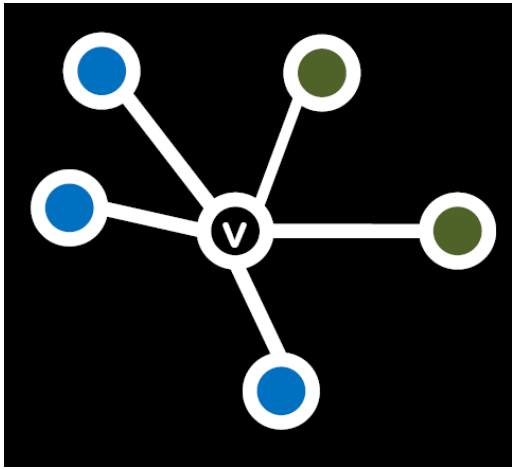
Announcements:

- Please fill HW Survey
- Weekend Office Hours starting this weekend (Hangout only)
- Proposal: Can use 1 late period

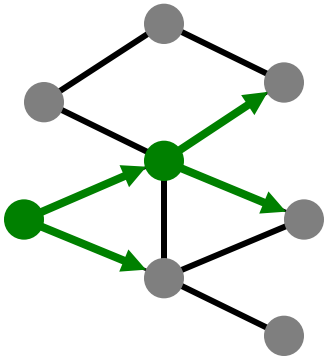
Probabilistic Contagion and Models of Influence

CS224W: Analysis of Networks
Jure Leskovec, Stanford University
http://cs224w.stanford.edu

Models of Cascading Behavior



- **So far:**
  Decision Based Models
  - Utility based
  - Deterministic
  - "Node" centric: A node observes decisions of its neighbors and makes its own decision
  - Require us to know too much about the data
- **Next: Probabilistic Models**
  - Lets you do things by observing data
  - We lose "why people do things"

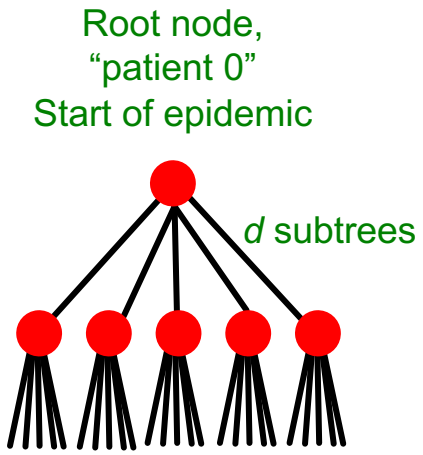


Epidemic Model Based on Trees

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

Probabilistic Spreading Models

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

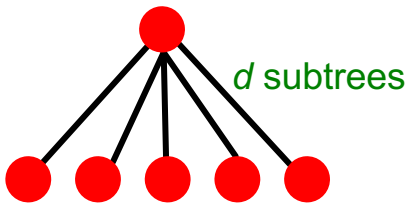


- **Q: For which values of *d* and *q* does the epidemic run forever?**
  - **Run forever:** $\lim_{h\to\infty} P\left[\begin{array}{l}\text{At least 1 infected}\\ \text{node at depth h}\end{array}\right] > 0$
  - **Die out:** -- || -- $= 0$

Probabilistic Spreading Models

- p_h = **prob. there is an infected node at depth h**
- **We need:** $\lim_{h\to\infty} p_h = ?$ (based on q and d)
  - We are reasoning about a behavior at the root of the tree. Once we get a level out, we are left with identical problem of depth h-1.
- **Need recurrence for p_h**

$$p_h = 1 - \underbrace{(1 - q \cdot p_{h-1})^d}_{\text{No infected node at depth } h \text{ from the root}}$$



- $\lim_{h\to\infty} p_h$ = **result of iterating**

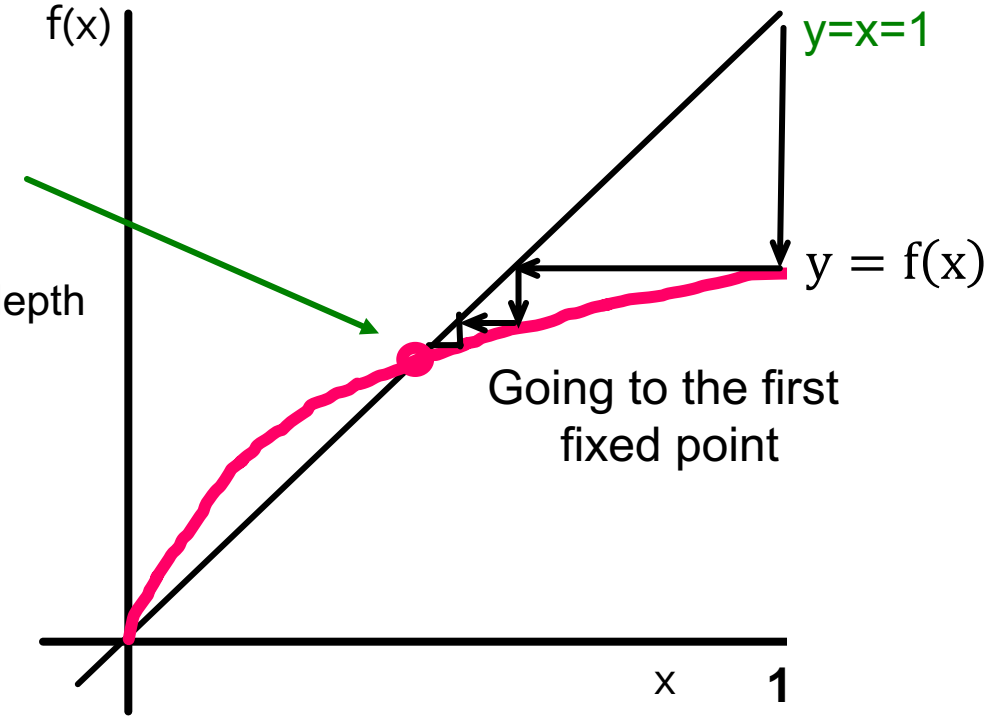
$$f(x) = 1 - (1 - q \cdot x)^d$$

  - Starting at the root: $x = 1$ (since $p_1 = 1$)

We iterate:
x_1=f(1)
x_2=f(x_1)
x_3=f(x_2)

Fixed Point: $f(x) = 1 - (1 - qx)^d$

Fixed point:
f(x)=x
This means that prob. there is an infected node at depth h *is constant (>0)*

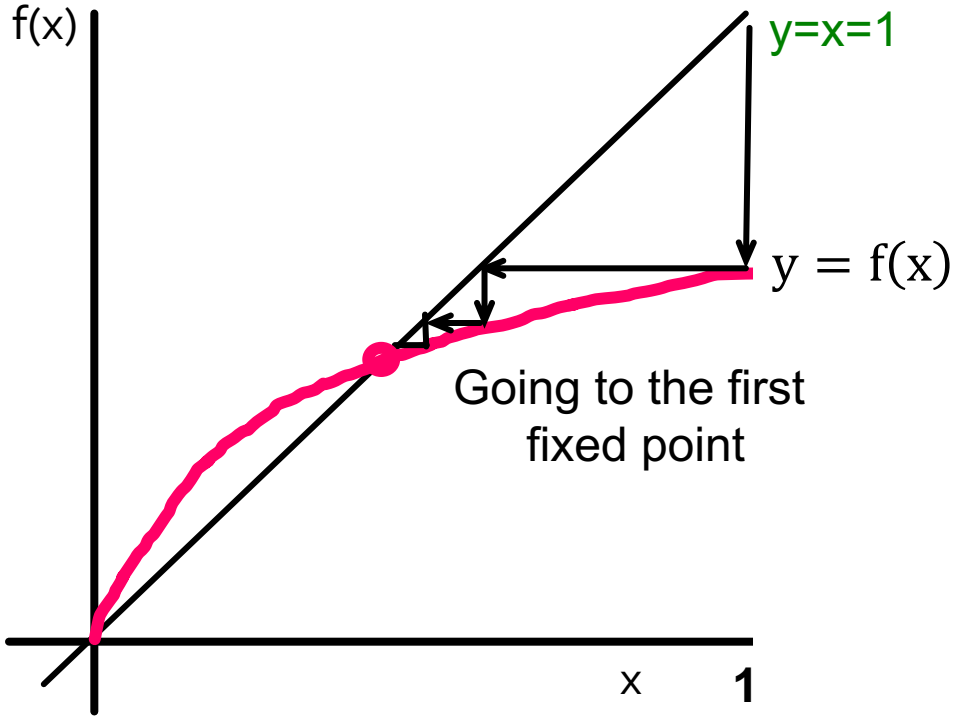


x … prob. there is an infected node at level *h-1*. We start at x=1 because p_1=1.
f(x) … prob. there is an infected node at level *h*
q … infection prob.
d … degree

We iterate:
x_1=f(1)
x_2=f(x_1)
x_3=f(x_2)

- If we want to epidemic to die out, then iterating f(x) must go to zero.
- So, f(x) must be **below** y=x

Fixed Point: $f(x) = 1 - (1 - qx)^d$



x … prob. there is an infected node at level *h-1*
f(x) … prob. there is an infected node at level *h*
q … infection prob.
d … degree

