Information Diffusion part 2

CS 579 Online Social Network Analysis

Dr. Cindy Hood
11/11/25

Final Project Proposals – Due 11/8

The final project is an extension of your analysis from HW 4. You are required to include the following: Chicago Community Areas, Census data, Network modeling and analysis, and Geographic mapping/visualization. Beyond this, there are many different ways to extend your analysis from HW 4 (e.g. include all CAs, include additional census data, include data from other sources, include additional analysis methods, etc.). Your proposal should include the following aspects:

- What is the goal of your project?
- How will you extend what you did for HW 4?
- What motivates this extension?
- What data will you use?
- What will be the most challenging part of the project?

The final project may be done individually or in a group of up to 4 students. Groups should submit one proposal that includes technical content for each team member.

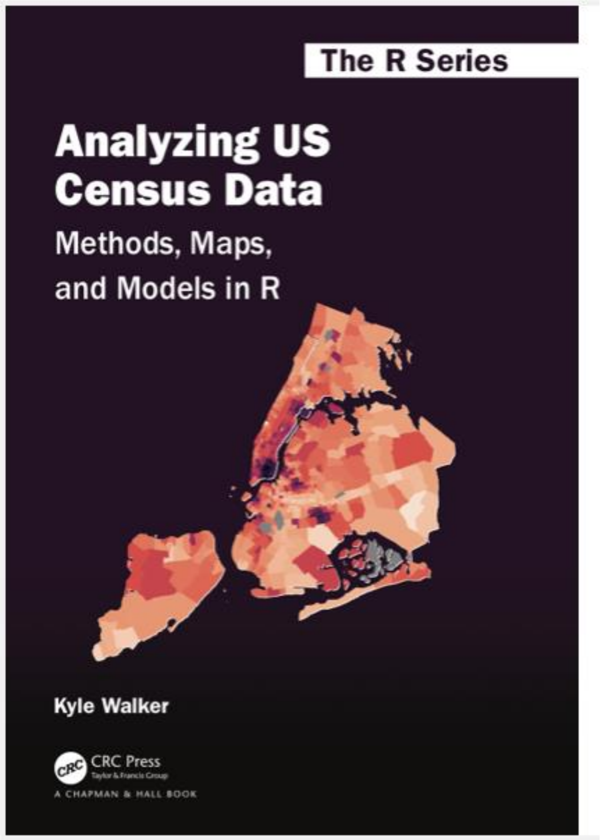
The proposal should be one page or less, group proposals may be longer as needed to convey the technical proposal for each group member. Groups should submit one proposal that includes the names of all group members.

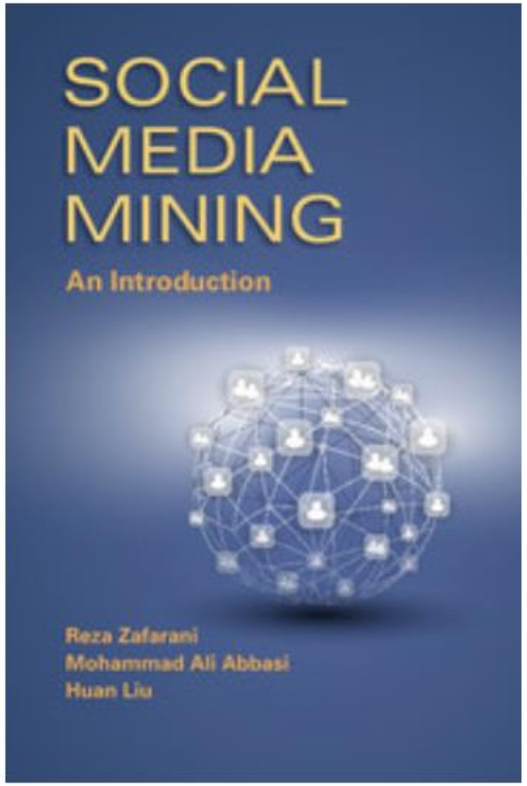
Review Underway – some have already received ok, a few revise and resubmits

Remaining Exam and Deliverables

- Final project progress report and video
  - Due 11/21
- HW #5
  - Problems that will help you prepare for Exam 2
  - Due 11/24 (No late days)
- Exam 2
  - 12/2
- Final project poster presentation/video (online students)
  - 12/4
- Final project report
  - Week of 12/8
    - Specific date tbd

References





https://walker-data.com/census-r/

http://www.socialmediamining.info

Some additional resources

- Myatt and Johnson (2014), *Making Sense of Data I*, 2nd Edition, Wiley, ISBN: 978-1-118-40741-7
  - https://onlinelibrary.wiley.com/doi/epdf/10.1002/9781118422007
- Networks, Crowds, and Markets: Reasoning About a Highly Connected World by David Easley and Jon Kleinberg.
  - http://www.cs.cornell.edu/home/kleinber/networks-book/
- Hanneman, Robert A. and Mark Riddle. 2005. Introduction to social network methods. Riverside, CA: University of California, Riverside
  - https://faculty.ucr.edu/~hanneman/nettext/

Information Diffusion

Chapter 7 in Social Media Mining Book

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.
- We discuss methods from
  - Sociology, epidemiology, and ethnography
  - All are useful for social media mining.
- We focus on techniques that can model information diffusion.

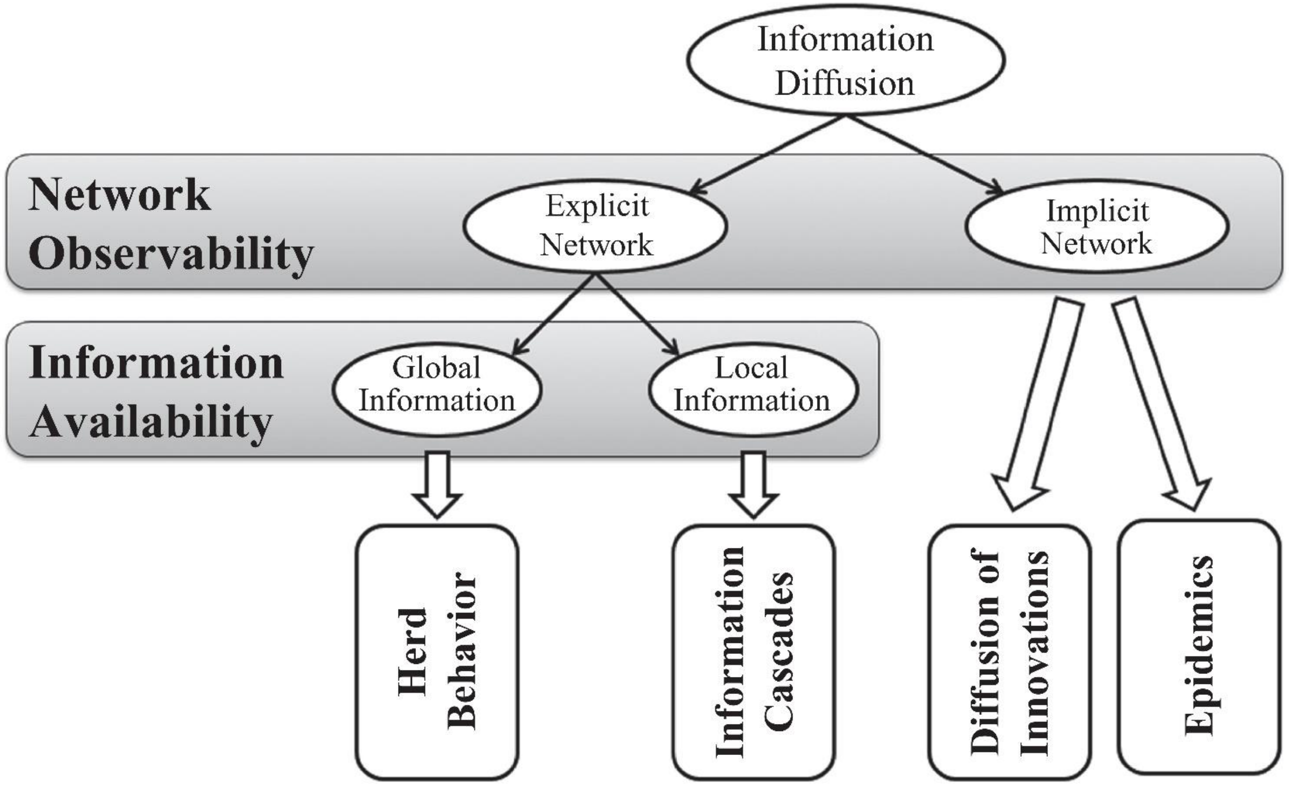
Information Diffusion

- **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



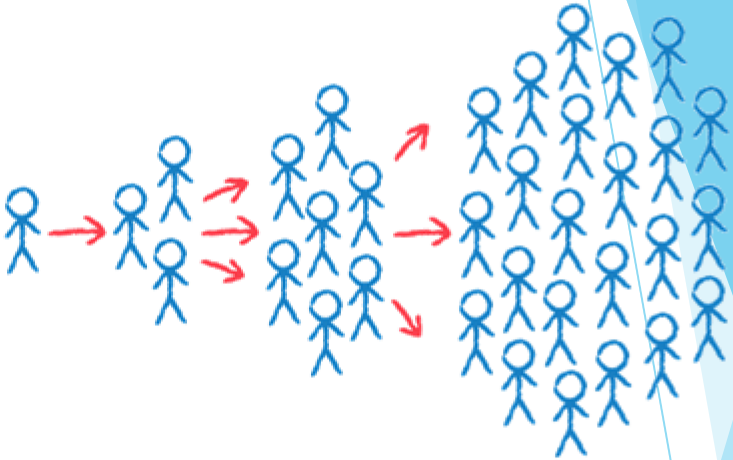
We define the process of interfering with information diffusion by expediting, delaying, or even stopping diffusion as Intervention

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).



