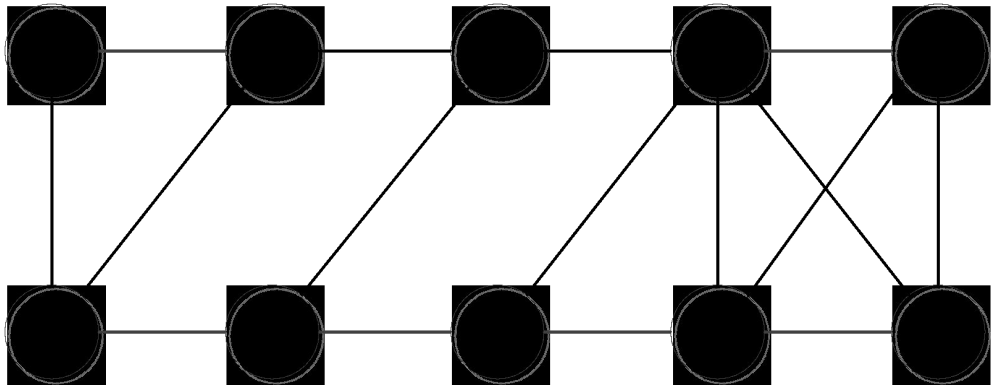
Maximizing the Spread of Cascades

Problem Setting

- **Given**
  - A limited budget **B** for initial advertising
    - Example: give away free samples of product
  - Connections 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: An Example

- We need to pick k nodes such that the maximum number of nodes can be activated

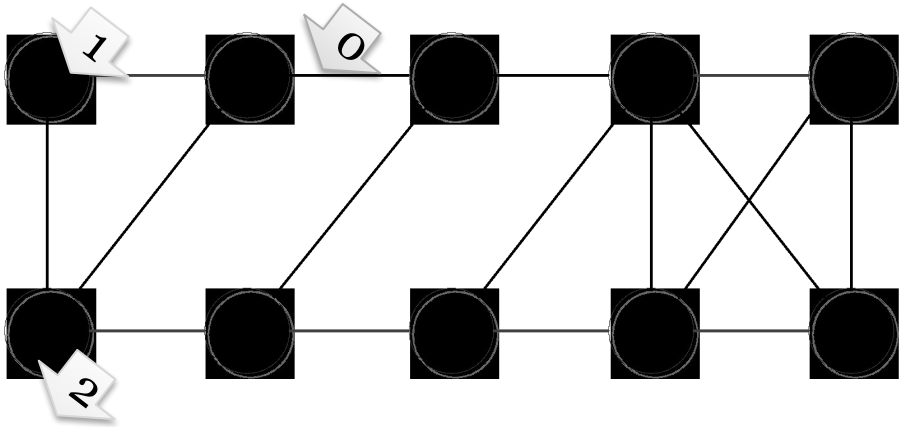


- What is the rule of activation?

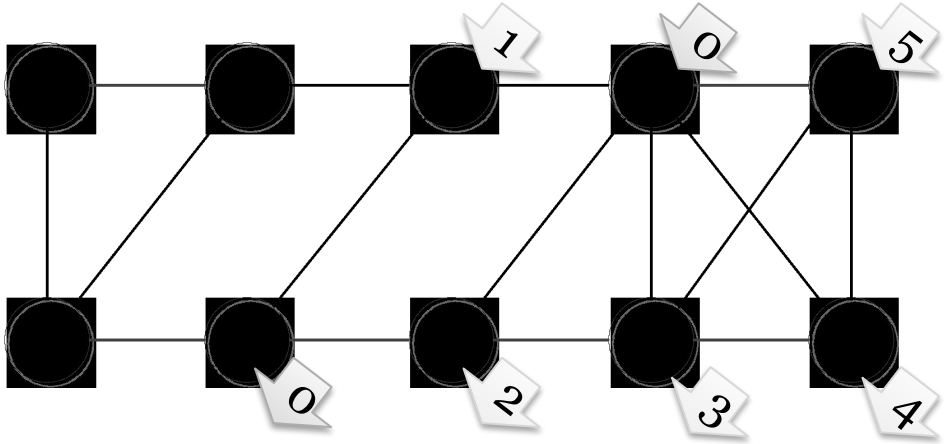
Maximizing the Spread of Cascade

In the first round, node 0 can infect or activate 1, subsequently, each inactive node requires two activated nodes to activate it

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

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

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

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

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.

