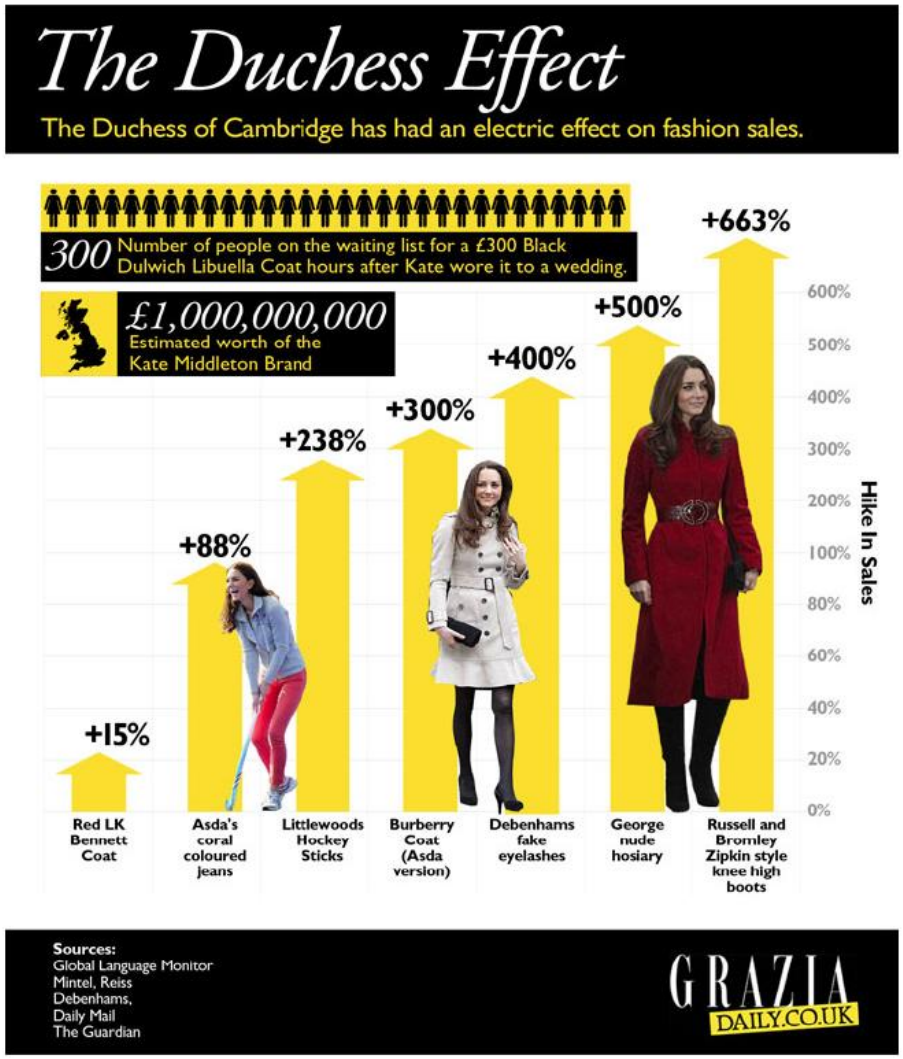
"Kate Middleton effect

The trend effect that Kate, Duchess of Cambridge has on others, from cosmetic surgery for brides, to sales of coral-colored jeans."