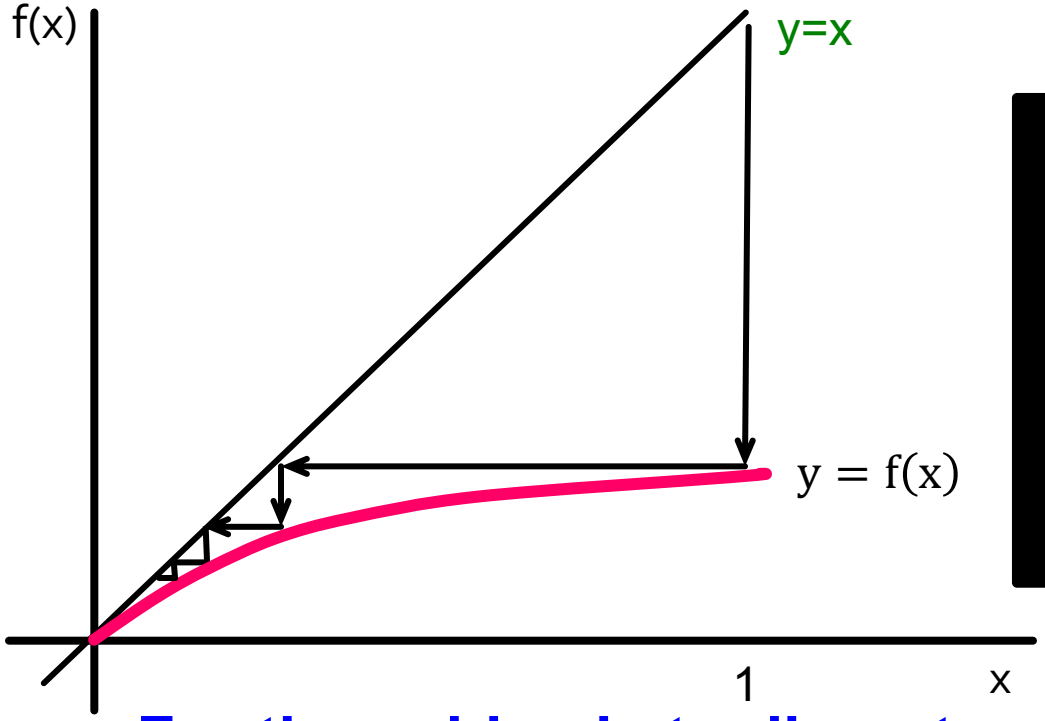
What do we know about the shape of f(x)?

$$f(0) = 0$$
$$f(1) = 1 - (1 - q)^d < 1$$
$$f'(x) = q \cdot d(1 - qx)^{d-1}$$
$$f'(0) = q \cdot d$$

f(x) is monotone: If g'(y)>0 for all y then g(y) is monotone.
In our case, 0≤x,q≤1, d>1 so f'(x)>0 so f(x) is monotone.
f'(x) non-increasing: since term $(1-qx)^{d-1}$ in *f'(x)* is decreasing as x decreases.

$f'(x)$ is monotone non-increasing on [0,1]!

Fixed Point: When is this zero?



Reproductive number $R_0 = q \cdot d$:
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: $f'(0) = q \cdot d < 1$

$$\lim_{h \to \infty} p_h = 0 \ \ when \ \ q \cdot d < 1$$

$q \cdot d$ = expected # of people that get infected

Important Points

- Reproductive number $\boldsymbol{R_0 = q \cdot d}$:
- There is an epidemic if $\boldsymbol{R_0 \geq 1}$

- **Only R_0 matters:**
  - $\boldsymbol{R_0 \geq 1}$: epidemic never dies and the number of infected people increases exponentially
  - $\boldsymbol{R_0 < 1}$: Epidemic dies out exponentially quickly

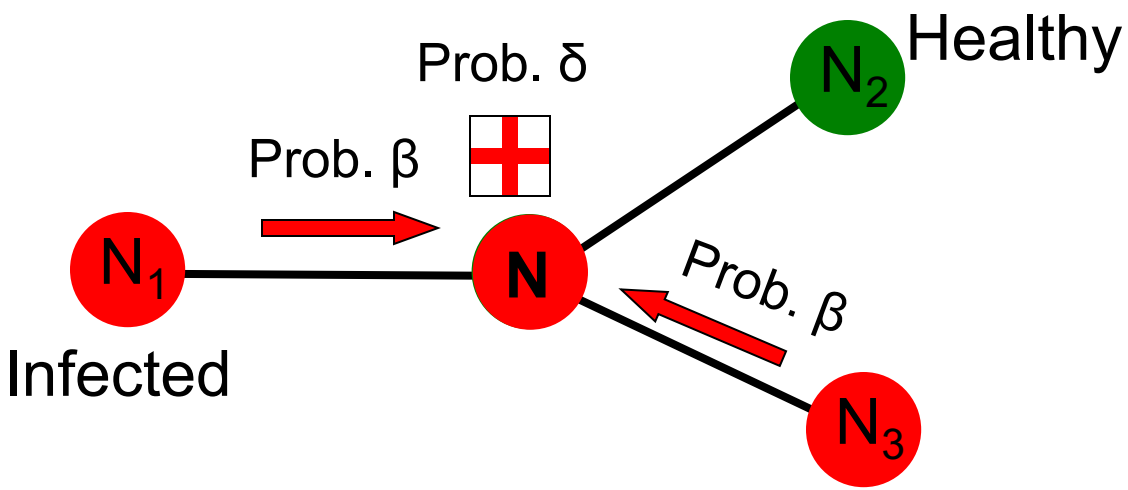
Models of Disease Spreading

We will learn about the
epidemic threshold

Spreading Models of Viruses

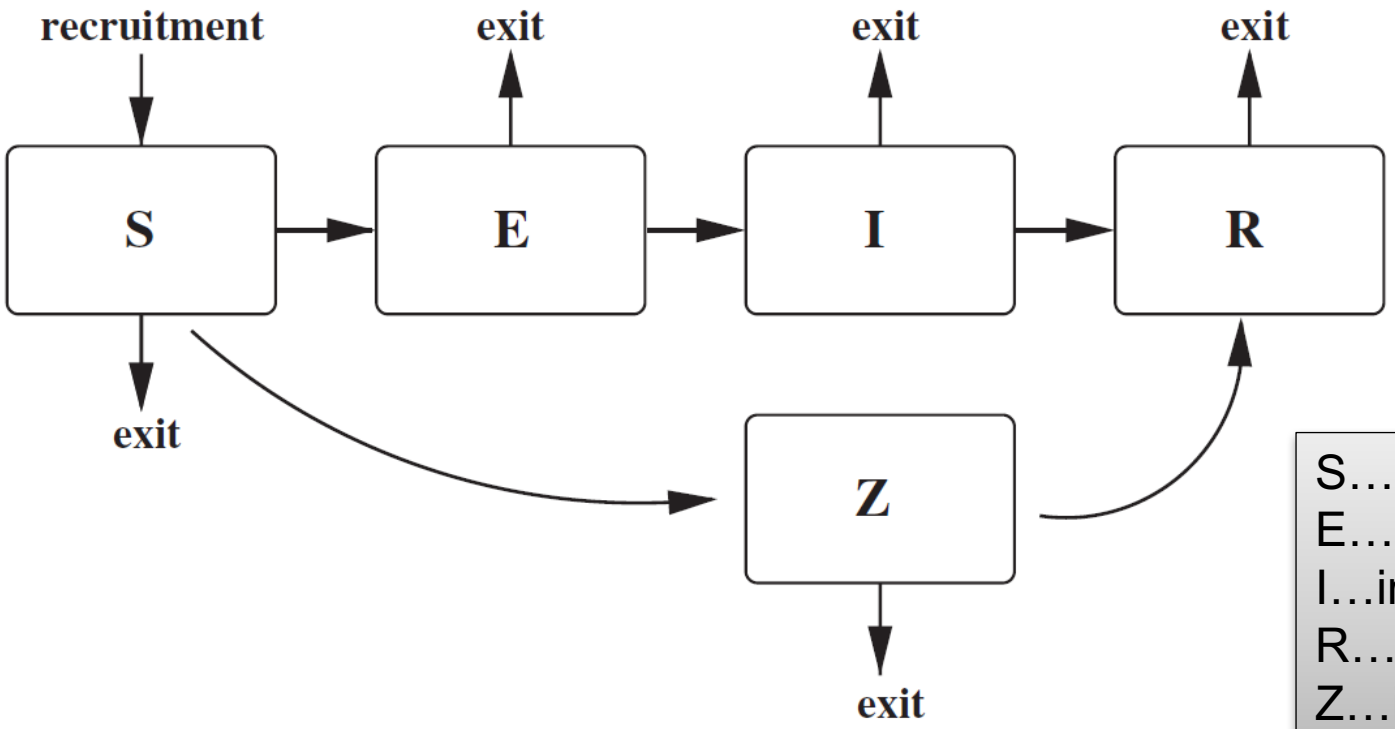
Virus Propagation: 2 Parameters:

- **(Virus) Birth rate β:**
  - probability than 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