Cascade Maximization: A Greedy Algorithm

The Algorithm

- Start with $B = \emptyset$
- *Evaluate f(v) for each node, and pick the node with maximum σ as the first node v1 to form B = {v1}*
- Select a node which will increase *f(B) most if the node is included in B.*

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

$$v = \arg\max_{v \in V \setminus B} f(B \cup \{v\})$$

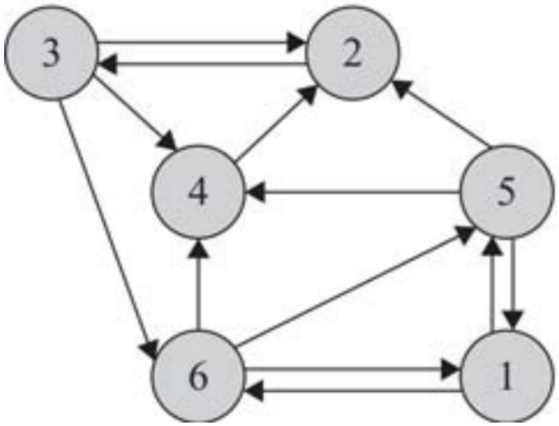
Cascade Maximization: A Greedy Algorithm

Algorithm 7.2 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;

Cascade Maximization Example



Algorithm 7.2 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;