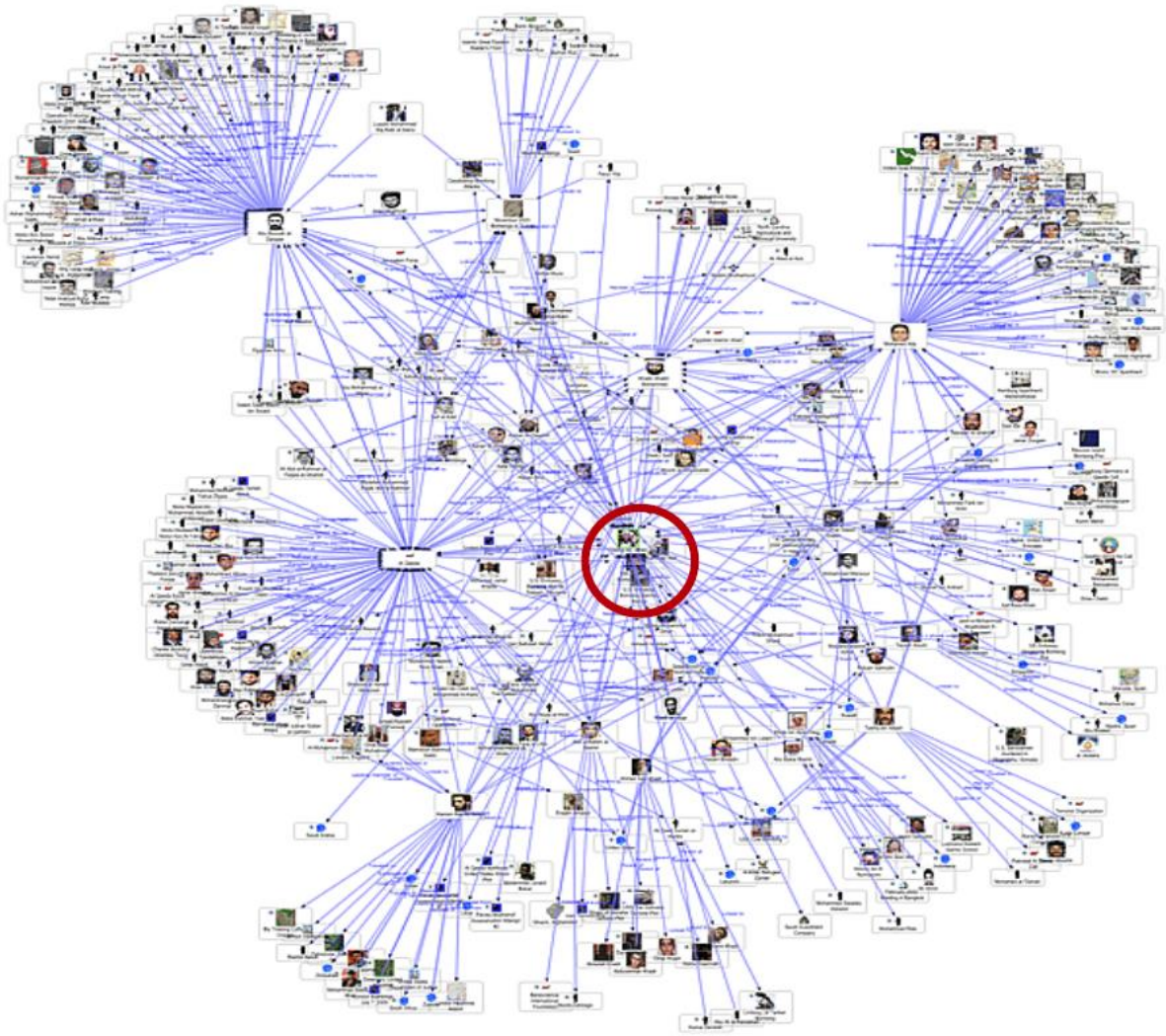
WIKIPEDIA
The Free Encyclopedia

Kate Middleton Effect

- According to Newsweek, "The Kate Effect may be worth £1 billion to the UK fashion industry."
- Tony DiMasso, L. K. Bennett's US president, stated in 2012, "...when she does wear something, it always seems to go on a waiting list."