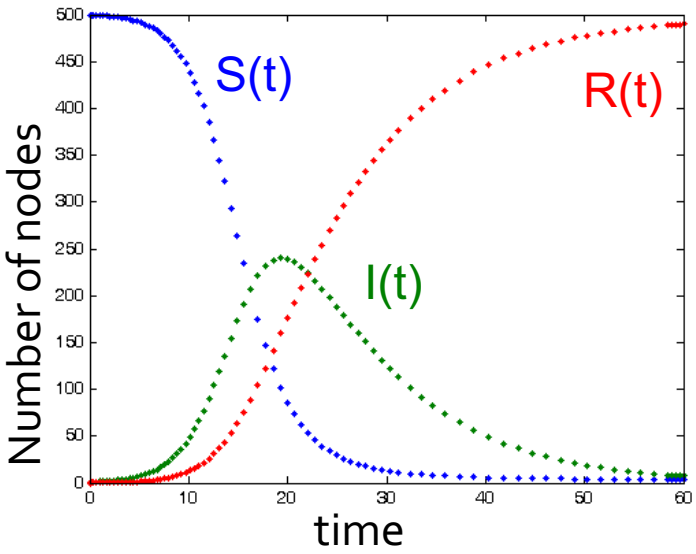
SIR Model

- **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 is:**

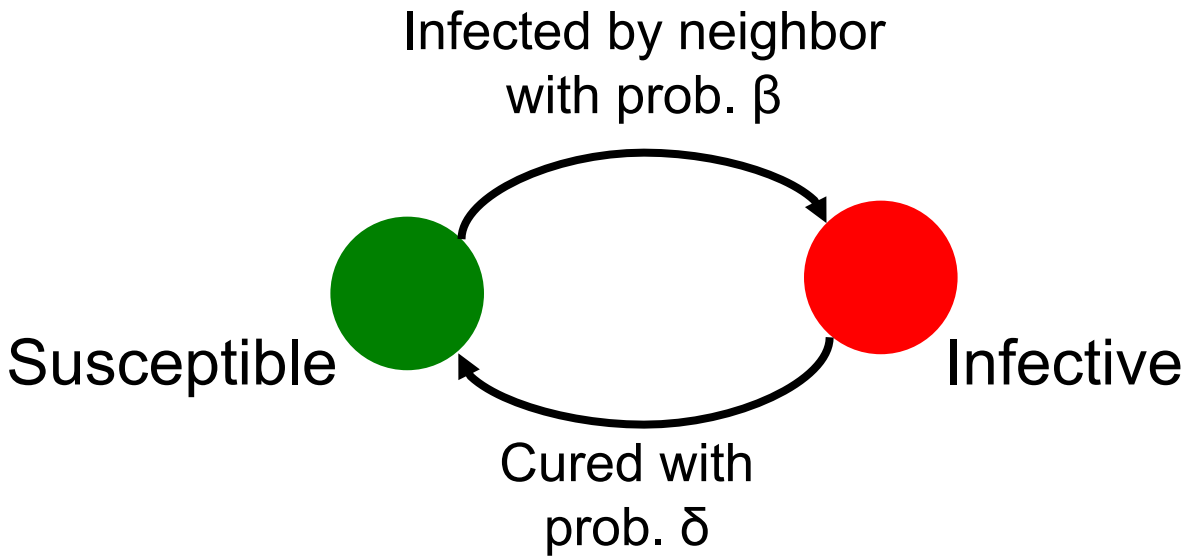
$$\frac{dS}{dt} = -\beta SI \qquad \frac{dR}{dt} = \delta I$$

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

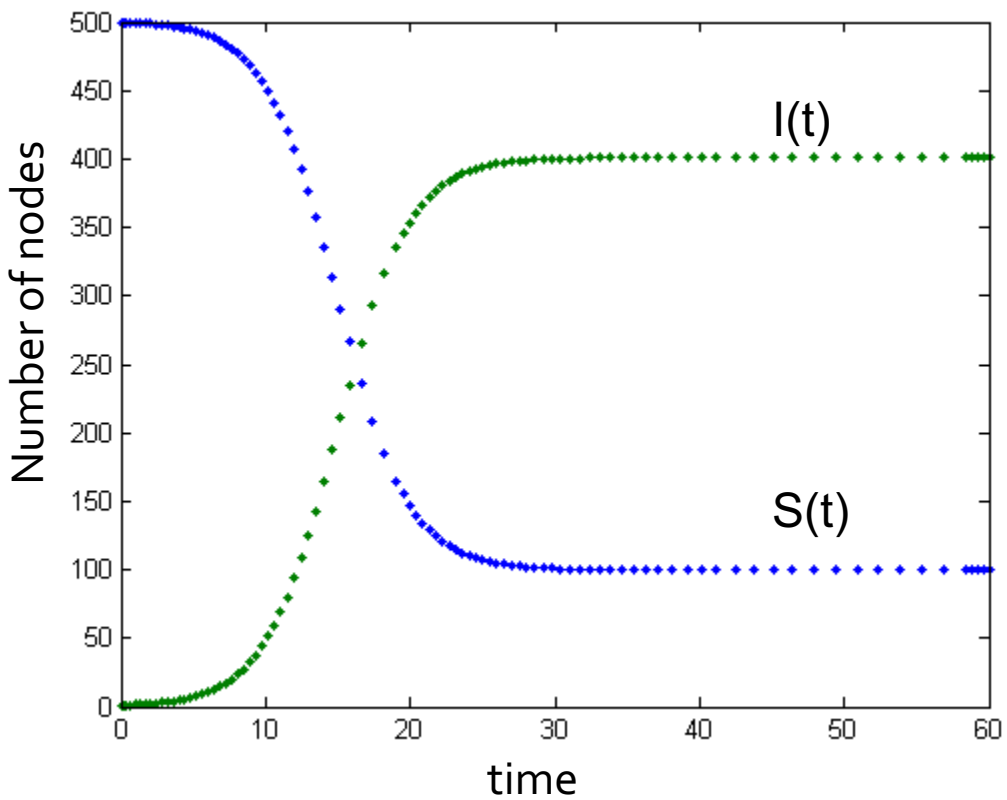


SIS Model

- **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*

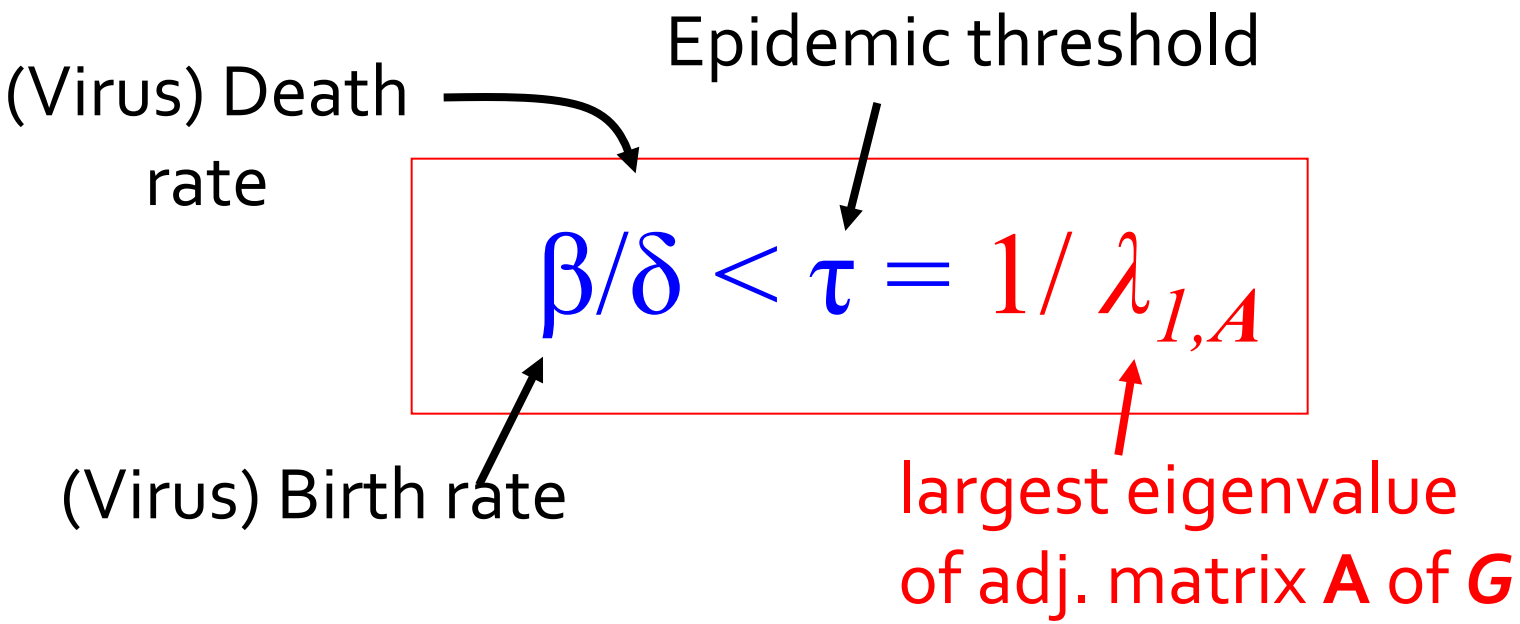
- **SIS Model:**
  Epidemic threshold of an arbitrary
- **graph *G* is τ, such that:**
  - **If virus "strength" $s = \beta / \delta < \tau$ the epidemic can not happen (it eventually dies out)**

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

[Wang et al. 2003]