- **Activation rule:** $|i - j| \equiv 2 \pmod 3$
- **Budget:** 2

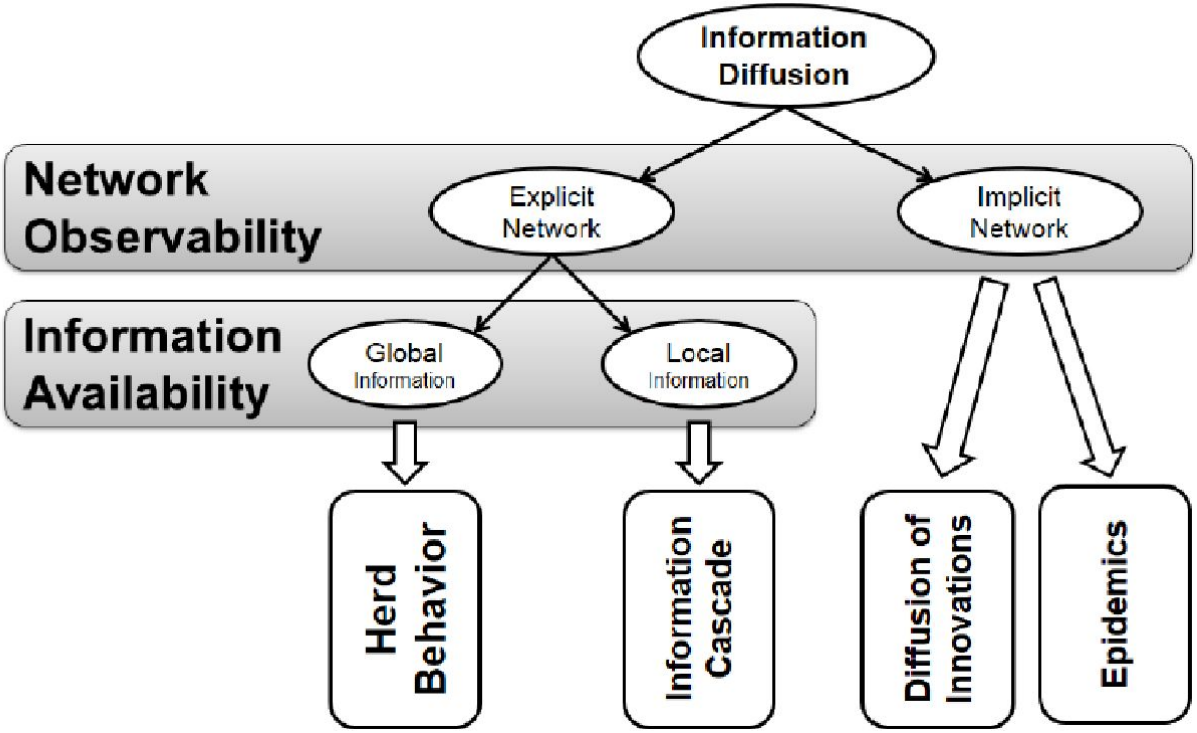
Intervention

- 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 probability p_{vw}
  - Reducing the chance of being activated by others

Diffusion of Innovations

- **The network is not observable**
- **Only public information is observable**

Information Diffusion Types



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

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
- **Compatible**
  - The degree to which the innovation is perceived to be consistent with socio- cultural values, previous ideas, and/or perceived needs
- **Trialable**
  - The degree to which the innovation can be experienced on a limited basis
- **Not too 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 (adoption types)

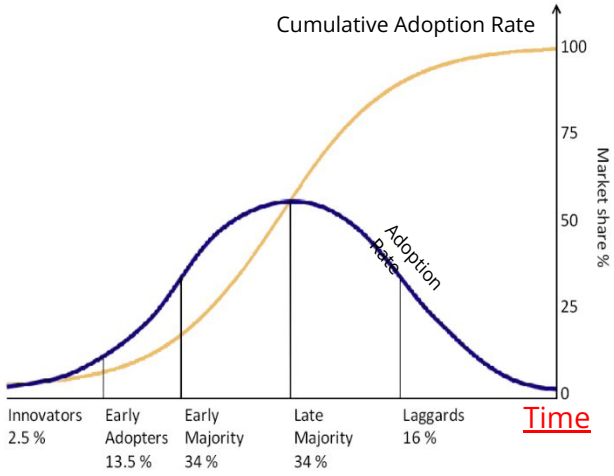
Farmers received information through two main channels

- **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.

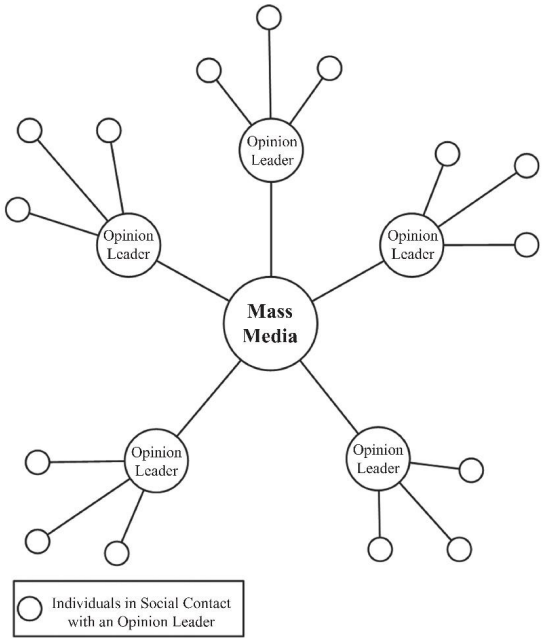
- The study showed that the adoption rate follows an *S*-shaped curve and that there are 5 different types of adopters based on the order that they adopt the innovations, namely:

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

- **A 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

III. EM Rogers (1962)'s Process of Diffusion of Innovations

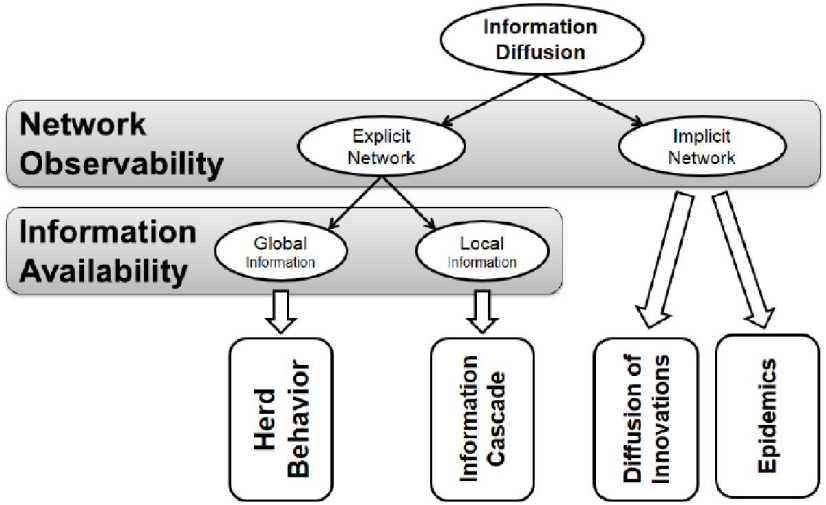
Adoption process:

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

Diffusion of Innovation: Intervention

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

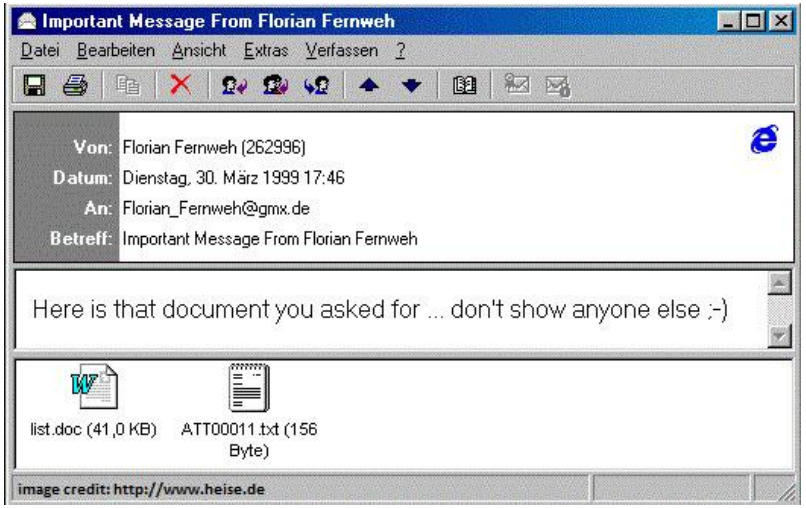
2. **Reducing interest in the product being sold**
   - A company can inform adopters of the faulty status of the product.
3. **Reducing interactions within the population**
   - Reduced interactions result in less imitations on product adoptions and a general decrease in the trend of adoptions.


Information Diffusion
Network Observability
Explicit Network
Implicit Network
Information Availability
Global Information
Local Information
Herd Behavior
Information Cascade
Diffusion of Innovations
Epidemics

Epidemics

Epidemics: Melissa computer worm

- Started on March 1999
- Infected MS Outlook users

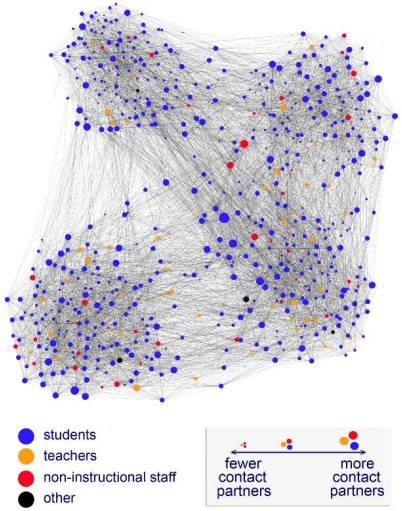


- The user
  - Receives email with a word document that has a virus
  - Once opened, the virus sends itself to the first 50 users in the outlook address book
- First detected on Friday, March 26
  - On Monday, March 29, the virus had infected more than 100K computers

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, retweeting, ...