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 have almost all information about decisions

Notable 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 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 can be
  - **Active**: the node has adopted the behavior/innovation/decision;
  - **Inactive**
- A activated node can activate its neighboring nodes; and
- 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 (more info in Chapter 8).

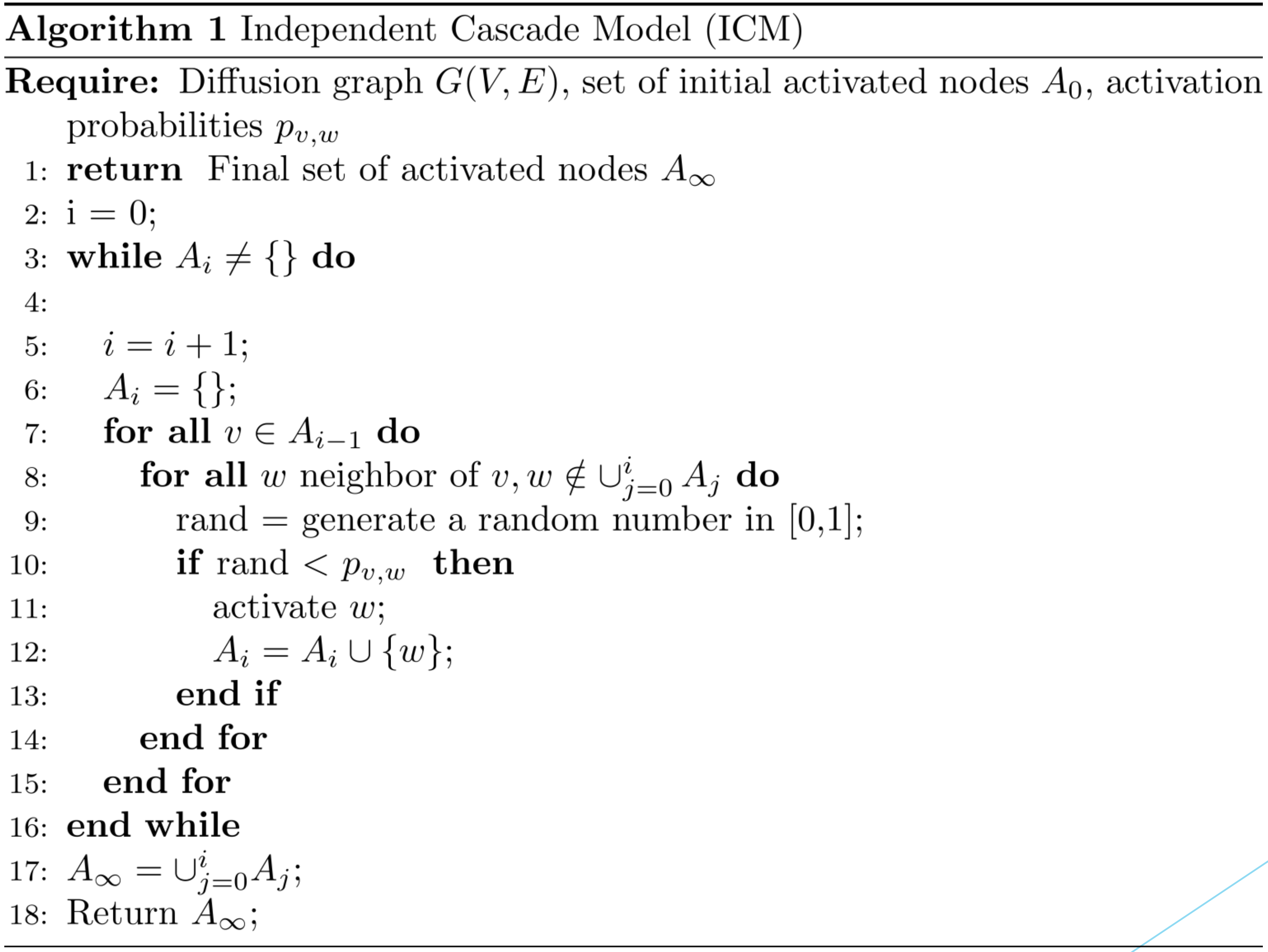
ICM Algorithm

- 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 $t + 1$.

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

ICM Algorithm

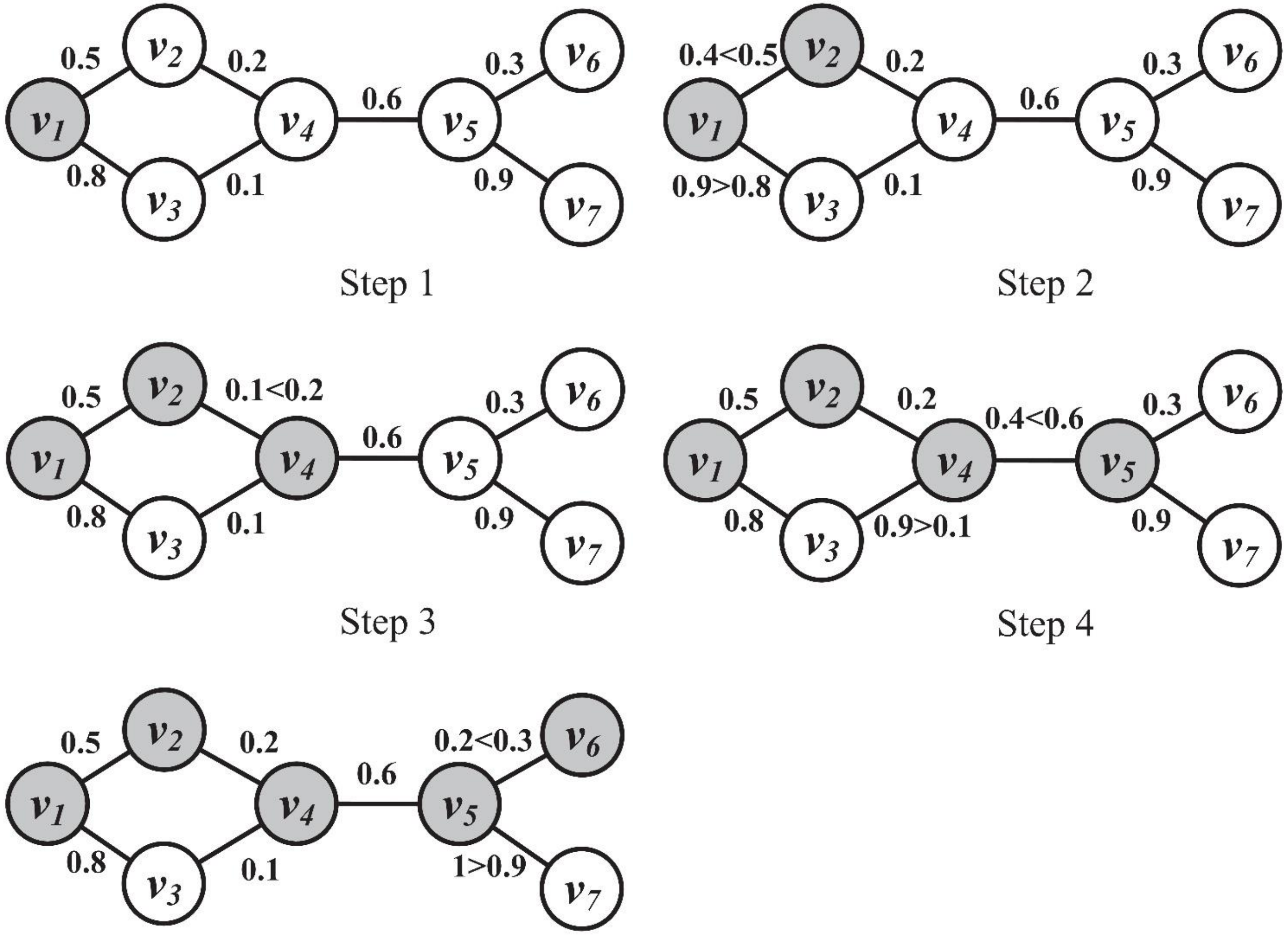


Algorithm 1 Independent Cascade Model (ICM)

```
Require: Diffusion graph $G(V, E)$, set of initial activated nodes $A_0$, activation
    probabilities $p_{v,w}$
return Final set of activated nodes $A_\infty$
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 $A_\infty$;
```

