# Probabilistic Contagion and Models of Influence

Thanks to Jure Leskovec, Stanford and Panayiotis Tsaparas, Univ. of Ioannina for slides

# Agenda

- Epidemic Model Based on Trees
- Models of Disease Spreading
- Independent Cascade Model
- Modeling Interactions Between Contagions (Optional)

# Epidemics

Understanding the spread of viruses and epidemics is of great interest to

- Health officials
- Sociologists
- Mathematicians
- Hollywood


The underlying contact network clearly affects the spread of an epidemic

# Epidemics

- Model epidemic spread as a random process on the graph and study its properties
- Questions that we can answer:
  - What is the projected growth of the infected population?
  - Will the epidemic take over most of the network?
  - How can we contain the epidemic spread?

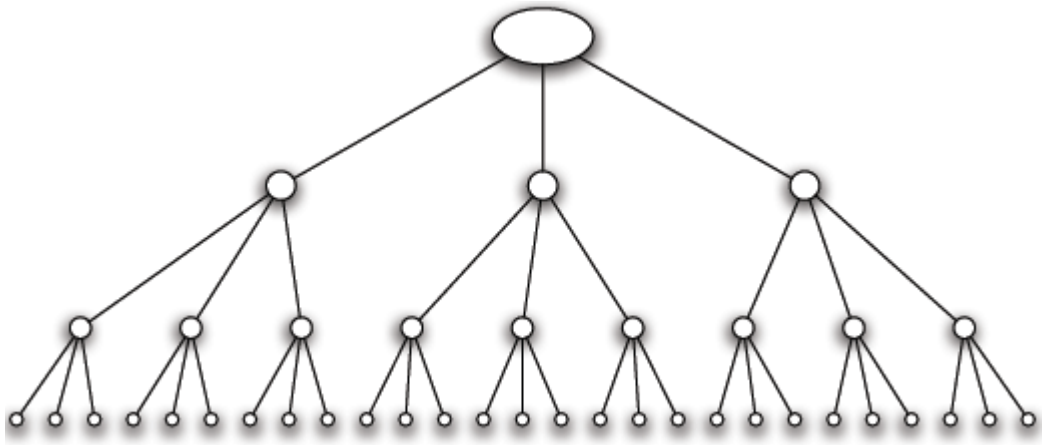
Diffusion of ideas and the spread of influence can also be modeled as epidemics

# Epidemic Model Based on Trees

Simple probabilistic model of cascades where we will learn about the **reproductive number**

# A Simple Model

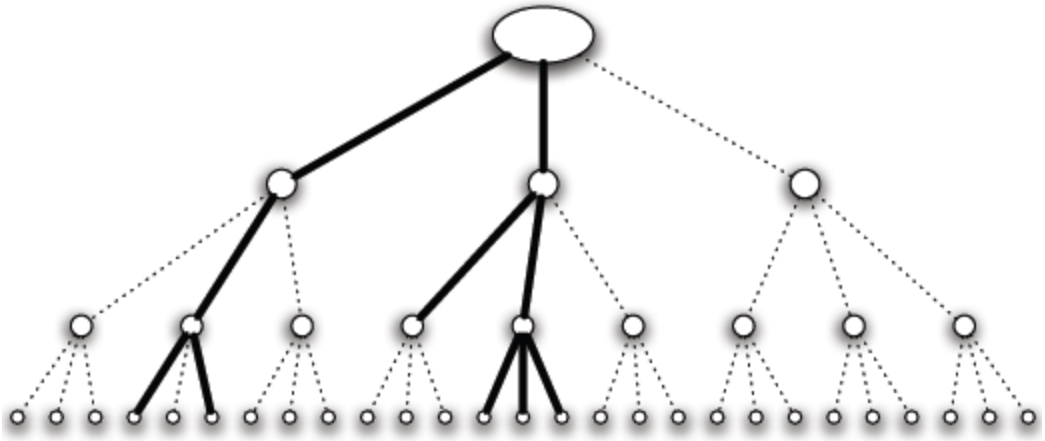
- **Branching process**: A person transmits the disease to each people she meets independently with a probability ***p***
- An infected person meets ***k*** (new) people while she is contagious
- Infection proceeds in **waves**



Contact network is a tree with branching factor *k*

# Infection Spread

- We are interested in the number of people infected (spread) and the duration of the infection
- This depends on the infection probability ***p*** and the branching factor ***k***

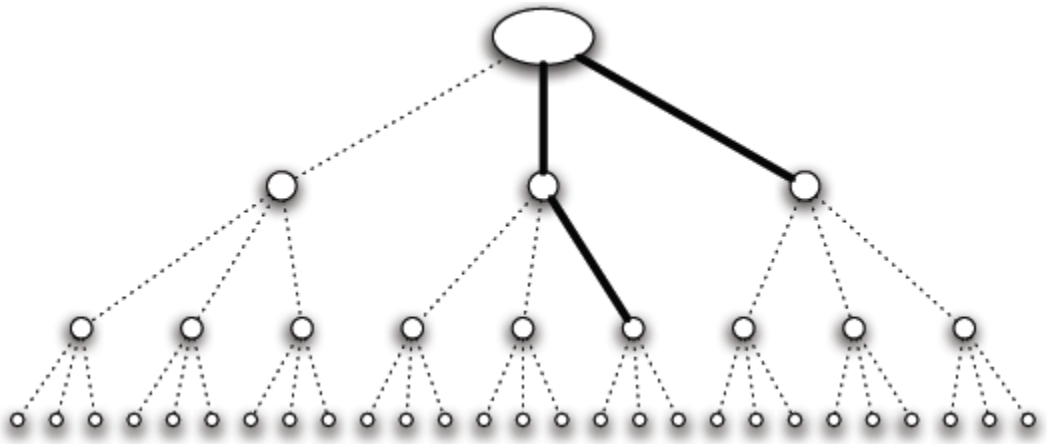


An aggressive epidemic with high infection probability

The epidemic survives after three steps

# Infection Spread

- We are interested in the number of people infected (spread) and the duration of the infection
- This depends on the infection probability ***p*** and the branching factor ***k***



A mild epidemic with low infection probability

The epidemic dies out after two steps

# Basic Reproductive Number

- Basic Reproductive Number ($\boldsymbol{R_0}$): the expected number of new cases of the disease caused by a single individual

$$\boldsymbol{R_0 = kp}$$

- Claim: (a) If $R_0 < 1$, then with probability 1, the disease dies out after a finite number of waves. (b) If $R_0 > 1$, then with probability greater than 0 the disease persists by infecting at least one person in each wave

1. If $\boldsymbol{R_0 < 1}$ each person infects less than one person in expectation. The infection eventually ***dies out***
2. If $\boldsymbol{R_0 > 1}$ each person infects more than one person in expectation. The infection ***persists***

# Analysis

- $X_n$: random variable indicating the number of infected nodes after ***n*** steps
- $q_n = \Pr[X_n \geq 1]$: probability that there exists at least 1 infected node after ***n*** steps
- $q^* = \lim q_n$: the probability of having infected nodes as $n \to \infty$

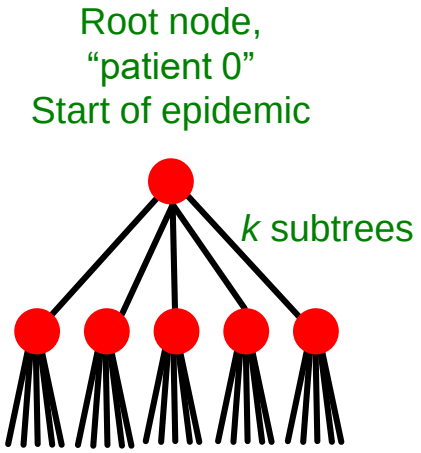
It can be shown that

(a) $R_0 < 1 \Rightarrow q^* = 0$

(b) $R_0 > 1 \Rightarrow q^* > 0$.

# Probabilistic Spreading Models

- **Epidemic Model based on Random Trees**
  - **(a variant of branching processes)**
  - A patient meets ***k*** other people
  - With probability ***p > 0*** infects each of them




- **Q: For which values of *k* and *p* does the epidemic run forever?**
  - **Run forever:** $\lim_{n\to\infty} P\left[\begin{array}{l}\text{At least 1 infected}\\ \text{node at depth n}\end{array}\right] > 0$
  - **Die out:** -- || -- $= 0$

# Probabilistic Spreading Models

- $q_{nj}$ = **prob. there is an infected node at depth $n$ *starting from a specific child node***

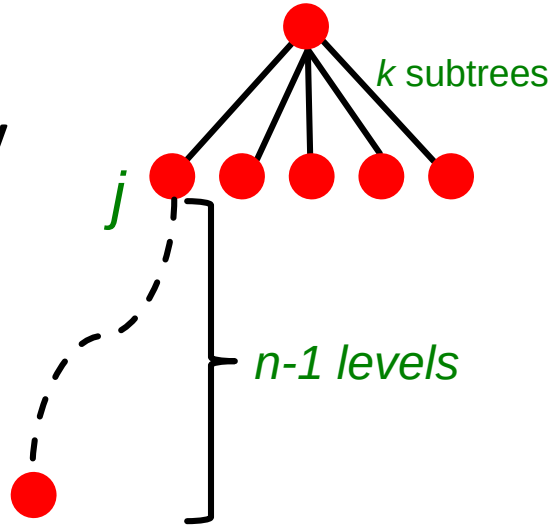
$$q_{nj} = p \cdot q_{n-1}$$

- Fails with probability (the complementary view)

$$1 - p \cdot q_{n-1}$$

- All **k** subtrees fail with probability

$$(1 - p \cdot q_{n-1})^k$$



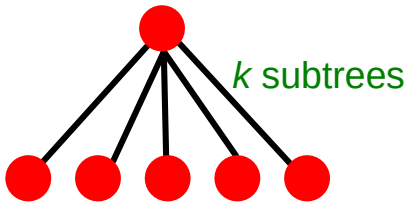
# Probabilistic Spreading Models

- $q_n$ = **prob. there is an infected node at depth $n$**
- **We need:** $\lim_{n\to\infty} q_n = ?$ (based on $p$ and $k$)
- All **k** subtrees fail with probability

$$(1 - p \cdot q_{n-1})^k$$

- **Taking the complement:**

$$q_n = 1 - \underbrace{(1 - p \cdot q_{n-1})^k}_{\text{No infected node at depth } n \text{ from the root}}$$




- $\lim_{n\to\infty} q_n$ = **result of iterating**

$$f(x) = 1 - (1 - p \cdot x)^k$$

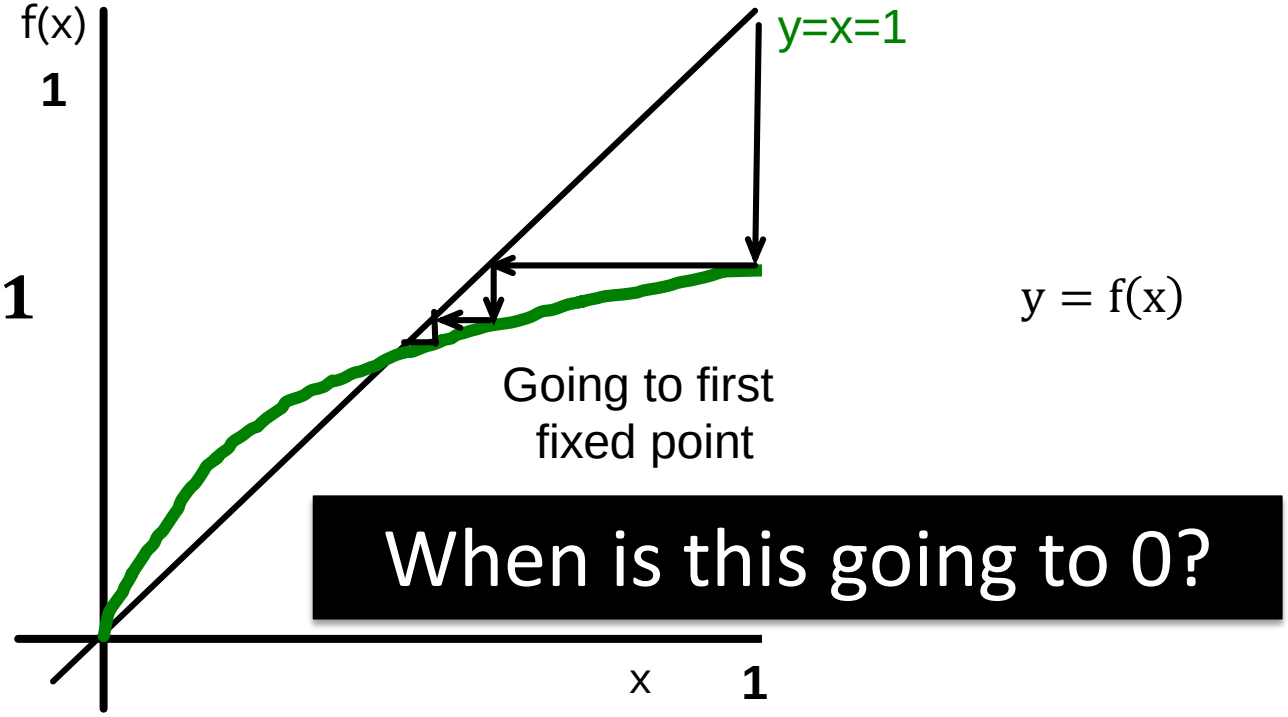
  - Starting at $x = 1$ (since $q_1 = 1$)

# Properties of $f(x) = 1 - (1 - px)^{k}$

- $f(0) = 0$ , *so intercepts at point (0,0)*
- $f(1) = 1 - (1 - p)^{k} < 1$, *so at x=1, f(1) is below the y=x line*
- $f'(x) = p \cdot k(1 - px)^{k-1}$, positive and f' monotonically decreasing on [0,1], so concave curve
- $f'(0) = p \cdot k = R_0$, so
  - for $\boldsymbol{R_0 > 1}$ f starts above the y=x line
  - for $\boldsymbol{R_0 < 1}$ f starts below the y=x line

# Fixed Point: $f(x) = 1 - (1 - px)^k$

**Case 1:**
$\boldsymbol{R_0 = pk > 1}$

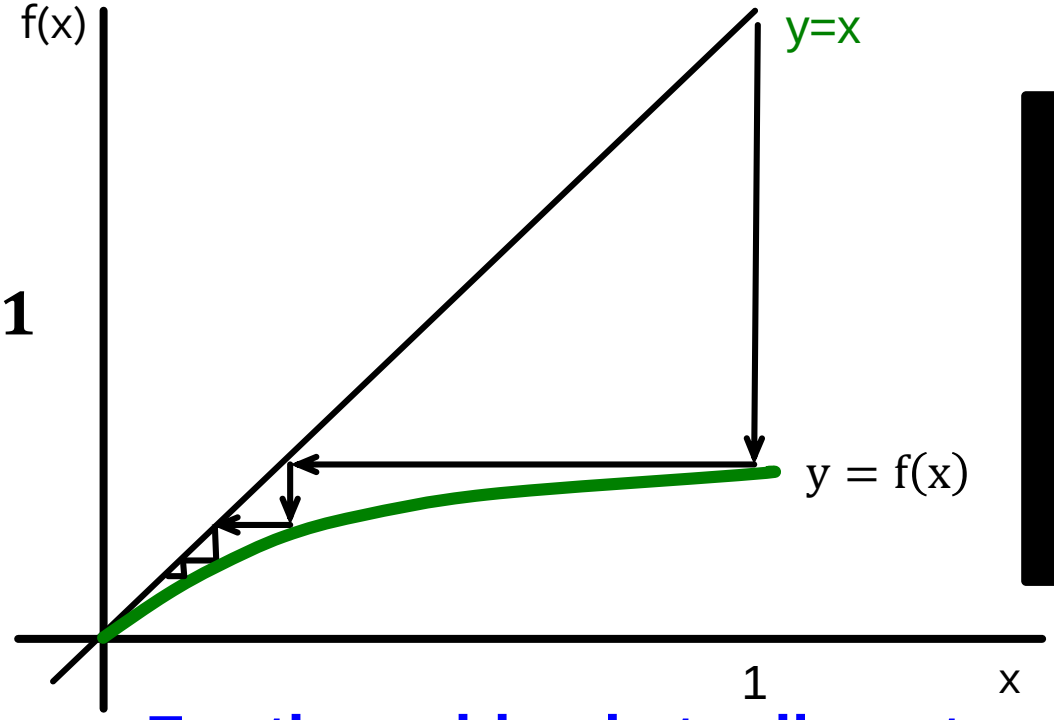



**What do we know about f(x)?**

$f(0) = 0, f(1) = 1 - (1 - p)^k < 1$, *so at x=1, f(1) is below the y=x line*
$f'(x) = p \cdot k(1 - px)^{k-1}$, so concave on [0,1]
$f'(0) = p \cdot k = R_0$, so for $R_0 > 1$ f starts above the y=x line

