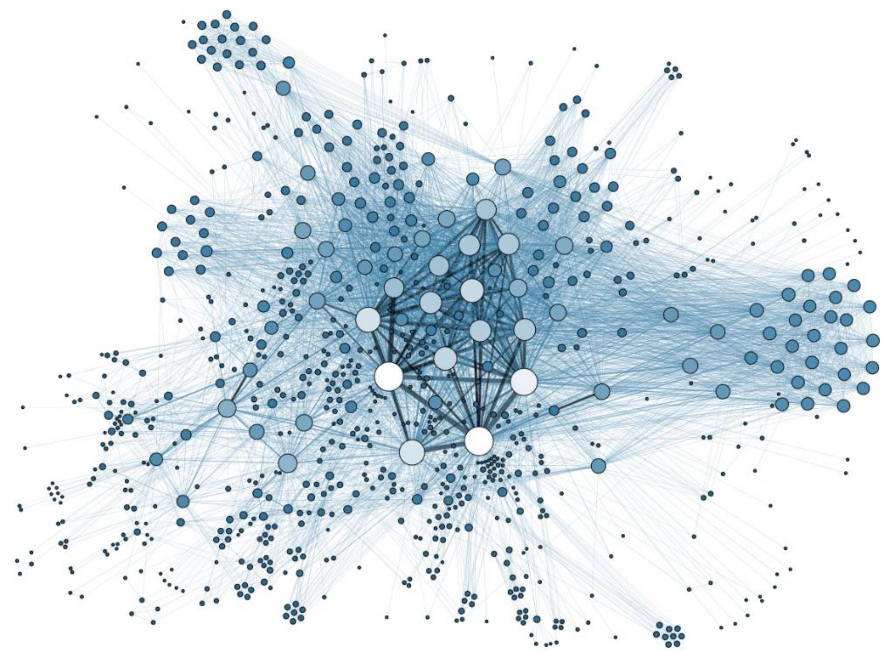
Social Media & Network Analytics



Information Diffusion

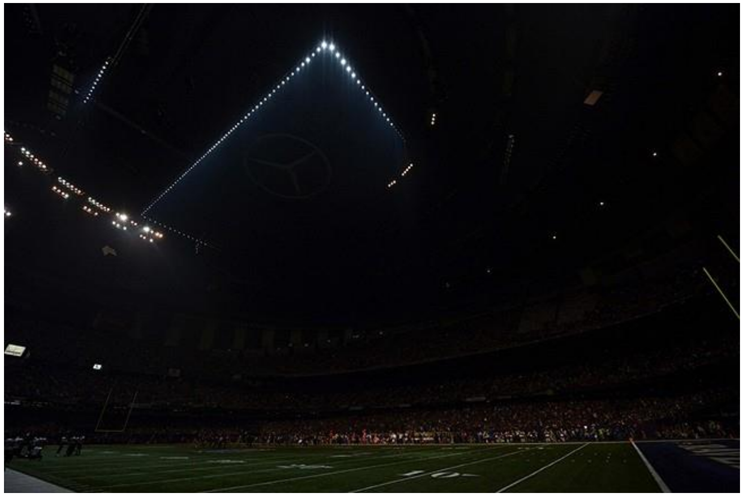
Information Diffusion Example

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 caught on almost immediately, reaching
  - 15,000 retweets 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.



Information Diffusion

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

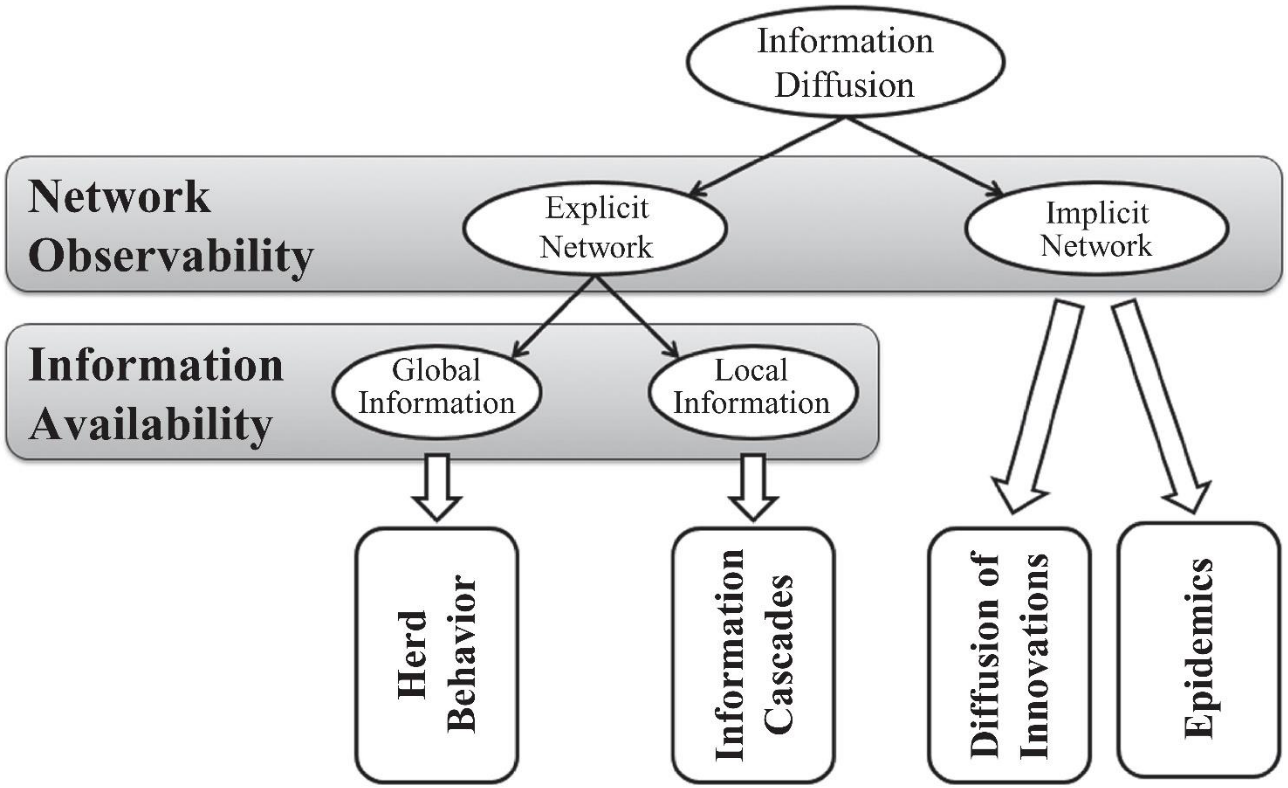
- Studied in a plethora of sciences.
  - Sociology, epidemiology, and ethnography
  - All are useful for social media analysis.