Epidemic Threshold in SIS Model

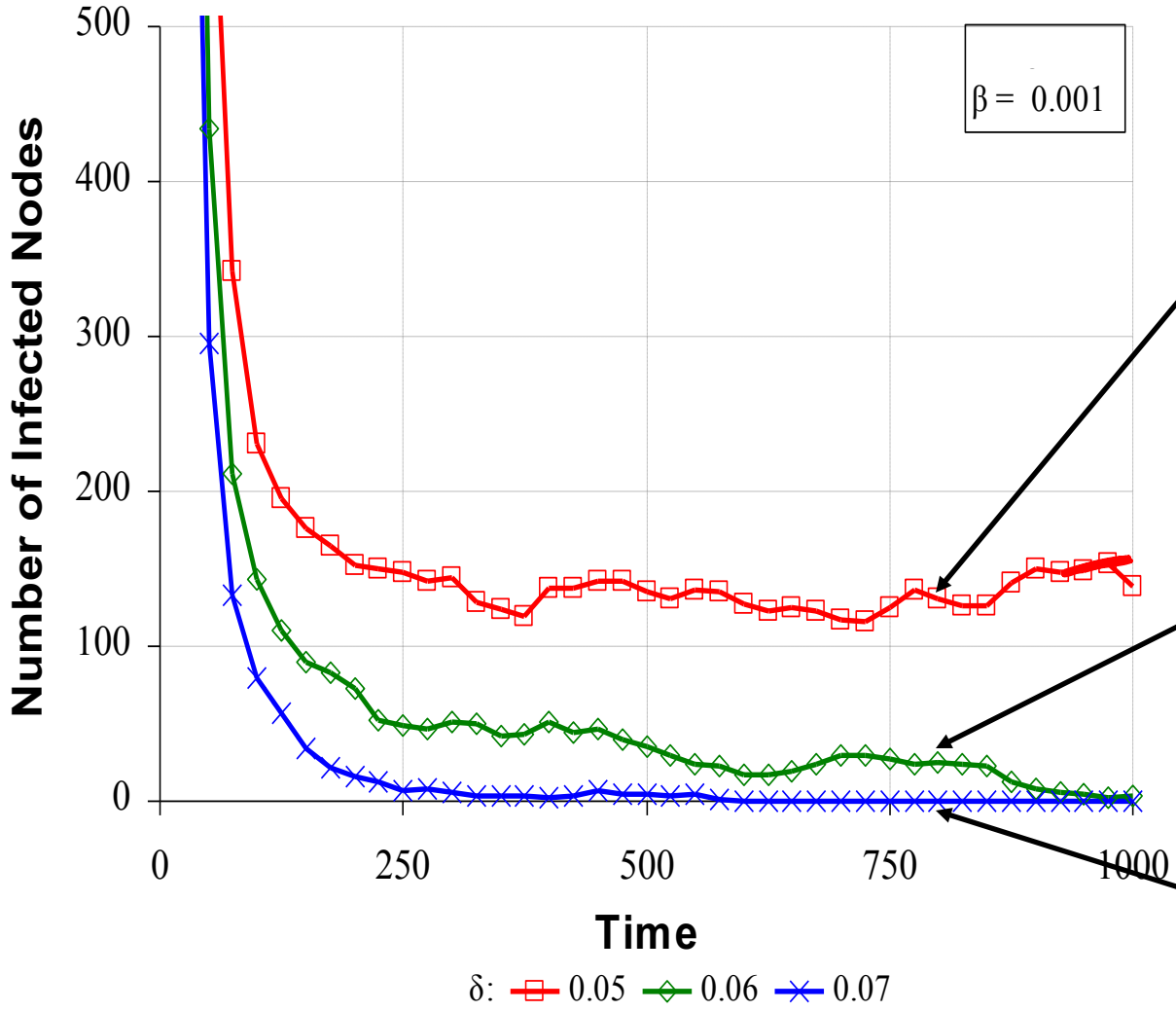
- **Fact: We have no epidemic if:**



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

[Wang et al. 2003]

Experiments (AS graph)



10,900 nodes and 31,180 edges

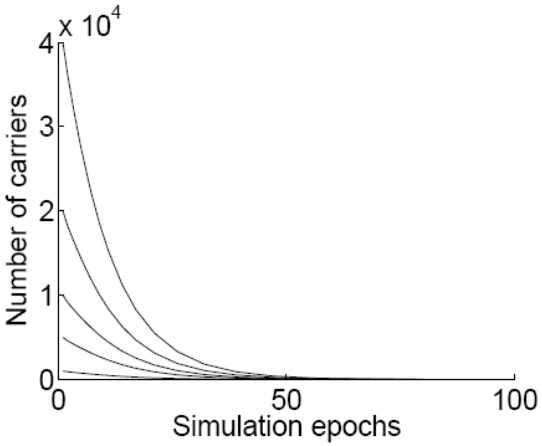
$s=\beta/\delta > \tau$ (above threshold)

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

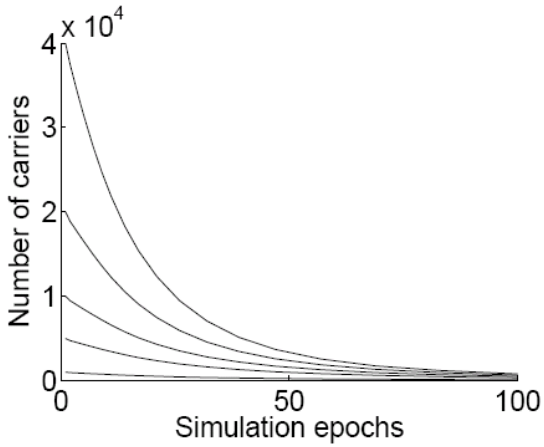
$s=\beta/\delta < \tau$ (below threshold)

Experiments

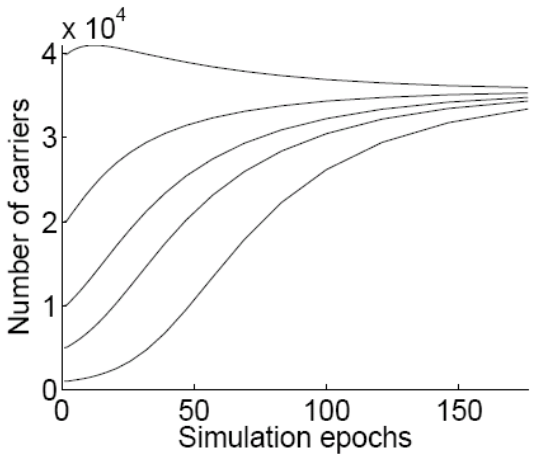
- **Does it matter how many people are initially infected?**