# Fixed Point: When is this zero?

**Case 2:**
$R_0 = pk < 1$




**Reproductive number $R_0 = p \cdot k$:**
There is an epidemic if $R_0 \geq 1$

**For the epidemic to die out we need *f(x)* to be below *y=x*!**

**So:** $f'(0) = p \cdot k < 1$

$$\lim_{n \to \infty} q_n = 0 \ \ when \ \ p \cdot k < 1$$

**$p \cdot k$ = expected # of people that we infect**

# Branching process

- Assumes no network structure, no triangles or shared neighbors

# Models of Disease Spreading

# The SIR model

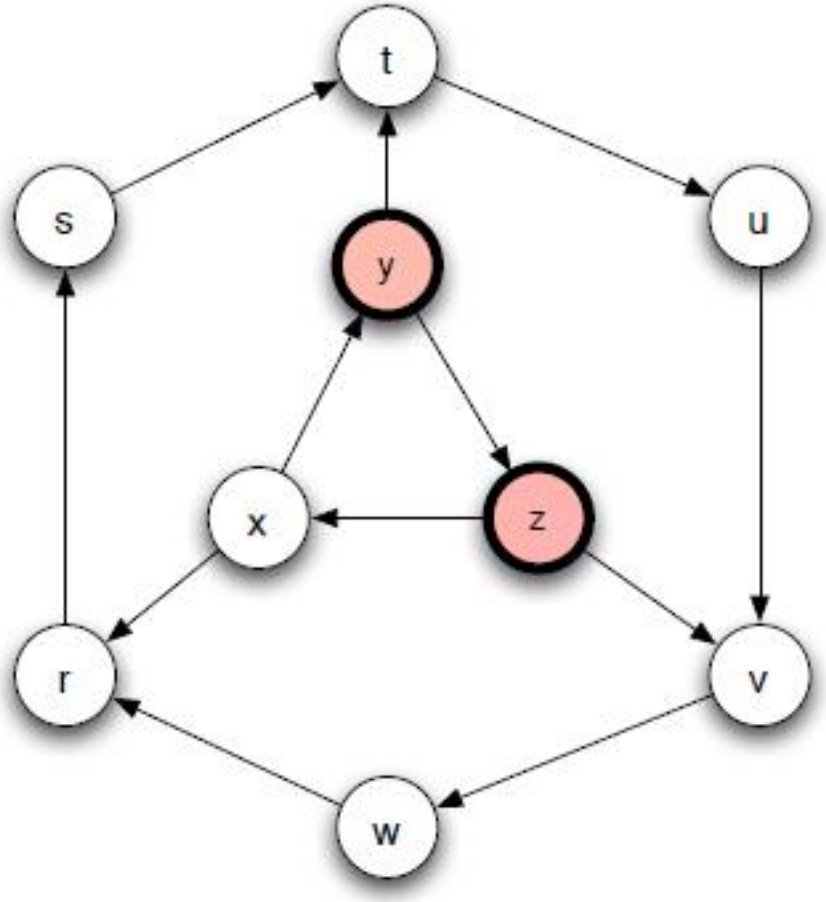
- Each node may be in the following states
  - Susceptible: healthy but not immune
  - Infected: has the virus and can actively propagate it
  - Removed: (Immune or Dead) had the virus but it is no longer active

- Parameter *p*: the probability of an Infected node to infect a Susceptible neighbor

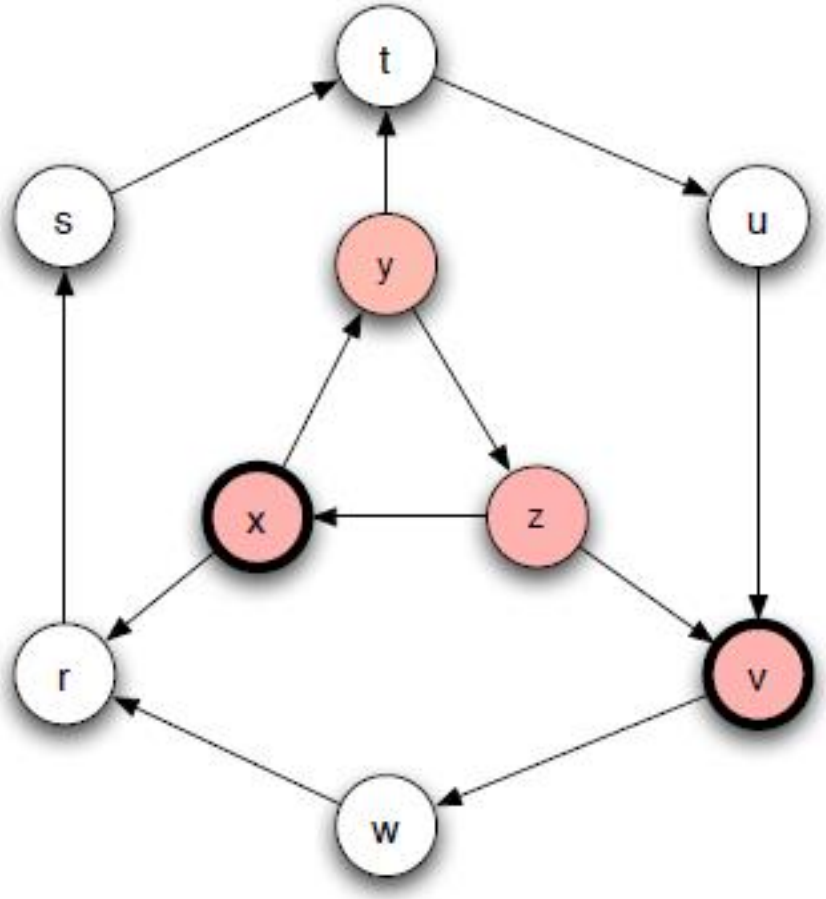
# The SIR process

- Initially all nodes are in state S(usceptible), except for a few nodes in state I(nfected).
- An infected node stays infected for $t_I$ steps.
  - Simplest case: $t_I = 1$
- At each of the $t_I$ steps the infected node has probability *p* of infecting any of its susceptible neighbors
  - *p*: Infection probability
- After $t_I$ steps the node is Removed

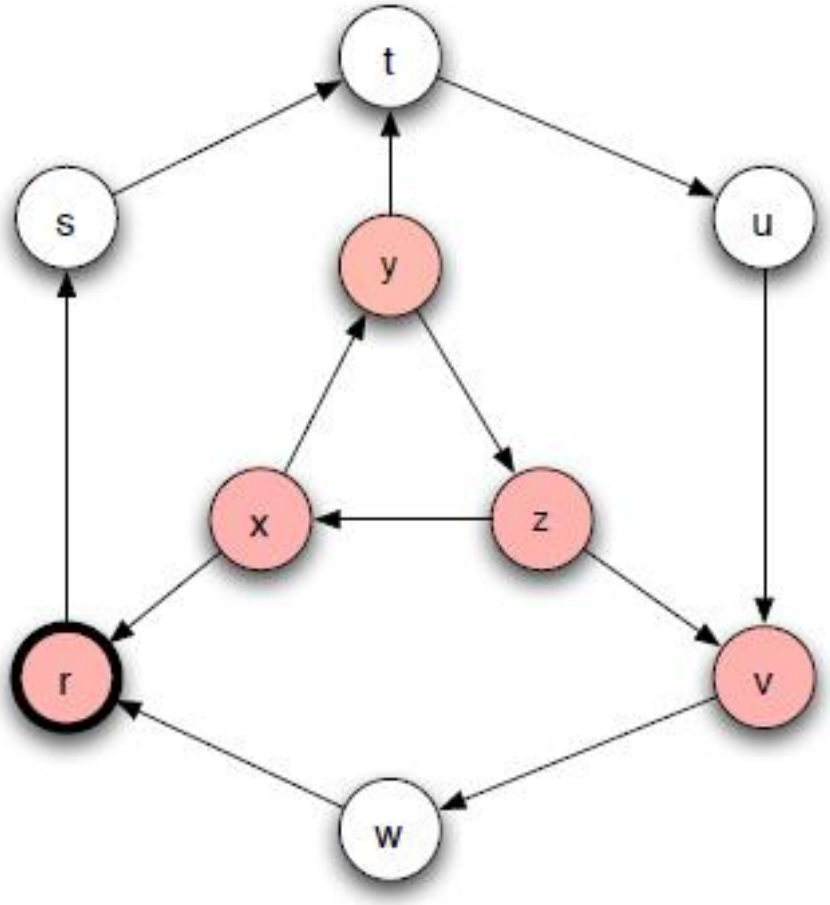
# Example



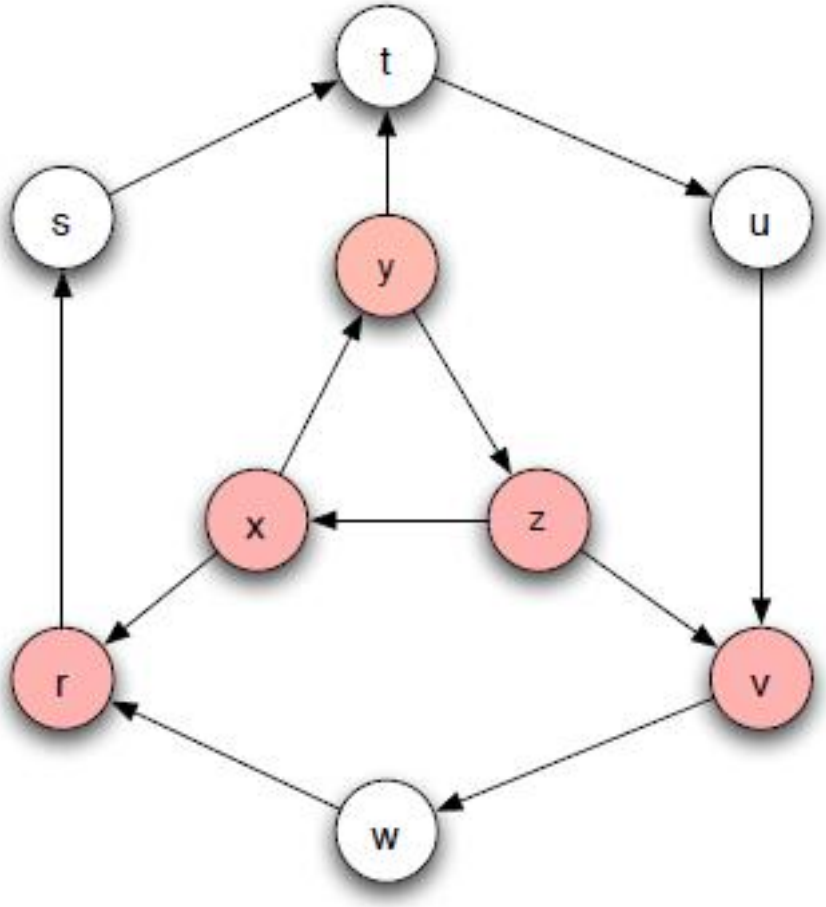
# Example



# Example



# Example



# Example SIR Epidemic

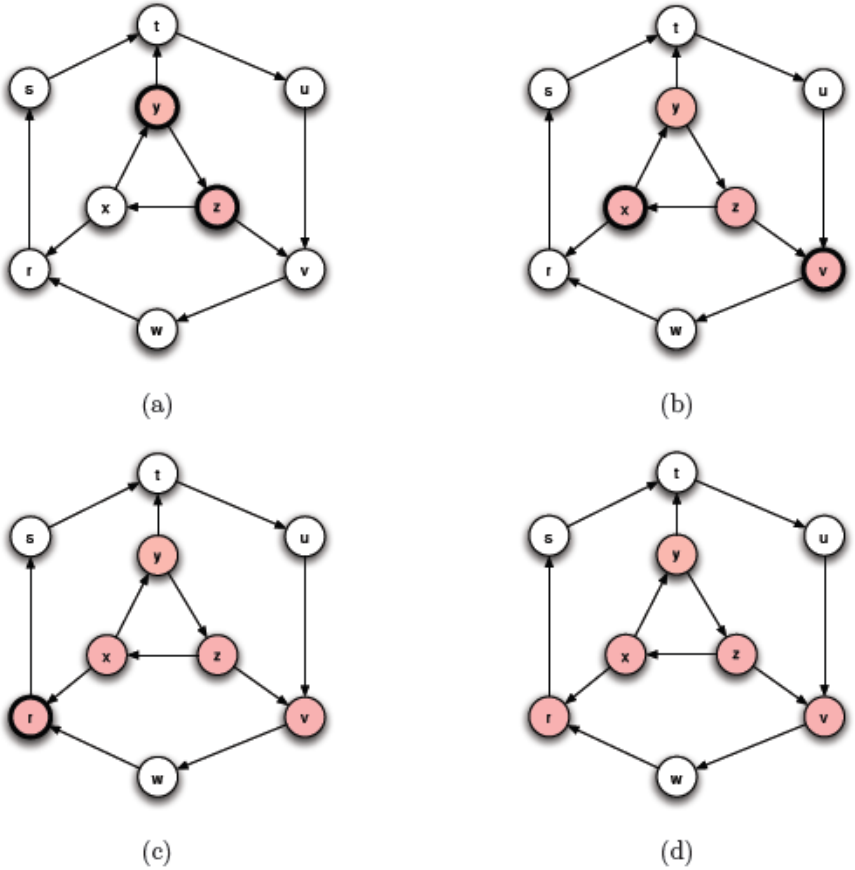



Figure 21.2: The course of an SIR epidemic in which each node remains infectious for a number of steps equal to $t_I = 1$. Starting with nodes $y$ and $z$ initially infected, the epidemic spreads to some but not all of the remaining nodes. In each step, shaded nodes with dark borders are in the Infectious ($I$) state and shaded nodes with thin borders are in the Removed ($R$) state.

# SIR and the Branching process

- The branching process is a special case where the graph is a tree (and the infected node is the root)
  - The existence of triangles shared neighbors makes a big difference
- The basic reproductive number is not necessarily informative in the general case

# SIR and the Branching process

## Example

$R_0$ the expected number of new cases caused by a single node

assume $p = 2/3$, $R_0 = 4/3 > 1$

Probability to fail at each level and stop $(1/3)^4 = 1/81$

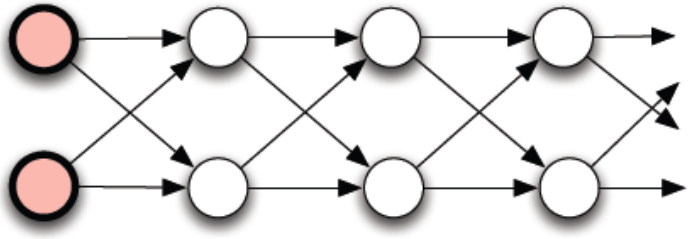


Figure 21.3: In this network, the epidemic is forced to pass through a narrow "channel" of nodes. In such a structure, even a highly contagious disease will tend to die out relatively quickly.

# Percolation

- Percolation: we have a network of “pipes” which can carry liquids, and they can be either open, or closed
  - The pipes can be pathways within a material
- If liquid enters the network from some nodes, does it reach most of the network?
  - The network percolates

# SIR and Percolation

- There is a connection between SIR model and percolation
- When a virus is transmitted from u to v, the edge (u,v) is activated with probability p
- We can assume that all edge activations have happened in advance, and the input graph has only the active edges
- Which nodes will be infected?
  - The nodes reachable from the initial infected nodes
- In this way we transformed the dynamic SIR process into a static one
  - This is essentially percolation in the graph

# Example

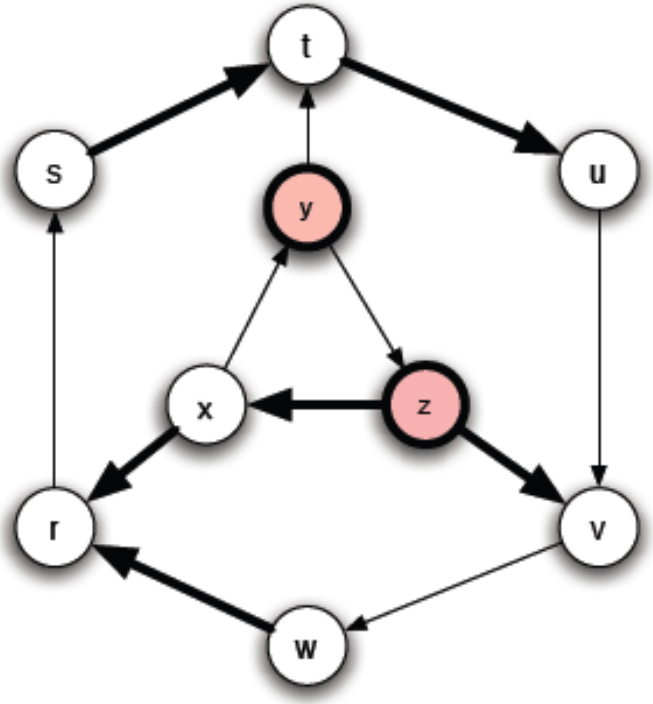



Figure 21.4: An equivalent way to view an SIR epidemic is in terms of *percolation*, where we decide in advance which edges will transmit infection (should the opportunity arise) and which will not.