- We focus on techniques that can model information diffusion.

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 can overlap with the set of senders; and
- **Medium.** This is the medium through which the diffusion takes place. For example, when a rumor is spreading, the medium can be the personal communication between individuals

Information Diffusion Types



Herding: Asch Elevator Experiment

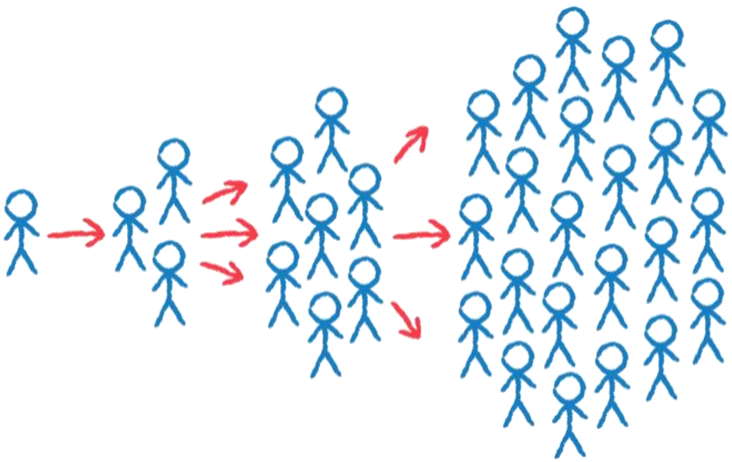
https://www.youtube.com/watch?v=BgRoiTWkBHU

Information Cascade

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

Information Cascade

- Users often repost content posted by others in the network.
  - Content is often received via immediate neighbors (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).

Examples

- Rumour spreading via social media & networks

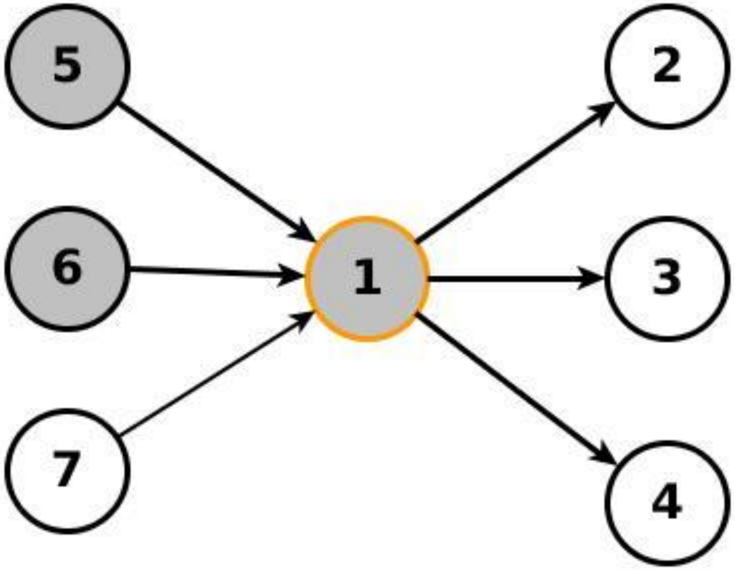
- Disaster & Evacuation Modelling

Viral Marketing Example

- Between 1996/1997,
  - Hotmail was one of the first internet business's 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 the 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.

Underlying Assumptions for Cascade Models

- The network is (typically) a directed graph.
  - Nodes are actors
  - Edges depict the communication channels between them.
- A node can only influence nodes that it is connected to
- Decisions are binary. Nodes are either:
  - **Active**: the node has adopted the behavior/innovation/decision;
  - **Inactive**



- A activated node can activate its neighboring nodes
- Activation is a progressive process, where nodes change from inactive to active, but not vice versa

Independent Cascade Model (ICM)

- **Independent Cascade Model (ICM)**
  - Sender-centric model of cascade
  - Each node has **one chance** to activate its neighbors

- In ICM, nodes that are active are senders and nodes that are being activated as receivers
  - The *linear threshold model* concentrates on the receiver (to be discussed next lecture).

