中国科学院软件研究所
Institute of Software, Chinese Academy of Sciences

Big Data on Social Media Mining and Analytics -Information Diffusion in Social Media

April 17, 2015

张晨

Examples: A Successful Product Placement

- 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 nearly 15,000 retweets and 20,000 likes on Facebook in less than 2 days.

- **A simple tweet diffused into a large population of individuals**.
- It helped Oreo gain fame with minimum budget in an environment where companies spent as much as 4 million dollars to run a 30 second ad. during the super bowl.

"且行且珍惜"

吃饭虽易，减肥不易，且吃且珍惜

学习虽易，考试不易，且学且珍惜

创业虽易，赚钱不易，且行且珍惜

网站虽易，赚钱不易，且行且珍惜

读研虽易，就业不易，且行且珍惜

失业虽易，求职不易，且行且珍惜

拜师虽易，情分难分，且行且珍惜

创业虽易，坚守不易，且行且珍惜

醒来虽易，起来不易，且行且珍惜

选择虽易，坚持不易，且行且珍惜

失业虽易，就业不易，且行且珍惜

生存虽易，生活不易，且行且珍惜

上课虽易，听懂不易，且行且珍惜

Definition

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

- Information diffusion is studied in a plethora of sciences.
- We discuss methods from fields such as sociology, epidemiology, and ethnography, which can help social media mining.

- Our focus is 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

- **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

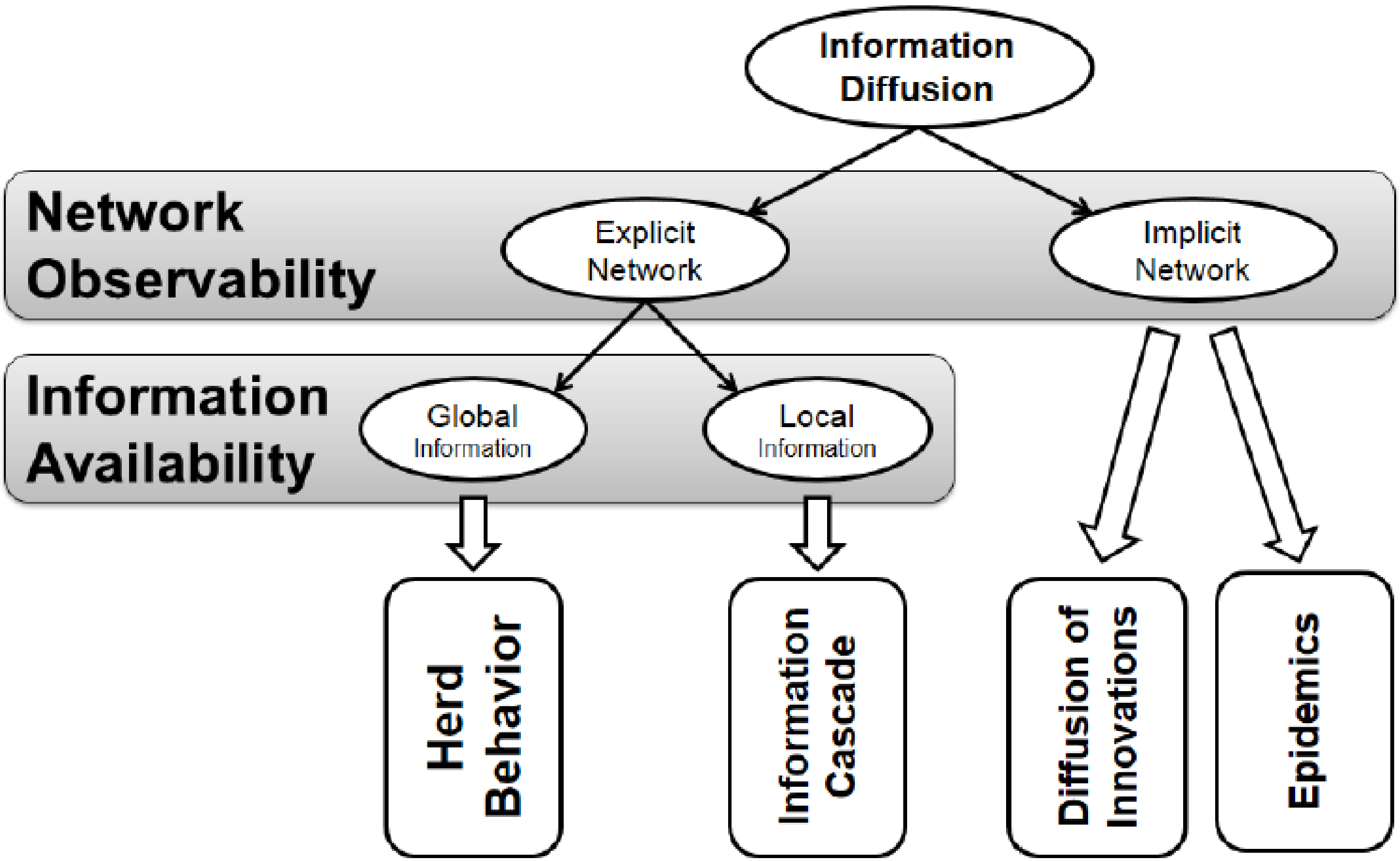
Information Diffusion

- Individuals in online social networks are situated in a network where they interact with others
- Individuals facilitate information diffusion by making individual decisions that allow information to flow
- They can make this decision
  - either independently (depending on their own knowledge/ experience)
  - or dependently (depending on the information they receive from others)
    - local dependence: immediate neighbors (friends)
    - global dependence: all individuals in the network

Information Diffusion: Intervention

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

Information Diffusion Types



Herd Behavior

- **Network is observable**
- **Only public information is available (global dependence)**
- **Private information is unavailable**

Herd Behavior Example

- Consider people participating in an online auction.
- individuals can observe the behavior of others by monitoring the bids that are being placed on different items.
- Individuals are connected via the auction's site where they can not only observe the bidding behaviors of others, but can also often view profiles of others to get a feel for their reputation and expertise.
- In these online auctions, it is common to observe individuals participating actively in auctions, where the item being sold might otherwise be considered unpopular.
- This is due to individuals trusting others and assuming that the high number of bids that the item has received is a strong signal of its value. In this case, Herd Behavior has taken place.

Herd Behavior: 11-11 Singles' Day

2009-11-11

2011-11-11

Herd Behavior: Rumor Spreading

抢盐帝：因日本核辐射事件引起的中国谣“盐”风波中，大量囤积食盐的一位武汉市民郭先生，他2011年3月17日花高价购入食盐13000斤。

Herd Behavior: Popular Restaurant Experiment

- 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 diners have also had the chance of going to **A**, is an example of herd behavior

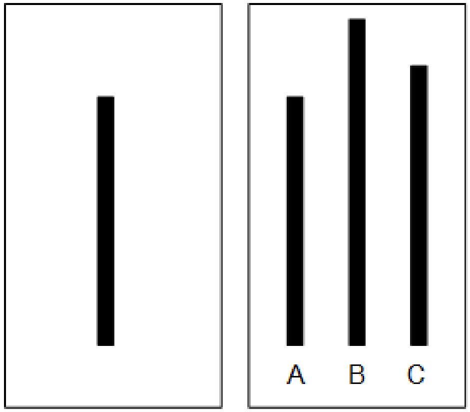
Herd Behavior: Milgram's Experiment

- Stanley Milgram asked one person to stand still on a busy street corner in New York City and stare straight up at the sky,
  - 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,
  - 50% of all passersby stopped to look up.

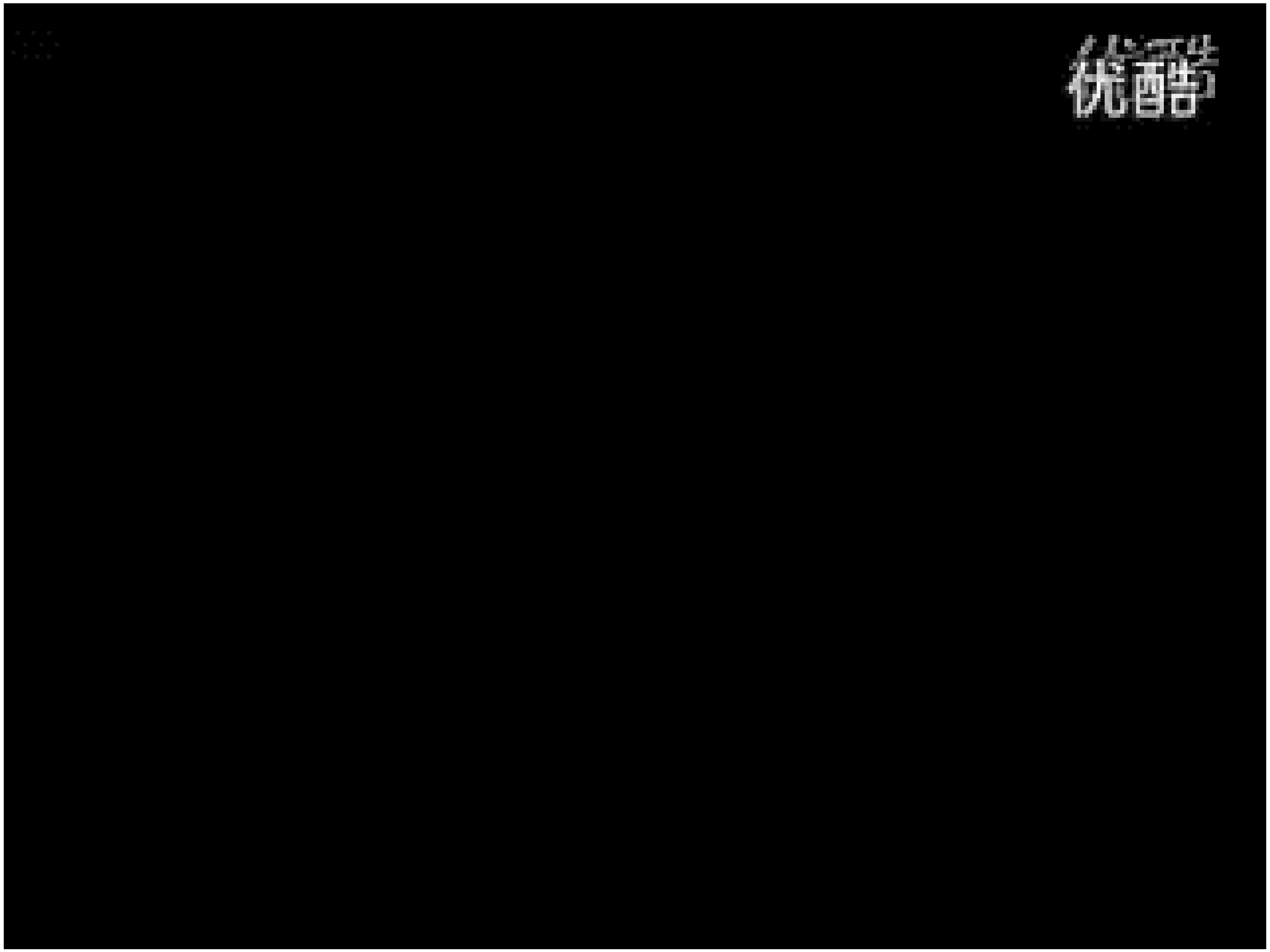
Herd Behavior: Solomom Asch's Experiment