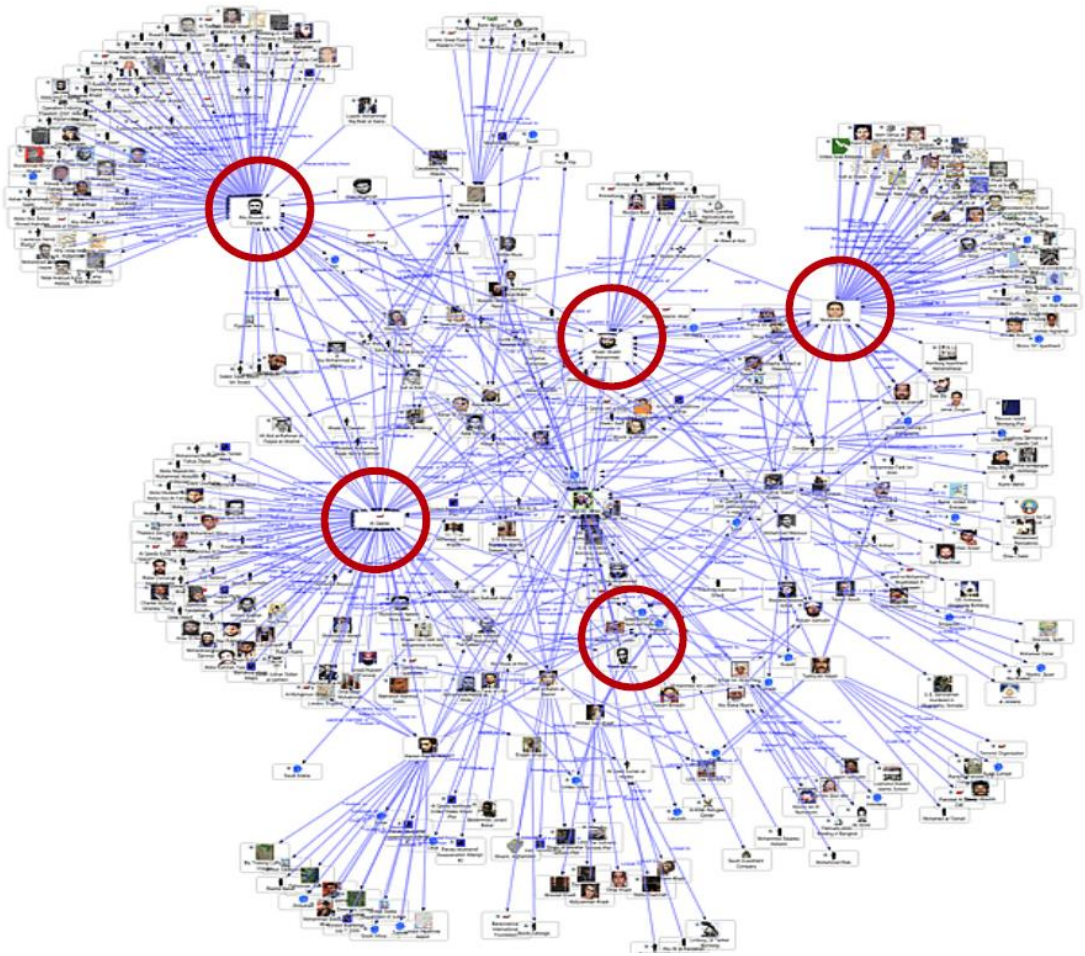


How to Find Kate-like Persons?



- Influential persons often have many friends
- Kate is one of the persons that have many friends in this social network
- For more Kates, it's not as easy as you might think!