ICM Algorithm

- 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 $t + 1$.
  - Only has single chance
  - Stochastic process
  - Mechanism:
    - Generate random number between 0 and 1
    - If this number randomNum < p_{vw} then w is active, otherwise not

ICM Algorithm

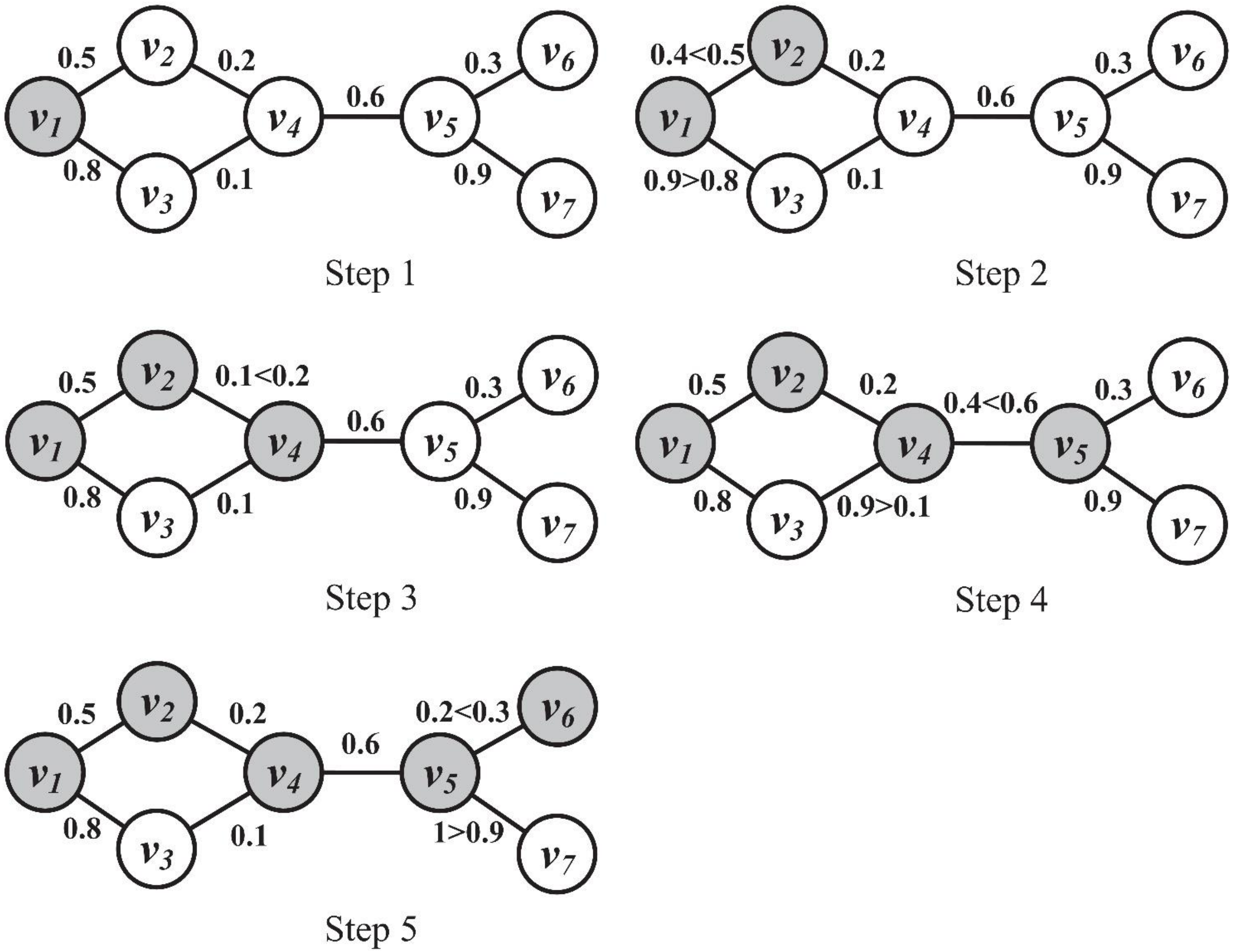
- Input:
  - Diffusion (directed) graph G(V,E), set of initially activated nodes A_0, activation probabilities for each edge $p_{v,w}$
- Output:
  - Set of activated nodes A_∞

- Steps:

```
i = 0
While A_i not empty
    A_{i+1} = empty set
    For each v in A_i
        For each neighbour w of v not active
            If rand() < p_{v,w}
                Activate w
                Add w to A_{i+1}
    i = i + 1
```

$$A_\infty = \bigcup_{j=0}^{i} A_j$$

Independent Cascade Model: An Example

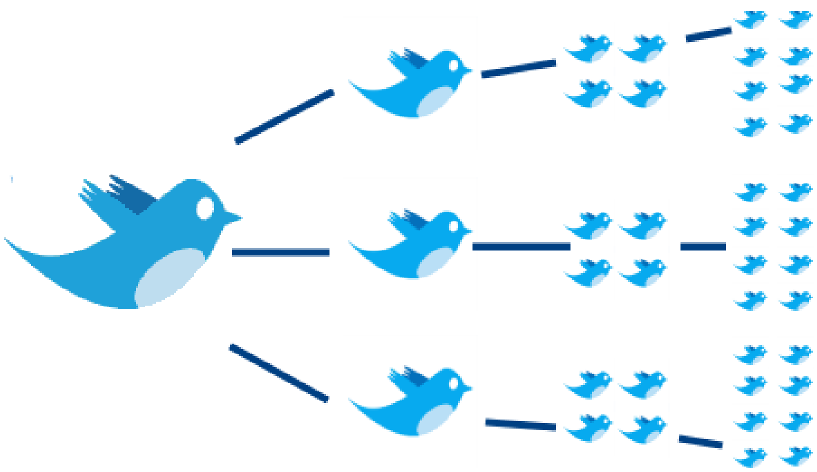


Maximizing
the Spread of Cascades

Maximizing the spread of cascades

- **Maximizing the Spread of Cascades** is the problem of finding a (small) set of nodes in a social network such that their aggregated spread in the network is maximized