# The SIS model

- Susceptible-Infected-Susceptible
  - Susceptible: healthy but not immune
  - Infected: has the virus and can actively propagate it
- An Infected node infects a Susceptible neighbor with probability ***p***
- An Infected node becomes Susceptible again with probability ***q*** (or after $t_I$ steps)
  - In a simplified version of the model q = 1
- Nodes alternate between Susceptible and Infected status

# Example

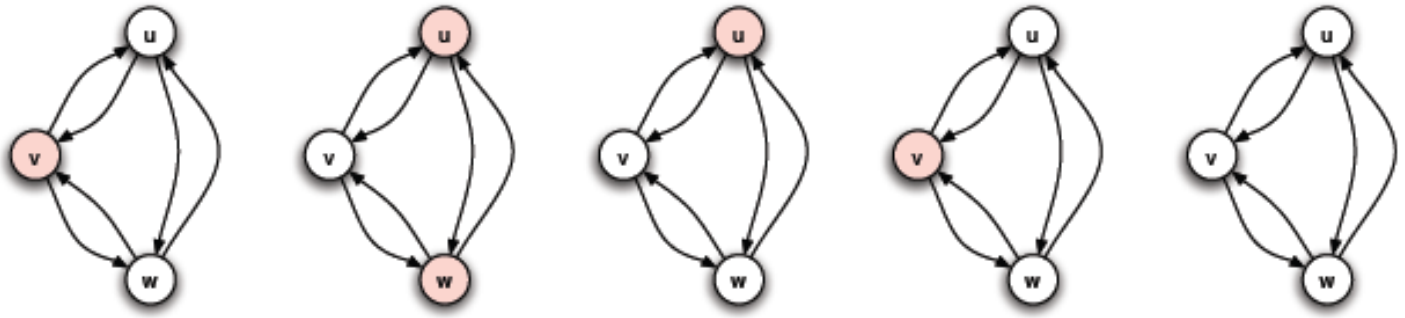



Figure 21.5: In an SIS epidemic, nodes can be infected, recover, and then be infected again. In each step, the nodes in the Infectious state are shaded.

- When no Infected nodes, virus dies out
- Question: will the virus die out?

# An eigenvalue point of view

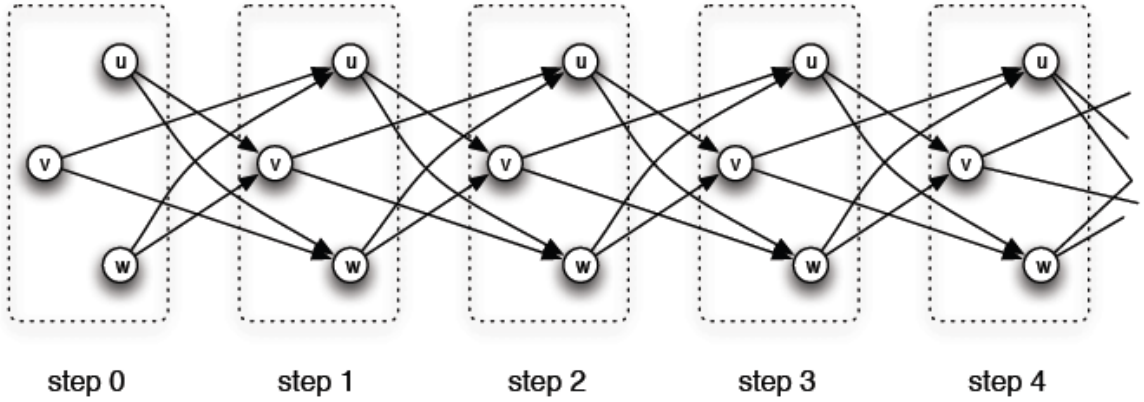
- If A is the adjacency matrix of the network, then the virus dies out if

$$\lambda_1(A) \leq \frac{q}{p}$$

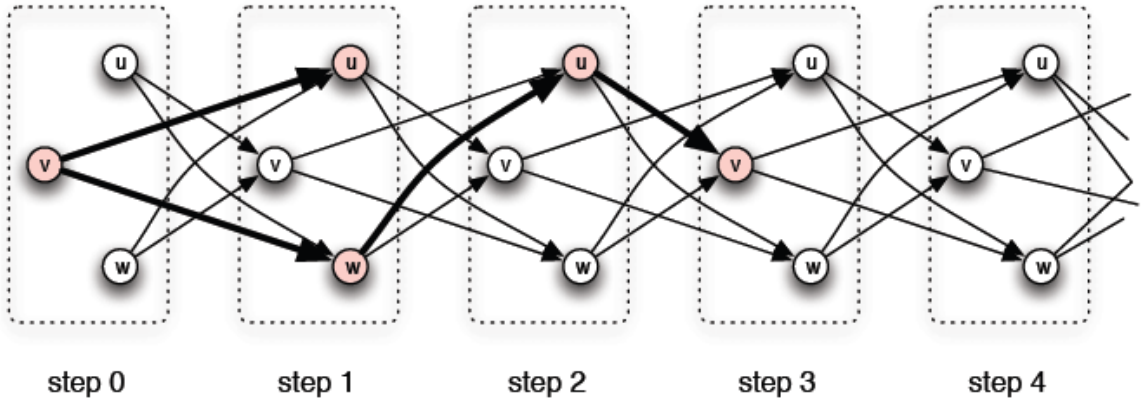
- Where $\lambda_1(A)$ is the first eigenvalue of A

Y. Wang, D. Chakrabarti, C. Wang, C. Faloutsos. *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*. SRDS 2003

# SIS and SIR




(a) *To represent the SIS epidemic using the SIR model, we use a "time-expanded" contact network*

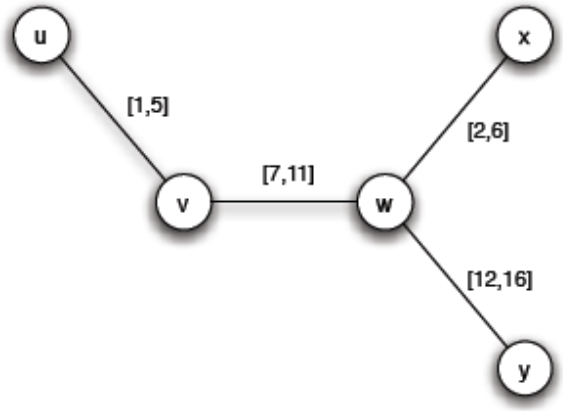



(b) *The SIS epidemic can then be represented as an SIR epidemic on this time-expanded network.*

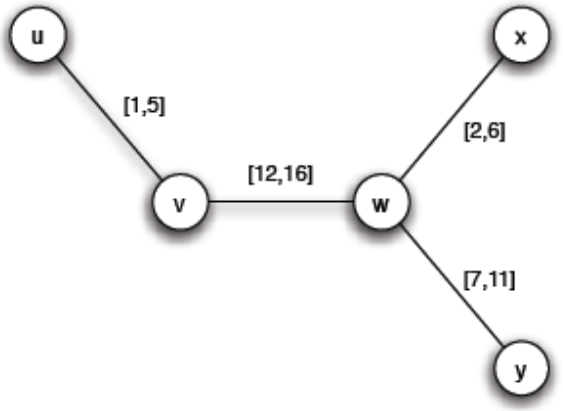
Figure 21.6: An SIS epidemic can be represented in the SIR model by creating a separate copy of the contact network for each time step: a node at time $t$ can infect its contact neighbors at time $t + 1$.

# Including time

- Infection can only happen within the active window




(a) *In a contact network, we can annotate the edges with time windows during which they existed.*

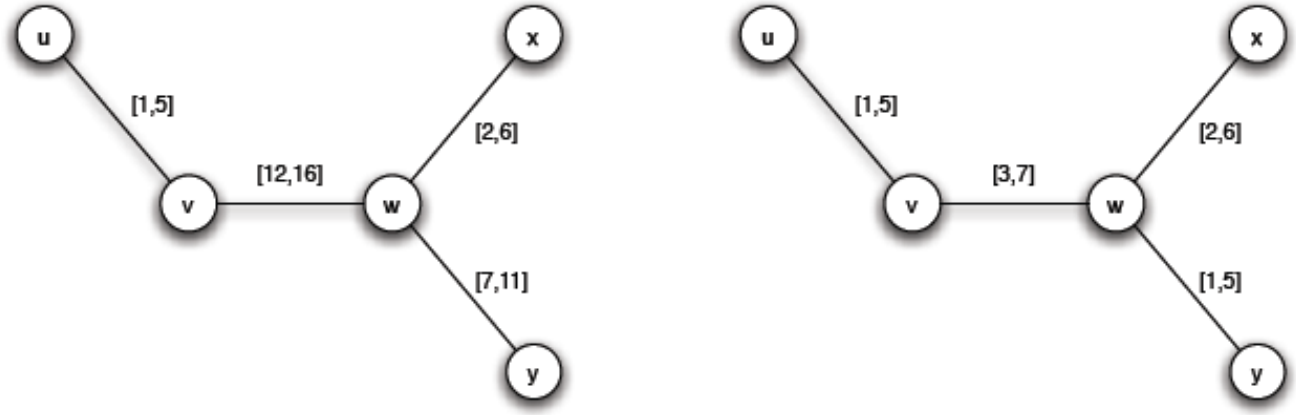



(b) *The same network as in (a), except that the timing of the w-v and w-y partnerships have been reversed.*

Figure 21.8: Different timings for the edges in a contact network can affect the potential for a disease to spread among individuals. For example, in (a) the disease can potentially pass all the way from $u$ to $y$, while in (b) it cannot.

# Concurrency

- Importance of concurrency – enables branching




(a) *No node is involved in any concurrent partnerships*

(b) *All partnerships overlap in time*

Figure 21.10: In larger networks, the effects of concurrency on disease spreading can become particularly pronounced.

# SIRS

- Initially, some nodes e in the ***I*** state and all others in the ***S*** state
- Each node ***u*** that enters the ***I*** state remains infectious for a fixed number of steps $t_I$. During each of these $t_I$ steps, ***u*** has a probability ***p*** of infecting each of its susceptible neighbors
- After $t_I$ steps, ***u*** is no longer infectious. Enters the ***R*** state for a fixed number of steps $t_R$. During each of these $t_R$ steps, ***u*** cannot be infected nor transmit the disease
- After $\mathbf{t_R}$ steps in the ***R*** state, node ***u*** returns to the ***S*** state

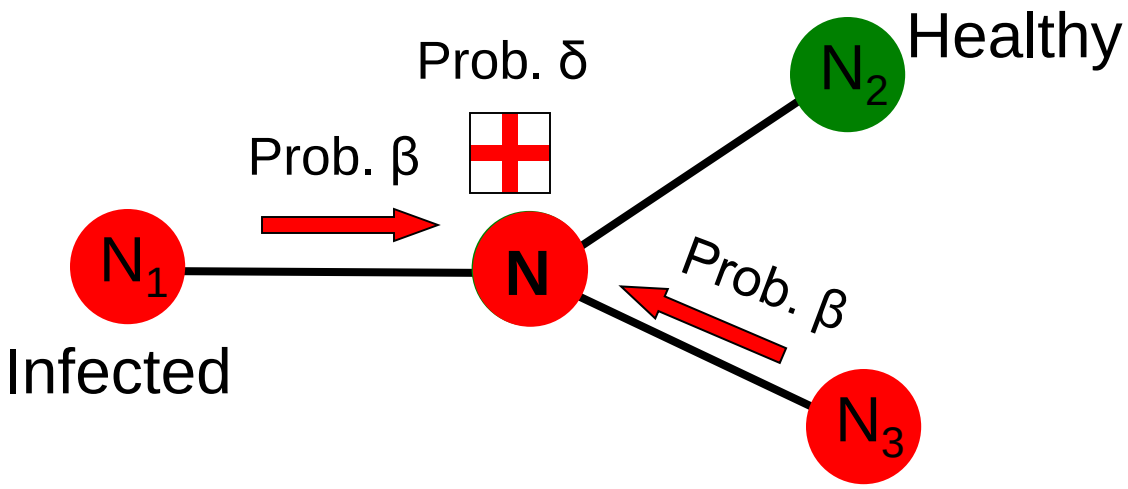
# Models of Disease Spreading

We will learn about the
**epidemic threshold**

# Spreading Models of Viruses

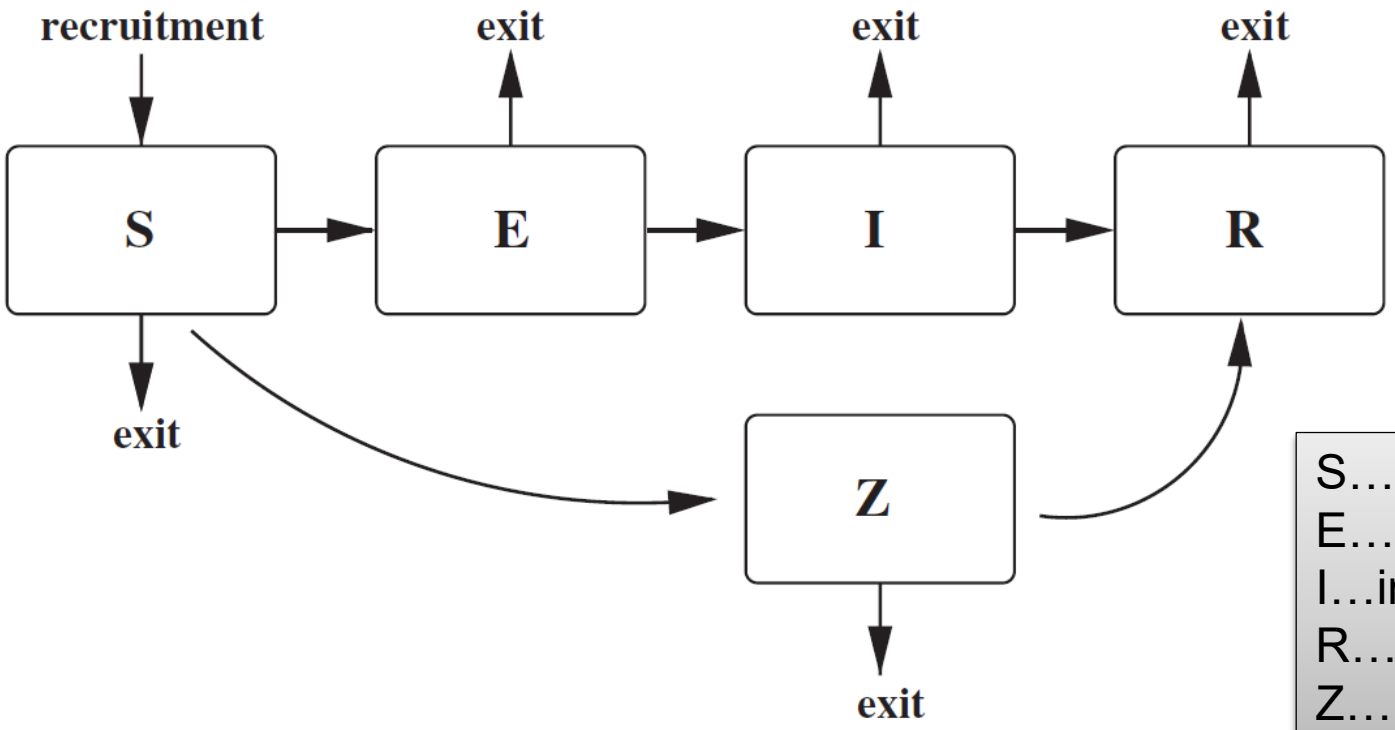
**Virus Propagation: 2 Parameters:**

- **(Virus) Birth rate β:**
  - probability that an infected neighbor attacks
- **(Virus) Death rate δ:**
  - Probability that an infected node heals



# More Generally: S+E+I+R Models

- **General scheme for epidemic models:**
  - **Each node can go through phases:**
    - Transition probs. are governed by the model parameters