Independent Cascade Model: An Example



Step 1

Step 2

Step 3

Step 4

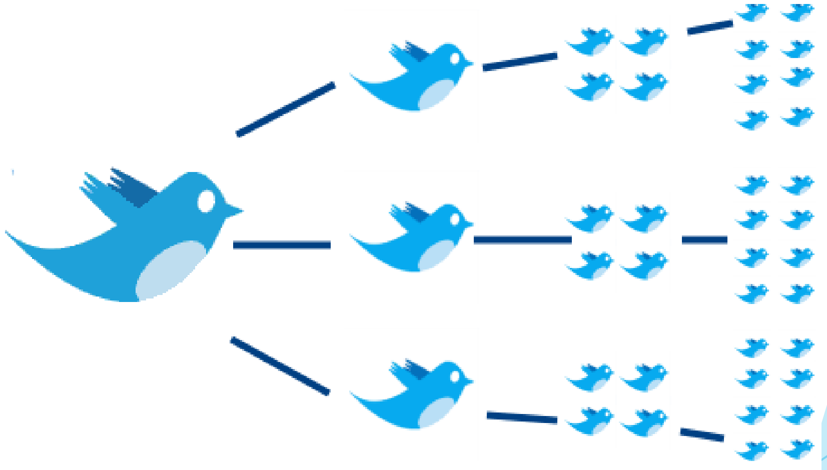
Step 5

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

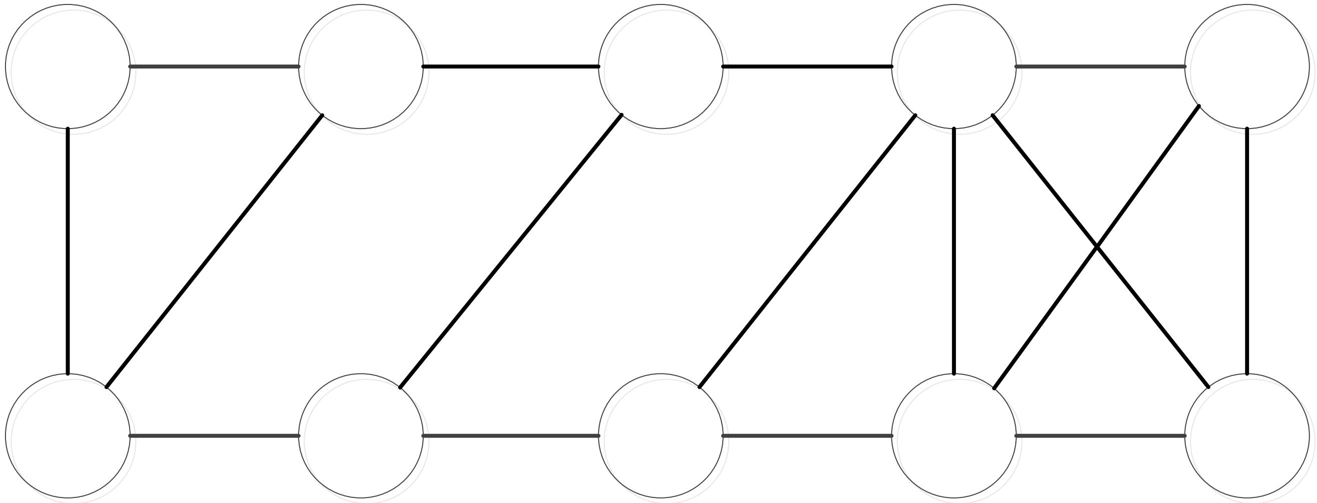
- **Given**
  - A limited budget **B** for initial advertising
    - Example: give away free samples of product
  - Estimating spread between individuals

- **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?

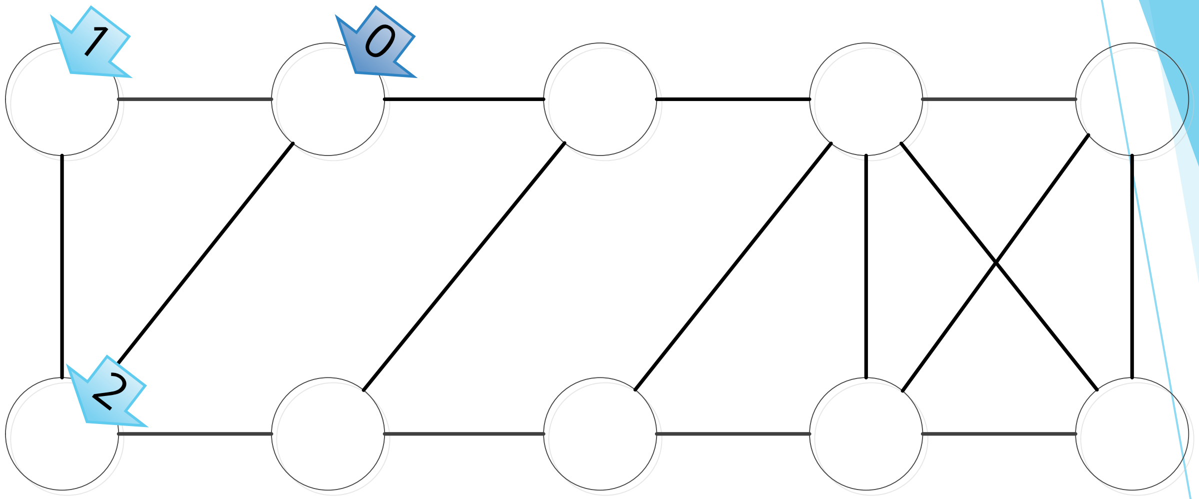
Maximizing the Spread of Cascade: Example

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

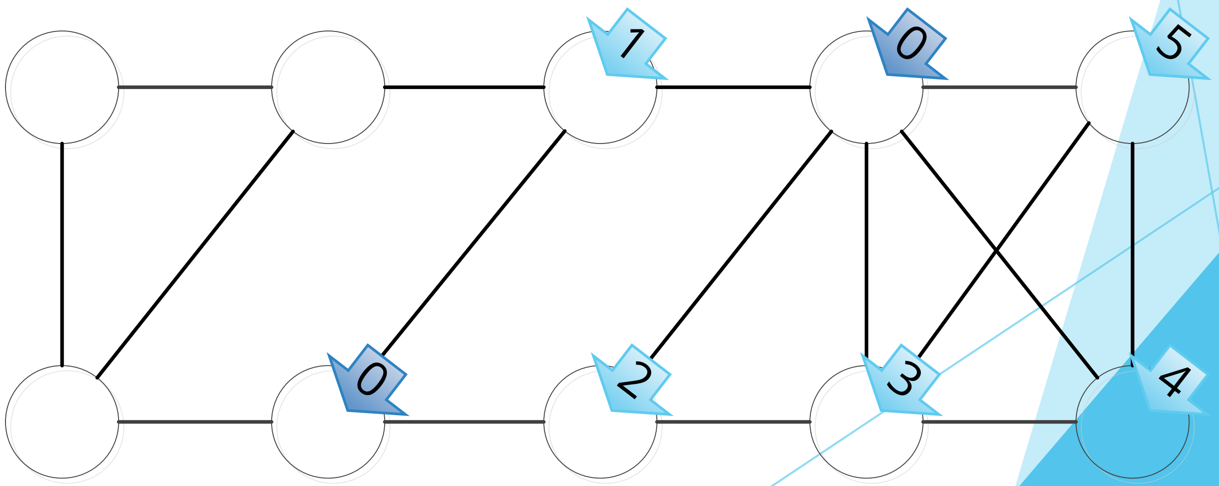


Maximizing the Spread of Cascade

Select one seed



Select two seeds



Problem Statement

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

f(S): Properties

1. **Non-negative** (obviously)