(a) Below the threshold, s=0.912

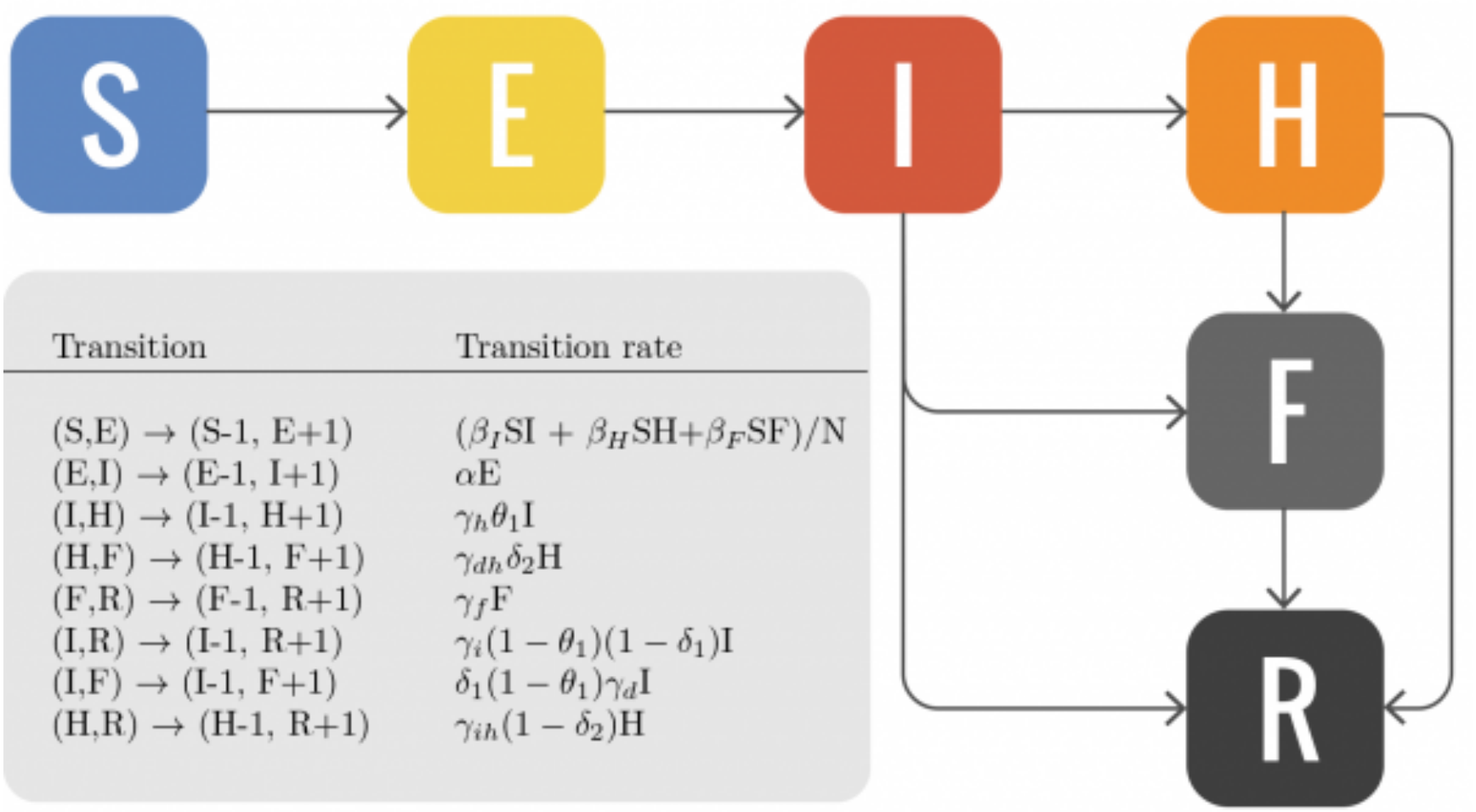


(b) At the threshold, s=1.003



(c) Above the threshold, s=1.1

Example: Ebola

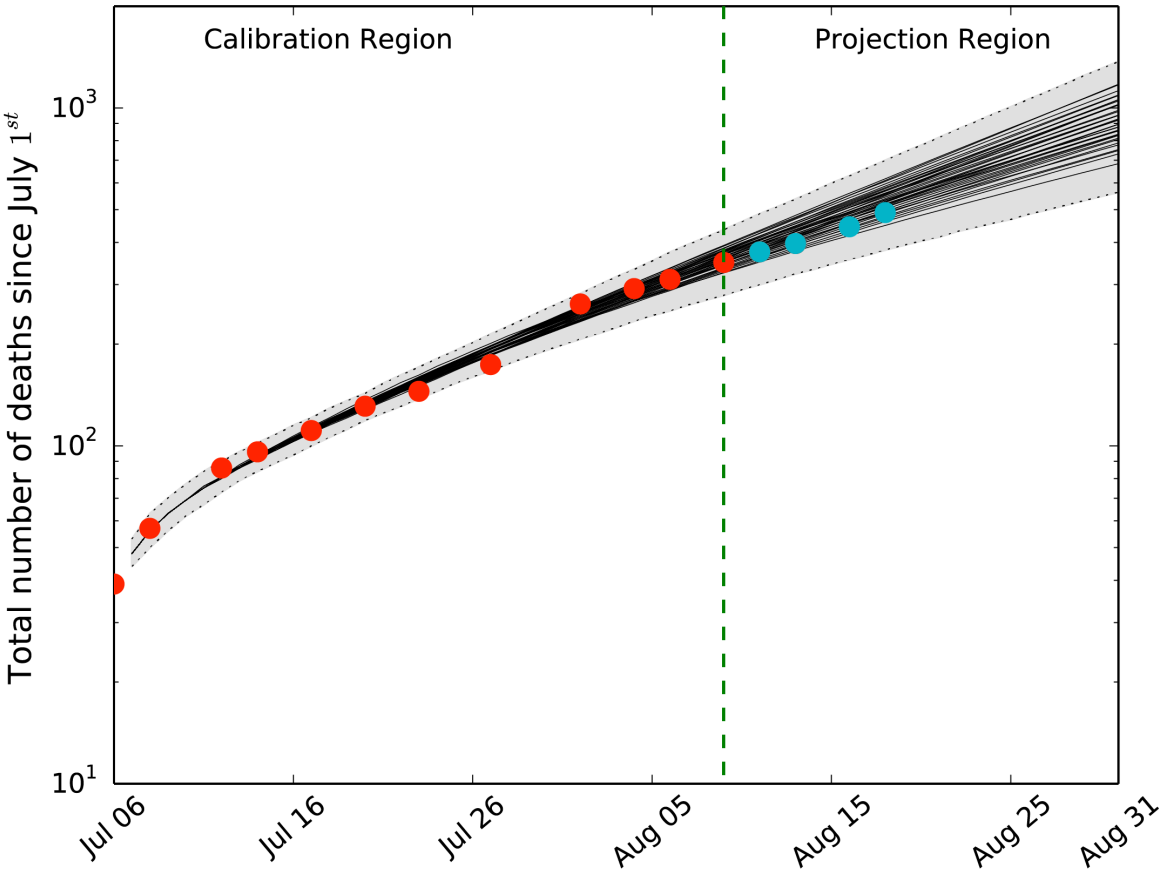


Transition	Transition rate
(S,E) → (S-1, E+1)	$(\beta_I SI + \beta_H SH + \beta_F SF)/N$
(E,I) → (E-1, I+1)	αE
(I,H) → (I-1, H+1)	$\gamma_h \theta_1 I$
(H,F) → (H-1, F+1)	$\gamma_{dh} \delta_2 H$
(F,R) → (F-1, R+1)	$\gamma_f F$
(I,R) → (I-1, R+1)	$\gamma_i (1-\theta_1)(1-\delta_1) I$
(I,F) → (I-1, F+1)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

[Gomes et al., 2014]

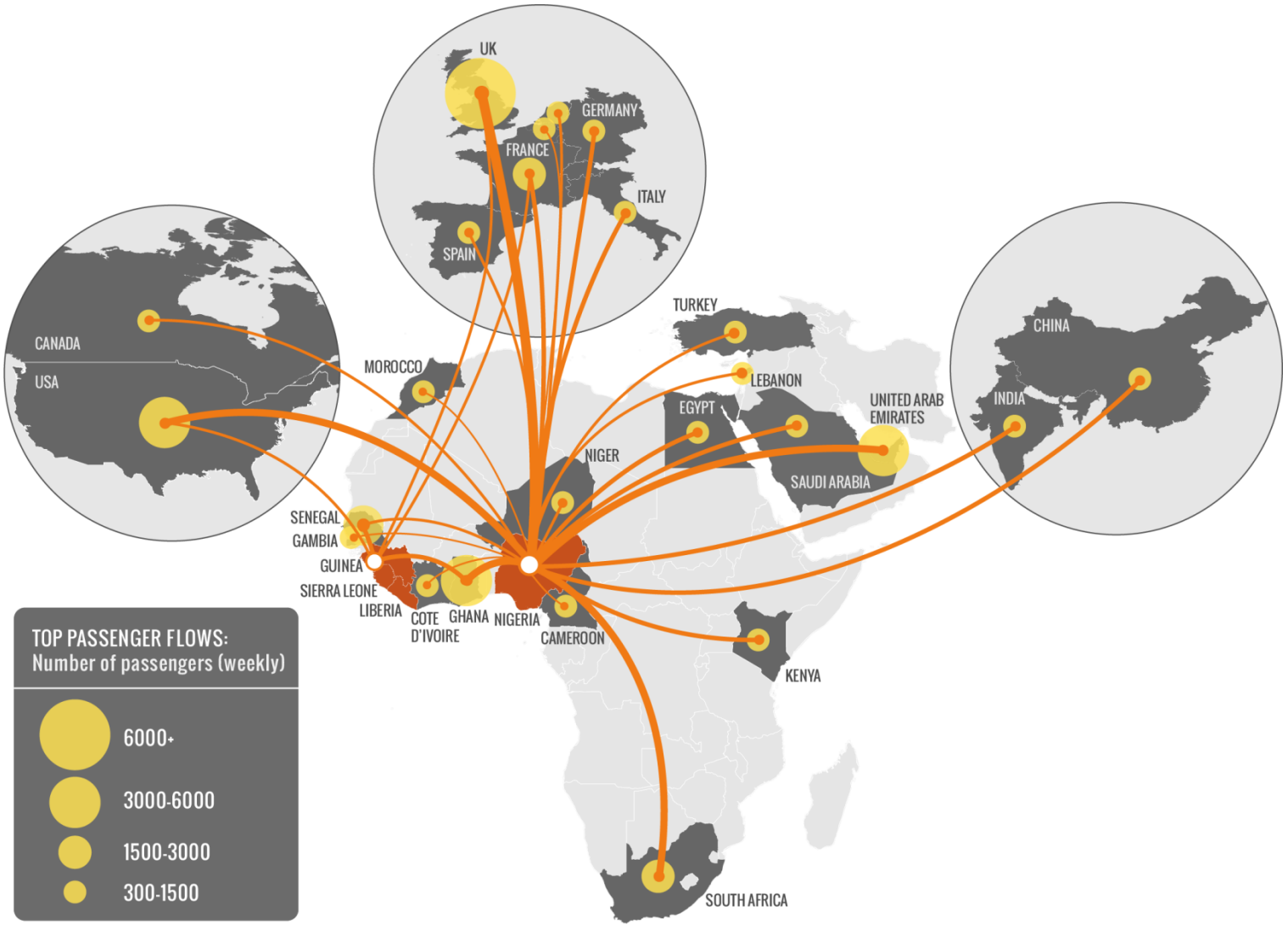
Example: Ebola, R_0=1.5-2.0



Read an article about how to estimate R_0 of ebola.

Example: Ebola

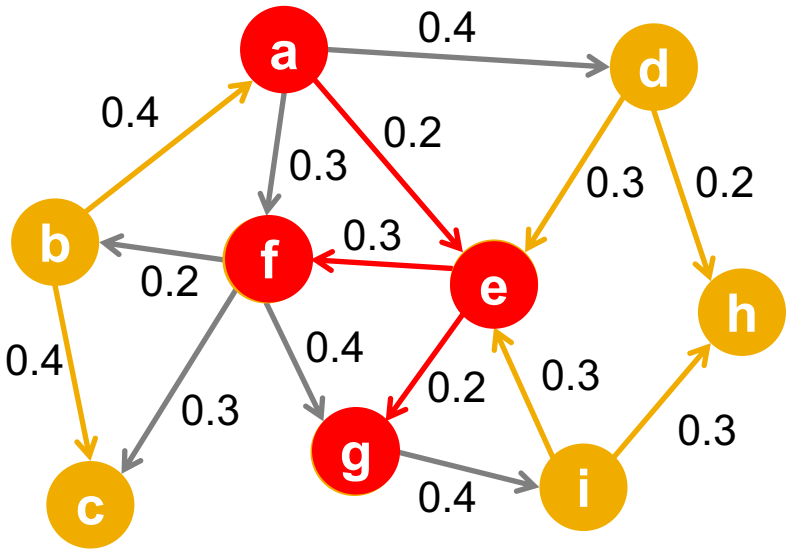
[Gomes et al., 2014]



Independent Cascade Model

Independent Cascade Model

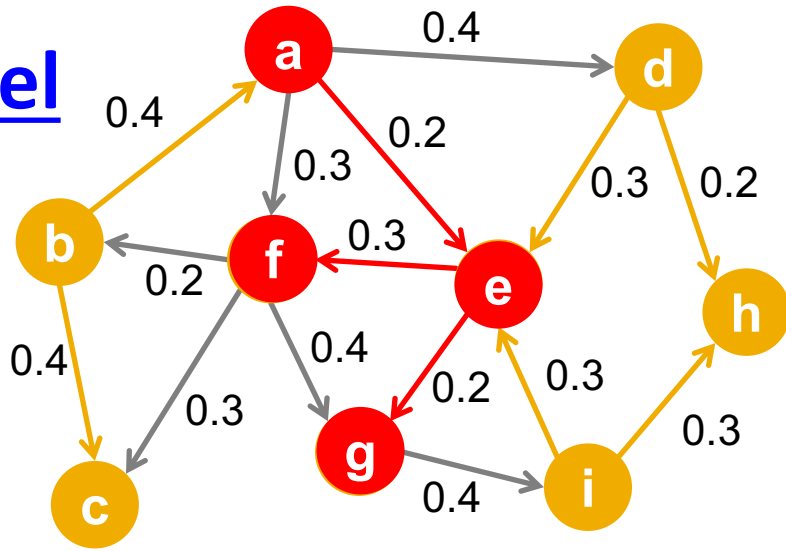
- **Initially some nodes *S* are active**
- Each edge ***(u,v)*** has probability (weight) p_{uv}



- **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]