S…susceptible
E…exposed
I…infected
R…recovered
Z…immune

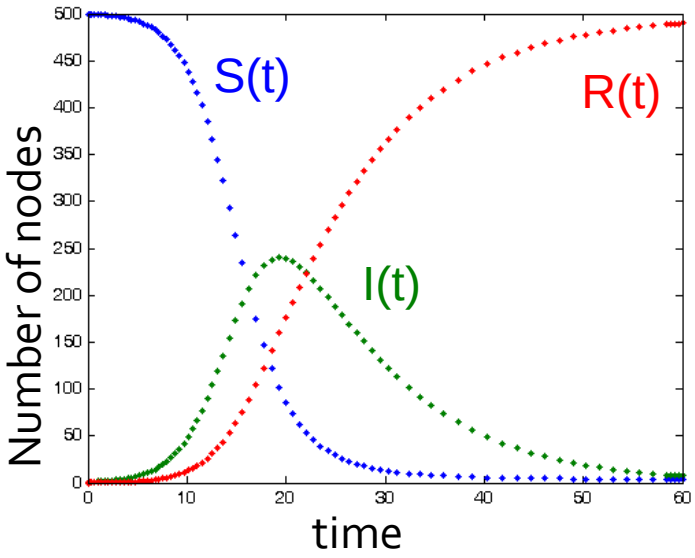
# SIR Model

- **SIR model:** Node goes through phases


  - Models chickenpox or plague:
    - Once you heal, you can never get infected again
- **Assuming perfect mixing** (The network is a complete graph) **the model dynamics are:**

$$\frac{dS}{dt} = -\,bSI \qquad \frac{dR}{dt} = dI$$

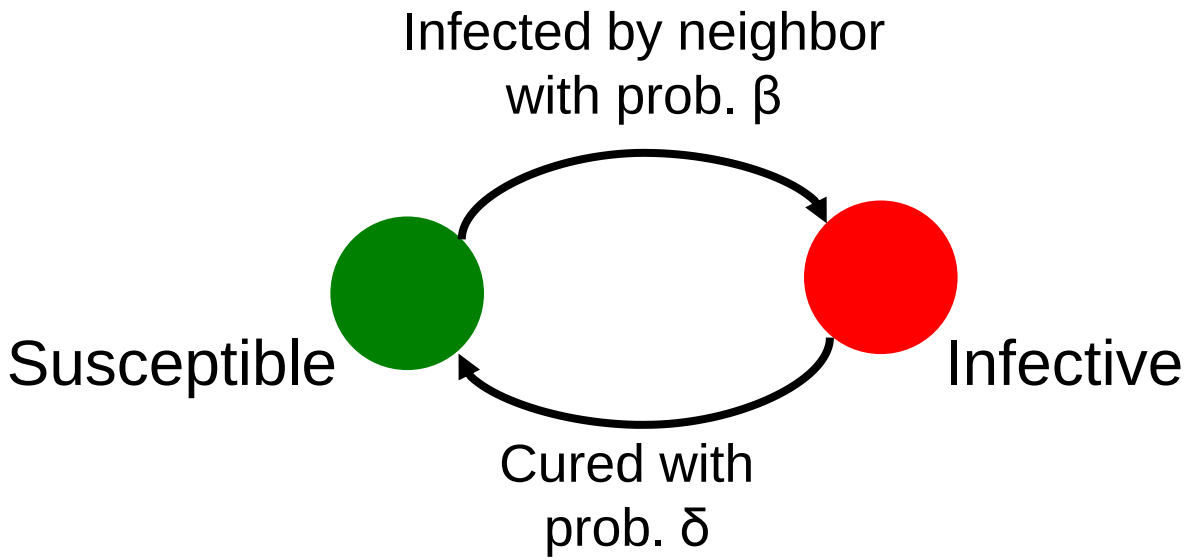
$$\frac{dI}{dt} = bSI - dI$$



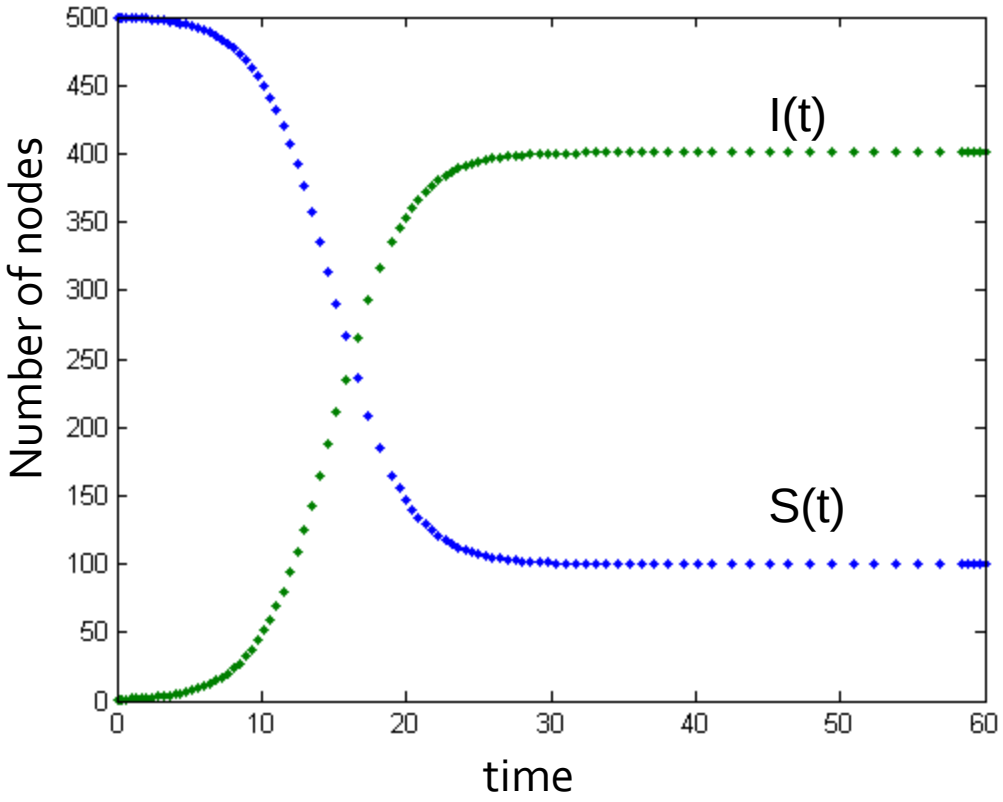

Kermack-McKendrick Model: http://mathworld.wolfram.com/Kermack-McKendrickModel.html

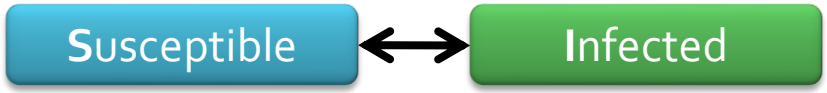
# SIS Model

- **Susceptible-Infective-Susceptible (SIS) model**
- Cured nodes immediately become susceptible
- **Virus "strength": $s = \beta / \delta$**
- **Node state transition diagram:**



# SIS Model




- **Models flu:**
  - Susceptible node becomes infected
  - The node then heals and become susceptible again
- **Assuming perfect mixing (complete graph):**

$$\frac{dS}{dt} = -\beta SI + \delta I$$

$$\frac{dI}{dt} = \beta SI - \delta I$$

# Question: Epidemic threshold *t*

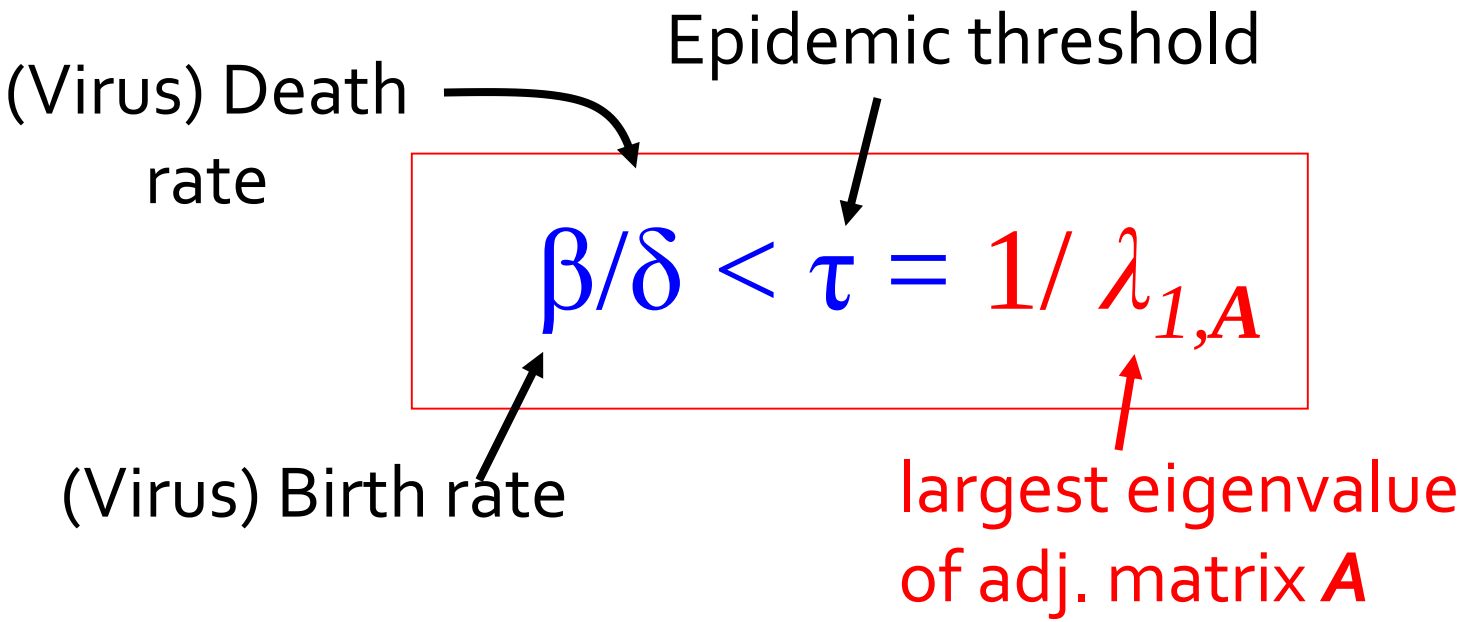
- **SIS Model:**
  **Epidemic threshold of an arbitrary graph *G* is τ, such that:**
  - **If virus strength $s = \beta / \delta < \tau$ the epidemic can not happen (it eventually dies out)**

- **Given a graph what is its epidemic threshold?**

[Wang et al. 2003]

# Epidemic Threshold in SIS Model

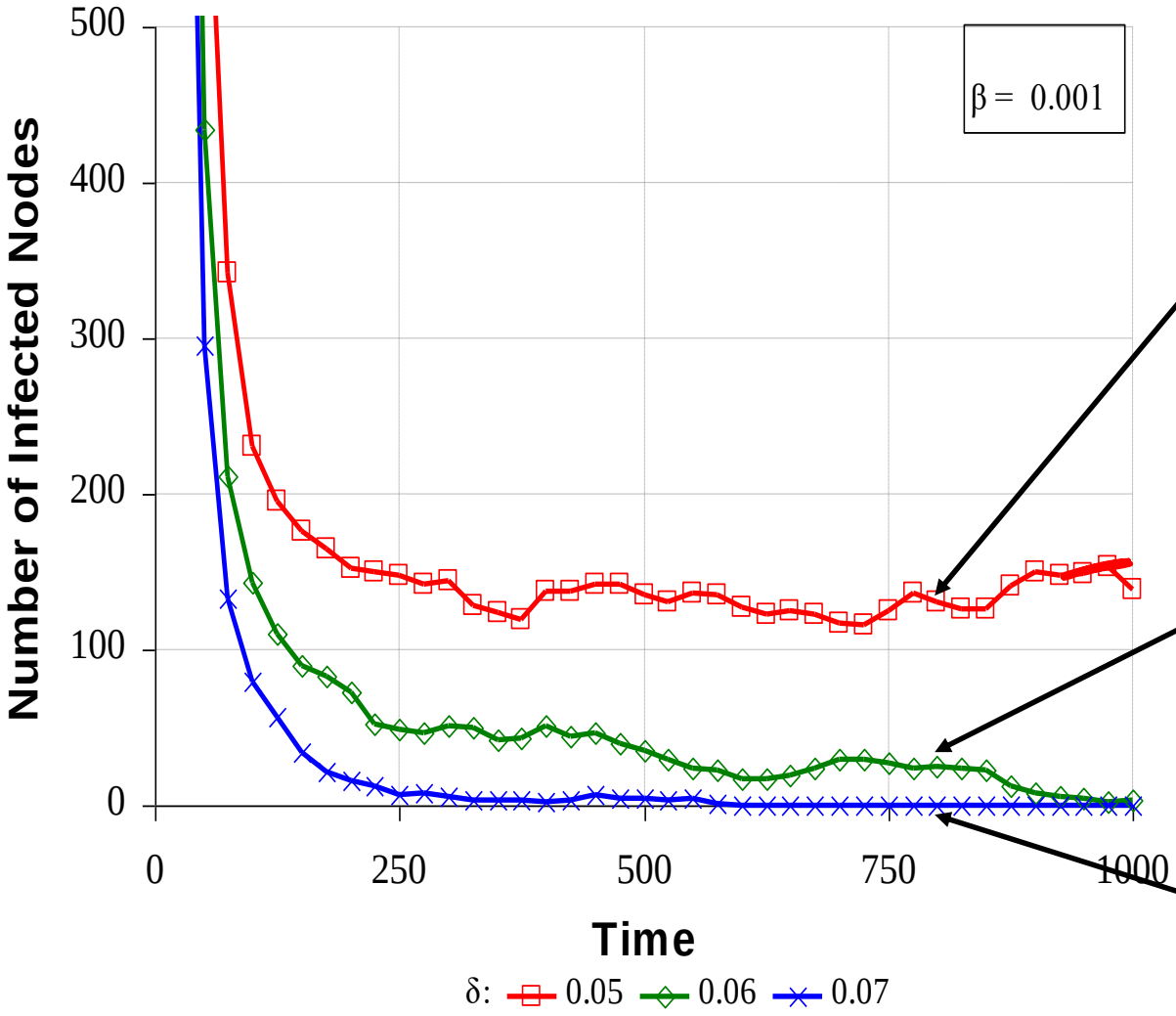
- **We have no epidemic if:**




► $\lambda_{1,A}$ ***alone captures the property of the graph!***

[Wang et al. 2003]

# Experiments (AS graph)




10,900 nodes and
31,180 edges

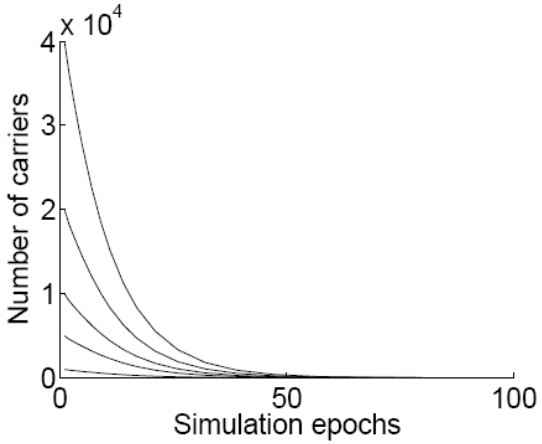
**$s=\beta/\delta > \tau$**
**(above threshold)**

**$s=\beta/\delta = \tau$**
**(at the threshold)**

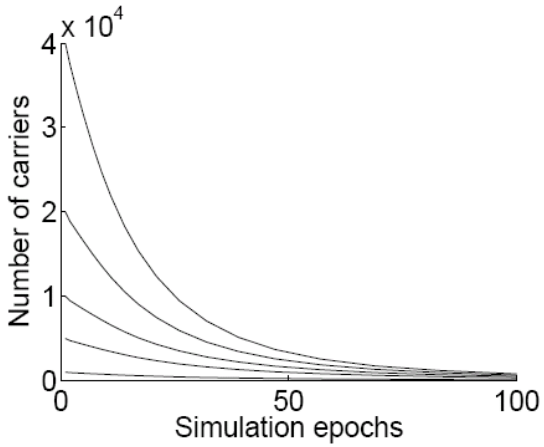
**$s=\beta/\delta < \tau$**
**(below threshold)**

# Experiments

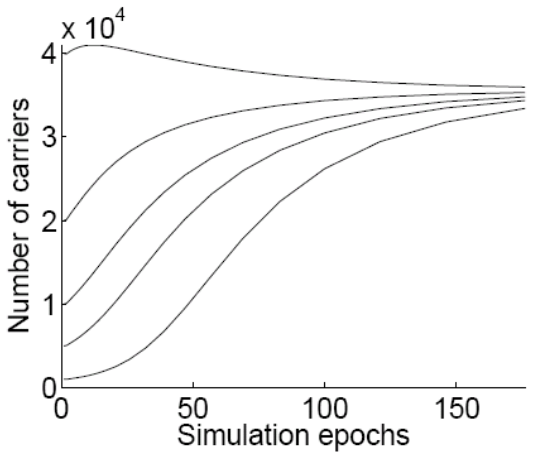
- **Does it matter how many people are initially infected?**