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

$$f:2^N \text{ a } \Re$$

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

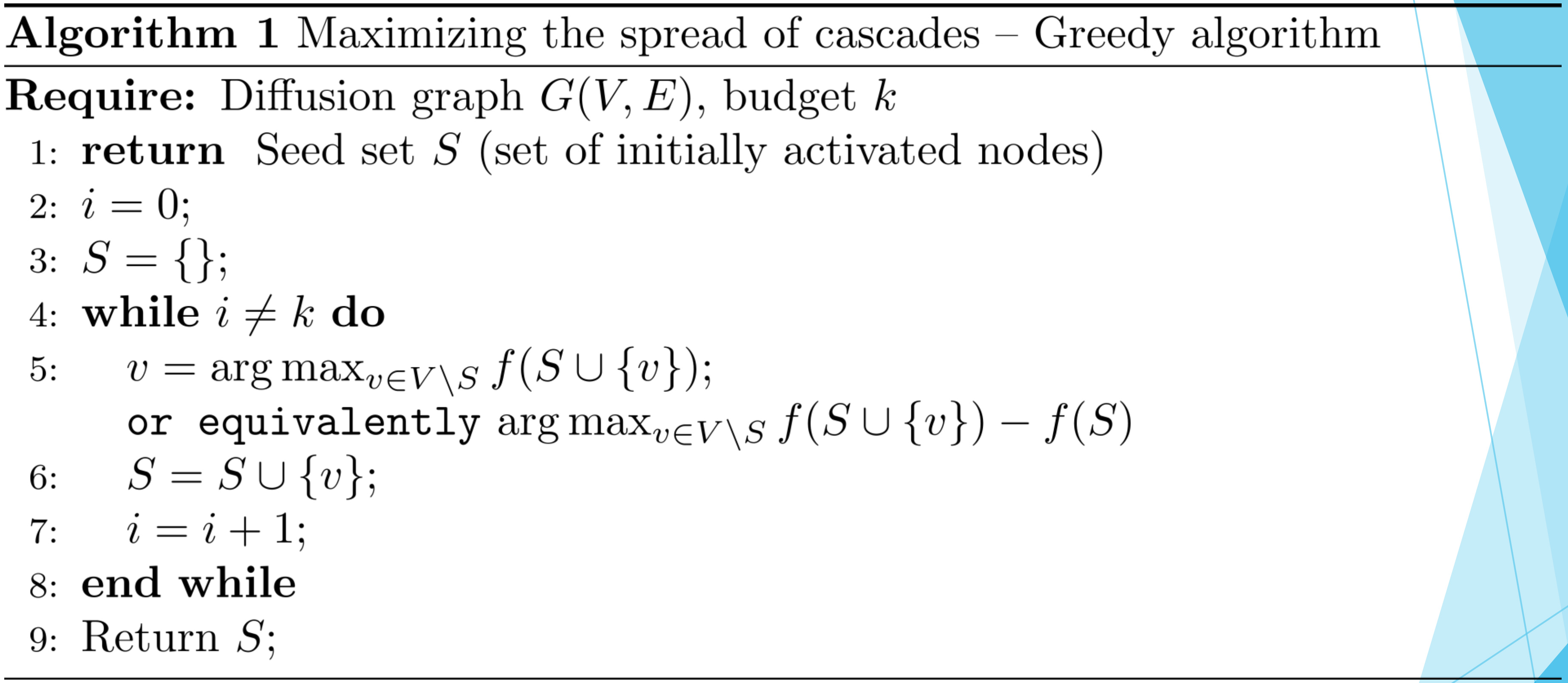
$$f(S+v) - f(S) \geq f(T+v) - f(T)$$

Some Facts Regarding this Problem

- **Bad News**
  - Consider a non-negative, monotone, submodular function f
  - 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 $f(S \cup \{v\}) - f(S)$.
  - How good (or bad) it is? (Kempe et al., before that Nemhauser 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.