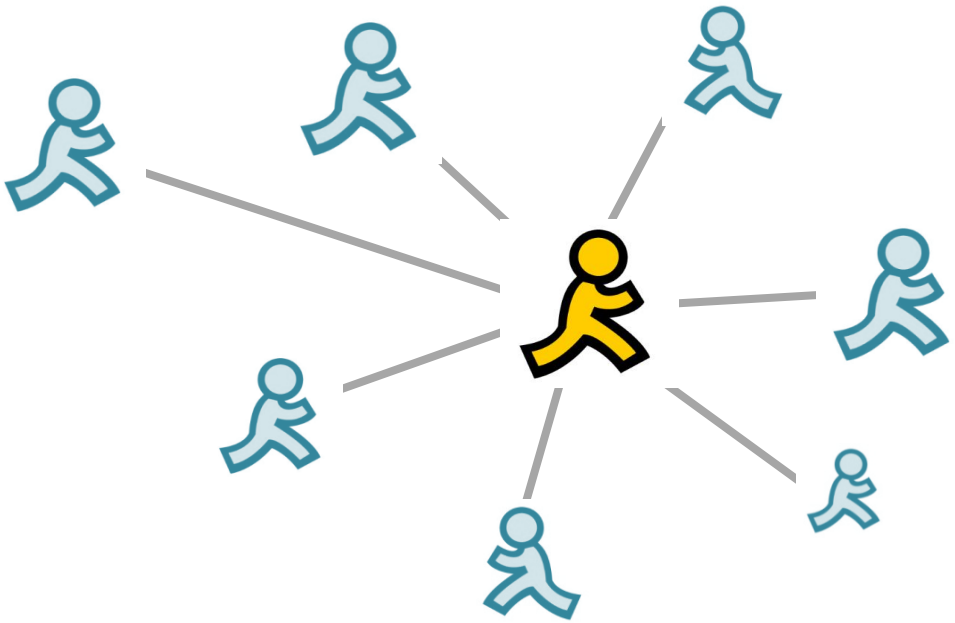


- **Solution:** Make all edges have the same weight (which brings us back to the SIR model)
  - Simple, but too simple
- **Can we do something better?**

Exposures and Adoptions

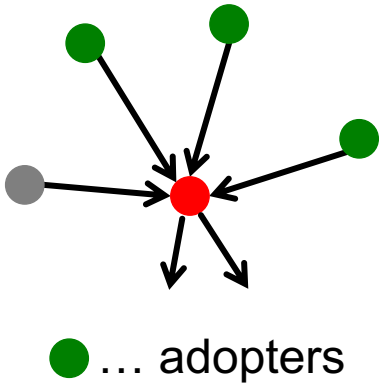
[KDD '12]

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

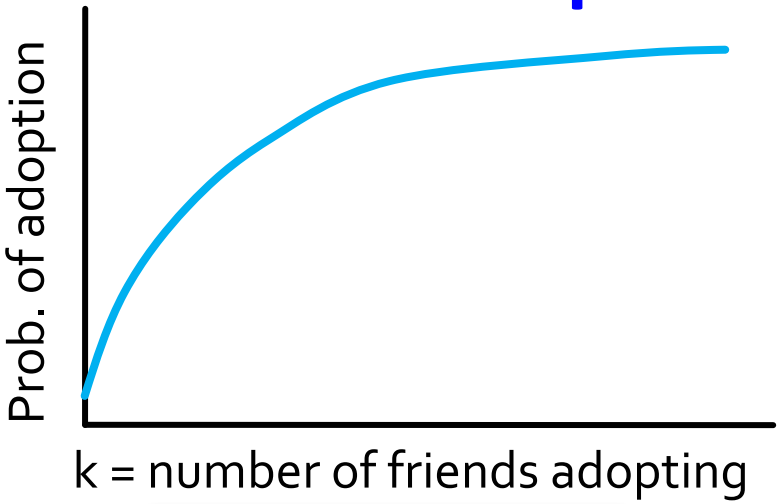


Exposure Curves

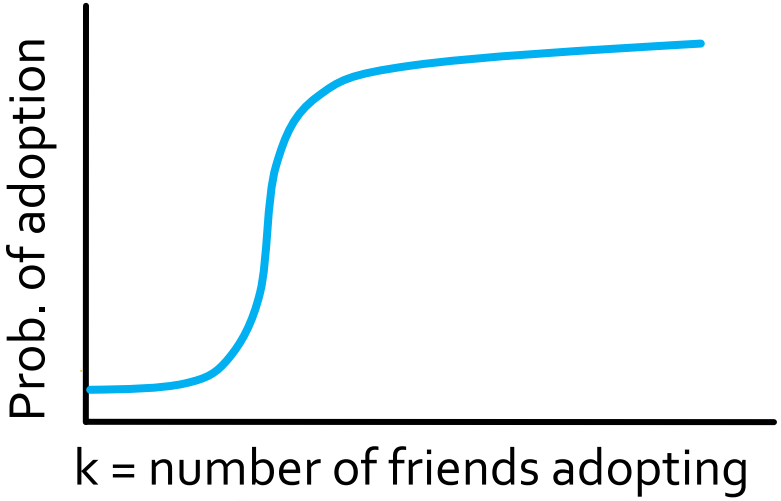
- **Exposure curve:**
  - **Probability of adopting new behavior depends on the total number of friends who have already adopted**



- **What's the dependence?**



Diminishing returns: Viruses, Information

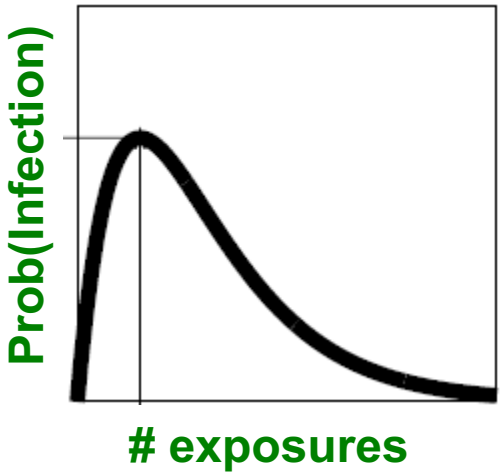


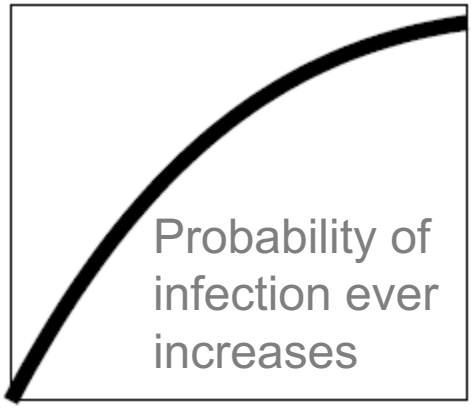
Critical mass: Decision making

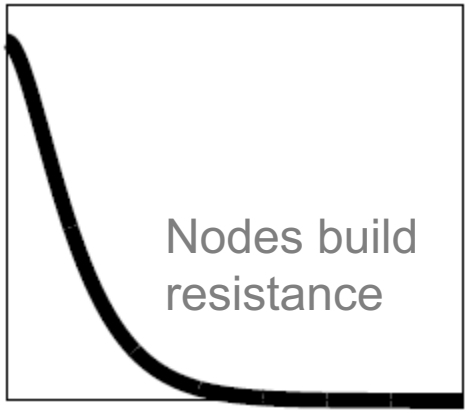
[KDD '12]

Exposure Curves

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

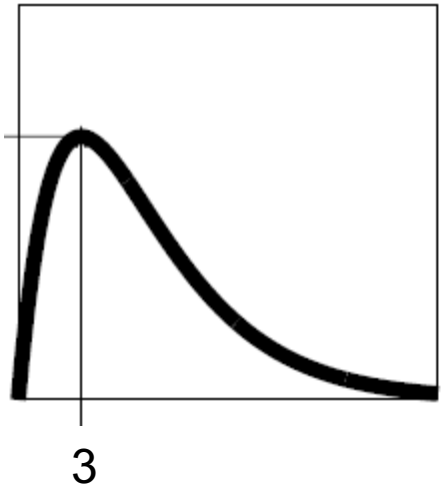






Example Application

- **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]

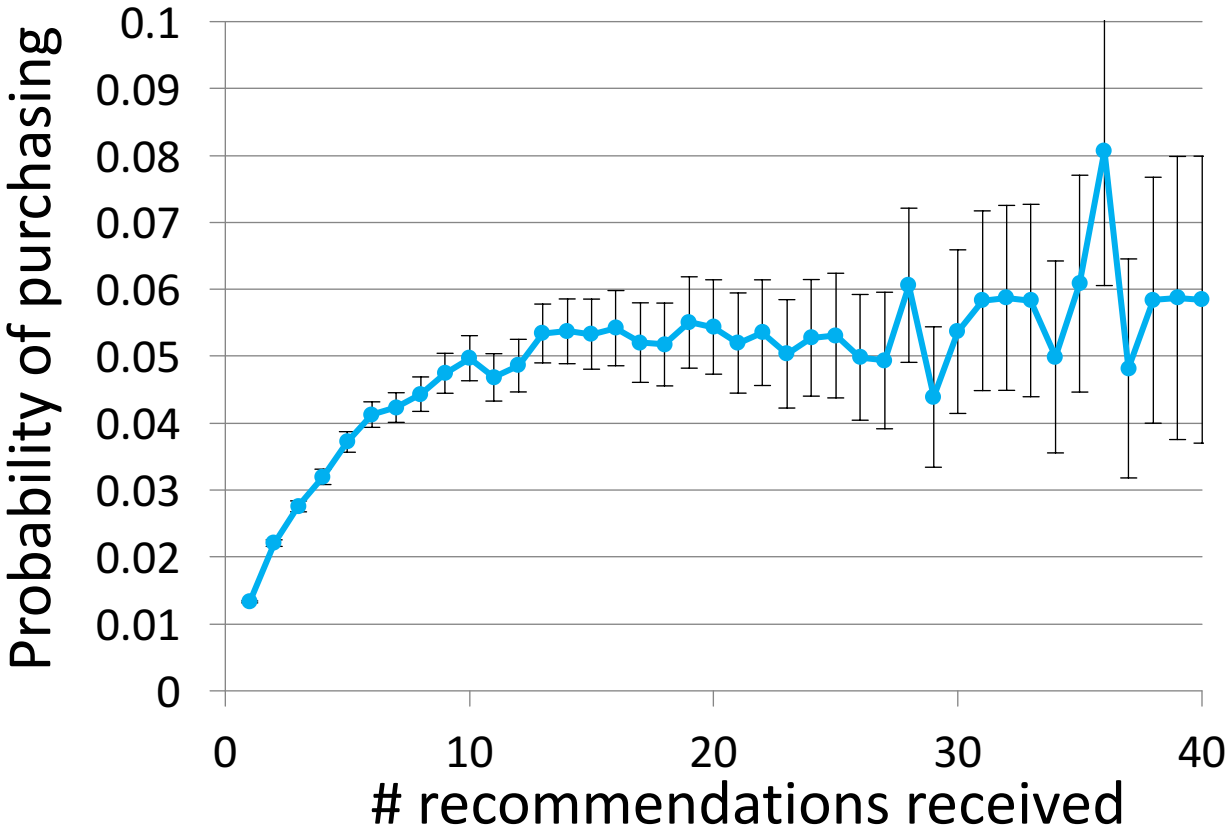
Diffusion in Viral Marketing

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

- **Data: Incentivized Viral Marketing program**
  - 16 million recommendations
  - 4 million people, 500k products