(a) Below the threshold, $s$=0.912

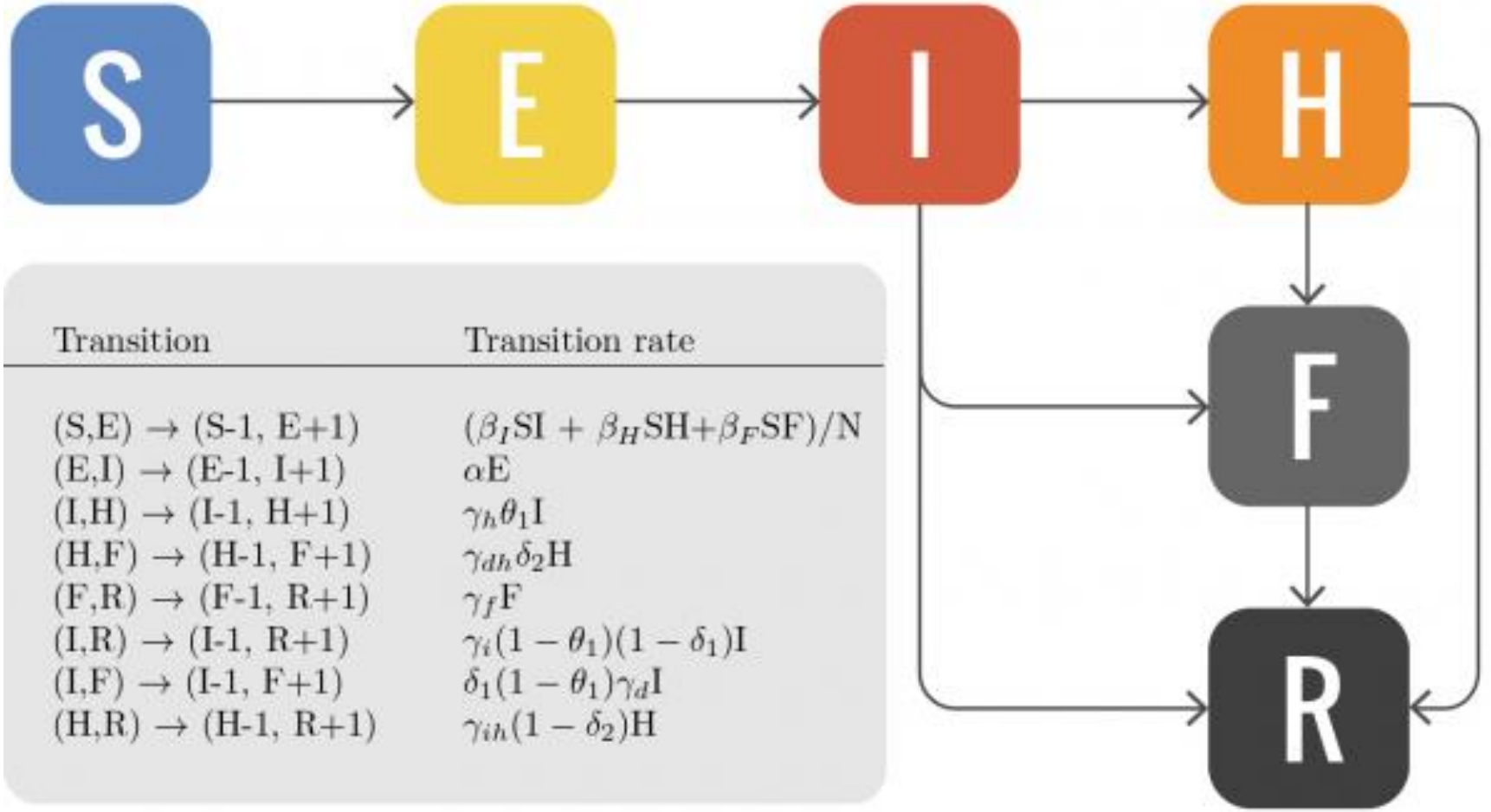



(b) At the threshold, $s$=1.003




(c) Above the threshold, $s$=1.1

# Example: Ebola




| Transition | Transition rate |
|---|---|
| (S,E) → (S-1, E+1) | $(\beta_I SI + \beta_H SH + \beta_F SF)/N$ |
| (E,I) → (E-1, I+1) | $\alpha E$ |
| (I,H) → (I-1, H+1) | $\gamma_h \theta_1 I$ |
| (H,F) → (H-1, F+1) | $\gamma_{dh} \delta_2 H$ |
| (F,R) → (F-1, R+1) | $\gamma_f F$ |
| (I,R) → (I-1, R+1) | $\gamma_i (1-\theta_1)(1-\delta_1) I$ |
| (I,F) → (I-1, F+1) | $\delta_1 (1-\theta_1) \gamma_d I$ |
| (H,R) → (H-1, R+1) | $\gamma_{ih} (1-\delta_2) H$ |

[Gomes et al., Assessing the International Spreading Risk Associated with the 2014 West African Ebola Outbreak, *PLOS Current Outbreaks*, 2014]
http://currents.plos.org/outbreaks/article/assessing-the-international-spreading-risk-associated-with-the-2014-west-african-ebola-outbreak/

# Ebola: Model States & Parameters

## Model States

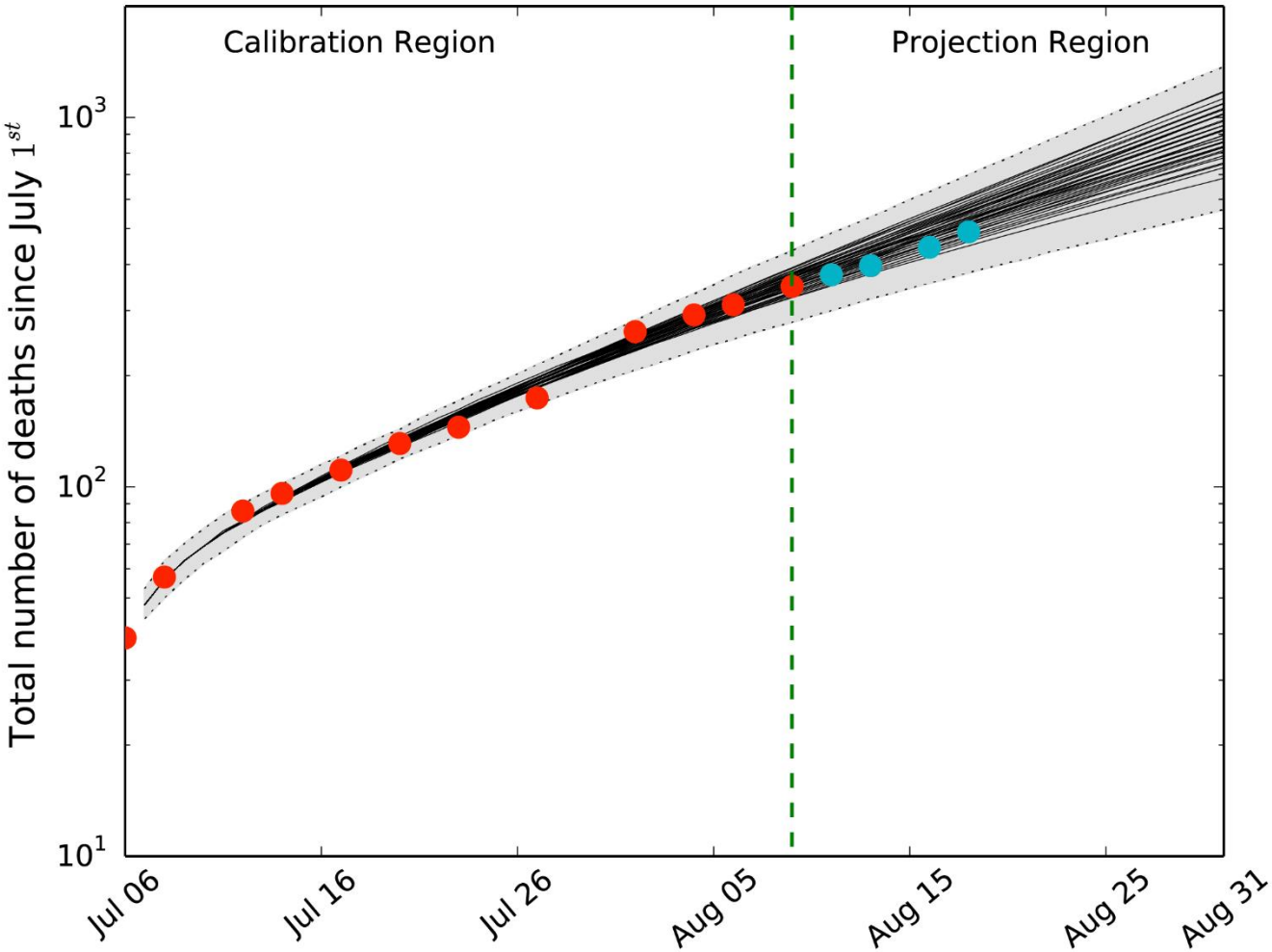
**S**: susceptible individuals
**E**: exposed individuals
**I**: infectious cases in the community
**H**: hospitalized cases
**F**: dead but not yet buried
**R**: individuals no longer transmitting the disease

## Model Parameters

$\beta_I$: transmission coefficient in the community
$\beta_H$: transmission coefficient at the hospital
$\beta_F$: transmission coefficient during funerals
$\theta_1$: computed so that **θ%** of infectious cases are hospitalized
**δ**: Compartment specific $\delta_1$ and $\delta_2$ so that overall case-fatality ratio is **δ**
$\alpha^{-1}$: the mean incubation period
$\gamma_h^{-1}$: the mean duration from symptom onset to hospitalization
$\gamma_{dh}^{-1}$: the mean duration from hospitalization to death
$\gamma_i^{-1}$: the mean duration of the infectious period for survivors
$\gamma_{ih}^{-1}$: the mean duration from hospitalization to end of infectiousness for survivors
$\gamma_f^{-1}$: the mean duration from death to burial

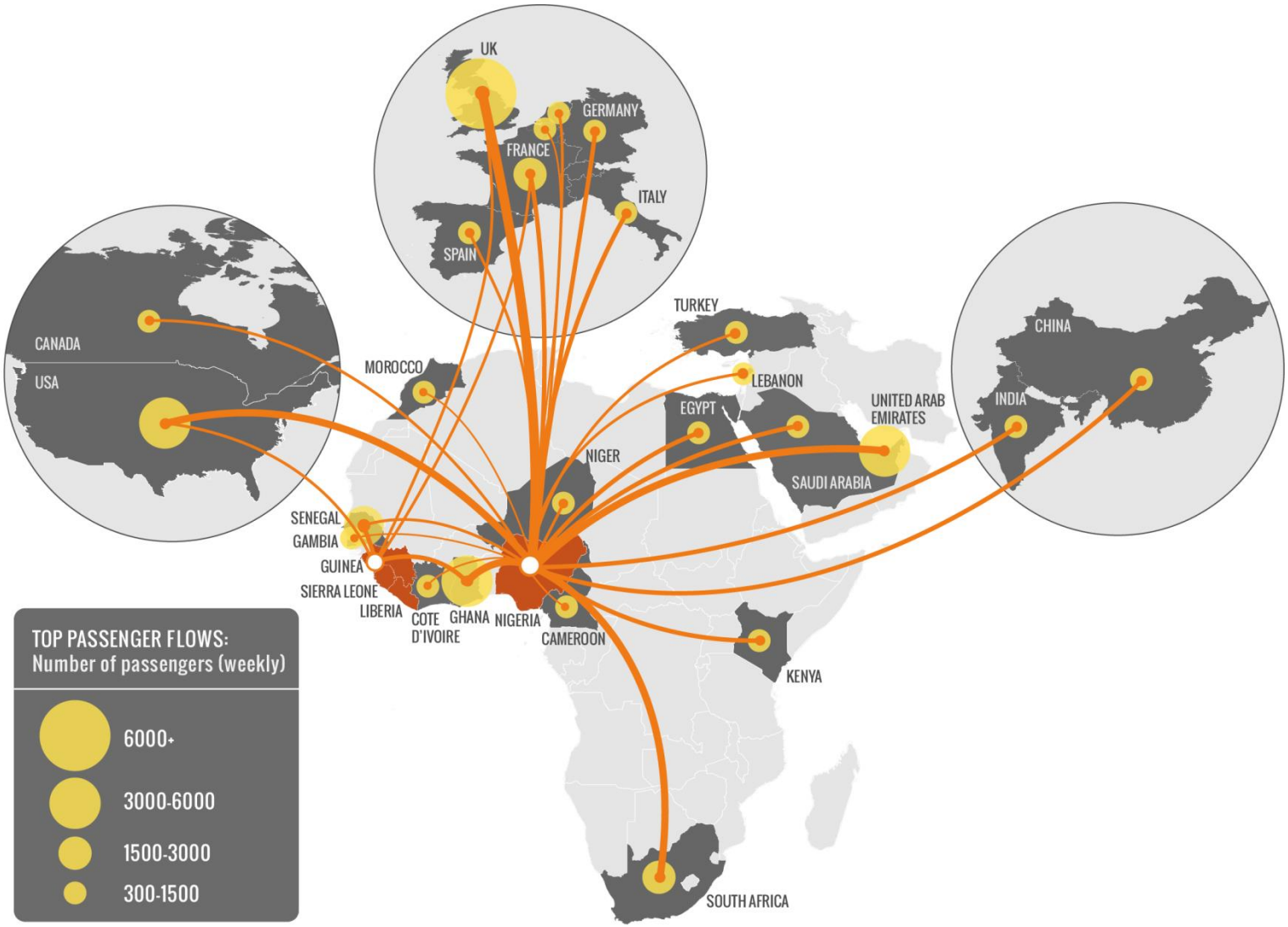
# Example: Ebola

Gomes et al., 2014]



# Example: Ebola

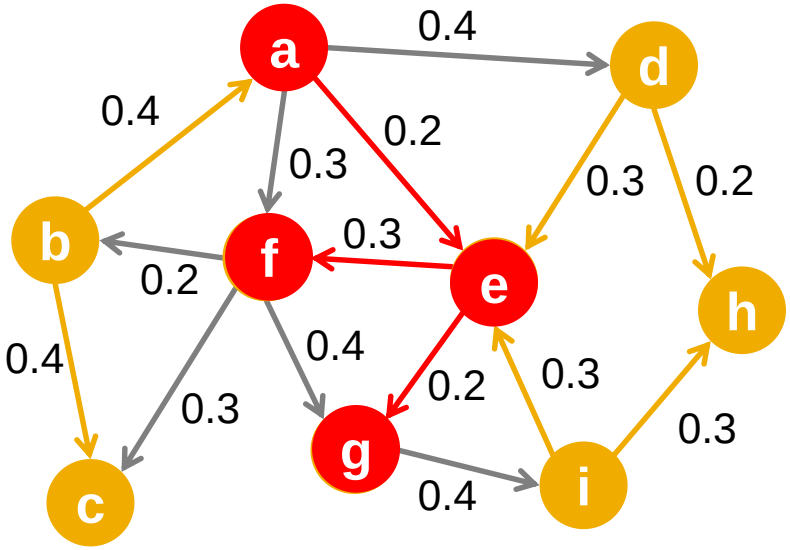
Gomes et al., 2014]



# Independent Cascade Model

# Independent Cascade Model

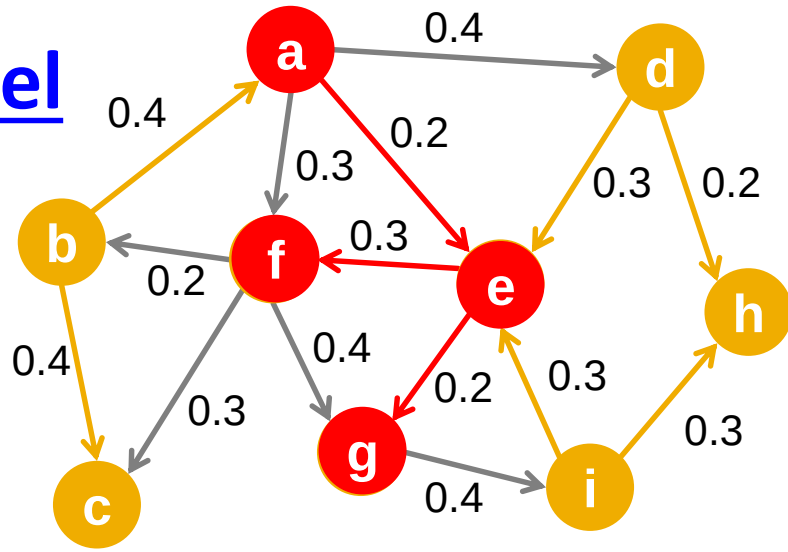
- **Initially some nodes *S* are active**
- Each edge ***(u,v)*** has probability (weight) $p_{uv}$




- **When node *u* becomes active/infected:**
  - It activates each out-neighbor ***v*** with prob. $p_{uv}$
- **Activations spread through the network!**

# Independent Cascade Modal

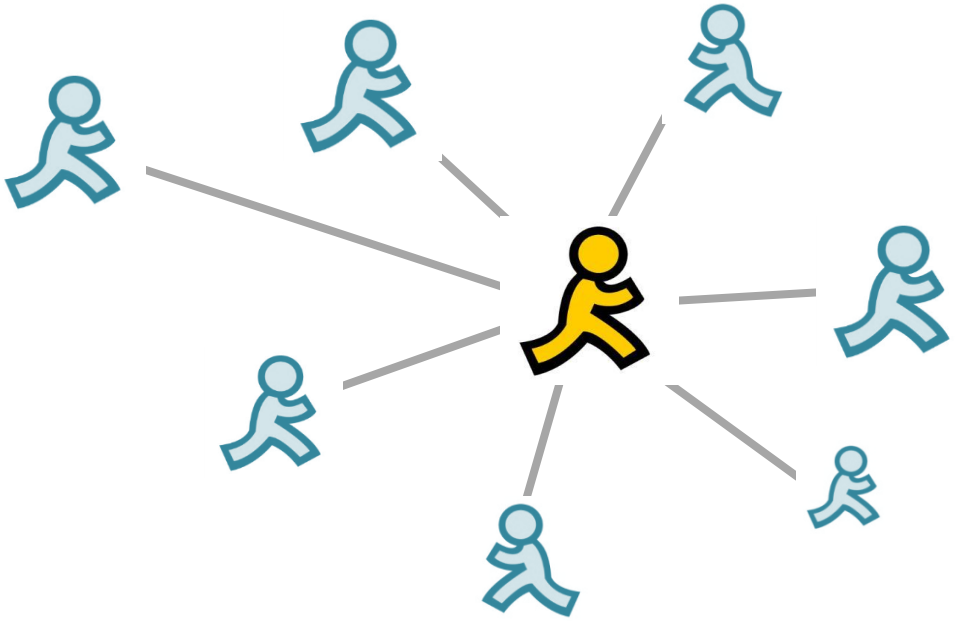
- **Independent cascade model is simple but requires many parameters!**
  - Estimating them from data is very hard [Goyal et al. 2010]