- In one experiment, he asked groups of students to participate in a vision test where they were shown two cards, one with a single line segment and one with 3 lines, and 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 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, up to **32% of the responses were incorrect**.



Herding: Elevator Example



Herd Behavior

Herd behavior describes when a group of individuals performs actions that are highly correlated without any plans

Main Components of Herd Behavior

- A connection between individuals
- A method to transfer behavior among individuals or to observe their behavior

Examples of Herd Behavior

- Flocks, herds of animals, and humans during sporting events, demonstrations, and religious gatherings

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 <u>*complete graph*</u> where nodes can observe at least most other nodes and they can only observe public information
  – For example, they can see the crowd

Designing a Herd Behavior Experiment

- There needs to be a decision made;
  - Choose a restaurant, bid an item…
- Decisions need to be in sequential order;
- Decisions are not mindless and people have private information (observation) that helps them decide;
- Individuals don’t know the private information of others, but can infer what others know from what is observed from their behavior (Decisions, public information).

Anderson and Holt: Urn Experiment

- There is an urn in a large class with three marbles in it

50% 50%

- During the experiment, each student comes to the urn, picks one marble, and <u>**checks its color in private**</u>.
- The student <u>**predicts majority blue or red**</u>, writes her <u>**prediction on the blackboard**</u>, and puts the marble back in the urn.
- Students can't see the color of the marble taken out and can only see the predictions made by different students regarding the majority color on the board

Urn Experiment: First and Second Student

- First Student:
  – *Board: -*
    - Observed: B → Guess: B

    -or-
    - Observed: R → Guess: R
- Second Student:
  – *Board: B*
    - Observed: B → Guess: B

    -or-
    - Observed: R → Guess: R/B (flip a coin)

Urn Experiment: Third Student

- *If board: B, R*
  - Observed: B → Guess: B, or
  - Observed: R → Guess: R
- *If board: B, B*
  - Observed: B → Guess: B, or
  - **Observed: R → Guess: B (Herding Behavior)**

The forth student and onward

  - Board: B,B,B
  - **Observed: B/R → Guess: B (Herding Behavior)**

Bayes's Rule in the Herding Experiment

$$P(Y_i \mid X) = \frac{P(Y_i)P(X \mid Y_i)}{P(X)} = \frac{P(Y_i)P(X \mid Y_i)}{\sum_{j=1}^{n} P(X \mid Y_j)P(Y_j)}$$

Prediction/Decision External/Internal Information

Bayesian modeling is an effective technique for demonstrating why this herd behavior occurs. Simply put, computing conditional probabilities and selecting the most probable majority color result in herding over time.

Bayes's Rule in the Herding Experiment

Each student tries to estimate the conditional probability that the urn is majority-blue or majority-red, given what she has seen

- She would guess majority-blue if:

  Pr[majority-blue | what she has seen] > 1/2

- She would guess majority-red if:

  Pr[majority-red | what she has seen] > 1/2

  $P(Y_i \mid X)$

- From the setup of the experiment we know:

  Pr[majority-blue] = Pr[majority-red]=1/2

  $P(Y_i)$

  Pr[blue|majority-blue] = Pr[red|majority-red]=2/3

  Pr[red|majority-blue] = Pr[blue|majority-red]=1/3

  $P(X \mid Y_i)$

Bayes's Rule in the Herding Experiment

$$P(Y_i \mid X) = \frac{P(Y_i)P(X \mid Y_i)}{P(X)} = \frac{P(Y_i)P(X \mid Y_i)}{\sum_{j=1}^{n} P(X \mid Y_j)P(Y_j)}$$

Pr[majority-blue|blue] = Pr[blue|majority-blue] * Pr[majority-blue] / Pr[blue]

Pr[blue] = Pr[blue|majority-blue] * Pr[majority-blue]
+ Pr[blue|majority-red] * Pr[majority-red]
= 2/3 * 1/2 + 1/3 * 1/2 = 1/2

Pr[majority-blue|blue] = (2/3 * ½)/(1/2)=2/3

- So the first student should guess "**blue**" when she sees "**blue**"
- The same calculation holds for the second student

Bayes's Rule in the Herding Experiment: Third Student

Pr[majority-blue|blue, blue, red] =
Pr[blue, blue, red|majority-blue] * Pr[majority-blue] /
Pr[blue, blue, red]

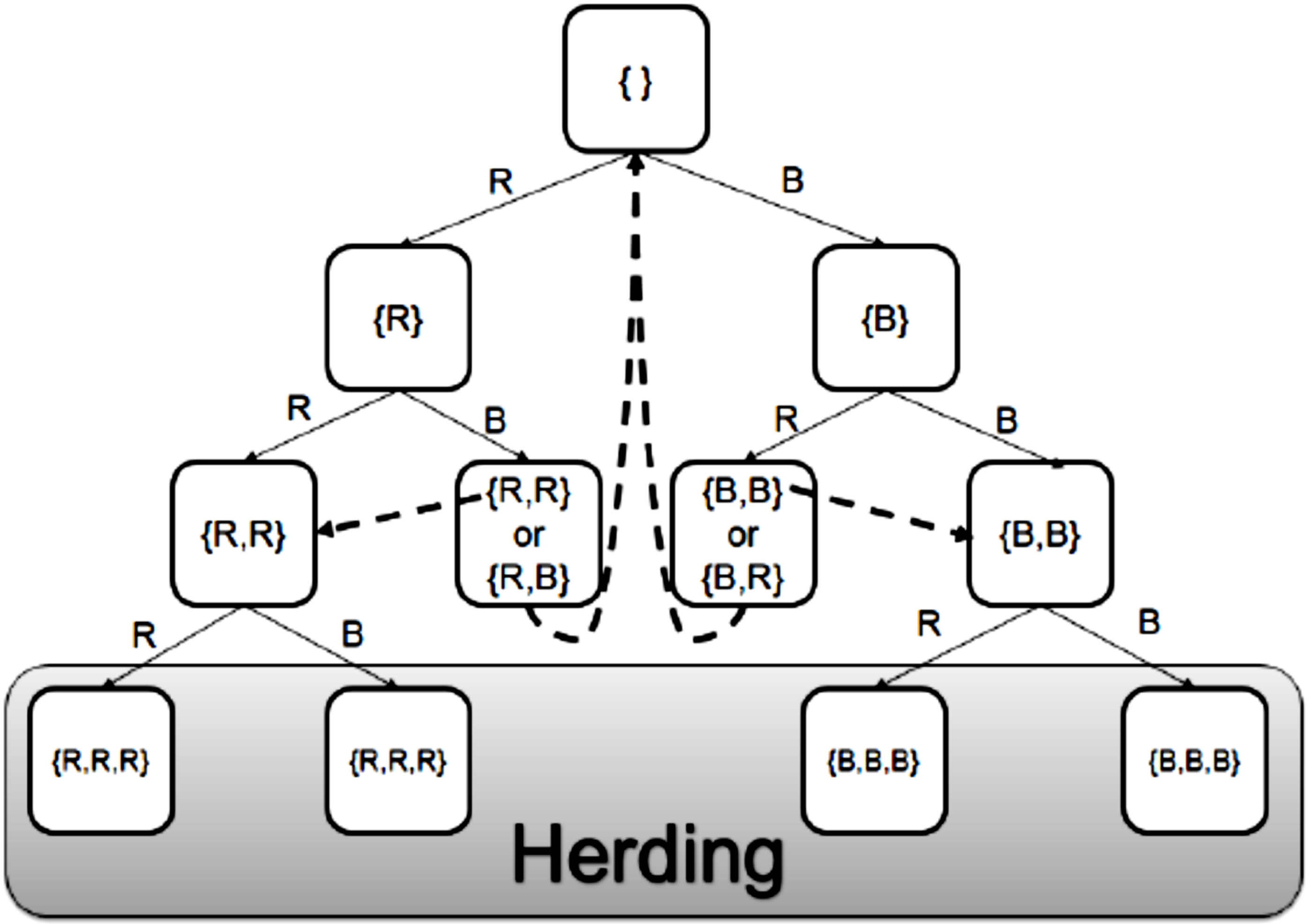
Pr[blue, blue, red|majority-blue] = 2/3 * 2/3 * 1/3 = 4/27

Pr[blue, blue, red] = Pr[blue, blue, red|majority-blue] * Pr[m-blue]
+ Pr[blue, blue, red|majority-red] * Pr[majority-red]
= (2/3 * 2/3 * 1/3) * 1/2 + (1/3 * 1/3 * 2/3) * 1/2 = 1/9

Pr[majority-blue|blue, blue, red] = (4/27 * 1/2) / (1/9) = 2/3

- So the third student should guess "**blue**" even when she sees "**red**"
- All future students will have the same information as the third student

Urn Experiment



Herding Intervention

In herding, the society only has 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 "he's got no clothes on"

Herding Intervention

Milgram Experiment: To intervene the herding effect, we need one person to tell the herd that there is nothing in the sky

How Does Intervention Work?