Influence Maximization

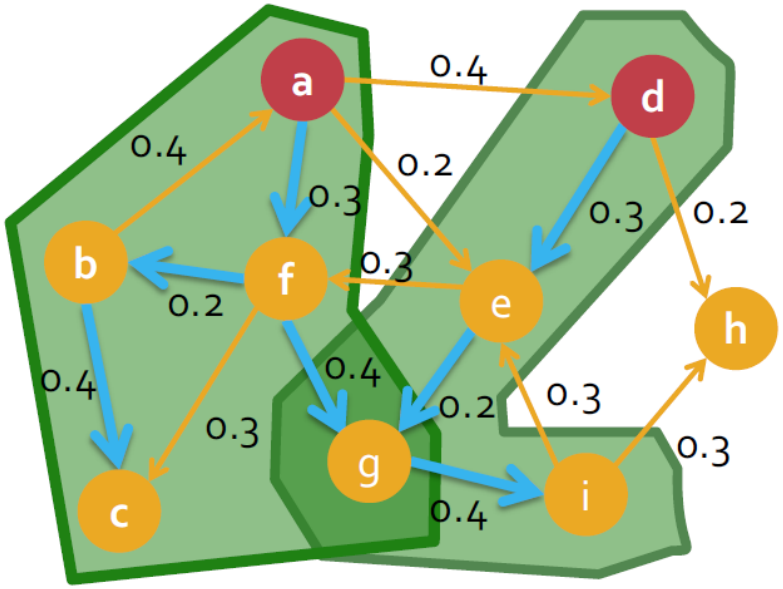


- Given a directed graph and $k>0$,
- Find k seeds (Kates) to maximize the number of influenced people (possibly in many steps)

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:** $\max_{S \text{ of size } k} f(S)$



*Influence set X_a of **a***

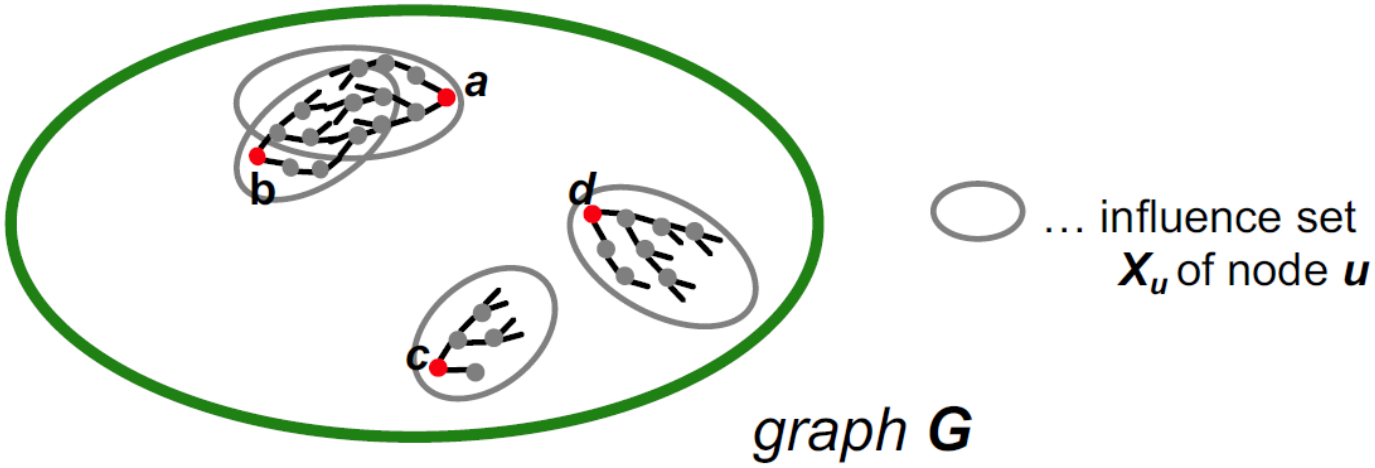
*Influence set X_d of **d***

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: $f(S) = |\cup_{u \in S} X_u|$

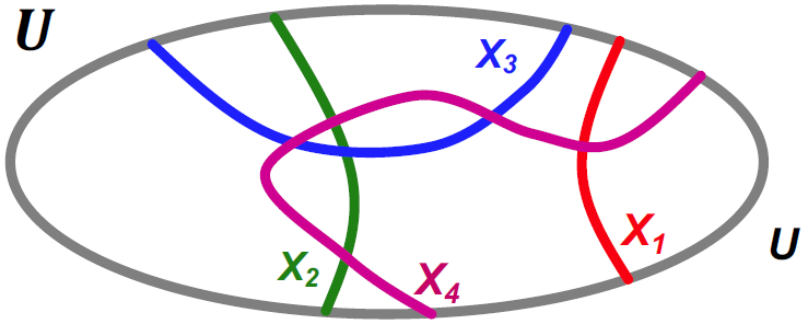


- **Set *S* is more influential if *f(S)* is larger**

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

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 $U = \{u_1, \ldots, u_n\}$ and sets $X_1, \ldots, X_m \subseteq U$