- Applications
  - Product marketing
  - Influence



Problem Setting

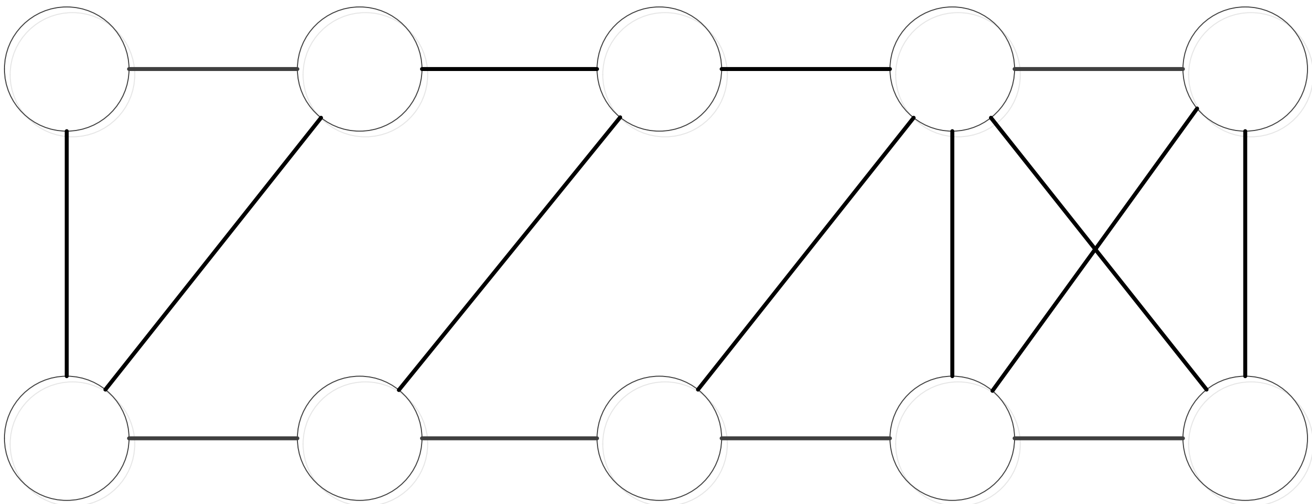
- **Given**
  - A limited budget **B** for initial advertising
    - Example: give away free samples of product
  - Estimating spread between individuals

- **Goal**
  - To trigger a large spread
    - i.e., further adoptions of a product

- **Question**
  - Which set of individuals should be targeted at the very beginning?

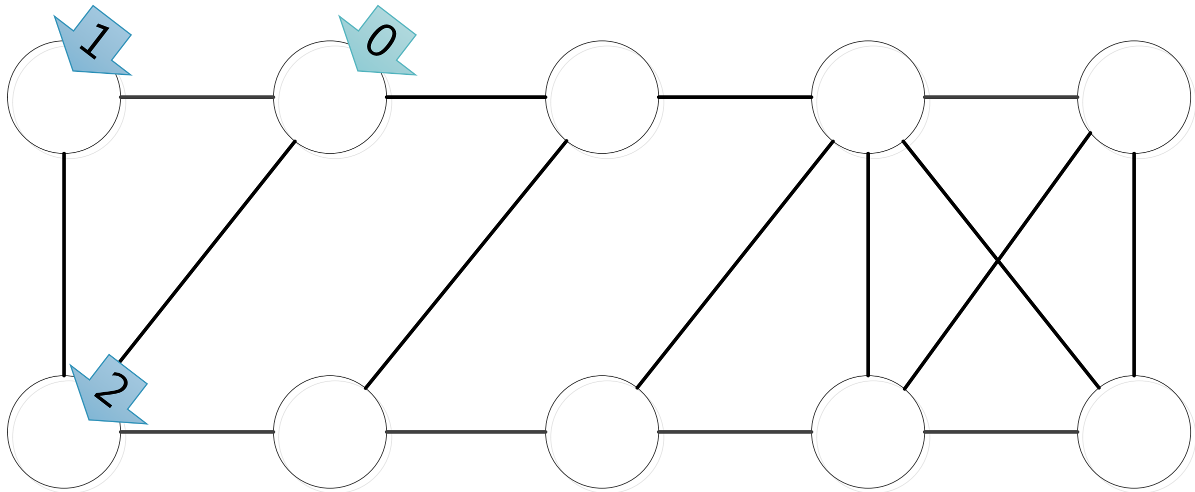
Maximizing the Spread of Cascade: Example

- We need to pick k nodes such that maximum number of nodes are activated

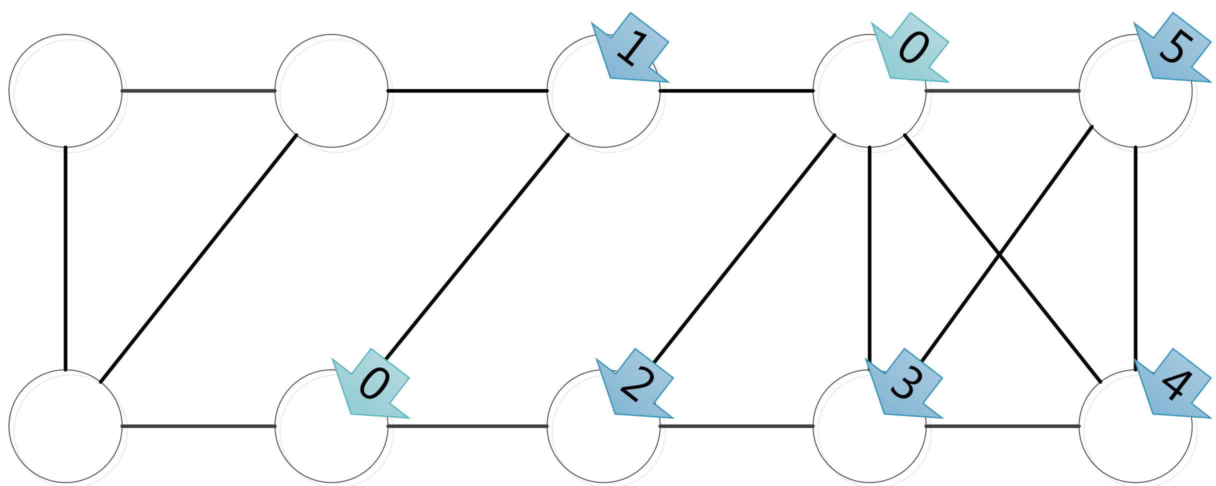


Maximizing the Spread of Cascade

Select one seed



Select two seeds



Problem Statement

- Spread of node set S: $f(S)$
  - An expected <u>number</u> of activated nodes at the end of the cascade, if set S is the initial active set