- When a new piece of private information releases, the herd re-evaluate 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

- In general, intervention is possible by providing private information to individuals not previously available. Consider an urn experiment where individuals decide on majority red over time. Either
  - 1) a private message to individuals informing them that the urn is majority blue or
  - 2) writing the observations next to predictions on the board stops the herding and changes decisions.

Information Cascade

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

Information Cascade (IC)

- In social media, individuals commonly repost content posted by others in the network. This content is often received via immediate neighbors (friends).
- An IC occurs as information propagates through friends.

- An information cascade is defined as a piece of information or decision being cascaded among a set of individuals, where
  - 1) individuals are connected by a network and
  - 2) individuals are only observing decisions of their immediate neighbors

- Therefore, cascade users have less information available to them compared to herding users, where almost all information about decisions are available.

In IC, local information is available to the users, but in herding the global information about the population is available.

Underlying Assumptions for Cascade Models

- The network is represented using a directed graph. Nodes are actors and edges depict the communication channels between them. A node can only influence nodes that it is connected to;

- Decisions are binary - nodes can be either active or inactive. An active nodes means that the node decided to adopt the behavior, innovation, or decision;

- A node, once activated, can activate its neighboring nodes; and

- Activation is a progressive process, where nodes change from inactive to active, but not vice versa 1.

Independent Cascade Model (ICM)

- Considering nodes that are active as *senders* and nodes that are being activated as *receivers*,
  - The *linear threshold model* concentrates on the receiver (to be discussed later).
  - The independent cascade model concentrates on the sender
- **Independent Cascade Model** is a sender centric model of cascade
  - In this model each activated node has <u>one chance</u> to activate its neighbors independently

Independent Cascade Model (ICM)

- In *ICM*, the node that is activated at time t, has **one** chance, at time step $t + 1$, to activate its neighbors
- Let v be an active node at time t, for any neighbor w of it, there's a probability p_{vw} that node w gets activated at time $t + 1$.
- A node v activated at time t has a single chance of activating its neighbors and that activation can only happen at $t + 1$

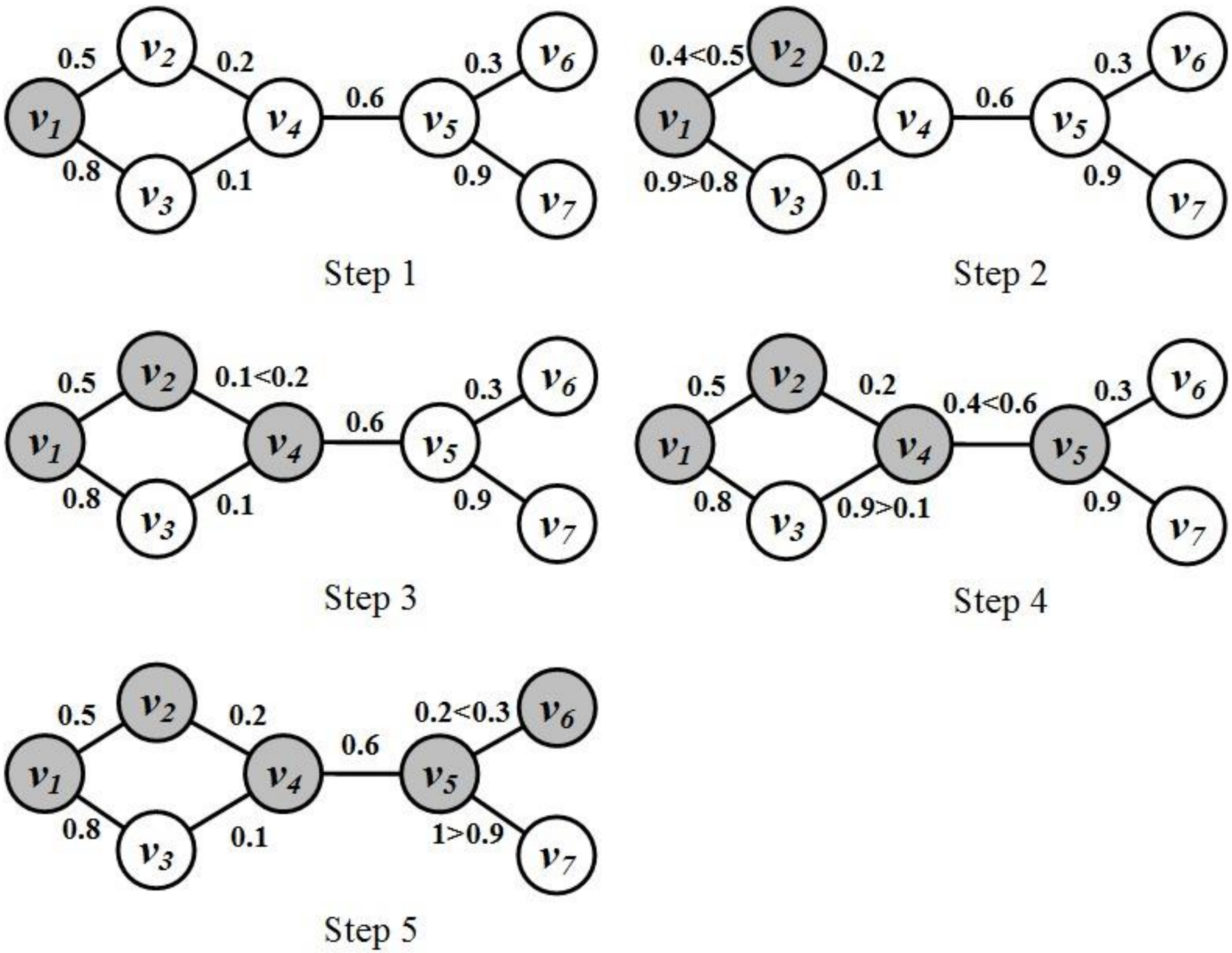
ICM Algorithm

Algorithm 7.1 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**
    $i = i + 1;$
    $A_i = \{\};$
    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



Maximizing the Spread of Cascades

Maximizing the spread of cascades

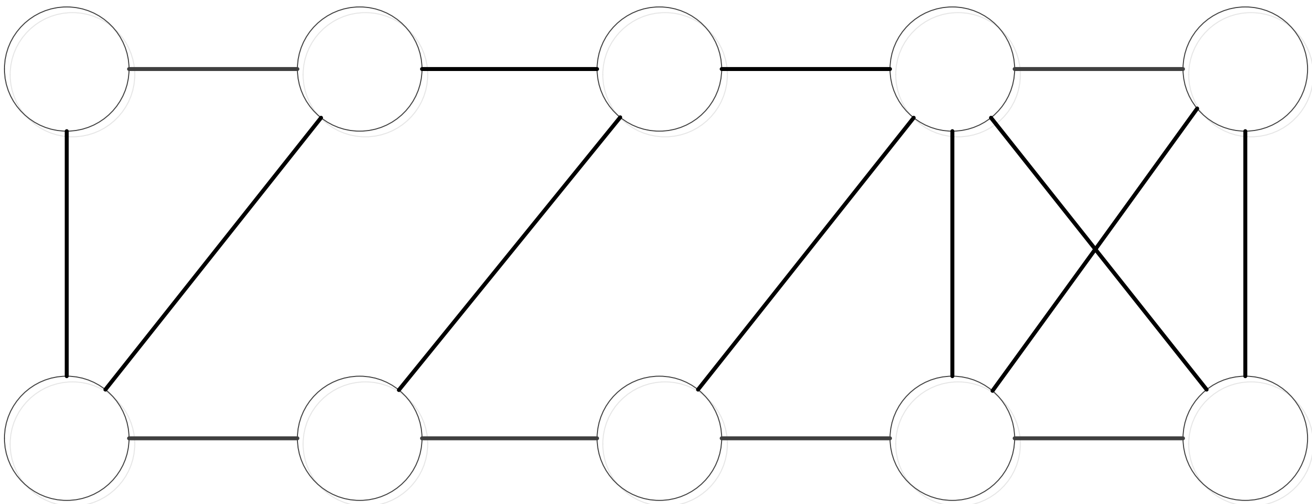
- **Maximizing the Spread of Cascades** is the problem of finding a small set of nodes in a social network such that their aggregated spread in the network is maximized
- Applications
  - Product marketing
  - Influence

Problem Setting

- Given
  - A limited budget B for initial advertising (e.g., give away free samples of product)
  - Estimating spread between individuals
- Goal
  - To trigger a large spread (e.g., further adoptions of a product)
- Question
  - Which set of individuals should be targeted at the very beginning?