[Leskovec et al., TWEB '07]

Exposure Curve: Validation

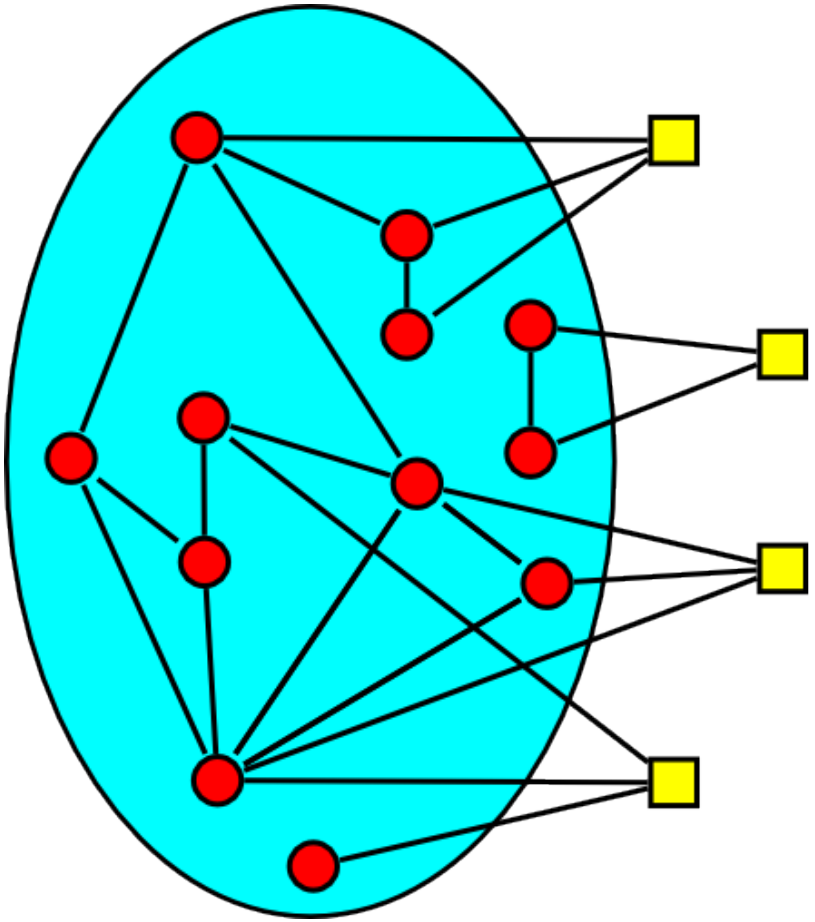


DVD recommendations
(8.2 million observations)

[Backstrom et al. KDD '06]

Exposure Curve: LiveJournal

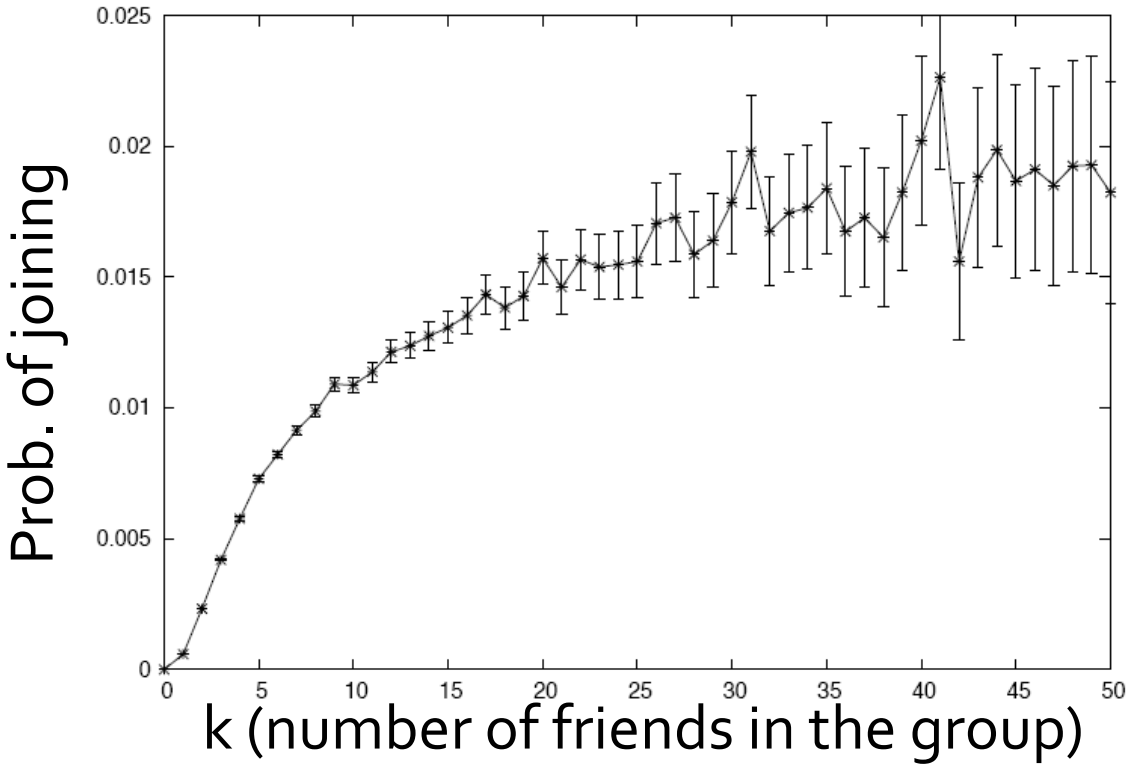
- **Group memberships spread over the network:**
  - **Red** circles represent existing group members
  - **Yellow** squares may join
- **Question:**
  - How does prob. of joining a group depend on the number of friends already in the group?



[Backstrom et al., KDD '06]

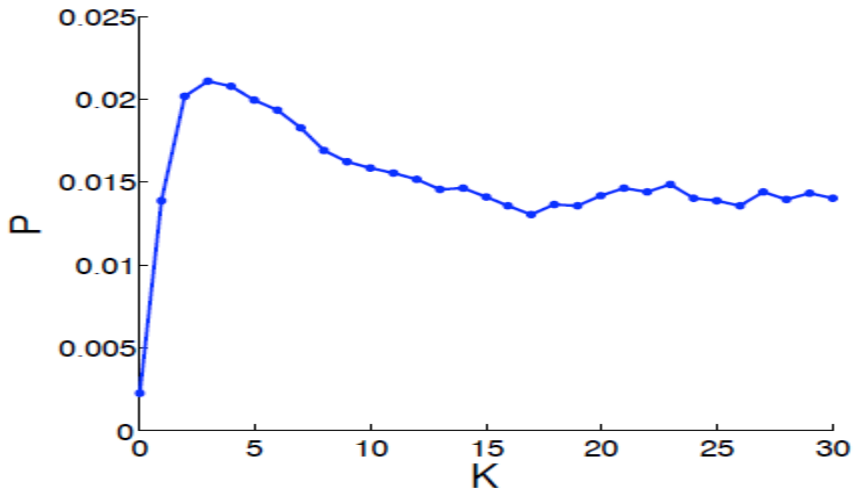
Exposure Curve: LiveJournal

- **LiveJournal group membership**



Exposure Curve: Information

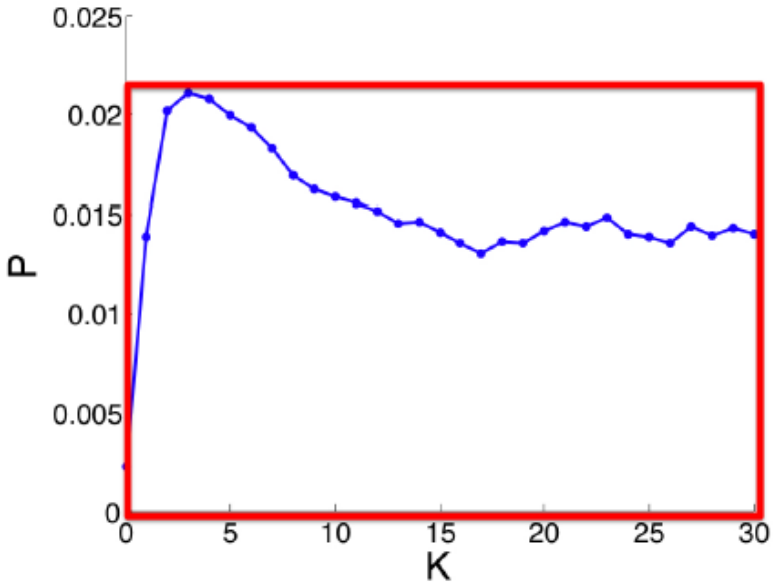
- **Twitter** [Romero et al. '11]
  - Aug '09 to Jan '10, 3B tweets, 60M users

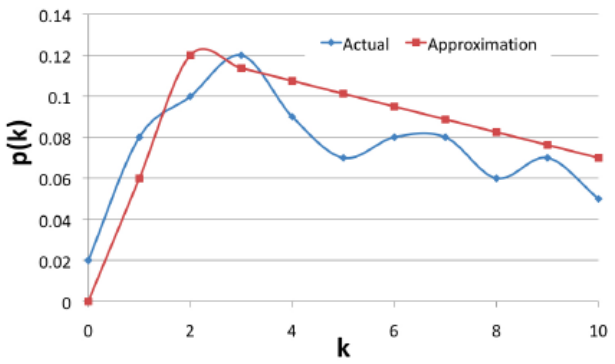


  - **Avg. exposure curve for the top 500 hashtags**
  - **What are the most important aspects of the shape of exposure curves?**
  - Curve reaches peak fast, decreases after!