- Problem:
  - Given a parameter k (budget), find a k-node set S to maximize $f(S)$
  - A discrete combinatorial optimization problem with $f(S)$ as the objective function
  - NP-Hard problem!

f(S): Properties

1. **Non-negative**

2. **Monotone**

$$f(S+v) \geq f(S)$$

3. **Submodular**
   - Let N be a finite set
   - A set function is submodular if and only if

$$f: 2^N \to \mathbb{R}$$

f(power set of N) to Real number

$$\forall S \subset T \subset N, \forall v \in N \setminus T,$$

S is a subset of T (both subset of N)

$$f(S+v) - f(S) \geq f(T+v) - f(T)$$

Difference from adding v to S is bigger than different from adding v to T

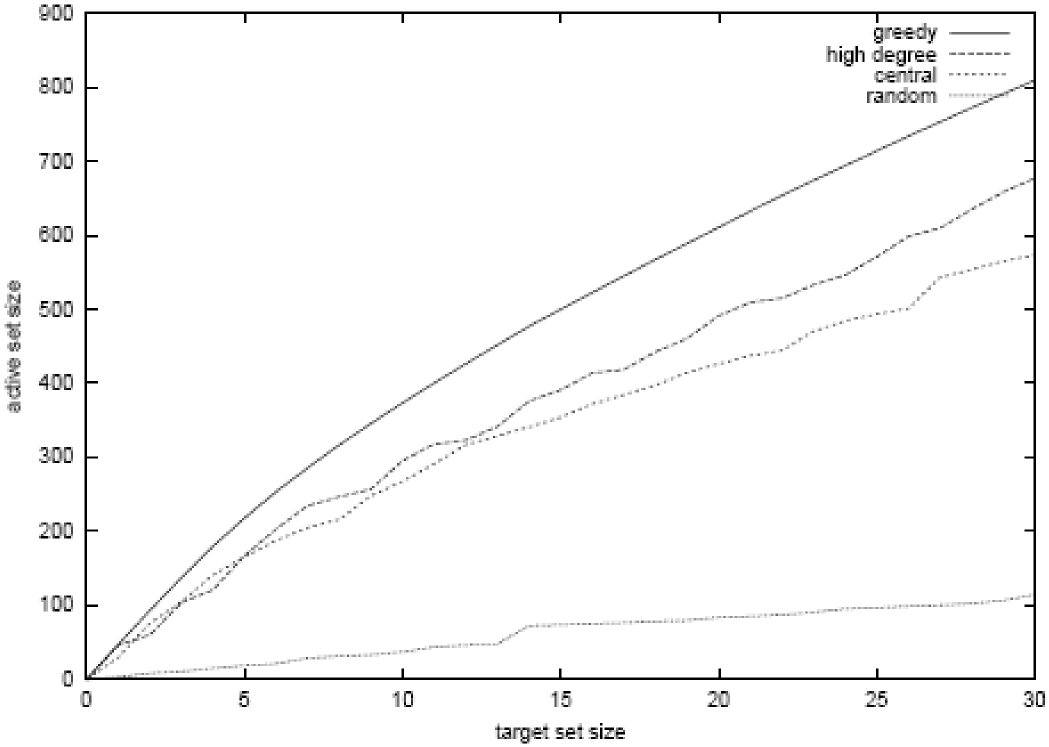
Some Facts Regarding this Problem

- **Bad News**
  - Consider a non-negative, monotone, submodular function f
    - Both estimation of Linear Threshold and ICM expected spread fall in this category
  - Finding a k-element set S for which $f(S)$ is maximized is **NP-hard**
    - It is NP-hard to determine the optimum initial set for cascade maximization
- **Good News**
  - We can use a greedy algorithm
    - Start with an empty set S
    - For k iterations:

      Add node v to S that maximizes $f(S \cup \{v\}) - f(S)$
  - We estimate f(S) by simulation, usually polynomial time in *n* (no. of vertices)