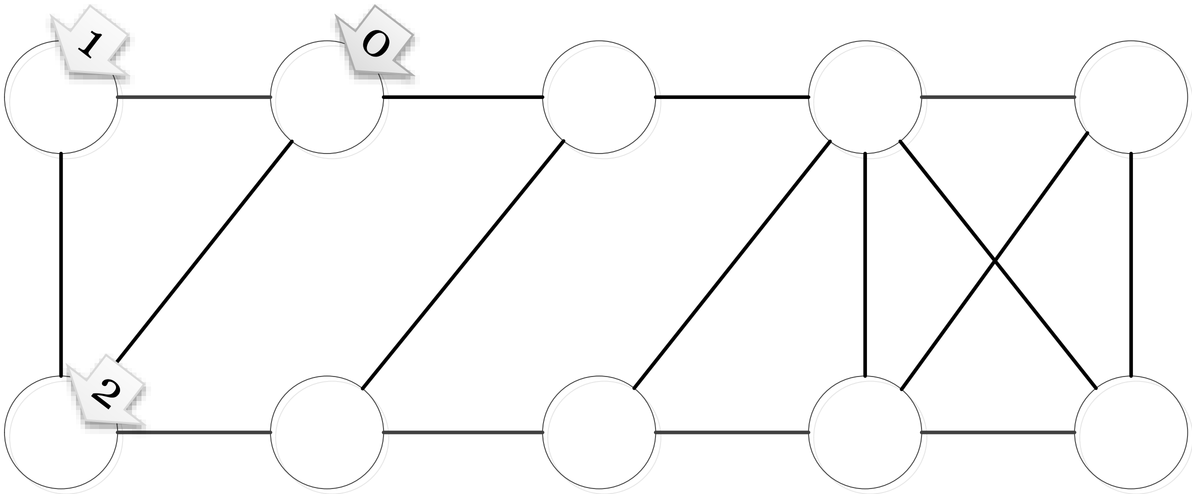
Maximizing the Spread of Cascade: Example

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

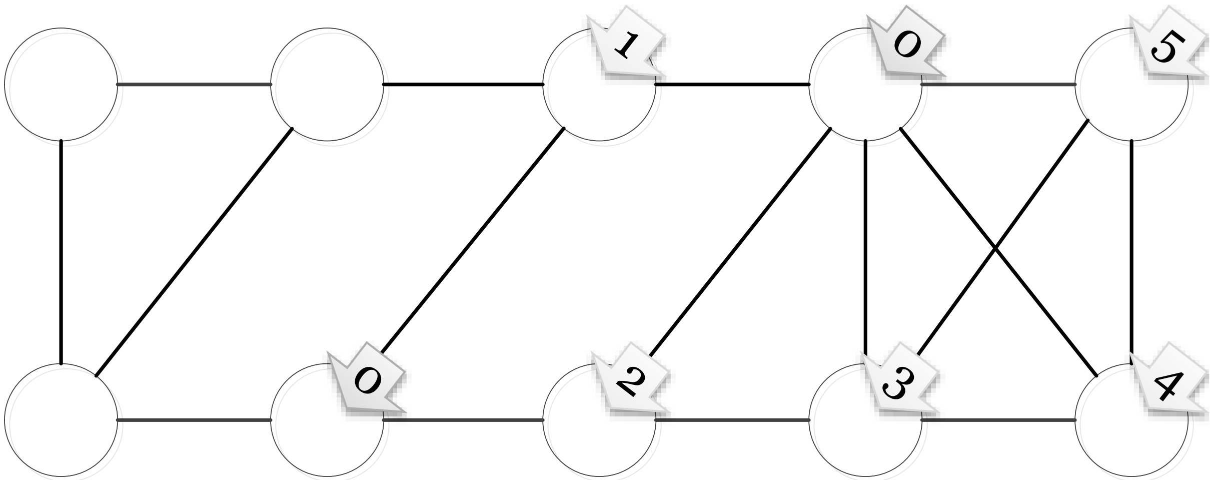


Maximizing the Spread of Cascade

Select one seed



Select two seeds



Problem Statement

- Spread of node set S: f(S)
  - An expected number of active nodes, 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 constrained optimization problem with f(S) as the objective function:

$$\arg\max_S [f(S)]$$

f(S): Properties

- Non-negative (obviously)
- Monotone: $f(S+v) \geq f(S)$
- Submodular:
  - Let N be a finite set
  - A set function is submodular iff

$$f: 2^N \mapsto \Re$$

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

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

If S is the set $\{x, y, z\}$, then the subsets of S are:

- $\{\}$ (also denoted $\varnothing$, the empty set)
- $\{x\}$
- $\{y\}$
- $\{z\}$
- $\{x, y\}$
- $\{x, z\}$
- $\{y, z\}$
- $\{x, y, z\}$

and hence the power set of S is $\{\{\}, \{x\}, \{y\}, \{z\}, \{x, y\}, \{x, z\}, \{y, z\}, \{x, y, z\}\}$.

Some Facts Regarding this Problem

- Bad News
  - For a submodular function monotone non-negative *f*, finding a *k*-element set *S* for which *f(S)* is maximized is an NP-hard optimization problem
- Good News
  - We can use Greedy Algorithm
    - Start with an empty set S
    - For k iterations:

      Add node v to S that maximizes $f(S+v) - f(S)$.