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 $S$;
```

Intervention

- By limiting the number of out-links
  - Disconnected nodes don’t get to activate anyone

- By limiting the number of in-links
  - Reducing the chance of getting activated by others

- By decreasing the activation probability p_{vw}
  - 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 **why** and **how** innovations spread.

- 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, 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

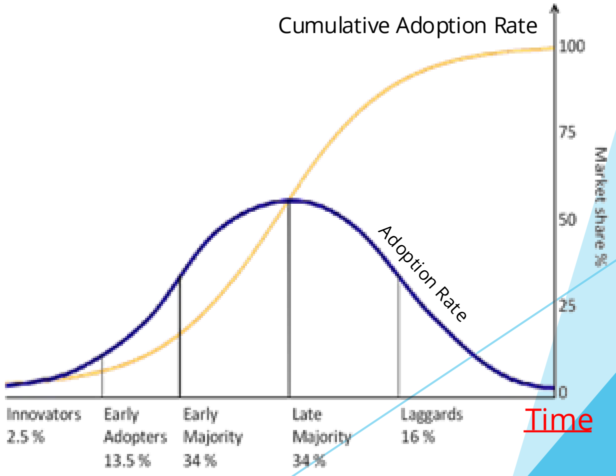
Farmers received information through two main channels

- **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:

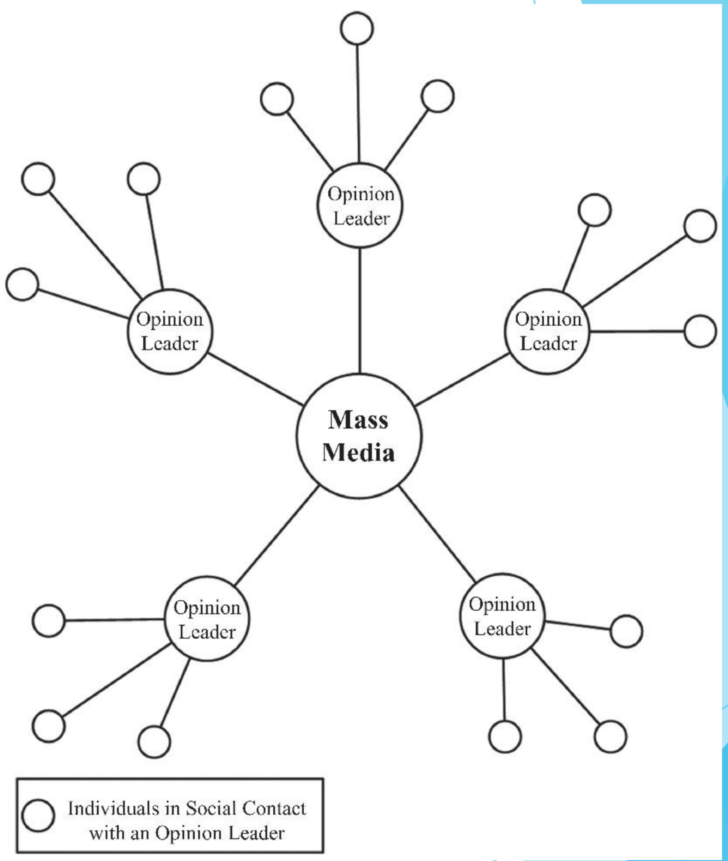
1) **Innovators** (top **2.5%**)
2) **Early Adopters** (next **13.5%**)
3) **Early Majority** (next **34%**)
4) **Late Majority** (next **34%**)
5) **Laggards** (last **16%**)



II. Katz Two-Step Flow Model

- **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 hub for other members of the society



III. Rogers: Diffusion of Innovations: The Process

Adoption process:

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

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 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
- The rate affects the potential adopters that have not yet adopted the product.

Modeling Diffusion of Innovations

We can rewrite $A(t)$ as $$A(t) = \int_{t_0}^{t} \underset{\uparrow}{a(t)}dt \xrightarrow{} \text{Let } A_0 = A(0)$$

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):

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

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

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

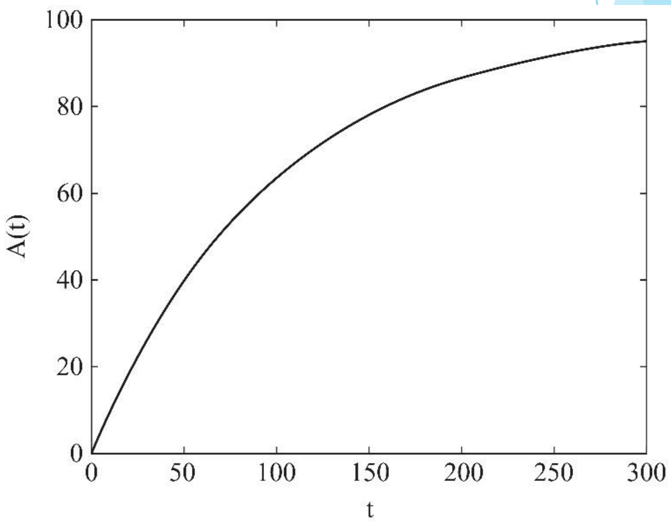
1. External-Influence Model

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)

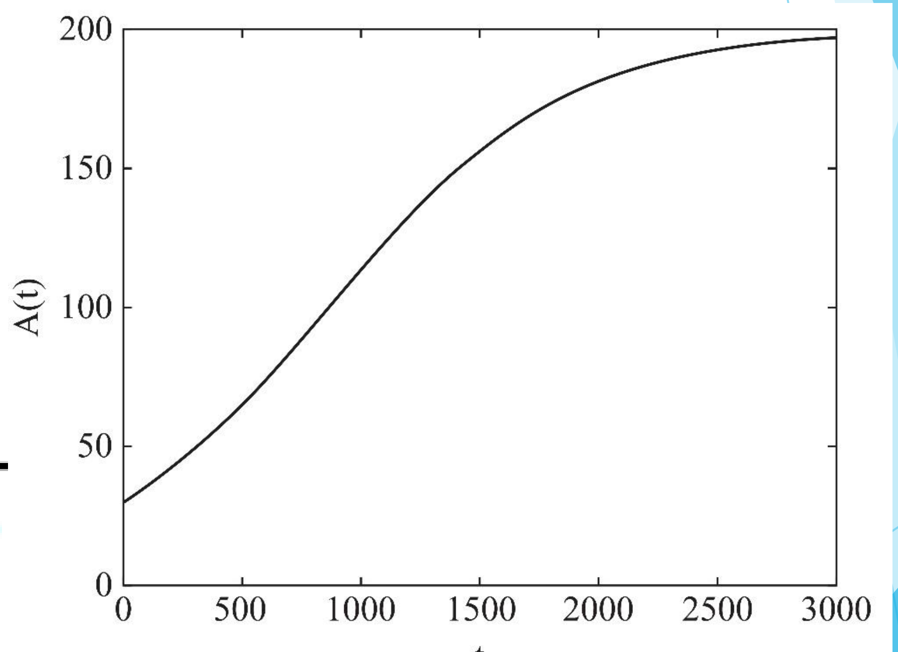
The adoption rate is a function that depends only on the number of already activated individuals

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

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

$$\Downarrow$$

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



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

3. Mixed-Influence Model

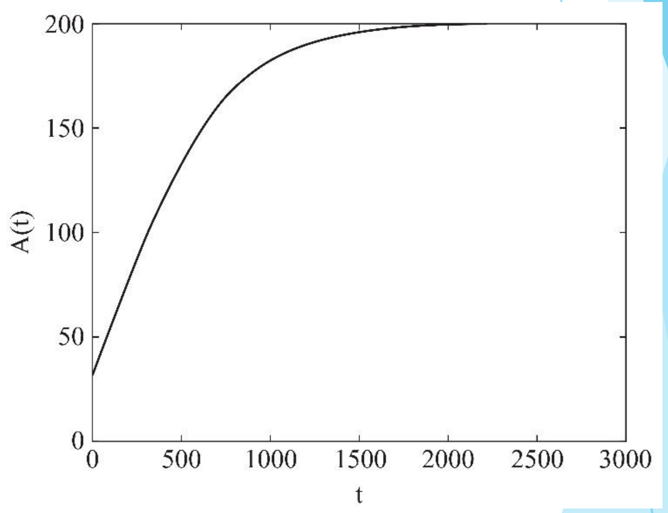