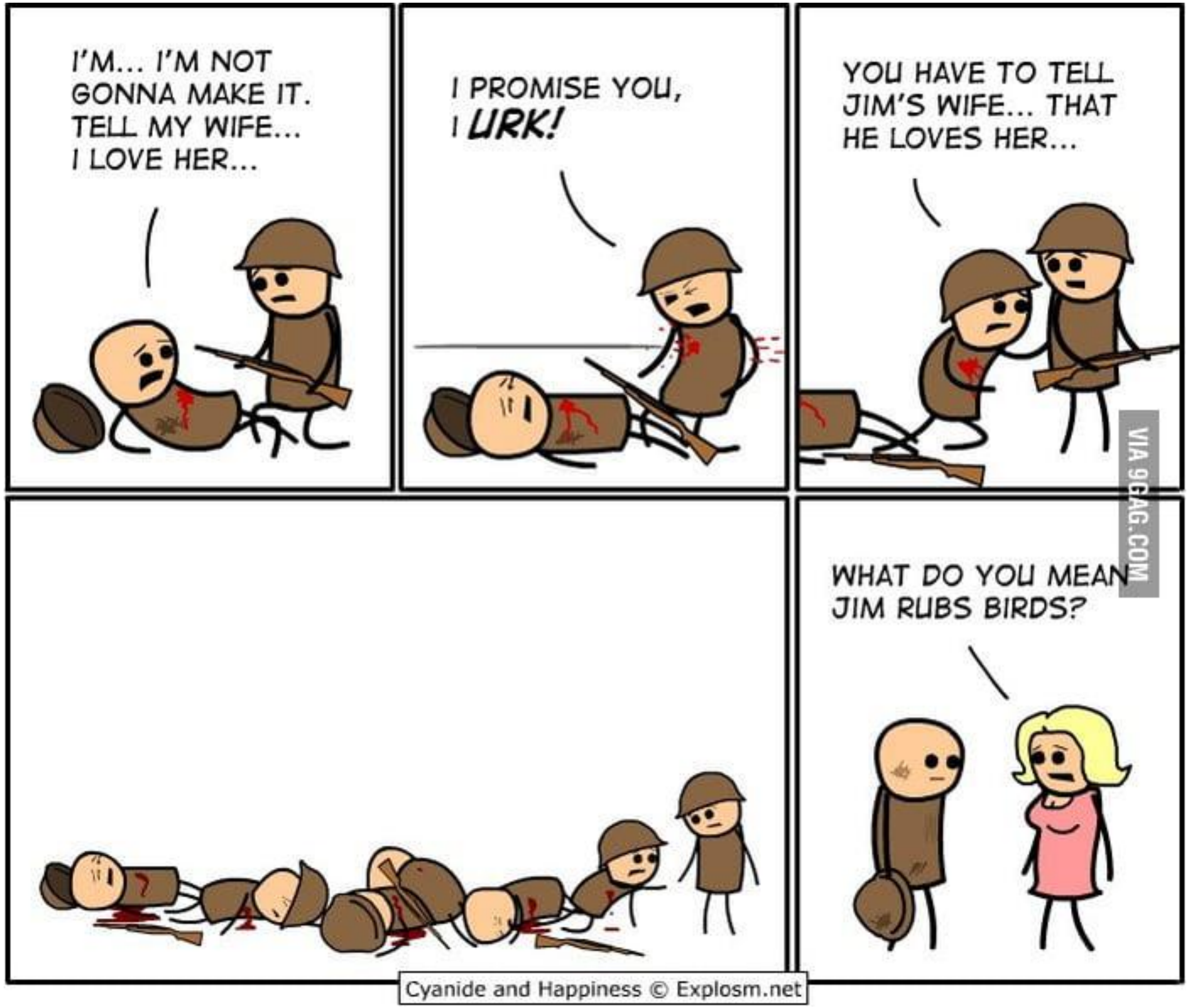
  - How good (or bad) it is? (Kempe et al.)
    - **Theorem**: the greedy algorithm provides a $(1 - 1/e)$ approximation.
    - The resulting set S activates **at least** $(1 - 1/e) \approx 63\%$ of the number of nodes that any size-k set S could activate.

Example Results



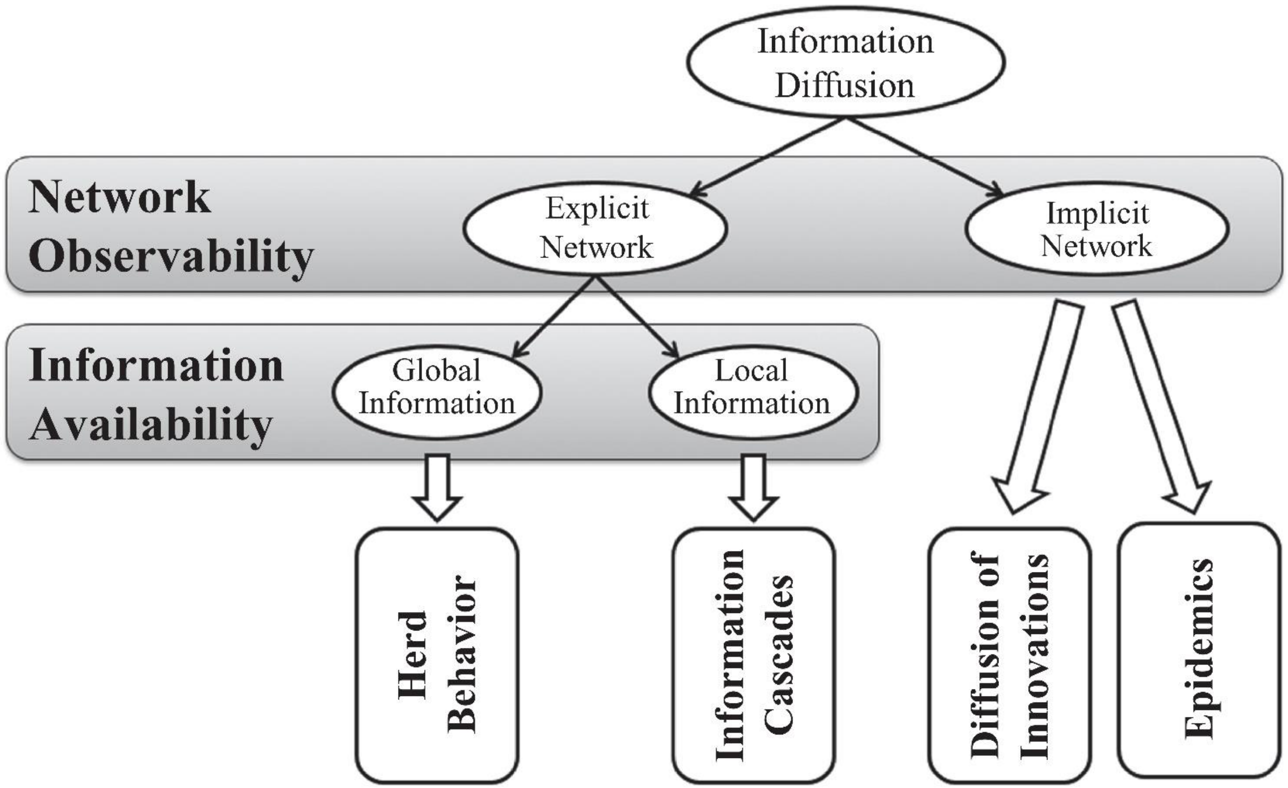
From Kempe et al.