- **Solution:** Make all edges have the same weight (which brings us back to the SIR model)
  - Simple, but too simple
- **Can we do something better?**

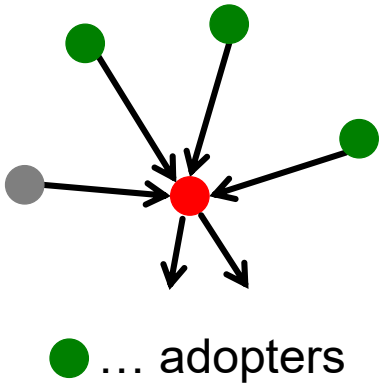
# Exposures and Adoptions

- **From exposures** to **adoptions**
  - **Exposure**: Node's neighbor exposes the node to the contagion
  - **Adoption**: The node acts on the contagion

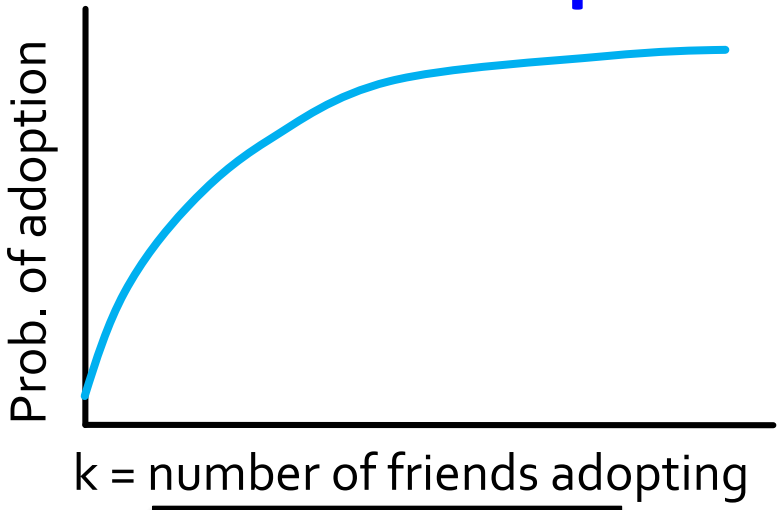


# Exposure Curves

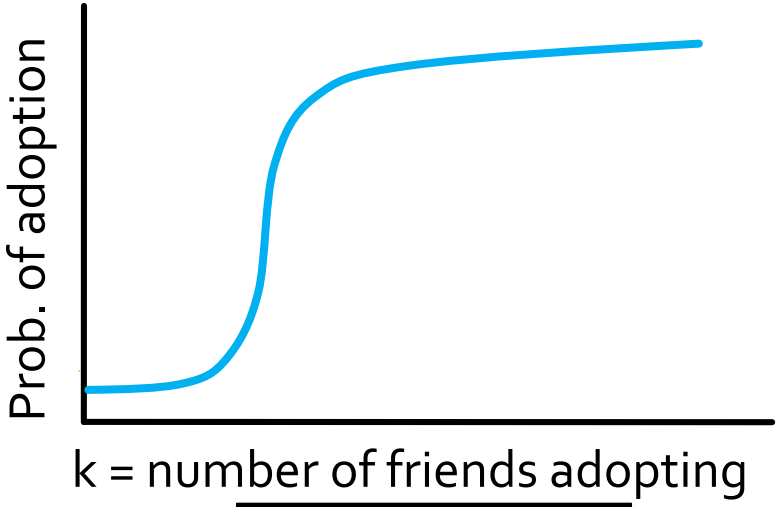
- **Exposure curve:**
  - **Probability of adopting new behavior depends on the total number of friends who have already adopted**




- **What's the dependence?**



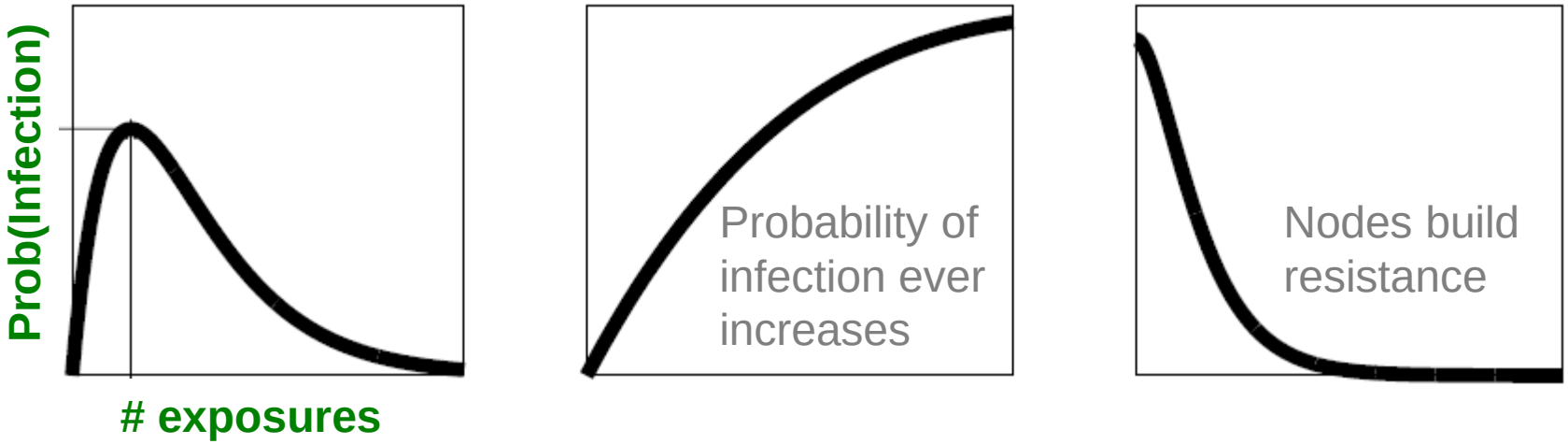

Diminishing returns:
Viruses, Information




Critical mass:
Decision making

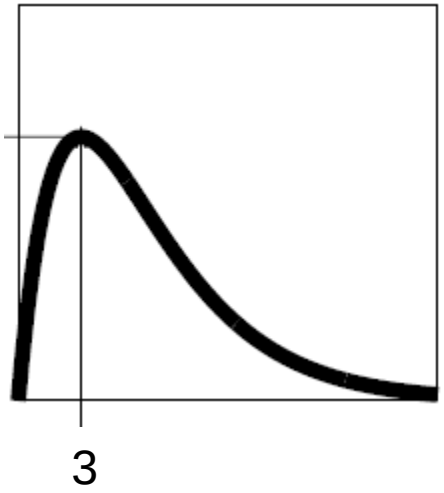
# Exposure Curves

- **From exposures to adoptions**
  - **Exposure**: Node's neighbor exposes the node to information
  - **Adoption**: The node acts on the information
- **Adoption curve:**



# Example Application

- **Marketing agency** would like you to adopt/buy product *X*
- They estimate the adoption curve




- **Should they expose you to *X* three times?**
- **Or, is it better to expose you *X*, then *Y* and then *X* again?**

[Leskovec et al., TWEB '07]

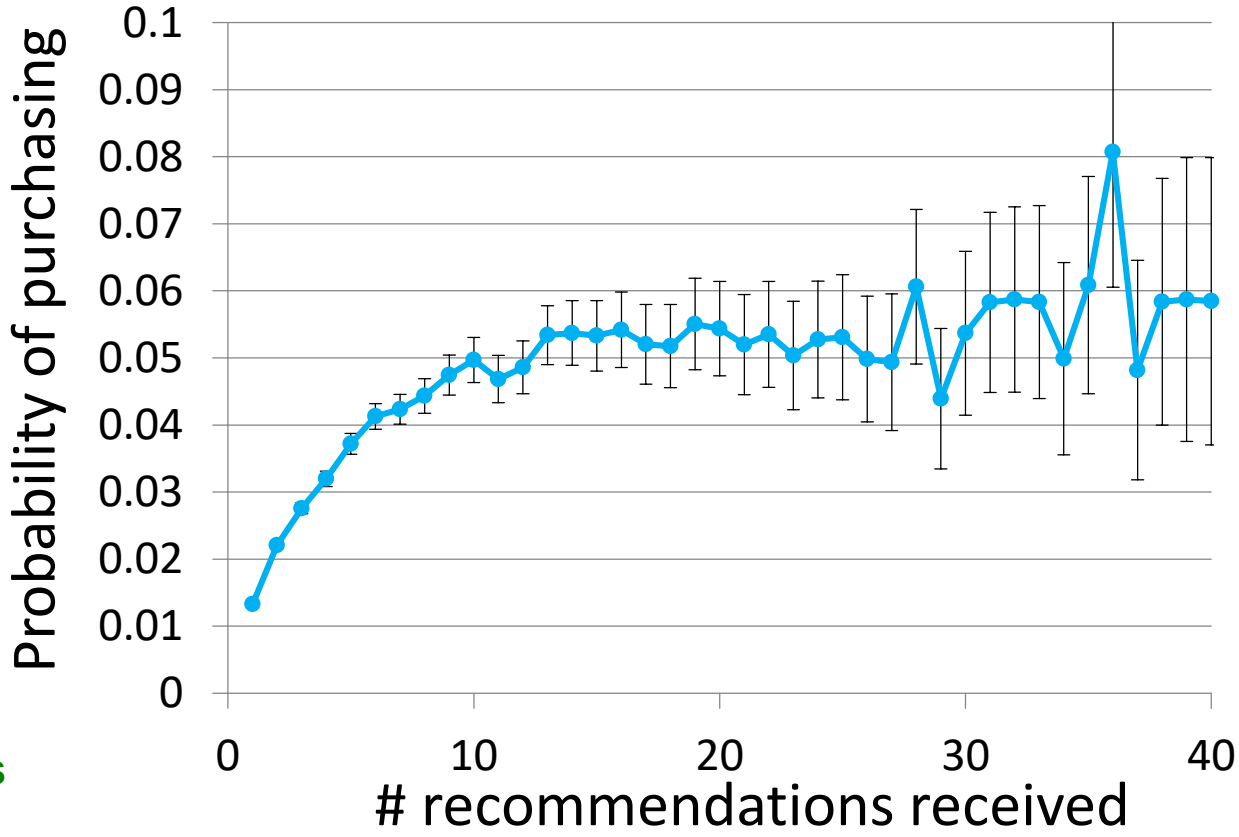
# Diffusion in Viral Marketing

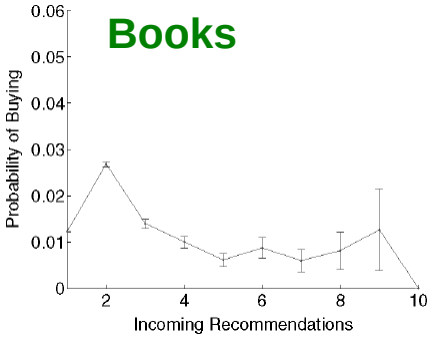
- **Senders and followers of recommendations receive discounts on products**


- **Data: Incentivized Viral Marketing program**
  - 16 million recommendations
  - 4 million people, 500k products
  - [Leskovec-Adamic-Huberman, 2007]

[Leskovec et al., TWEB '07]

# Exposure Curve: Validation



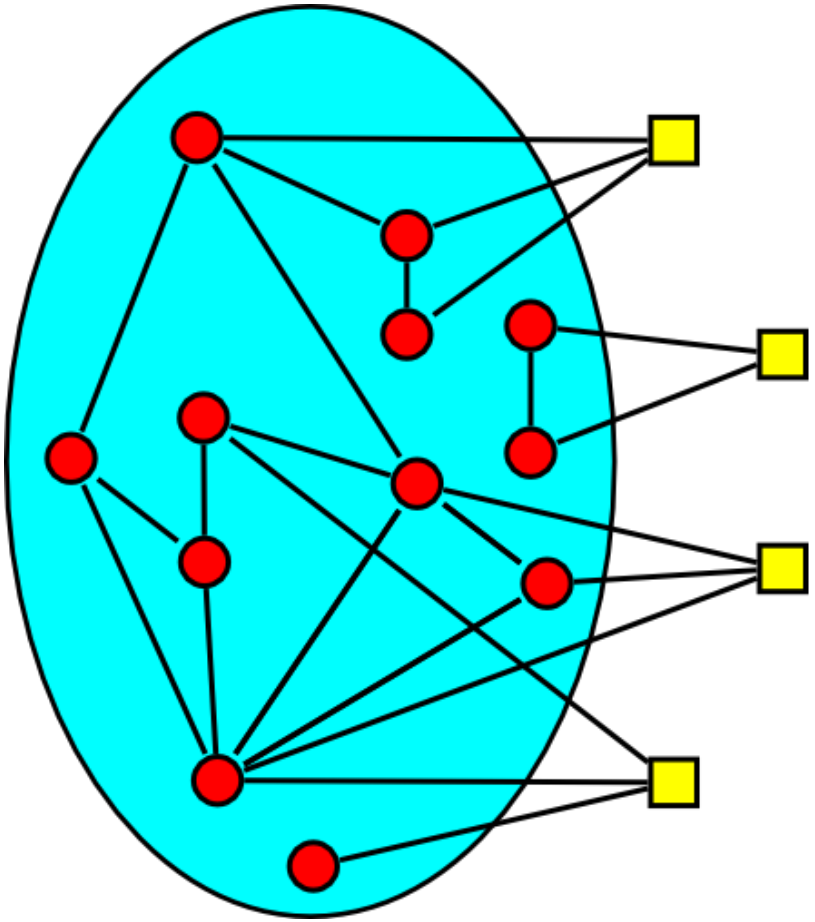




DVD recommendations
(8.2 million observations)

[Backstrom et al. KDD '06]

# Exposure Curve: LiveJournal

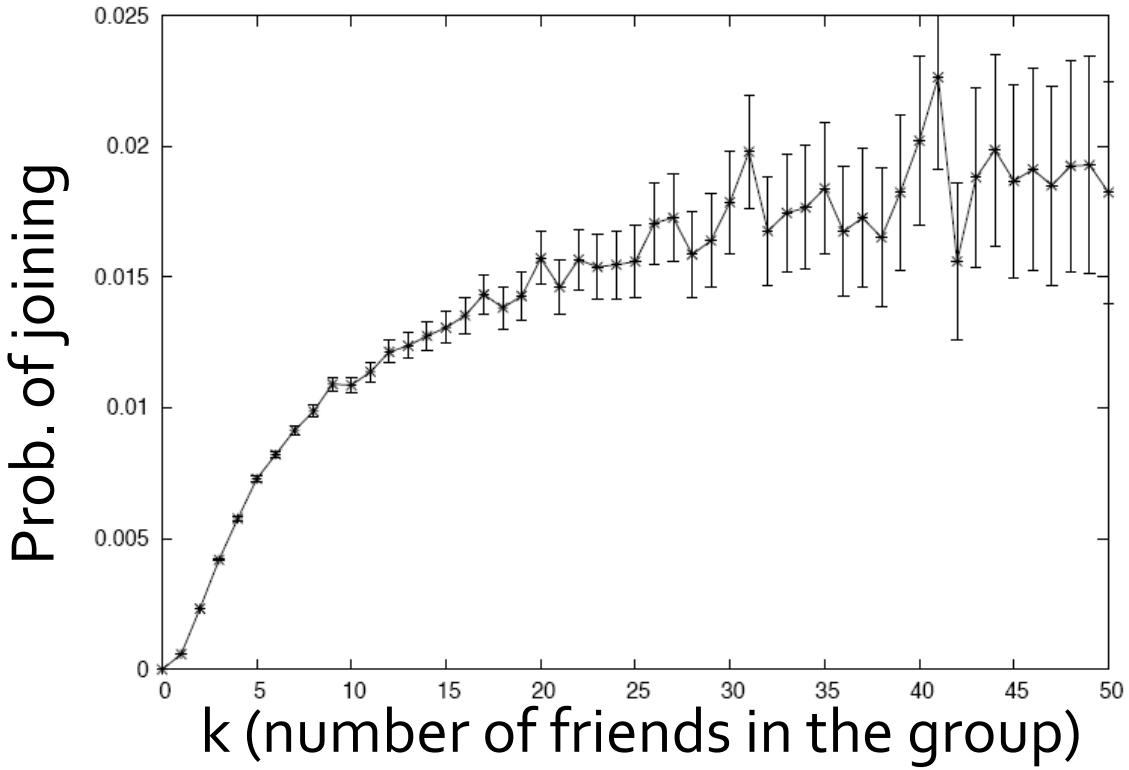
- **Group memberships spread over the network:**
  - **Red** circles represent existing group members
  - **Yellow** squares may join
- **Question:**
  - How does prob. of joining a group depend on the number of friends already in the group?