The adoption rate is a function that depends on both the number of already activated individuals and external forces,

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

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

↓

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



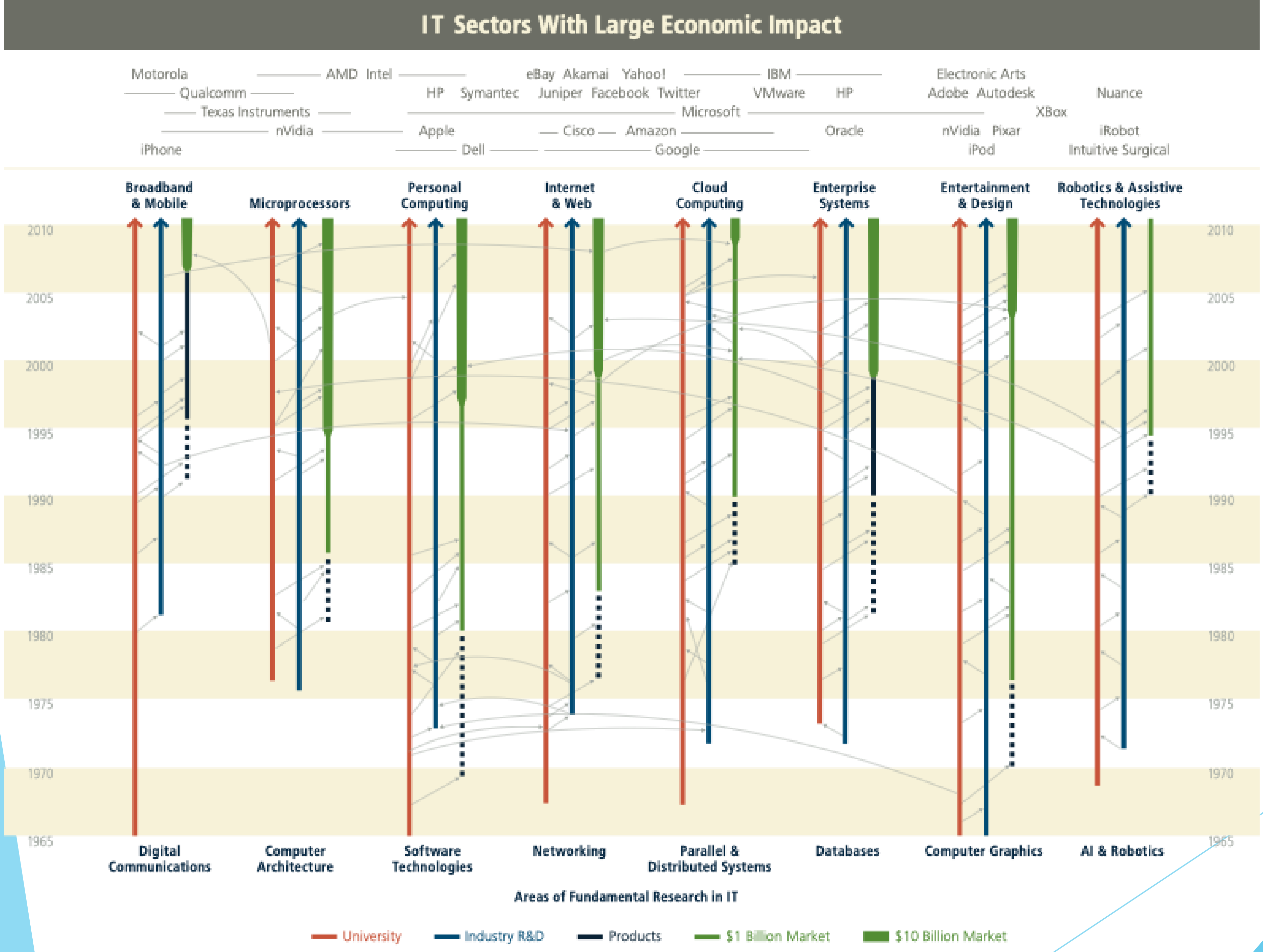
Simulation for
$P = 200$
$A_0 = 30$,
$\beta = 0.00001$ and $\alpha = 0.001$

Diffusion of Innovation: Intervention

1. **Limiting the distribution of the product or the audience that can adopt the product**.

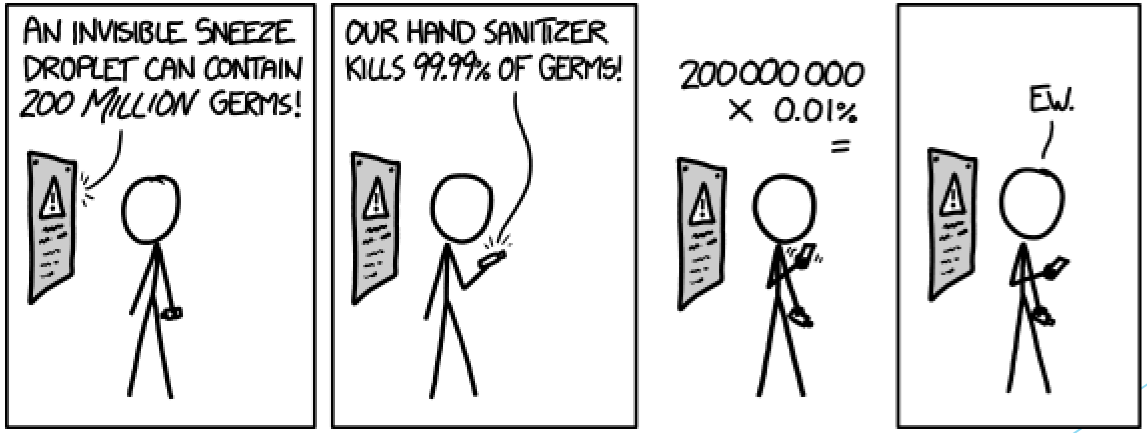
2. **Reducing interest in the product being sold.**
   - A company can inform adopters of the faulty status of the product.

3. **Reducing interactions within the population.**
   - Reduced interactions result in less imitations on product adoptions and a general decrease in the trend of adoptions.

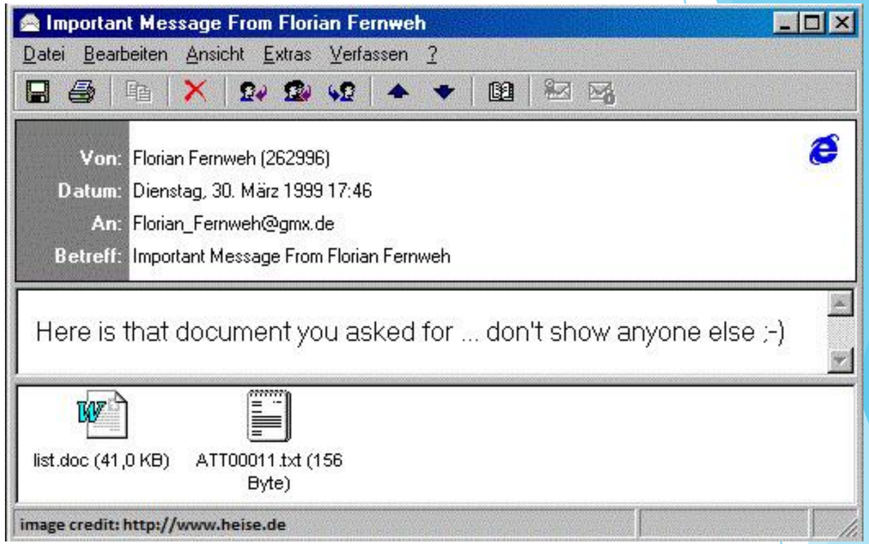


Tire Track Chart, produced by Computing Research Association, 2012

Epidemics


AN INVISIBLE SNEEZE DROPLET CAN CONTAIN 200 MILLION GERMS!
OUR HAND SANITIZER KILLS 99.99% OF GERMS!
200000000 × 0.01% =
EW.

Epidemics: Melissa computer worm


Important Message From Florian Fernweh
Datei Bearbeiten Ansicht Extras Verfassen ?

Von: Florian Fernweh (262996)
Datum: Dienstag, 30. März 1999 17:46
An: Florian_Fernweh@gmx.de
Betreff: Important Message From Florian Fernweh

Here is that document you asked for ... don't show anyone else ;-)

list.doc (41,0 KB) ATT00011.txt (156 Byte)

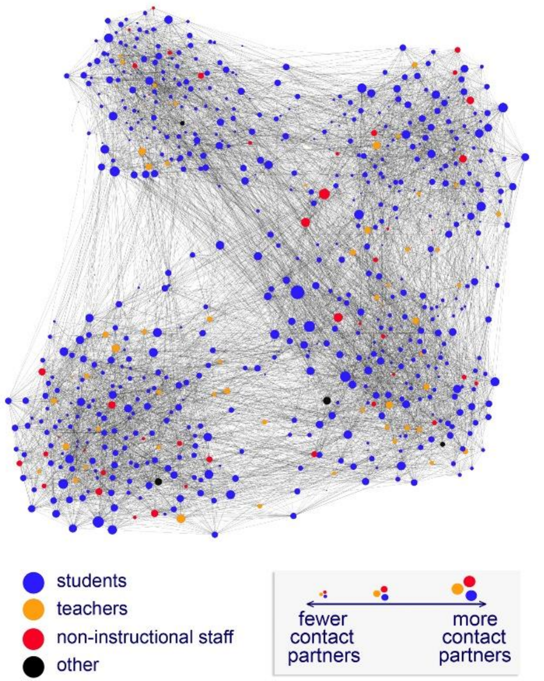
image credit: http://www.heise.de

- Started on March 1999
- Infected MS Outlook users
- The user
  - Receives email with a word document that has a virus
  - Once opened, the virus sends itself to the first 50 users in the outlook address book
- First detected on Friday, March 26
  - On Monday, March 29, the virus had infected more than 100K computers

Epidemics

Epidemics describes the process by which diseases spread. This process consists of

- A pathogen
  - The disease being spread
  - Tweet being retweeted
- A population of hosts
  - Humans, animals, plants, etc.
- A spreading mechanism
  - Breathing, drinking, sexual activity, etc.



Comparing Epidemics and Cascades

- Epidemic models assume an **implicit network** and unknown connections between users.