Cascade Maximization: A Greedy approach

Algorithm 7.2 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 ≠ k do
    v = arg max_{v∈V\S} f(S ∪ {v});
    or equivalently arg max_{v∈V\S} f(S ∪ {v}) − f(s)
    S = S ∪ {v};
    i = i + 1;
end while
Return S;
```

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

$$v = \arg\max_{v \in V \setminus S} \left[f(S + v) - f(S) \right]$$

About the Greedy approach

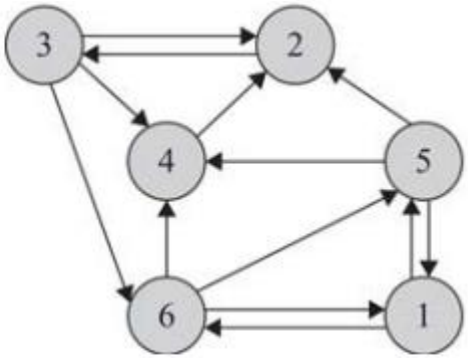
- How good (or bad) it is?
  - Theorem: The greedy algorithm is a $(1 - 1/e)$ approximation.
  - The resulting set S activates at least $(1- 1/e) > 63\%$ of the number of nodes that any size-k set S could activate.

Theorem 7.1 (Kempe et al. [146]). *Let f be a (1) non-negative, (2) monotone, and (3) submodular set function. Construct k-element set S, each time by adding node v, such that $f(S \cup \{v\})$ (or equivalently, $f(S \cup \{v\}) - f(s)$) is maximized. Let S^{Optimal} be the k-element set such that f is maximized. Then $f(S) \geq (1 - \frac{1}{e})f(S^{\text{Optimal}})$.*

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

Example- Cascade Maximization

Example 7.3. *For the following graph, assume that node i activates node j when* $|i - j| \equiv 2 \ (mod\ 3)$. *Solve cascade maximization for* $k = 2$.



Begin with 1:

$|1 - 6| \equiv 2 \ (mod\ 3),$

$|1 - 5| \not\equiv 2 \ (mod\ 3).$

$|6 - 4| \equiv 2 \ (mod\ 3)$

$|6 - 5| \not\equiv 2 \ (mod\ 3).$

$|4 - 2| \equiv 2 \ (mod\ 3)$

$f(\{1\}) = 4 \quad f(\{2\}) = 1,\ f(\{3\}) = 1,$

$f(\{4\}) = 2,\ f(\{5\}) = 1,\ f(\{6\}) = 4$

If 6 is initially activated, nodes 1, 2, 4, and 6 will become activated at the end.
From the set {1, 2, 3, 4, 5, 6} \ {1, 2, 4, 6} = {3, 5}, we need to select one more node, because:

$f(\{6,1\}) = f(\{6,2\}) = f(\{6,4\}) = f(\{6\}) = 4.$

$f(\{6,3\}) = f(\{6,5\}) = 5$

$\boxed{So,\ S = \{6,3\}\ and\ f(S) = 5.}$

Intervention of Information Cascade

- By limiting the number of out-links of the sender node and potentially reducing the chance of activating others.
- By limiting the number of in-links of receiver nodes and therefore reducing their chance of getting activated by others.
- By decreasing the activation probability of a node ($p_{v,w}$) and therefore reducing the chance of activating others.

Diffusion of Innovations

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

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 <u>why and how these innovations spread</u>. It also describes the reasons behind the diffusion process, individuals involved, as well as the rate at which ideas spread.

Innovation Characteristics

For an innovation to be adopted, various qualities associated with different parts of the process need to be present

- Observability,
  - The degree to which the results of an innovation are visible to potential adopters
- Relative Advantage
  - The degree to which the innovation is perceived to be superior to current practice
- Compatibility
  - The degree to which the innovation is perceived to be consistent with socio- cultural values, previous ideas, and/or perceived needs
- Triability
  - The degree to which the innovation can be experienced on a limited basis
- Complexity
  - 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 20^{th} century

The Iowa Study of Hybrid Corn Seed

- Ryan and Gross studied the adoption of hybrid seed corn by farmers in Iowa
- 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
  – The hybrid corn was highly resistant to diseases and other catastrophes such as droughts
  – However, farmers did not adopt it due to its high price and its inability to reproduce (e.g., new seeds have to be purchased from the seed provider)

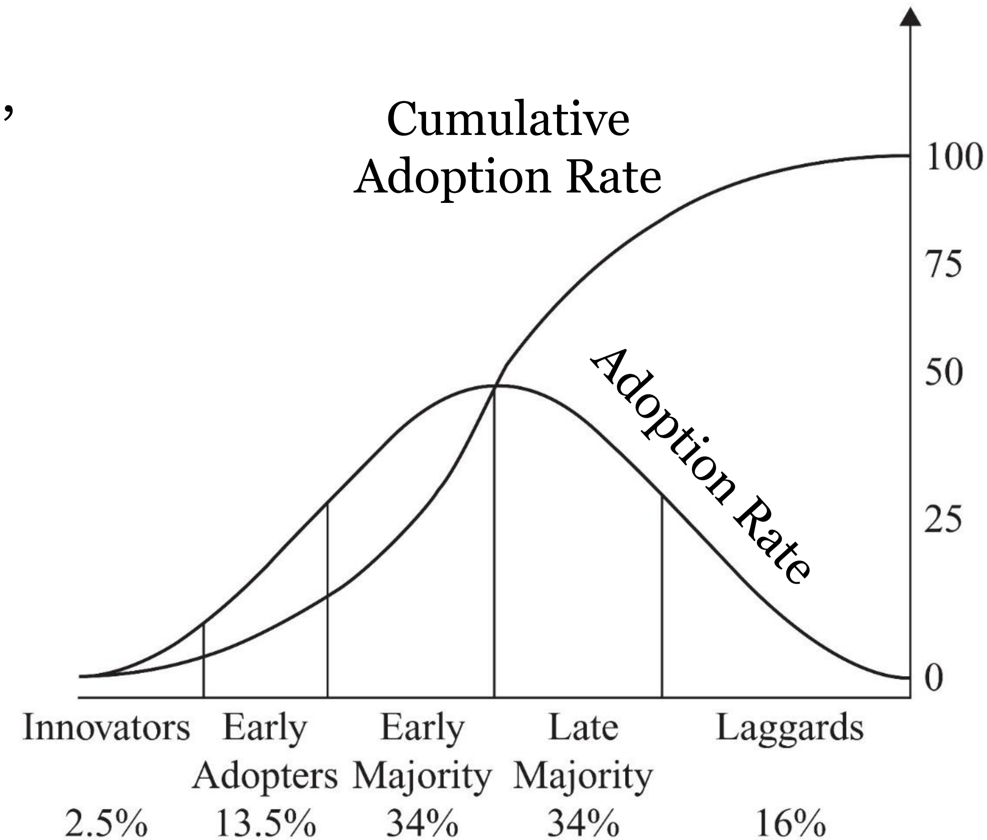
The Iowa Study of Hybrid Corn Seed, contd.

- farmers received information through two main channels:
  - mass communications from companies selling the seeds (information)
  - interpersonal communications with other farmers. (influence)

- Adoption is depended on a combination of both.

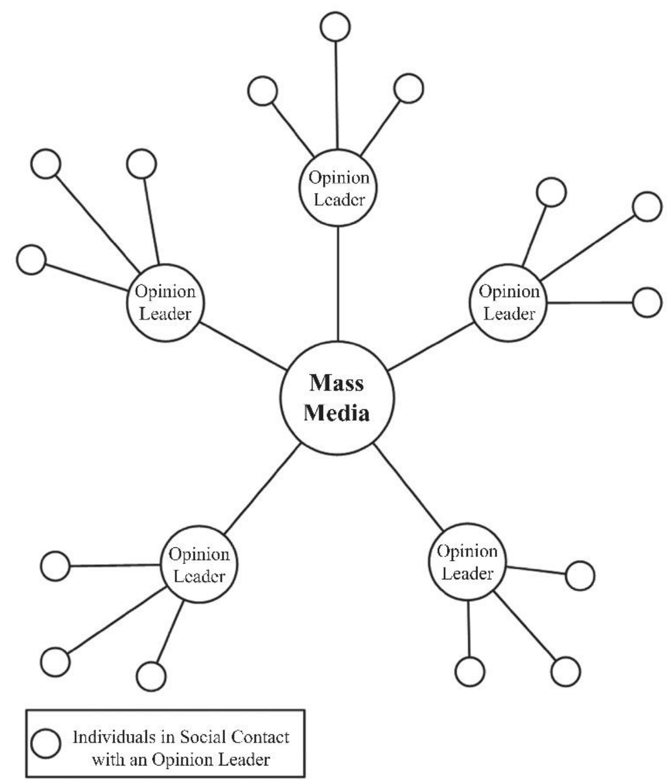
Adopter Categories

- 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 (13.5%),
  - 3) Early Majority (34%),
  - 4) Late Majority (34%),
  - 5) Laggards (16%).



Katz: Two-Step Flow Model

- According to the two-step flow model, most information comes from mass media, which is then directed toward influential figures called *opinion leaders*.
- These leaders then convey the information (or form opinions) and act as hubs for other members of the society



Rogers: Diffusion of Innovations: The Process

- **Awareness**
  - The individual becomes aware of the innovation, but her information regarding the product is limited
- **Interest**
  - The individual shows interest in the product and seeks more information
- **Evaluation**
  - The individual tries the product in his mind and decides whether or not to adopt it
- **Trial**
  - The individual performs a trial use of the product
- **Adoption**
  - The individual decides to continue the trial and adopts the product for full use

Modeling Diffusion of Innovations

This diffusion of innovation model describes the rate at which the number of adopters changes in terms of time:

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

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

- The rate depends on how innovative the product is
- The rate affects the potential adopters that have not yet adopted the product.

Information Diffusion: Mathematical Model

$$\frac{dA(t)}{dt} = i(t)[P - A(t)] \Rightarrow A(t) = \int_{t_0}^{t} a(t)dt,$$

$$A(t) = \sum_{t=t_0}^{t} a(t)$$

the adopters at time t

Defining the diffusion coefficient by defining i(t) as a function of number of adopters A(t), (A_0: the number of adopters at time t_0)

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

Diffusion Models

Three models of diffusion:

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

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

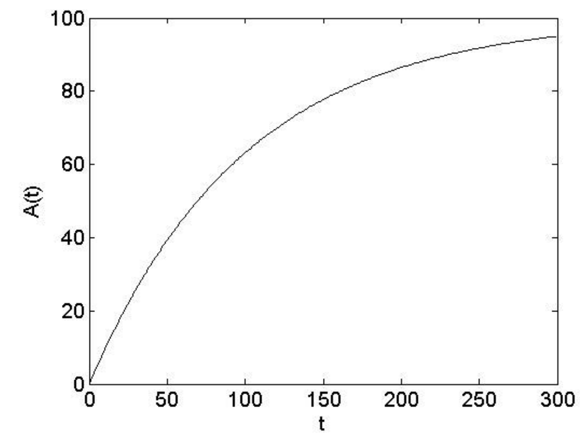
- **α: Innovativeness factor of the product**
- **β: Imitation factor**

External-Influence Model

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

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

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



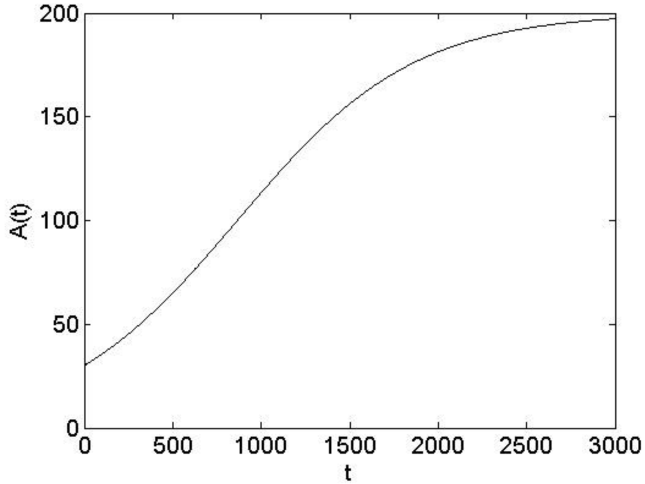
$P = 100$ and $\alpha = 0.01$

Internal-Influence Model

The adoption rate is a function that depends only on the number of already adopted individuals , $i(t) = \beta A(t)$

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

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



$A_0 = 30$, $\beta = 10^{-5}$, and $P = 200$

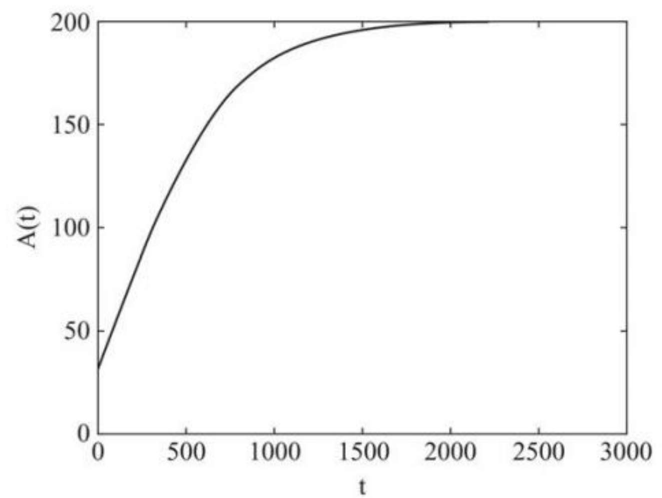
Mixed-Influence Model

The adoption rate is a function that depends on both the number of already activated individuals and external forces, <u>$i(t) = \alpha + \beta A(t)$</u>

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

$$\Downarrow$$

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



P=200, β=10e-5, A_0=30, α=10e-3

Diffusion of Innovation: Intervention

- Limiting the distribution of the product or the audience that can adopt the product (reducing P).

- Reducing interest in the product being sold. For instance, the company can inform adopters of the faulty status of the product (reducing α).

- Reducing interactions within the population. Reduced interactions result in less imitations on product adoptions and a general decrease in the trend of adoptions (reducing β).

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

Epidemics

Epidemics