[Backstrom et al., KDD '06]

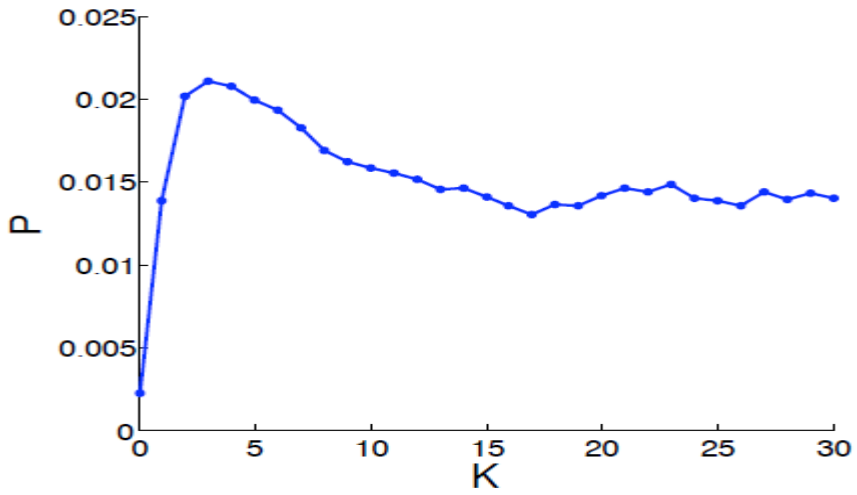
# Exposure Curve: LiveJournal

- **LiveJournal group membership**



# Exposure Curve: Information

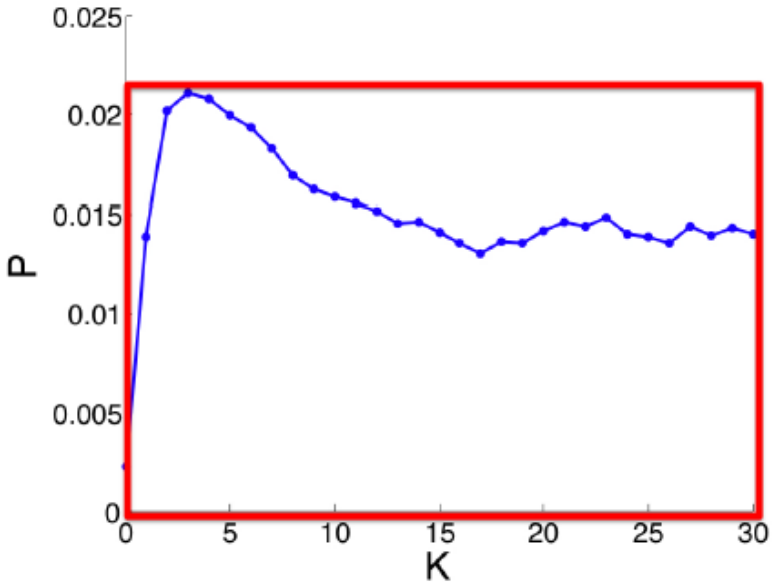
- **Twitter** [Romero et al. '11]
  - Aug '09 to Jan '10, 3B tweets, 60M users

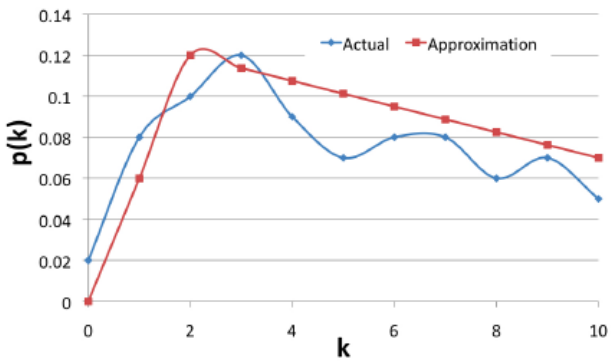



  - **Avg. exposure curve for the top 500 hashtags**
  - **P(K)** is the fraction of users who adopt the hashtag directly after their **Kth** exposure to it
  - Curve reaches peak fast, decreases after!

# Modeling the Shape of the Curve

- **Persistence of *P*** is the ratio of the area under the curve ***P*** and the area of the rectangle of length *max(P)*, width *max(D(P))*
  - *D(P)* is the domain of *P*
  - **Persistence measures the decay of exposure curves**
- **Stickiness of *P*** is ***max(P)***
  - **Stickiness is the probability of usage at the most effective exposure**

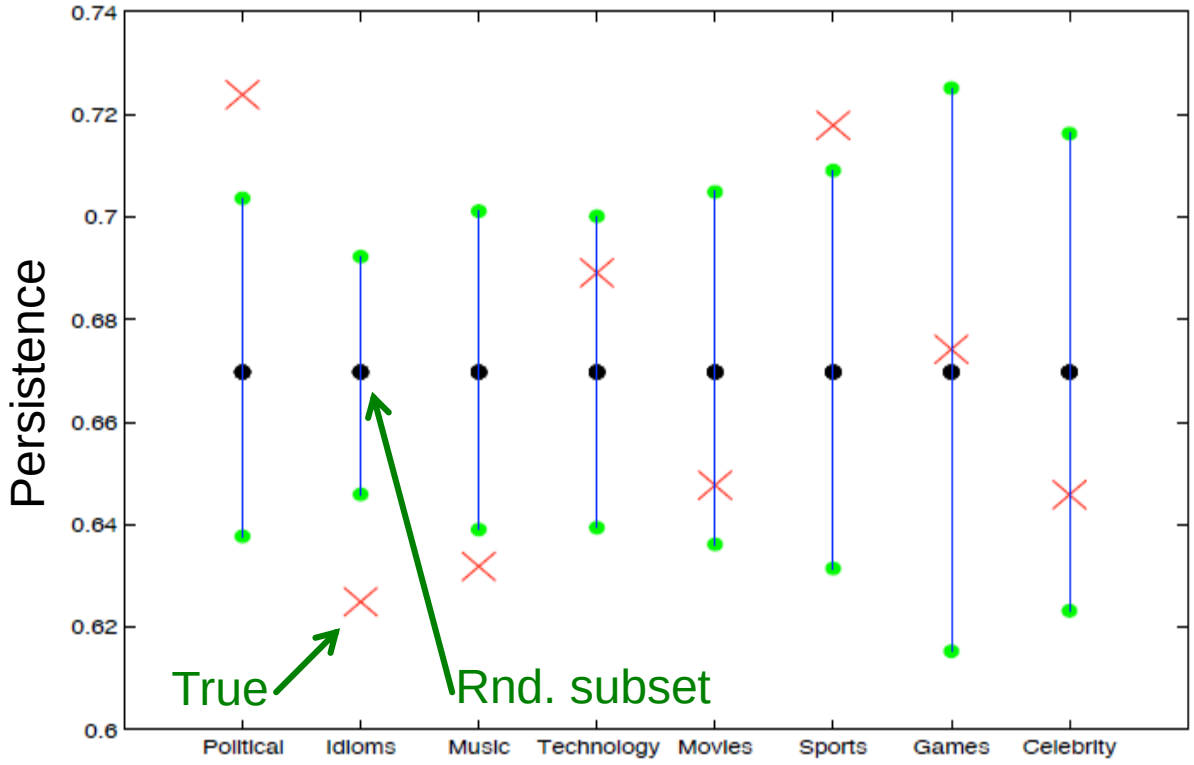


# Exposure Curve: Persistence

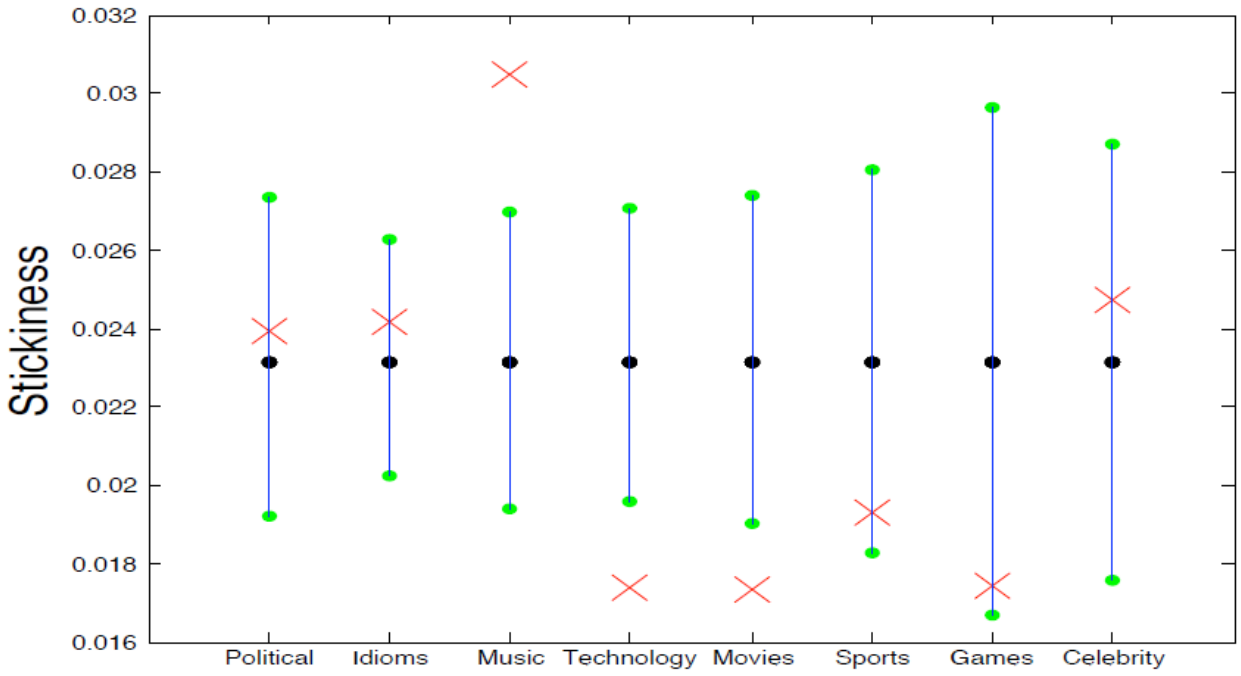
- **Manually identify 8 broad categories with at least 20 HTs in each**

| Category | Examples |
|---|---|
| Celebrity | mj, brazilwantsjb, regis, iwantpeterfacinelli |
| Music | thisiswar, mj, musicmonday, pandora |
| Games | mafiawars, spymaster, mw2, zyngapirates |
| Political | tcot, glennbeck, obama, hcr |
| Idiom | cantlivewithout, dontyouhate, musicmonday |
| Sports | golf, yankees, nhl, cricket |
| Movies/TV | lost, glennbeck, bones, newmoon |
| Technology | digg, iphone, jquery, photoshop |




- Idioms and Music have lower persistence than that of a random subset of hashtags of the same size
- Politics and Sports have higher persistence than that of a random subset of hashtags of the same size

# Exposure Curve: Stickiness




- Technology and Movies have lower stickiness than that of a random subset of hashtags
- Music has higher stickiness than that of a random subset of hashtags (of the same size)

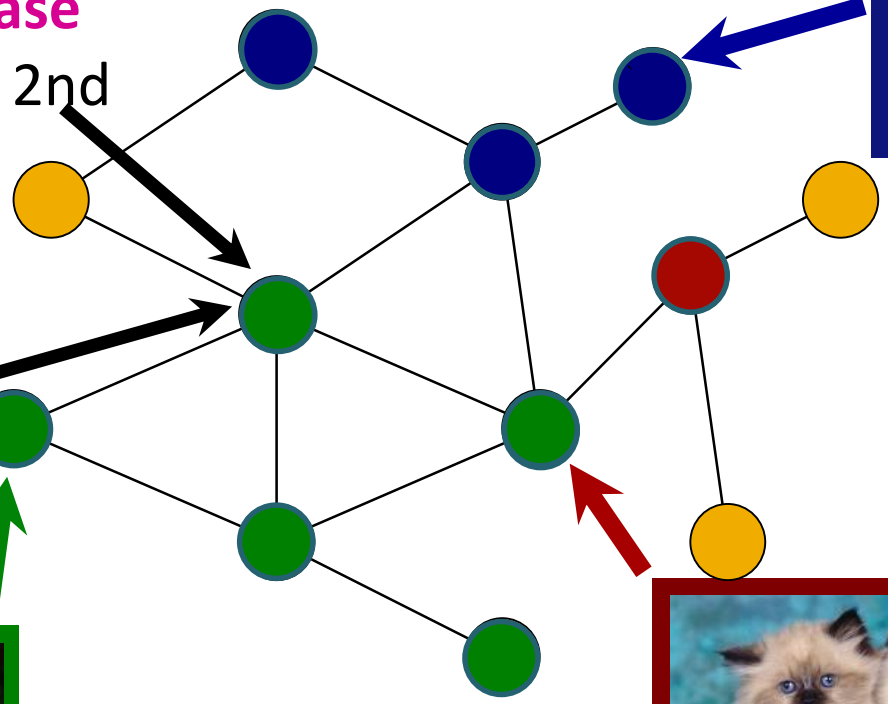
# Modeling Interactions Between Contagions

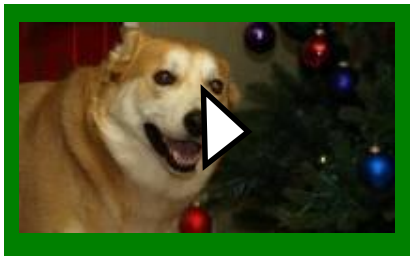
# Information Diffusion

So far we considered pieces of information as **independently** propagating. **Do pieces of information interact?**

Did 1st cat video **decrease** adoption probability of 2nd cat video?

Did cat videos **increase** adoption probability of dog video?

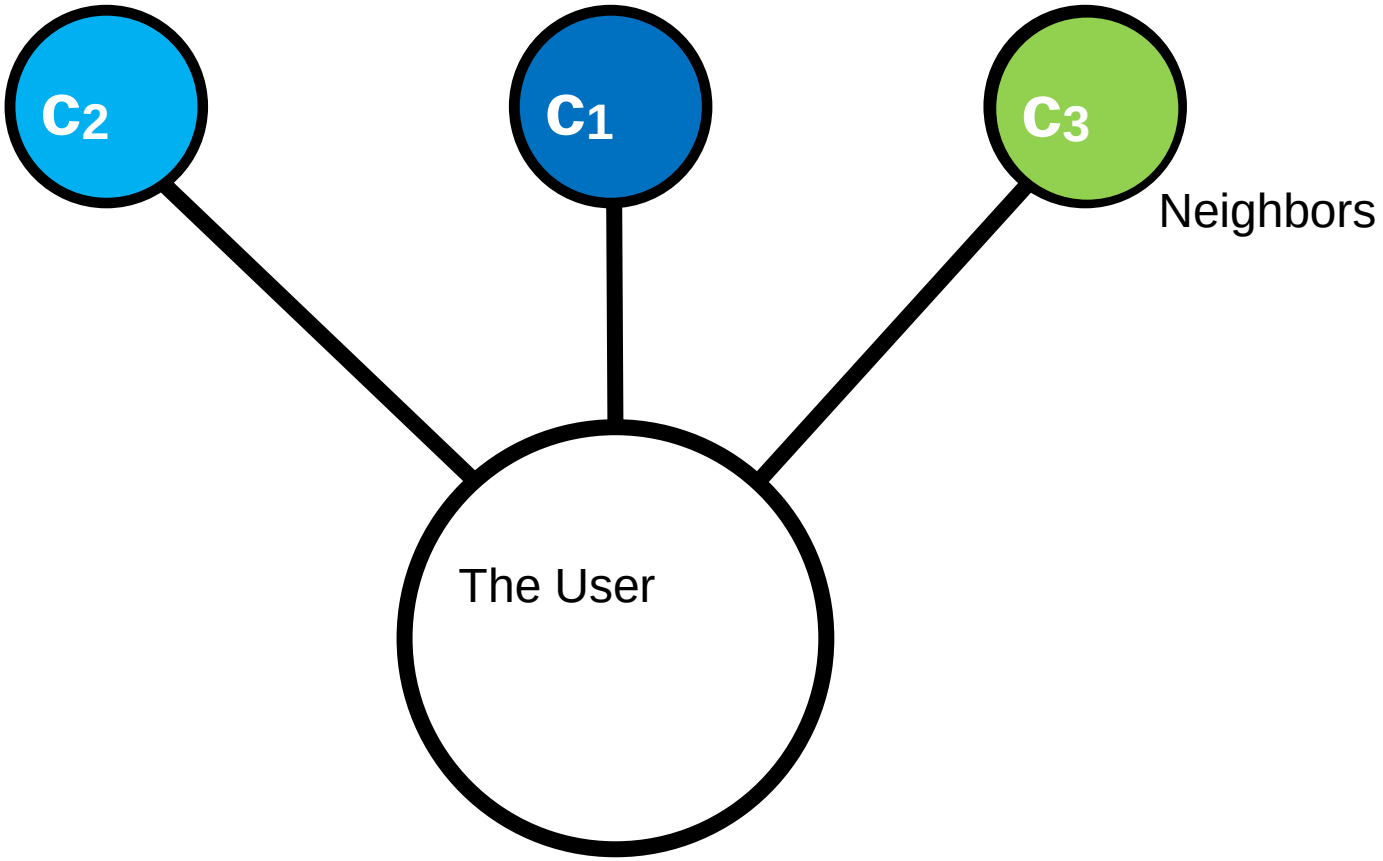




# Modeling Interactions

- **Goal: Model interaction between many pieces of information**
  - **Some pieces of information may help each other in adoption**
  - **Other may compete for attention**