  - Unlike information cascades and herding
  - Similar to diffusion of innovations models,
- Epidemic models are more suitable when we are interested in global patterns
  - Trends
  - Ratios of people getting infected
  - Not suitable for who infects whom

How to Analyze Epidemics?

I. Using Contact Network

- look at how hosts contact each other and devise methods that describe how epidemics happen in networks.
- **Contact network**: a graph where nodes represent the hosts and edges represent the interactions between these hosts.
  - E.g., In influenza contact network, hosts (nodes) that breathe the same air are connected

II. Fully-mixed Method

- 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

Basic Epidemic Models

- SI
- SIR
- SIS
- SIRS

SI Model: Definition

SI model:

- The *susceptible* individuals get infected
- Once *infected,* they will never get cured

Two Types of Users:

- **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

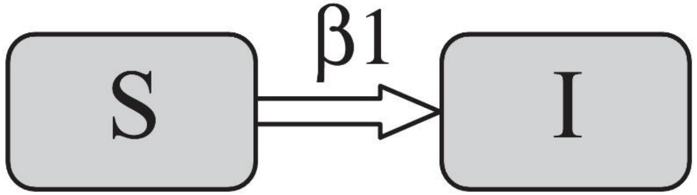
Notations

- 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
  - if β = 1 everyone comes to contact with everyone else
  - if β = 0 no one meets another individual

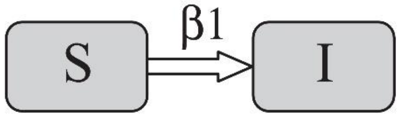
$$N = S(t) + I(t)$$

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

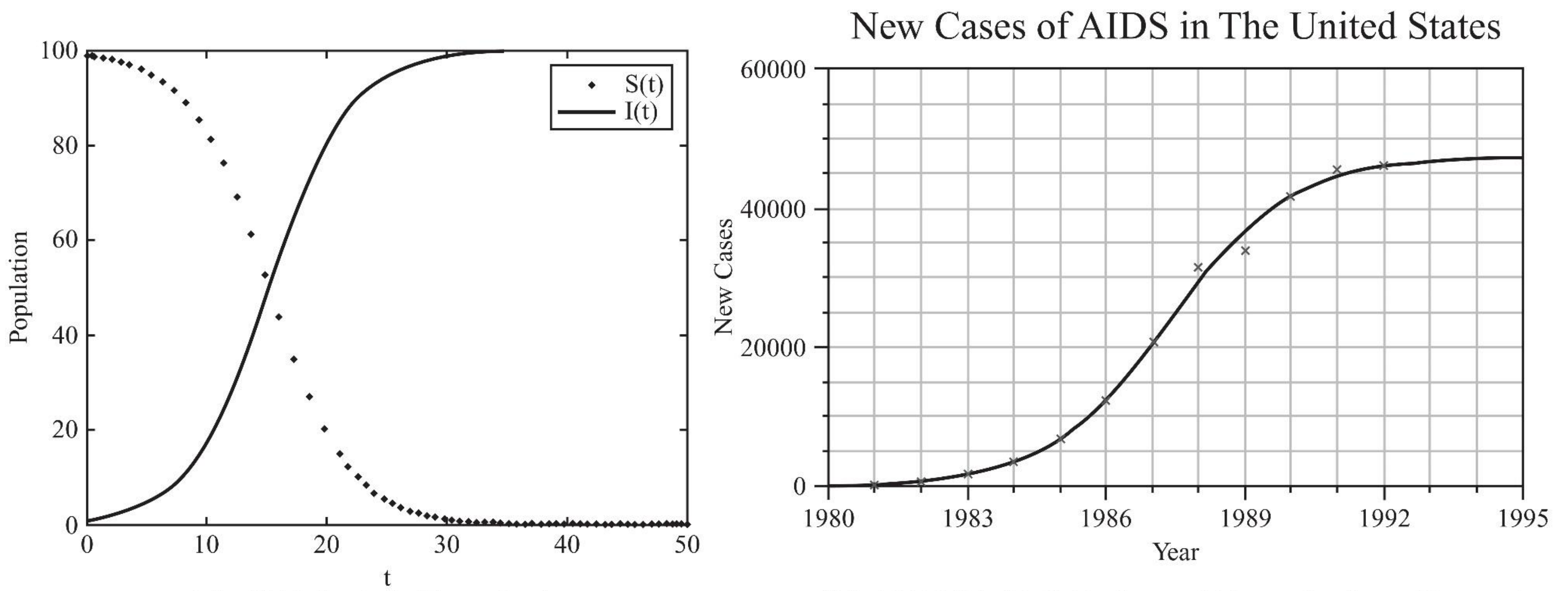


$$\frac{dS}{dt} = -\beta IS, \qquad \frac{dI}{dt} = \beta IS.$$

$$(S + I = N) \Longrightarrow \frac{dI}{dt} = \beta I(N - I) \Longrightarrow I(t) = \frac{NI_0e^{\beta tN}}{N + I_0(e^{\beta tN} - 1)}$$

I_0 is the number of individuals infected at time 0

SI Model: Example



(a) SI Model Simulation

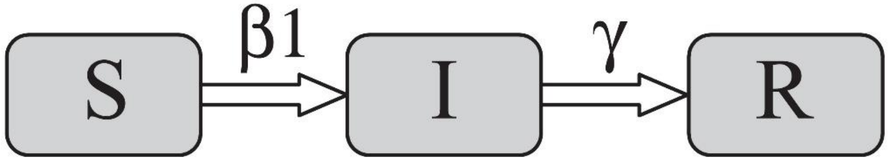
(b) HIV/AIDS Infected Population Growth

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

SIR Model

SIR model:

- In addition to the **I** and **S** states, a recovery state **R** is present
- Individuals get infected and some recover
- Once hosts recover (or are removed) they can no longer get infected and are not susceptible



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.

$$\begin{aligned}\frac{dS}{dt} &= -\beta IS,\\ \frac{dI}{dt} &= \beta IS - \gamma I,\\ \frac{dR}{dt} &= \gamma I.\end{aligned} \quad \Rightarrow \quad \frac{dS}{dR} = -\frac{\beta}{\gamma} S$$