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 Networks

- 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 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

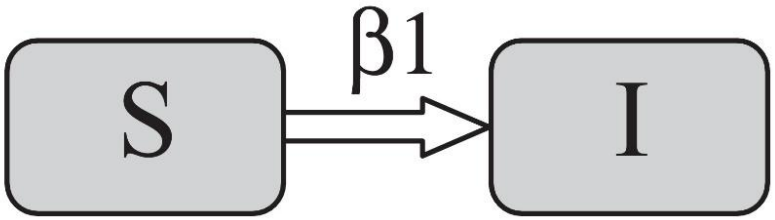
Notations

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

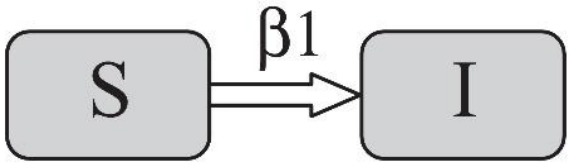
$$N = S(t) + I(t)$$

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

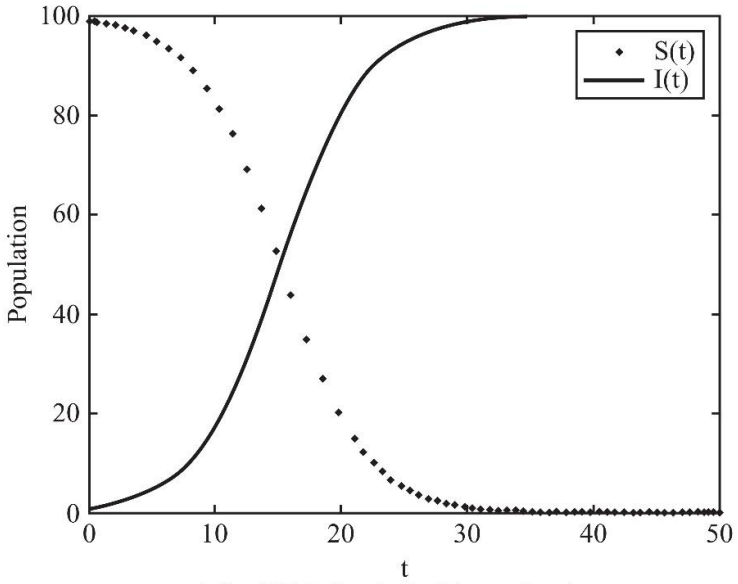


$$\frac{dS}{dt} = -\beta IS, \qquad \frac{dI}{dt} = \beta IS.$$

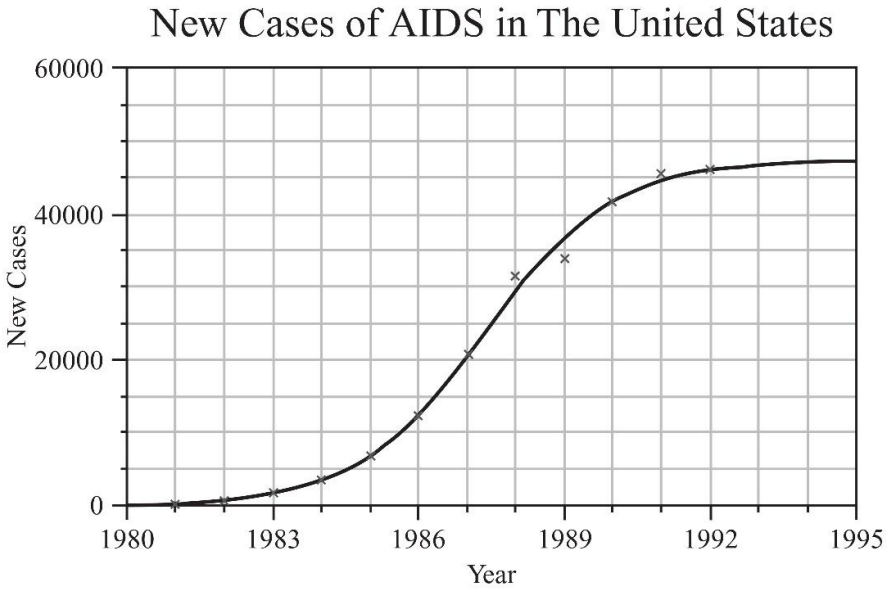
$$(S + I = N) \Rightarrow \frac{dI}{dt} = \beta I(N - I) \Rightarrow I(t) = \frac{NI_0 e^{\beta t N}}{N + I_0(e^{\beta t N} - 1)}$$

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