- Epidemics describes the process by which diseases spread. This process consists of
  - A pathogen (the disease being spread),
  - A population of hosts (humans, animals, plants, etc.)
  - A spreading mechanism (breathing, drinking, sexual activity, etc.)

Comparing Epidemics and Cascades

- Unlike information cascades and herding and similar to diffusion of innovations models, epidemic models assume an implicit network and unknown connections between individuals.

- This makes epidemic models more suitable when we are interested in global patterns, such as trends and ratios of people getting infected, and not in who infects whom.

How to Analyze Epidemics?

- Contact Network
  - look at how hosts contact each other and devise methods that describe how epidemics happen in networks.
  - A contact network is a graph where nodes represent the hosts and edges represent the interactions between these hosts. For instance, in the case of the HIV/AIDS, edges represent sexual interactions, and in the case of influenza, nodes that are connected represent hosts that breathe the same air.
- Fully-mixed
  - Analyze only the rates at which hosts get infected, recover, etc. and avoid considering network information

The models discussed here will assume:

- No contact network information is available
- The process by which hosts get infected is unknown

SI Model: Definition

- **Susceptible**
  - When an individual is in the susceptible state, he or she can potentially get infected by the disease.
- **Infected**
  - An infected individual has the chance of infecting susceptible parties
- ~~**Recovered (removed)**~~
  - individuals who have either recovered from the disease and have complete/partial immunity against the infection or were killed by the infection
- The *susceptible* individuals get infected by *infected* individuals and once *infected* they will never get *recovered* (removed)

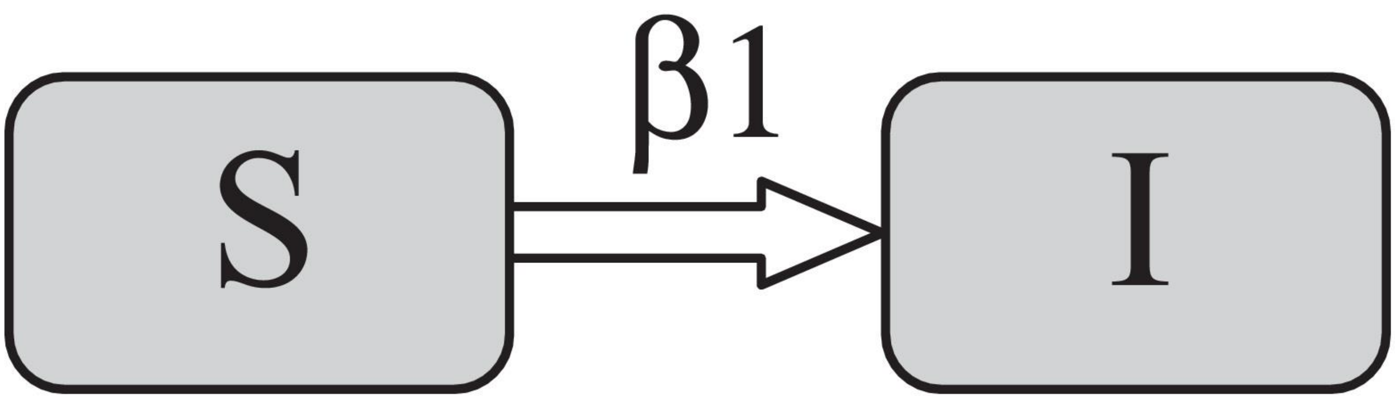
Notations

- N: size of the crowd
- S(t): number of *susceptible* individuals at time t
  - $s(t) = S(t)/N$
- I(t): number of *infected* individuals at time t
  - $i(t) = I(t)/N$
- β: Contact probability
  - probability of a pair of people meeting <u>in any time step</u>
  - if $\beta = 1$ everyone comes to contact with everyone else
  - if $\beta = 0$ no one meets another individual

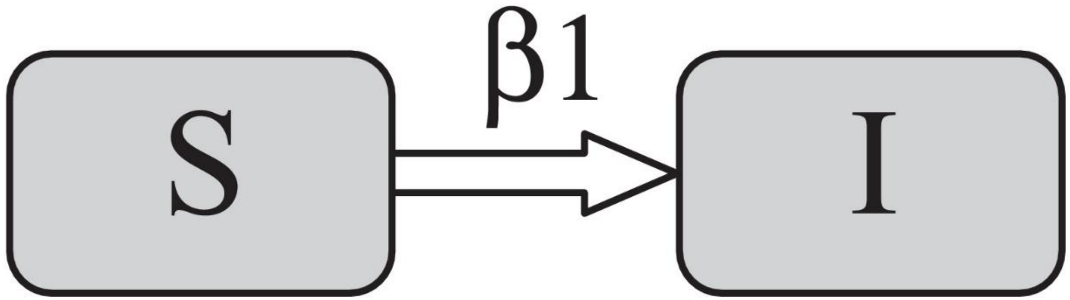
$$N = S(t) + I(t)$$

SI Model

- At each time stamp, an infected individual will meet βN people on average and will infect βS of them
- Since I are infected, βIS will be infected in the next time step



SI Model: Equations



$$\frac{dS}{dt} = -\beta I S, \qquad \frac{dI}{dt} = \beta I S.$$

(S + I = N)

$$\frac{dI}{dt} = \beta I (N - I)$$

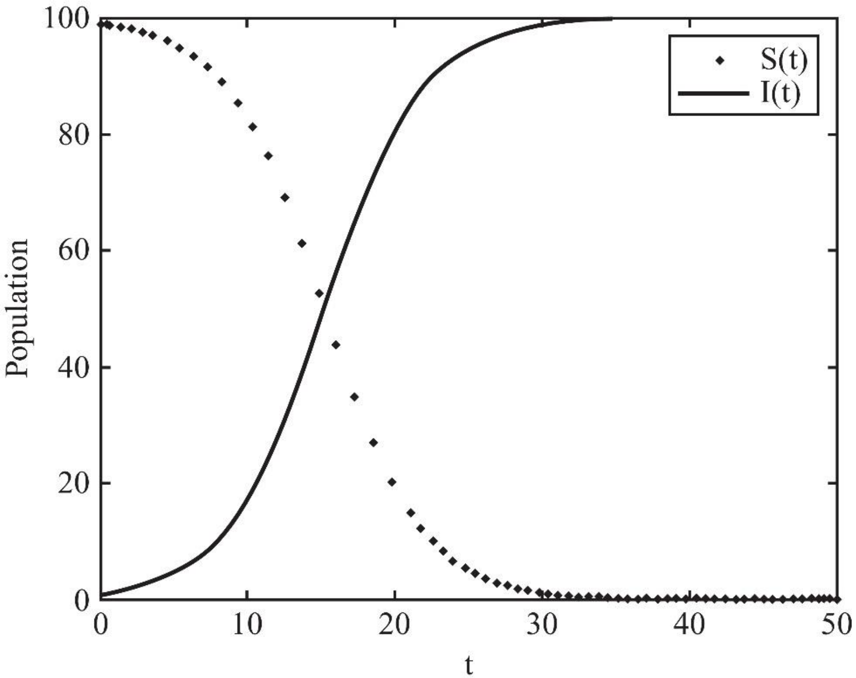
$$I(t) = \frac{N I_0 e^{\beta t}}{N + I_0 (e^{\beta t} - 1)}$$

$$i_0 = \frac{I_0}{N}$$

$$i(t) = \frac{i_0 e^{\beta t}}{1 + i_0 (e^{\beta t} - 1)}$$

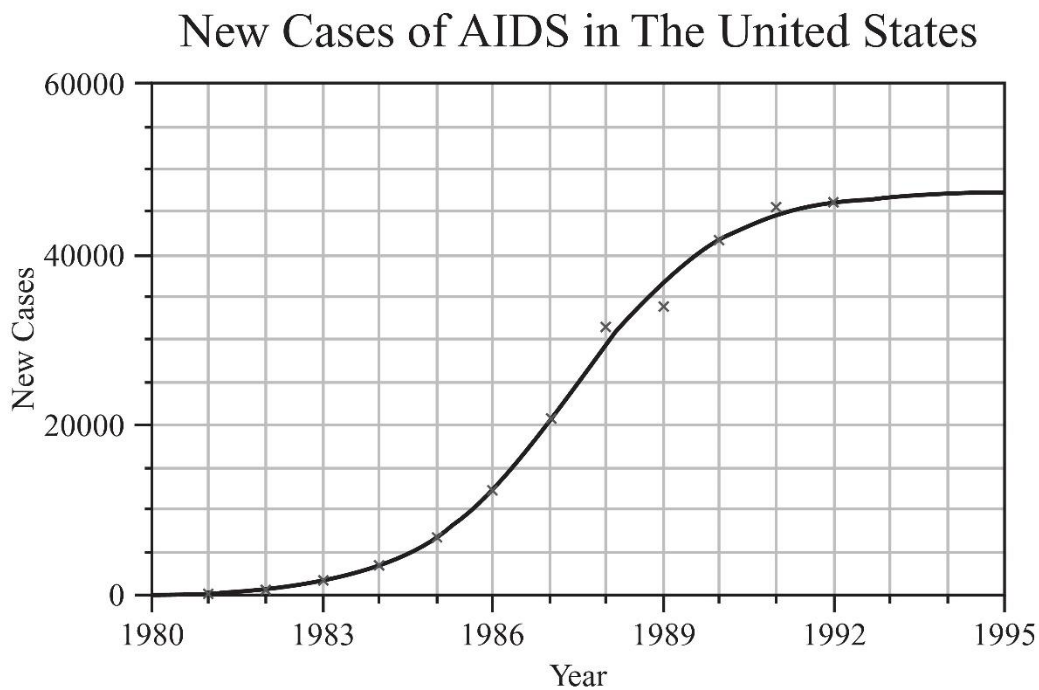
I_0 is the number of individuals infected at time 0

SI Model: Example



(a) SI Model Simulation

$N=100$, $I_0=1$, $\beta=0.003$

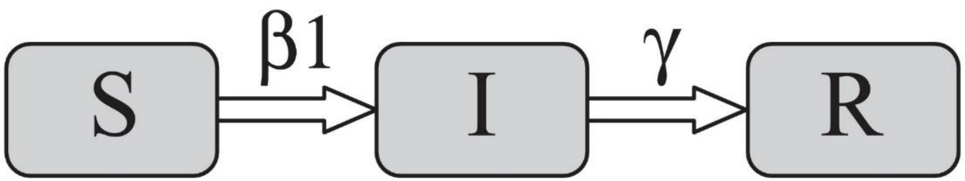


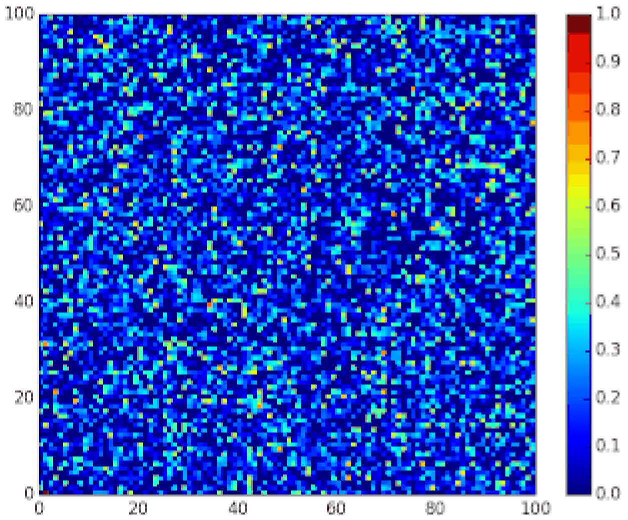
(b) HIV/AIDS Infected Population Growth

Logistic growth function compared to the HIV/AIDS growth in the United States

SIR Model

- In the SIR model, in addition to the I and S states, a recovery state R is present.
- In the SIR model, individuals get infected, then some recovered.
- Once hosts are recovered (or removed) they can <u>no longer get infected and are not susceptible any longer</u>.





SIR Model, Equations

$$I + S + R = N$$

$$\frac{dS}{dt} = -\beta IS,$$

$$\frac{dI}{dt} = \beta IS - \gamma I,$$

$$\frac{dR}{dt} = \gamma I.$$

γ defines the recovering probability of an infected individual at a time stamp

SIR Model, Equations, Cont.