  - **Q: Are there *k* sets among $X_1, \ldots, X_m$ 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 approximation 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 $\boldsymbol{S}$ that where $\boldsymbol{f(S) \geq 0.63 * f(OPT)}$, where *OPT* is the globally optimal set.

The Approximation Algorithm

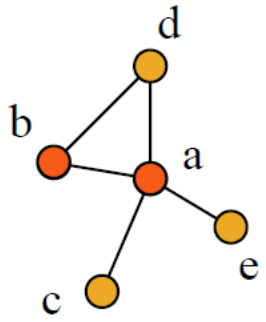
- **Consider a Greedy Hill Climbing algorithm to find *S*:**
  - **Input:**
    Influence set X_u of each node u: $X_u = \{v_1, v_2, \dots\}$
    - That is, if we activate u, nodes $\{v_1, v_2, \dots\}$ will eventually get active
  - **Algorithm:** At each iteration i activate the node u that gives largest marginal gain: $\max_u f(S_{i-1} \cup \{u\})$

S_i … Initially active set
$f(S_i)$ … Size of the union of X_u, $u \in S_i$

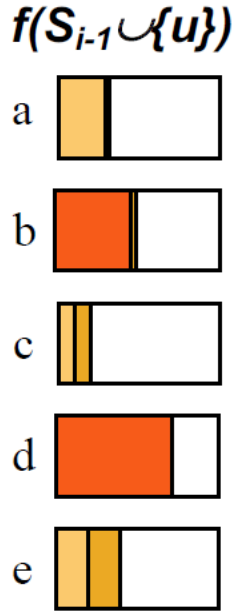
(Greedy) Hill Climbing

Algorithm:

- Start with $\boldsymbol{S_0} = \{\}$
- For $\boldsymbol{i} = \boldsymbol{1} \ldots \boldsymbol{k}$
  - Activate node $\boldsymbol{u}$ that max $\boldsymbol{f}(\boldsymbol{S_{i-1}} \cup \{\boldsymbol{u}\})$
  - Let $\boldsymbol{S_i} = \boldsymbol{S_{i-1}} \cup \{\boldsymbol{u}\}$



- **Example:**
  - Eval. $f(\{a\}), \ldots, f(\{e\})$, pick argmax of them
  - Eval. $f(\{\boldsymbol{d}, a\}), \ldots, f(\{\boldsymbol{d}, e\})$, pick argmax
  - Eval. $f(\boldsymbol{d}, \boldsymbol{b}, a\}), \ldots, f(\{\boldsymbol{d}, \boldsymbol{b}, e\})$, pick argmax



Greedy Algorithm

Algorithm 1 Maximizing the spread of cascades – Greedy algorithm

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

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

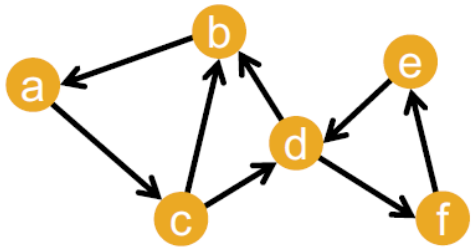
Solution Quality

- Hill climbing finds solution *S* which
  f(S) ≥ (1-1/e)*f(OPT) i.e., $f(S) \geq 0.63 * f(OPT)$

- This is a **data independent bound**
  - This is a worst case bound
  - No matter what is the input data, we know that the Hill-Climbing **will never do worse** than *0.63*f(OPT)*