# The Model

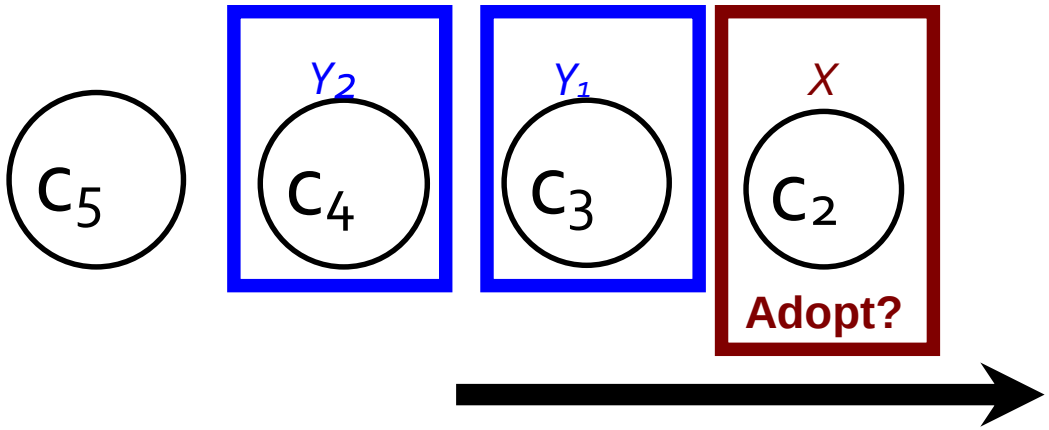



$P$(adopt $c_3$ | exposed *to* $c_2$ , $c_1$, $c_0$)

# The Model

- **You are reading posts on Twitter:**
  - You examine posts one by one
  - Currently you are examining ***X***
  - How does your probability of reposting ***X*** depend on what you have seen in the past?

**Contagions adopted by neighbors:**



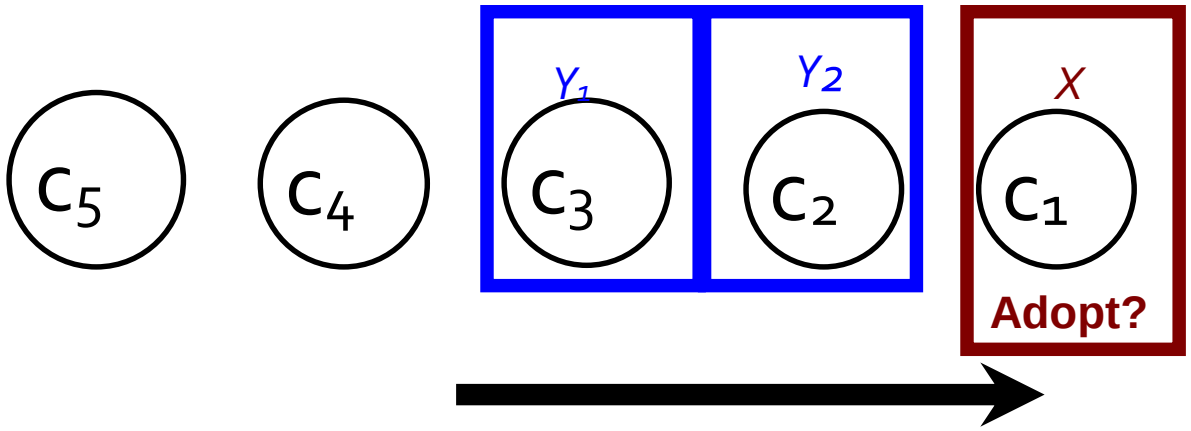
# The Model

- We assume **K** most recent exposures effect a user's adoption:
- **P(adopt $X=c_0$ | exposed $Y_1=c_1$, $Y_2=c_2$, … , $Y_K=c_k$)**

Contagion the user is viewing now.

Contagions the user previously viewed.

**Contagions adopted by neighbors:**



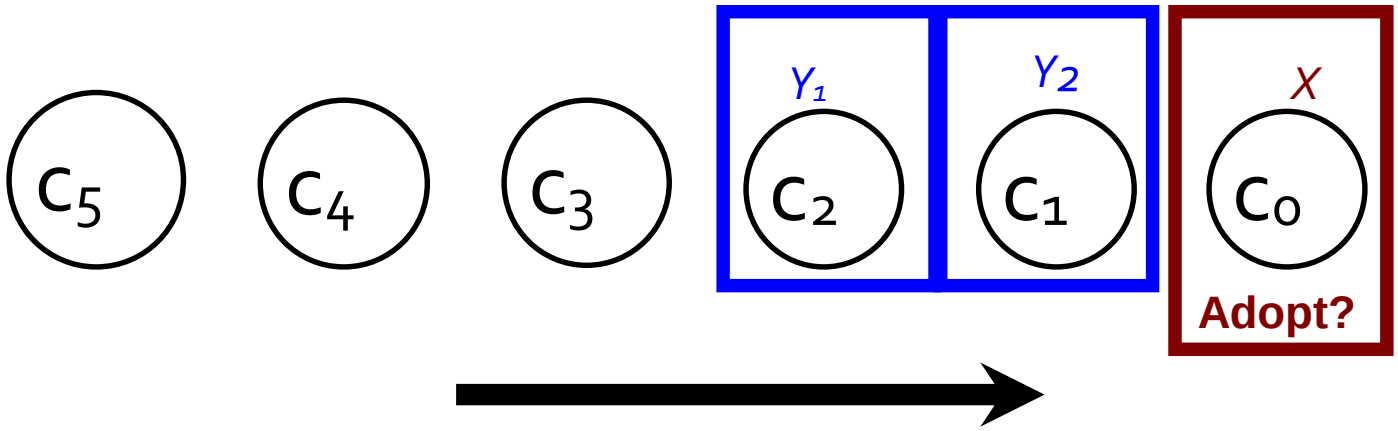
# The Model

- We assume **K** most recent exposures effect a user's adoption:
- **P(adopt $X=c_0$ | exposed $Y_1=c_1$, $Y_2=c_2$, … , $Y_K=c_k$)**

Contagion the user is viewing now.

Contagions the user previously viewed.

**Contagions adopted by neighbors:**



# The Model: Problem

- Imagine we want to estimate: **P(X | $Y_1$, … $Y_5$)**
- **What's the problem?**
  - **What's the size of probability table P(X | $Y_1$, … $Y_5$)?**
    **= (Num. Contagions)$^5$ ≈ 1.9x10$^{21}$**
- **<u>Simplification:</u> Assume $Y_i$ is independent of $Y_j$**

$$P(X|Y_1, \ldots, Y_K) = \frac{1}{P(X)^{K-1}} \prod_{k=1}^{K} P(X|Y_k)$$

- **How many parameters? $K \cdot w^2$ Too many!**
  - $K$ … history size
  - $w$ … number of contagions

# The Model

- **Goal:** Model $P(\text{adopt } X \mid Y_1, \ldots, Y_K)$
- **First, assume:**

$$P(X = u_j | Y_k = u_i) \approx \underbrace{P(X = u_j)}_{\text{Prior infection prob.}} + \underbrace{\Delta^{(k)}_{cont.}(u_i, u_j)}_{\text{Interaction term (still has } w^2 \text{ entries!)}}$$

- **Next, assume "topics":**

$$\begin{bmatrix} \\ \Delta^{(k)}_{cont.} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \mathbf{M} \\ \\ \end{bmatrix} \times \begin{bmatrix} \Delta^{(k)}_{clust} \end{bmatrix} \times \begin{bmatrix} \mathbf{M}^T \end{bmatrix}$$

# The Model

Details

- **Goal:** Model $P(adopt\ X \mid Y_1, \ldots, Y_K)$
- **First, assume:**

$$P(X = u_j | Y_k = u_i) \approx \underbrace{P(X = u_j)}_{\text{Prior infection prob.}} + \underbrace{\Delta^{(k)}_{cont.}(u_i, u_j)}_{\text{Interaction term (still has } w^2 \text{ entries!)}}$$

- **Next, assume "topics":**

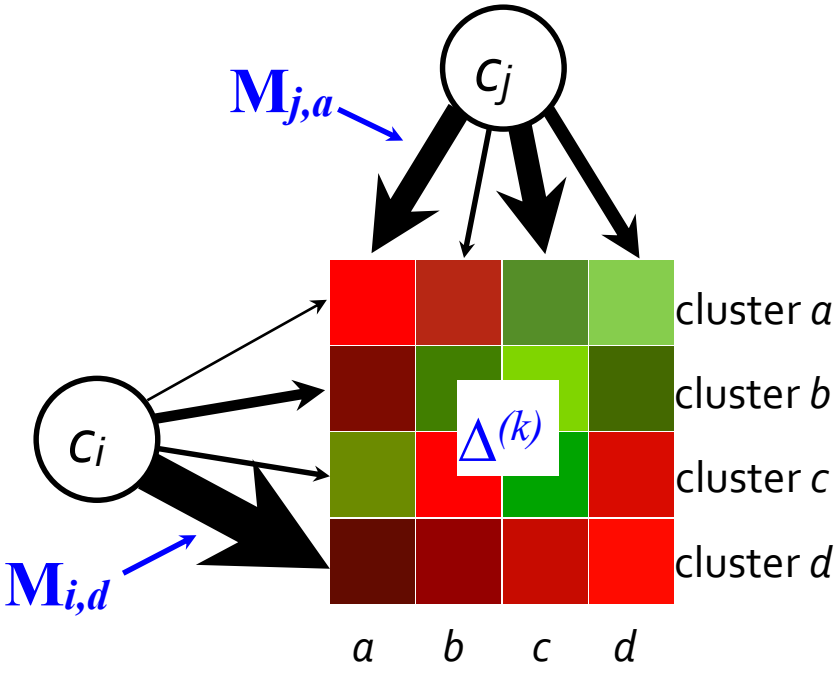
$$\Delta^{(k)}_{cont.}(u_i, u_j) = \sum_t \sum_s \mathbf{M}_{j,t} \cdot \Delta^{(k)}_{clust}(c_t, c_s) \cdot \mathbf{M}_{i,s}$$

  - Each contagion $\boldsymbol{u_i}$ has a vector $\boldsymbol{M_i}$
    - Entry $\boldsymbol{M_{is}}$ models how much $\boldsymbol{u_i}$ belongs to topic $\boldsymbol{s}$
  - $\boldsymbol{\Delta^{(k)}_{clust}(s,t)}$ models the change in infection prob. given that $\boldsymbol{u_i}$ is on topic $\boldsymbol{s}$ and exposure $\boldsymbol{k}$-steps ago was on topic $\boldsymbol{t}$

# The Model

Details

$$P(X = u_j | Y_k = u_i) = P(X = u_j) + \sum_t \sum_s \mathbf{M}_{i,t} \cdot \Delta^{(k)}_{t,s} \cdot \mathbf{M}_{j,s}$$




Memberships to clusters

Interactions between clusters

$$P(X = c_i | Y_k = c_j) = P(X = c_i) + \sum_{a,b} \mathbf{M}_{i,a} \times \mathbf{M}_{ib} \times \Delta^{(k)}(a,b)$$

# Inferring the Model

Details

- **Model parameters:**
  - $\Delta^k$ … topic interaction matrix
  - $M_{i,t}$ … topic membership vector
  - $P(X)$ … Prior infection prob.
- **Maximize data likelihood:**

$$\arg\max_{P(x),M,\Delta} \prod_{X \in R} P(X|X, Y_1 \dots Y_K) \prod_{X \notin R} 1 - P(X|X, Y_1 \dots Y_K)$$

  - $R$ … contagions X that resulted in infections
  - **Solve using stochastic coordinate ascent:**
    - Alternate between optimizing $\Delta$ and $M$

# Dataset: Twitter

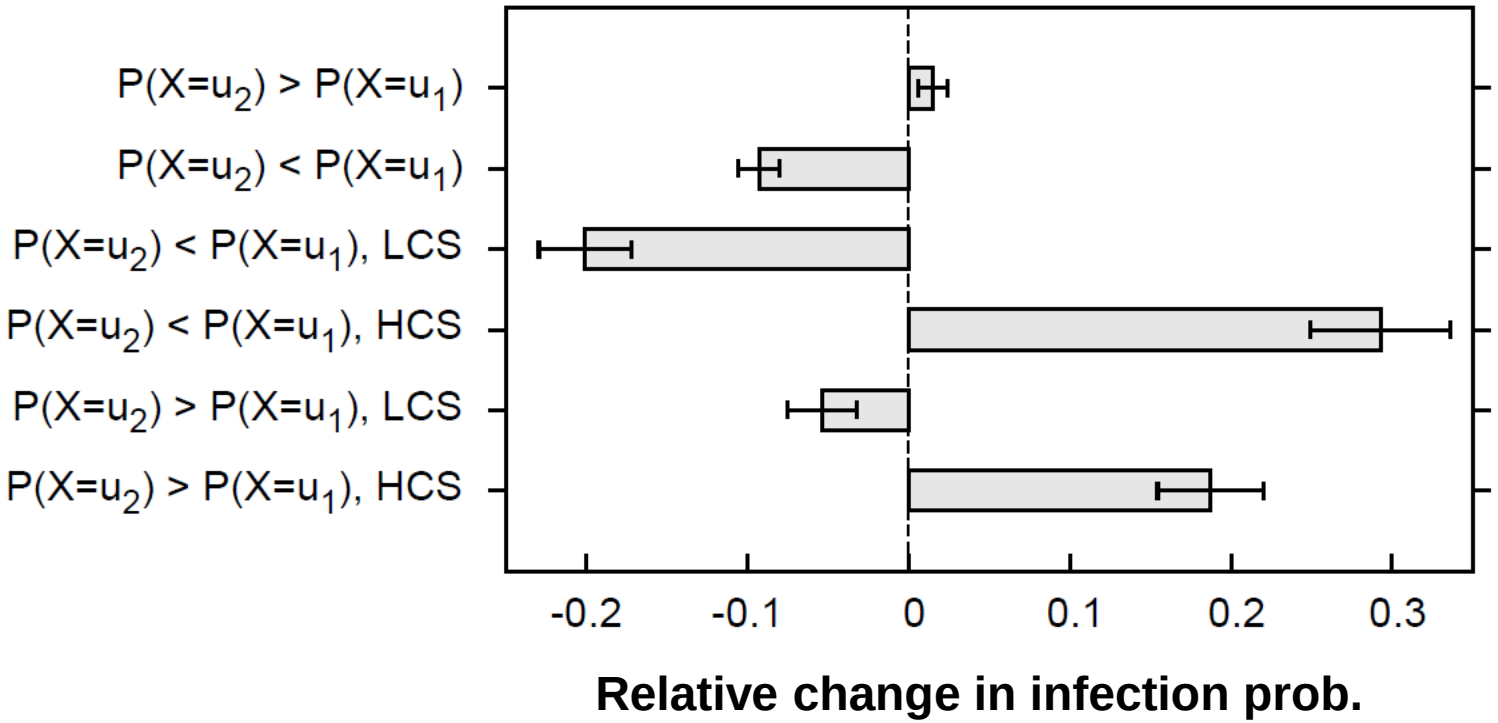
- **Data from Twitter**
  - *Complete* data from Jan 2011: 3 billion tweets
  - All URLs tweeted by at least 50 users: 191k
- **Task:**

  Predict whether a user will post URL X

- **What do we learn from the model?**

# How do Tweets Interact?

- **How *P(post $u_2$| exp. $u_1$)* changes if …**
  - $u_2$ and $u_1$ are similar/different in the content?
  - $u_1$ is highly viral?




**Observations:**
- If $u_1$ is not viral, this boost $u_2$
- If $u_1$ is highly viral, this kills $u_2$

**BUT:**
Only if $u_1$ and $u_2$ are of low content similarity (LCS) else, $u_1$ helps $u_2$

# Final Remarks

- **Modeling contagion interactions**
  - 71% of the adoption probability comes from the topic interactions!
  - Modeling user bias does not matter