$$\frac{dS}{dt} = -\beta IS,$$
$$\frac{dI}{dt} = \beta IS - \gamma I,$$
$$\frac{dR}{dt} = \gamma I.$$

$$\Rightarrow \frac{dS}{dR} = -\frac{\beta}{\gamma} S \xrightarrow{(R_0 = 0)} \log \frac{S_0}{S} = \frac{\beta}{\gamma} R$$

$$\Downarrow$$

$$S = S_0 e^{-\frac{\beta}{\gamma} R}$$

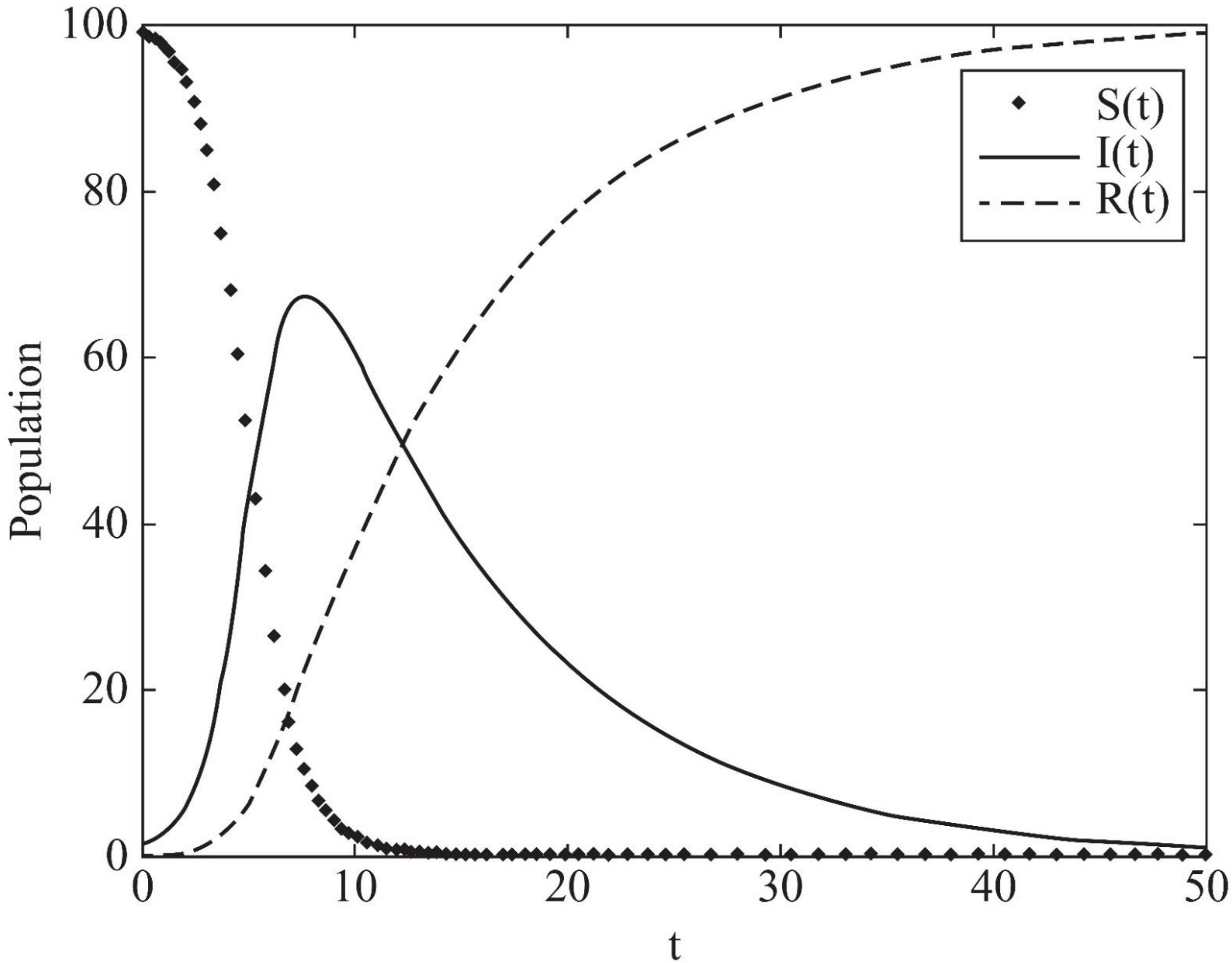
$$\frac{dR}{dt} = \gamma(N - S - R)$$

$$\Downarrow$$

$$\frac{dR}{dt} = \gamma(N - S_0 e^{-\frac{\beta}{\gamma} R} - R) \Rightarrow t = \frac{1}{\gamma} \int_0^R \frac{dx}{N - S_0 e^{-\frac{\beta}{\gamma} x} - x}$$

There is no closed form solution for this integration and only numerical approximation is possible.

SIR Simulation



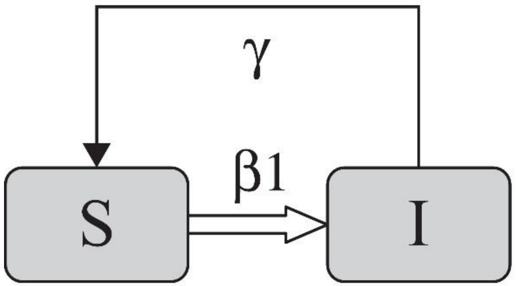
$S_0 = 99, I_0 = 1, R_0 = 0, \beta = 0.01, \gamma = 0.1$

SIS Model

- The SIS model is the same as the SI model with the addition of infected nodes recovering and becoming susceptible again

$$\frac{dS}{dt} = \gamma I - \beta IS,$$

$$\frac{dI}{dt} = \beta IS - \gamma I$$



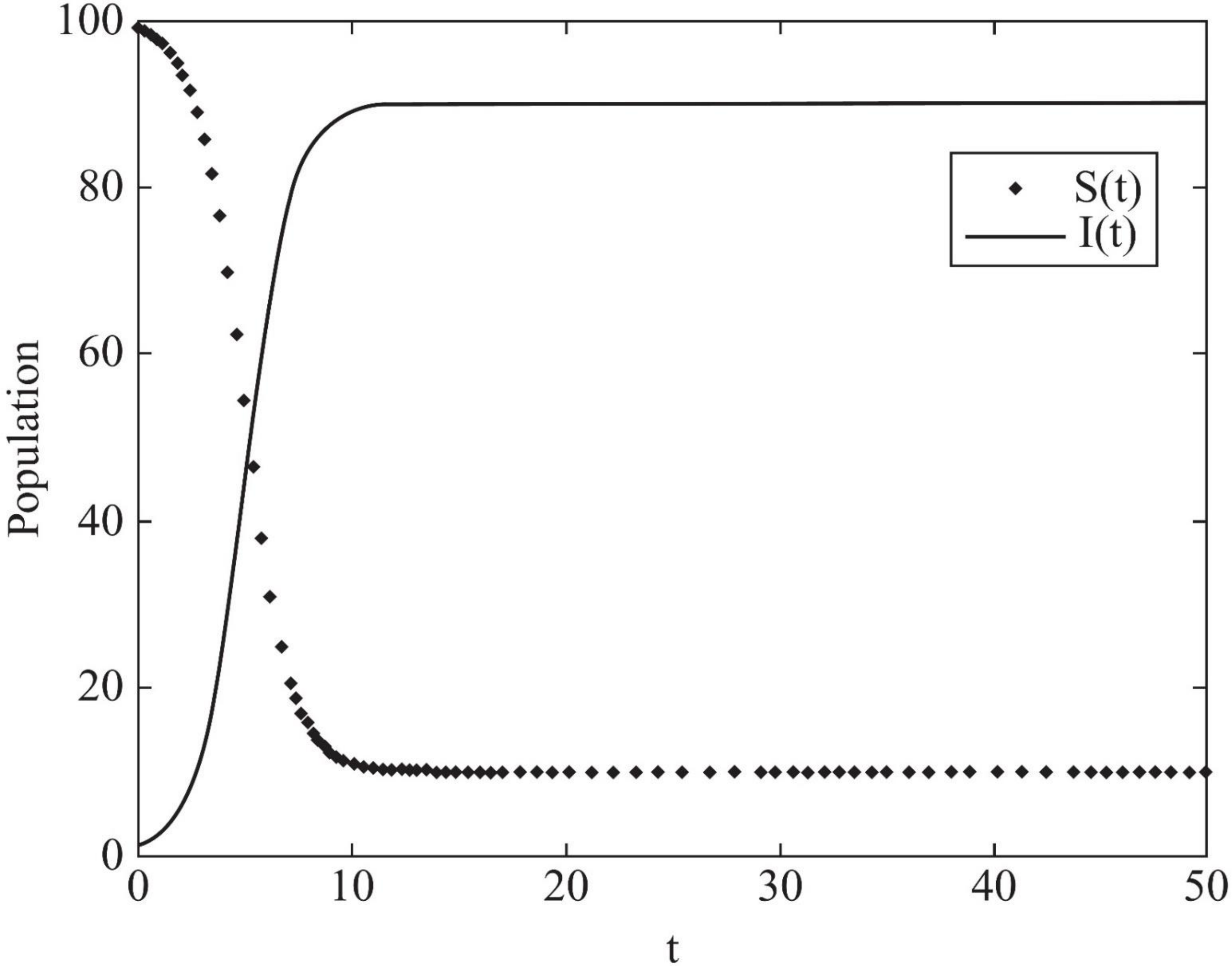
$$\Rightarrow \frac{dI}{dt} = \beta I(N - I) - \gamma I = I(\beta N - \gamma) - \beta I^2$$

SIS Model

$$\frac{dI}{dt} = \beta I(N - I) - \gamma I = I(\beta N - \gamma) - \beta I^2$$

- When $\beta N <= \gamma$:
  - the first term will be at most zero or negative hence the whole term becomes negative and therefore, in the limit, the value I(t) will decrease exponentially to zero
- When $\beta N > \gamma$:
  - We will have a logistic growth function like the SI model

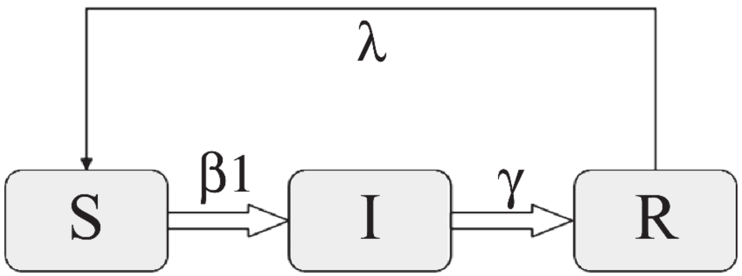
SIS Model Simulation



$S_0 = 99, I_0 = 1, \beta = 0.01, \text{and } \gamma = 0.1$

SIRS Model

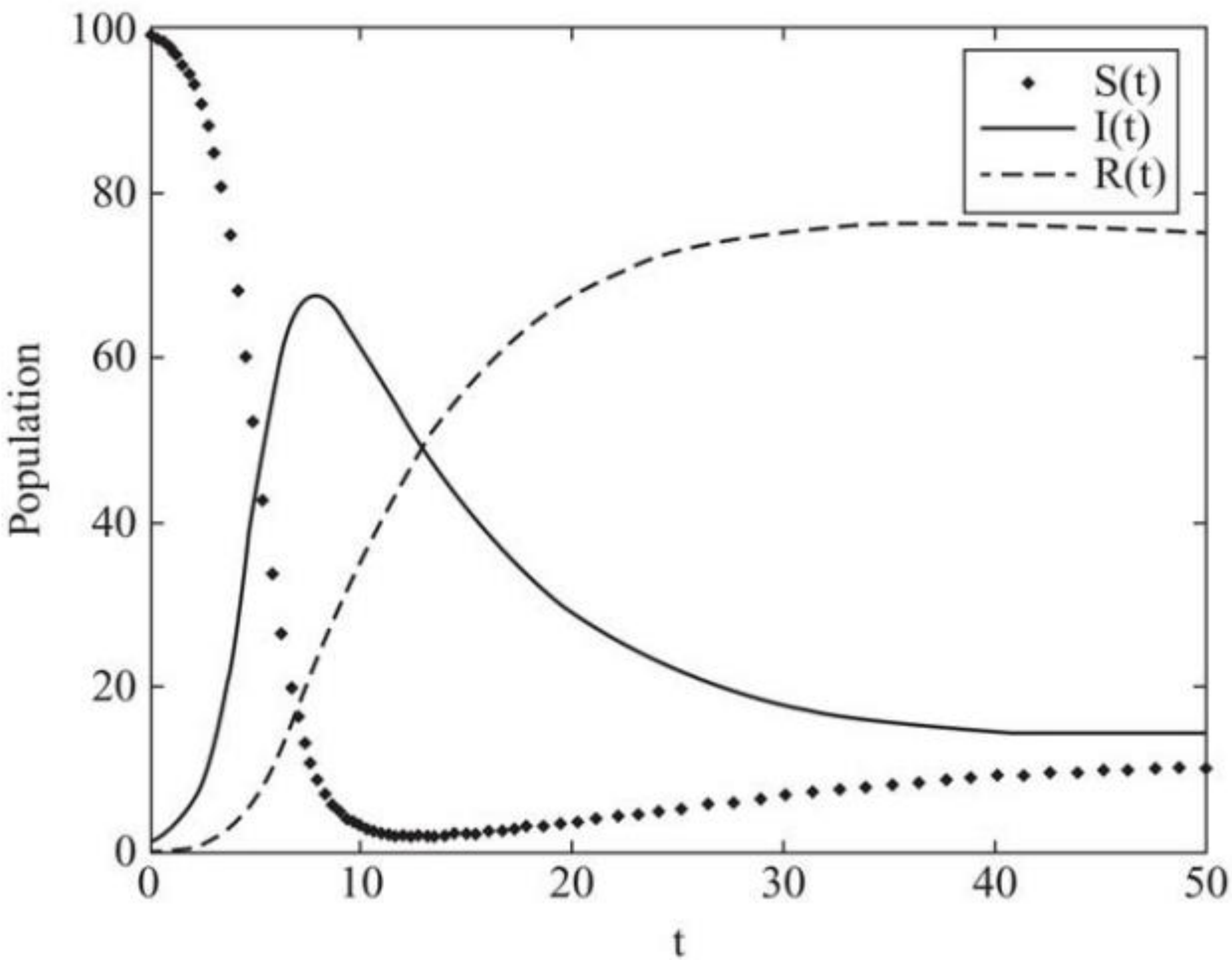
The individuals who have recovered will lose immunity after a certain period of time and will become susceptible again



$$\frac{dS}{dt} = \lambda R - \beta IS,$$
$$\frac{dI}{dt} = \beta IS - \gamma I,$$
$$\frac{dR}{dt} = \gamma I - \lambda R.$$

Like the SIR, model this model has no closed form solution, so numerical integration can be used

SIRS Model Simulation



$S_0 = 99$, $I_0 = 1$, $R_0 = 0$, $\gamma = 0.1$, $\beta = 0.01$, and $\lambda = 0.02$.

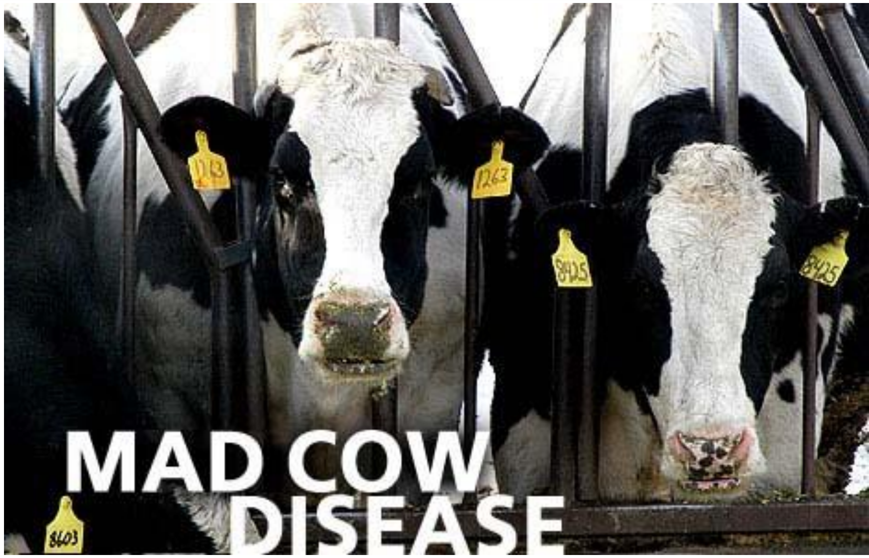
Epidemic Intervention

- Suppose that we have a susceptible society and want to prevent more spread by vaccinating the most vulnerable individuals (*highly connected nodes*) to gain herd immunity.
- How to find the most vulnerable individuals?

Randomly pick 96% nodes and vaccinate those individuals!

Randomly pick 30% nodes and ask them who is the most vulnerable (higher-degree and more-connected nodes) from their point of view, then vaccinate those individuals! (using social network)

Epidemic Intervention: Mad-cow disease



- Jan. 2001
  - First case observed in UK
- Feb. 2001
  - 43 farms infected
- Sep. 2001
  - 9000 farms infected

How to stop the disease (older methods):

- Banned movement
- Killed millions of animals

Summary

We discussed Information Diffusion in Social Media.

- In the herd behavior, individuals observe the behaviors of others and act similarly to them based on their own benefit.

- In the information cascade, the independent cascade model (ICM) is a sender-centric model, where the spread of cascades can be maximized in a network given a budget on how many initial nodes can be activated.
  - We introduced a greedy approximation algorithm that has guaranteed performance due to the sub-modularity of ICM's activation function.

- In the diffusion of innovations, we detailed mathematical models that account for internal, external, and mixed influences and their intervention procedures.

- In the epidemics, we discussed four epidemic models: SI, SIR, SIS, and SIRS; the two last models allow for re-infected individuals..

- We discussed how to intervene various information diffusion processes.