Expected Influence Maximization

- **Find most influential set *S* of size *k*:** largest expected cascade size ***f(S)*** if set ***S*** is activated

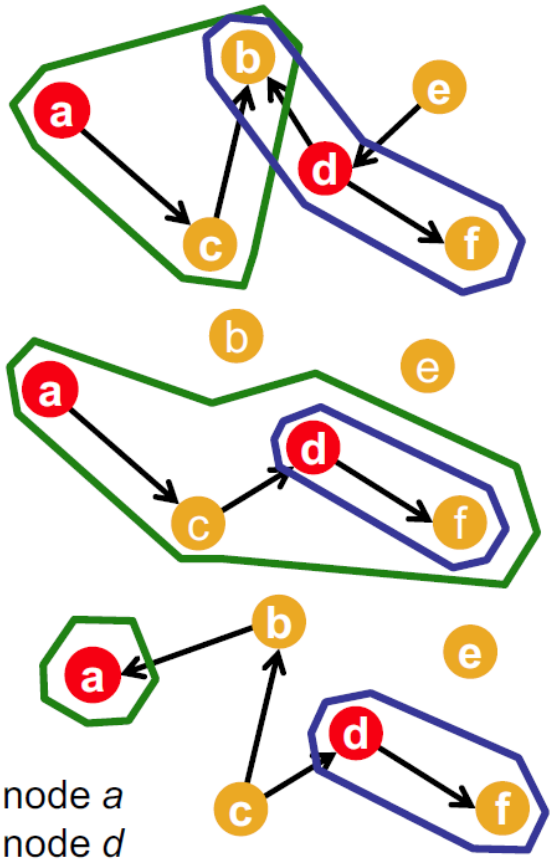


Network, each edge activates with prob. p_{uv}

Activate edges by coin flipping

Possible worlds

Multiple realizations ***i***. Each realization is a "parallel universe"



- **Want to solve:**

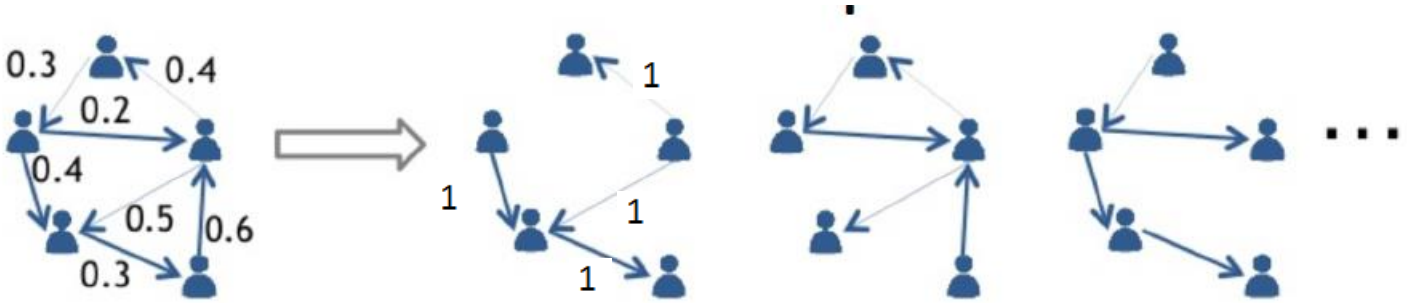
$$\arg\max_{|S|=k} f(S) = \frac{1}{|I|}\sum_{i\in I} f_i(S)$$

Consider ***S={a,d}*** then:
$f_1(S)=5$, $f_2(S)=4$, $f_3(S)=3$
and $f(S) = 1/3*(5+4+3)=4$

… influence set of node *a*
… influence set of node *d*

Generating Possible Worlds

- Generate R possible parallel worlds which are **independent** from specific seeds



- In each possible world, the activation is **deterministic** rather than **being probabilistic**.
  - A node will be **certainly** (instead of being **likely**) activated if one of its incoming neighbours is activated;

References

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