(Irrelevant) Detour: Chinese Whispers …



https://9gag.com/gag/awrnmZ8/chinese-whispers

Information Diffusion Types



Diffusion of Innovations

- The network is <u>not observable</u>
- Only public information is observable

Diffusion of Innovation

- Information Cascade
  - Influence/diffusion network is known
- What if unknown? (Implicit network)
  - Can infer the links, subject of next set of lectures
  - Alternative: without other explicit information, assume **full mixing**
    - i.e., everyone can influence everyone else, or a influence network that is a complete graph
    - In reality it isn't likely though
    - "Macro models"
- Few example models that uses this idea
  - Epidemic models
  - Diffusion of innovation

Diffusion of Innovation

- An innovation is "*an idea, practice, or object that is perceived as new by an individual or other unit of adoption*"

- The theory of diffusion of innovations aims to answer **why** and **how** innovations spread.

- It also describes the **reasons** behind the diffusion process, as well as the rate at which ideas spread.

Innovation Characteristics

For an innovation to be adopted, the individuals adopting it (adopters) and the innovation must have certain qualities

Innovations must:

- **Be Observable**,
  - The degree to which the results of an innovation are visible to potential adopters
- **Have Relative Advantage**
  - The degree to which the innovation is perceived to be superior to current practice
- **Be Compatible**
  - The degree to which the innovation is perceived to be consistent with socio-cultural values, previous ideas, and/or perceived needs
- **Have Trialability**
  - The degree to which the innovation can be experienced on a limited basis
- **Not be Complex**
  - The degree to which an innovation is difficult to use or understand.

Diffusion of Innovations Models

- First model was introduced by Gabriel Tarde in the early 20th century

I. The Iowa Study of Hybrid Corn Seed

- Ryan and Gross studied the adoption of hybrid seed corn by farmers in Iowa
  - The hybrid corn was highly resistant to diseases and other catastrophes such as droughts

- Despite the fact that the use of new seed could lead to an increase in quality and production, the adoption by Iowa farmers was slow
  - Farmers did not adopt it due to its high price and its inability to reproduce
    - i.e., new seeds have to be purchased from the seed provider

I. The Iowa Study of Hybrid Corn Seed

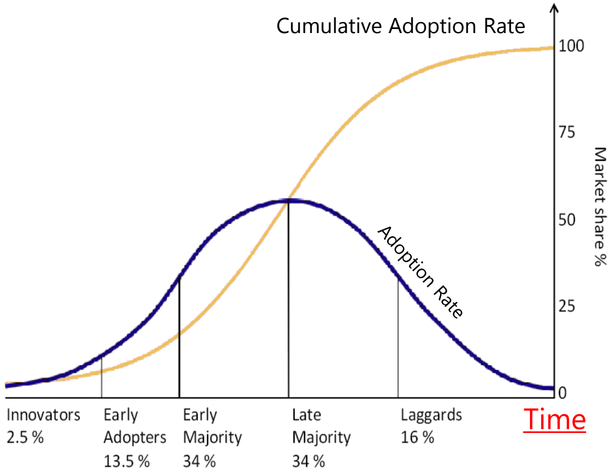
Farmers received information through two main channels

- **Mass communications** from companies selling the seeds (i.e., information)
- **Interpersonal communications** with other farmers. (i.e., influence)

- Adoption depended on a combination of information and influence.

- The study showed that the adoption rate follows an S-shaped curve and that there are 5 different types of adopters based on the order that they adopt the innovations, namely:

1) **Innovators** (top **2.5%**)
2) **Early Adopters** (next **13.5%**)
3) **Early Majority** (next **34%**)
4) **Late Majority** (next **34%**)
5) **Laggards** (last **16%**)



Modeling Diffusion of Innovations

The diffusion of innovations models can be concretely described as

$$\frac{dA(t)}{dt} = i(t)[P - A(t)]$$

- $A(t)$ is the total population that adopted the innovation
- $i(t)$ denotes the coefficient of diffusion corresponding to the innovativeness and influences of the product being adopted
- P is the total number of potential adopters (till time t)

- The rate depends on how innovative the product is and those yet to adopt the innovation
- The rate affects the potential adopters that have not yet adopted the product.

Modeling Diffusion of Innovations

We can rewrite $A(t)$ as

$$A(t) = \int_{t_0}^{t} a(t)dt \quad \cdots\!\!\rightarrow \quad \text{Let } A_0 = A(0)$$

The adopters at time t

We can define the diffusion coefficient $i(t)$ as a function of number of adopters $A(t)$

$$i(t) = \alpha + \alpha_0 A_0 + \ldots + \alpha_t A(t) = \alpha + \sum_{i=t_0}^{t} \alpha_i A(i)$$

We can simplify this linear combination

Diffusion Models

Three models of diffusion:
i.e., each having different ways to compute i(t):

$$\frac{dA(t)}{dt} = i(t)[P - A(t)]$$

$$i(t) = \alpha, \qquad \text{External-Influence Model}$$
$$i(t) = \beta A(t), \qquad \text{Internal-Influence Model}$$
$$i(t) = \alpha + \beta A(t). \qquad \text{Mixed-Influence Model}$$