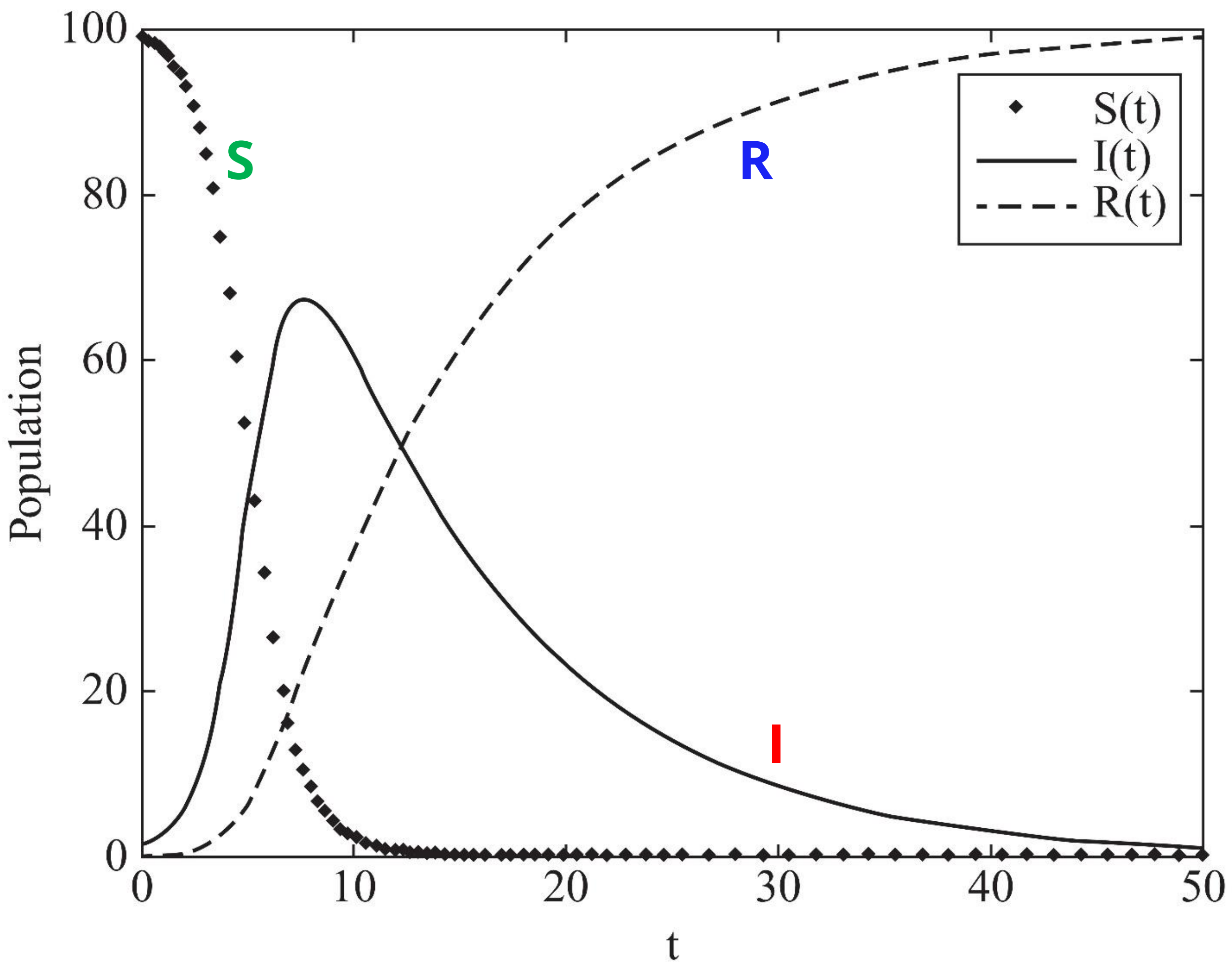
$$\Rightarrow \quad \log \frac{S_0}{S} = \frac{\beta}{\gamma} R \qquad (R_0 = 0)$$

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

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

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

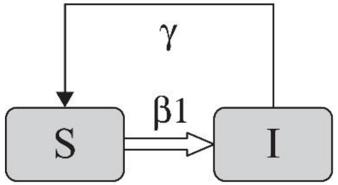
SIR Model: Example



SIR model simulated with $S_0 = 99$, $I_0 = 1$, $R_0 = 0$, $\beta = 0.01$, and $\gamma = 0.1$

SIS Model

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



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

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

SIS Model

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

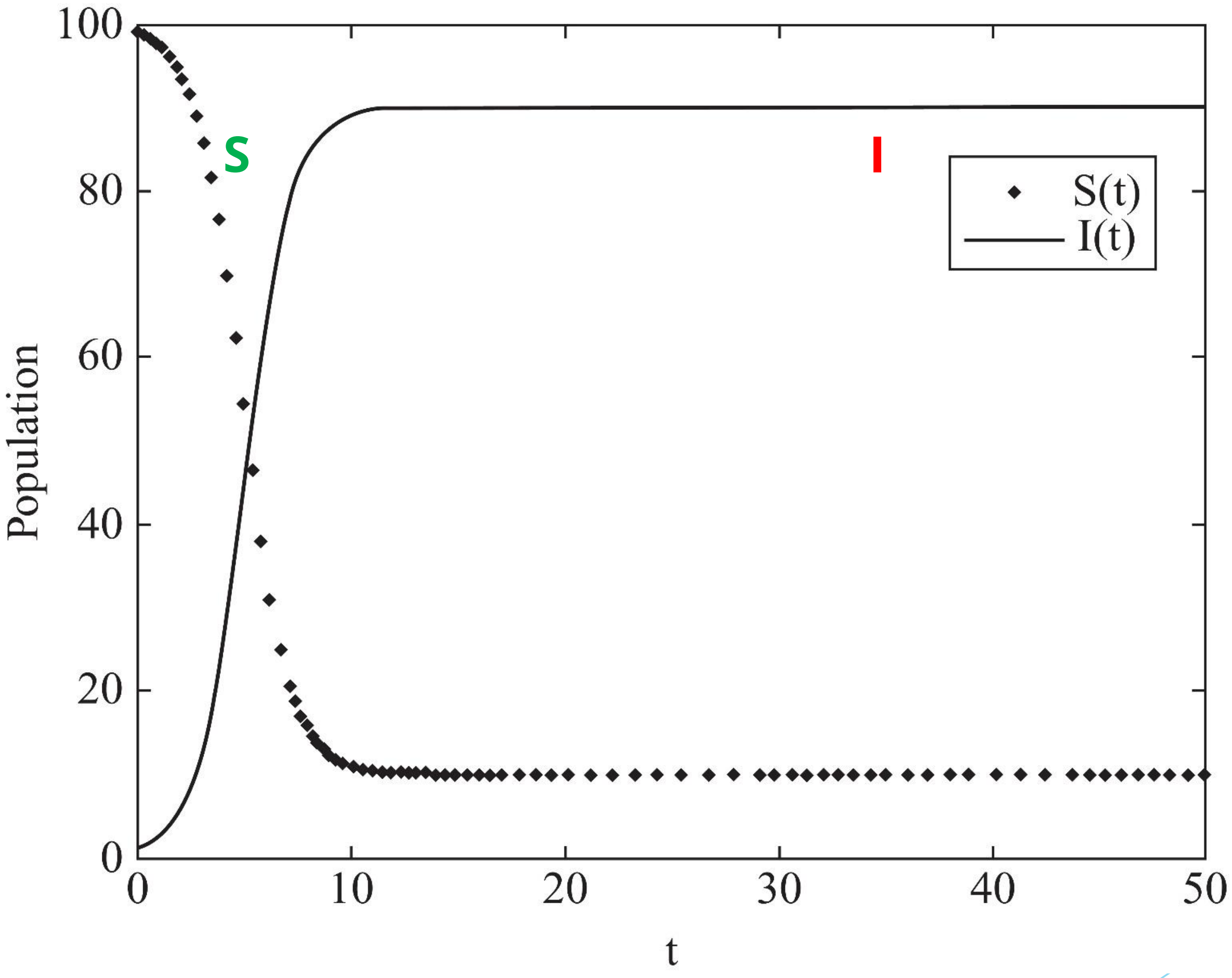
Case 1: When $\beta N \leq \gamma$ (or when $N \leq \frac{\gamma}{\beta}$):

- The first term will be at most zero or negative
- The whole term becomes negative
- In the limit, I(t) will decrease exponentially to zero

Case 2: When $\beta N > \gamma$ (or when $N > \frac{\gamma}{\beta}$):

- We will have a logistic growth function like the **SI** model

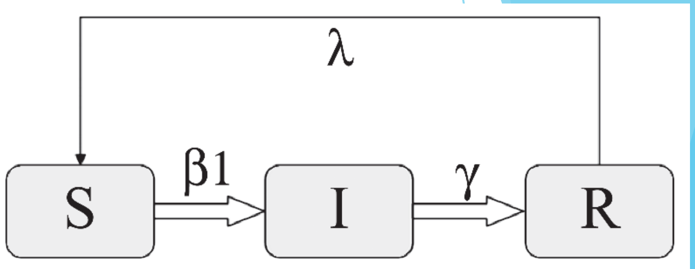
SIS Model



SIS model simulated with **S_0 = 99**, **I_0 = 1**, **β = 0.01**, and **γ = 0.1**

SIRS Model

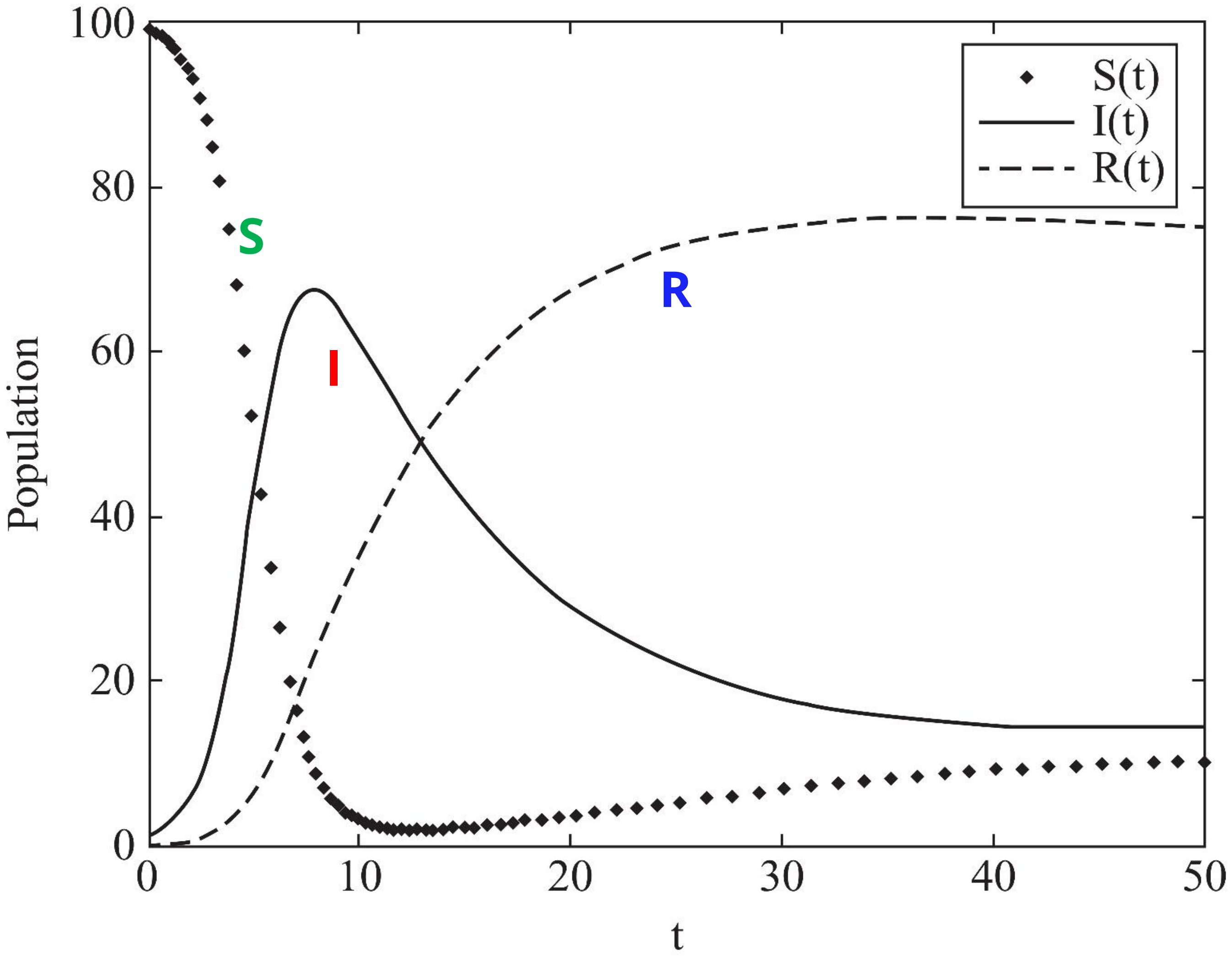
The individuals who have recovered will lose immunity after a certain period of time and will become susceptible again



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

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

SIRS Model



SIRS model simulated with S_0 = 99, I_0 = 1, R_0 = 1, β = 0.01, λ = 0.02, and γ = 0.1

Epidemic Intervention

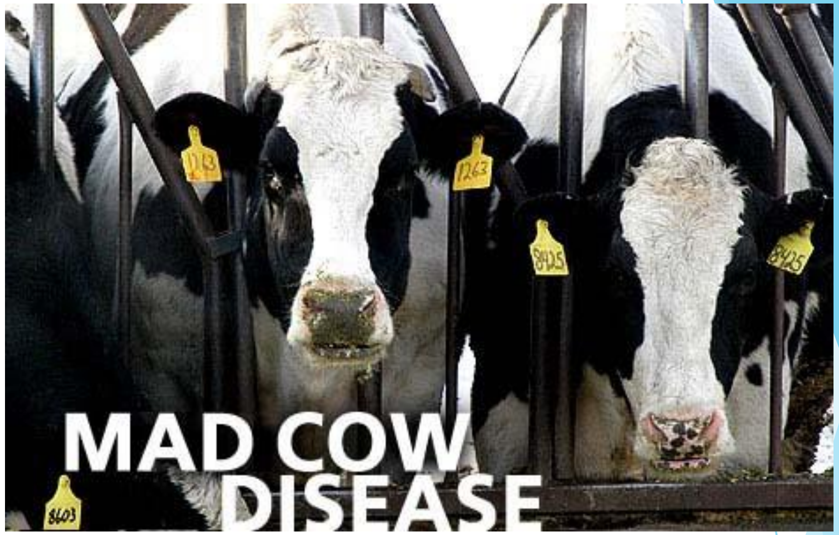
Epidemic Intervention

- Suppose that we have a susceptible society and want to prevent more spread by vaccinating the most vulnerable individuals
- How to find the most vulnerable individuals?

Randomly pick some nodes and ask them who is the most vulnerable from their point of view, then vaccinate those individuals!

Epidemic Intervention: *Mad Cow Disease*

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



In the mad cow disease case, we have weak ties

- Animals being bought and sold
- Soil from tourists, etc.

How to stop the disease:

- Banned movement (make contagion harder)
- Killed millions of animals (remove weak ties)