- α **- Innovativeness factor of the product**
- β **- Imitation factor**

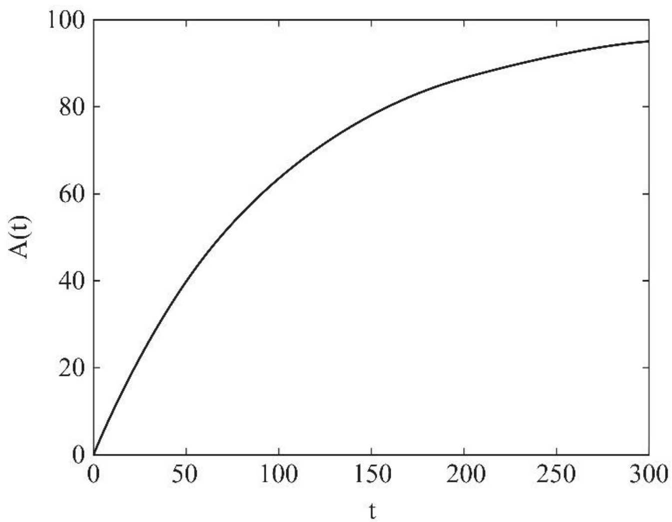
1. External-Influence Model

The adoption rate is a function that depends on external entities, $i(t) = \alpha$

- Assuming $A(t = t_0 = 0) = 0$

$$\frac{dA(t)}{dt} = \alpha[P - A(t)] \quad \Rightarrow \quad A(t) = P(1 - e^{-\alpha t})$$

The number of adopters increases exponentially and then saturates near *P*



Simulation for $P = 100$ and $\alpha = 0.01$

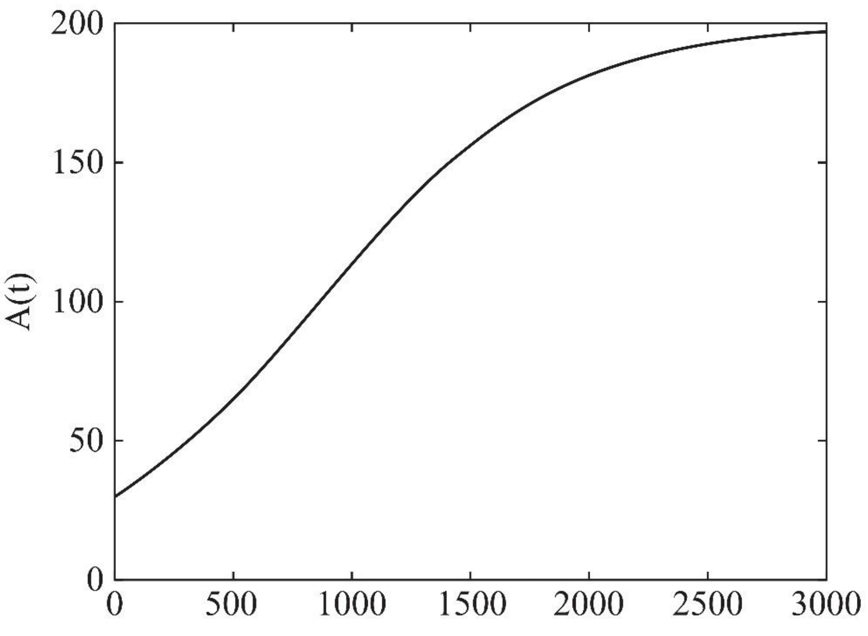
2. Internal-Influence Model (Pure imitation Model)

The adoption rate is a function that depends only on the number of already activated individuals

- $i(t) = \beta A(t)$

$$\frac{dA(t)}{dt} = \beta A(t)[P - A(t)]$$

$$A(t) = \frac{P}{1 + \frac{P - A_0}{A_0} e^{-\beta P(t - t_0)}}$$

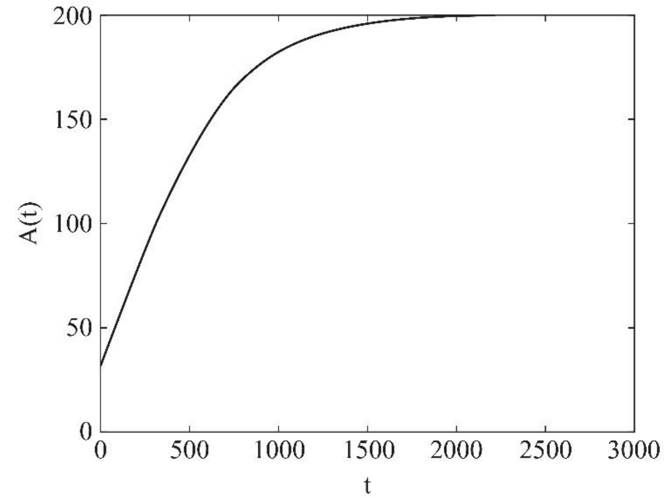


Simulation for $A_0 = 30$, $P = 200$ and $\beta = 0.00001$

3. Mixed-Influence Model

The adoption rate is a function that depends on both the number of already activated individuals and external forces,

- $i(t) = \alpha + \beta A(t)$



$$\frac{dA(t)}{dt} = \alpha + \beta A(t)A(t)[P - A(t)]$$

$$\Downarrow$$

$$A(t) = \frac{P - \frac{\alpha(P-A_0)}{\alpha+\beta A_0} e^{-(\alpha+\beta P)(t-t_0)}}{1 + \frac{\beta(P-A_0)}{\alpha+\beta A_0} e^{-(\alpha+\beta P)(t-t_0)}}$$

Simulation for
$\boldsymbol{P = 200}$
$\boldsymbol{A_0 = 30}$**,**
$\beta = \mathbf{0.00001}$ **and** $\alpha = \mathbf{0.001}$

Summary

- Information Diffusion
- Information cascade
  - Influence maximisation
- Diffusion of Innovations
  - 3 models