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 height *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**

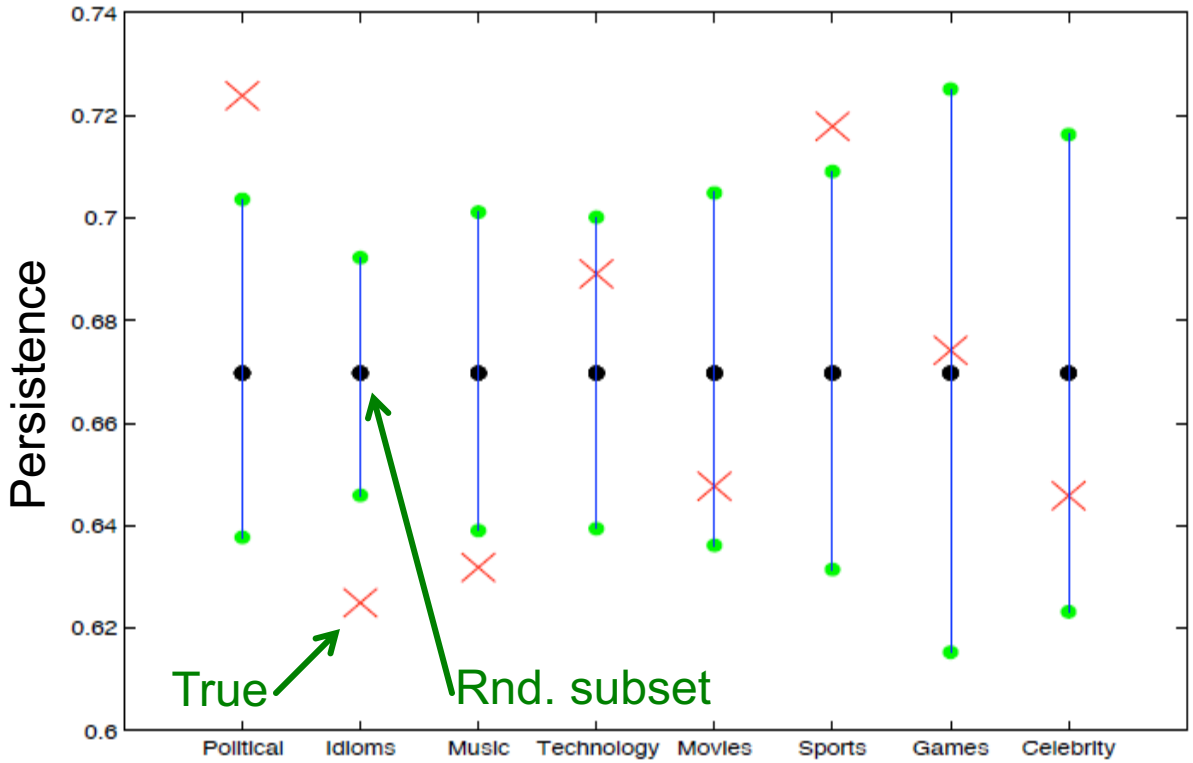




Exposure Curve: Persistence

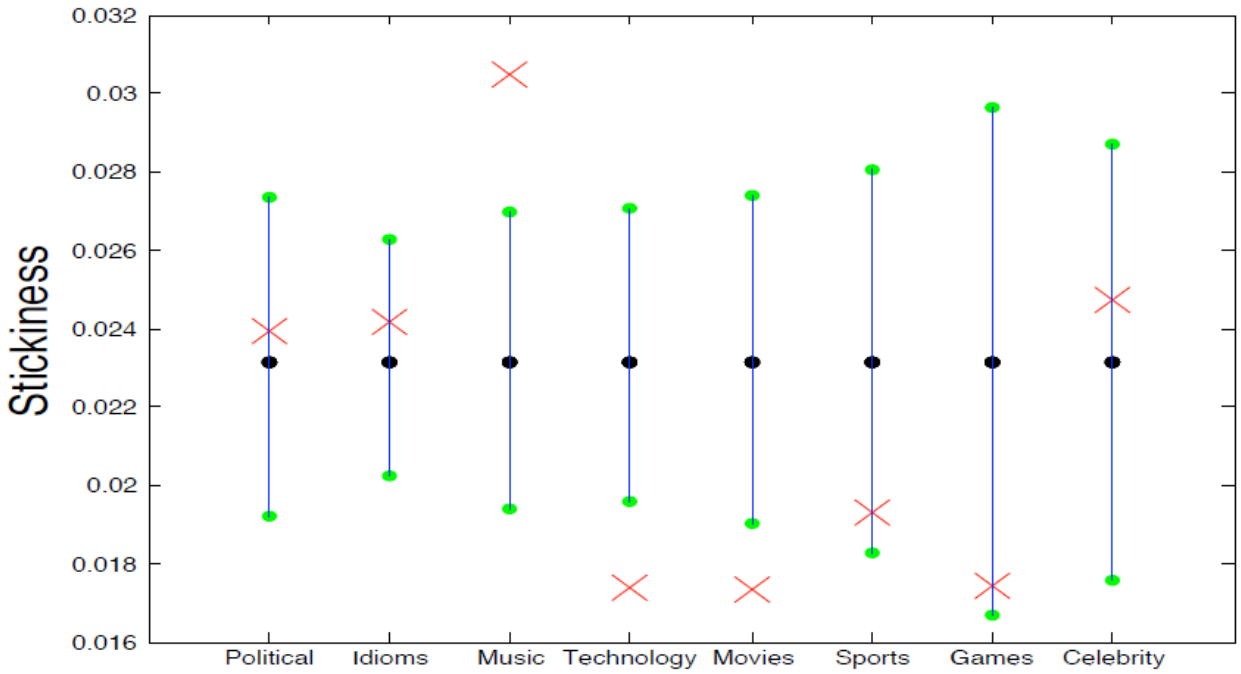
- **Manually identify 8 broad categories with at least 20 HTs in each**

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)

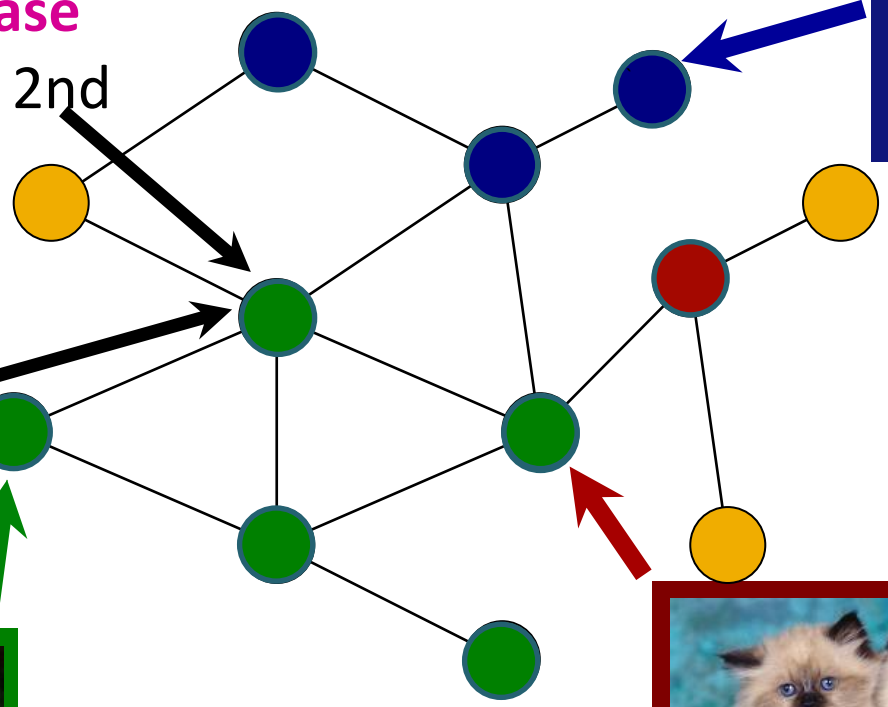
Modeling Interactions Between Contagions

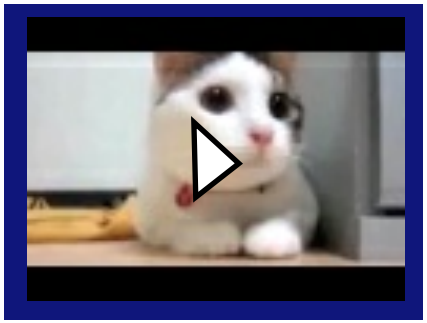
Information Diffusion

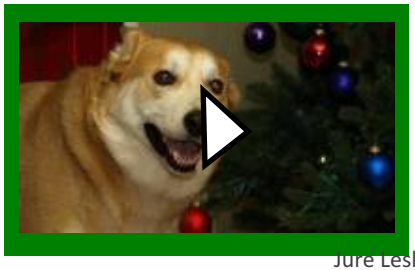
So far we considered pieces of information as **independently** propagating. **Do pieces of information interact?**

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

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



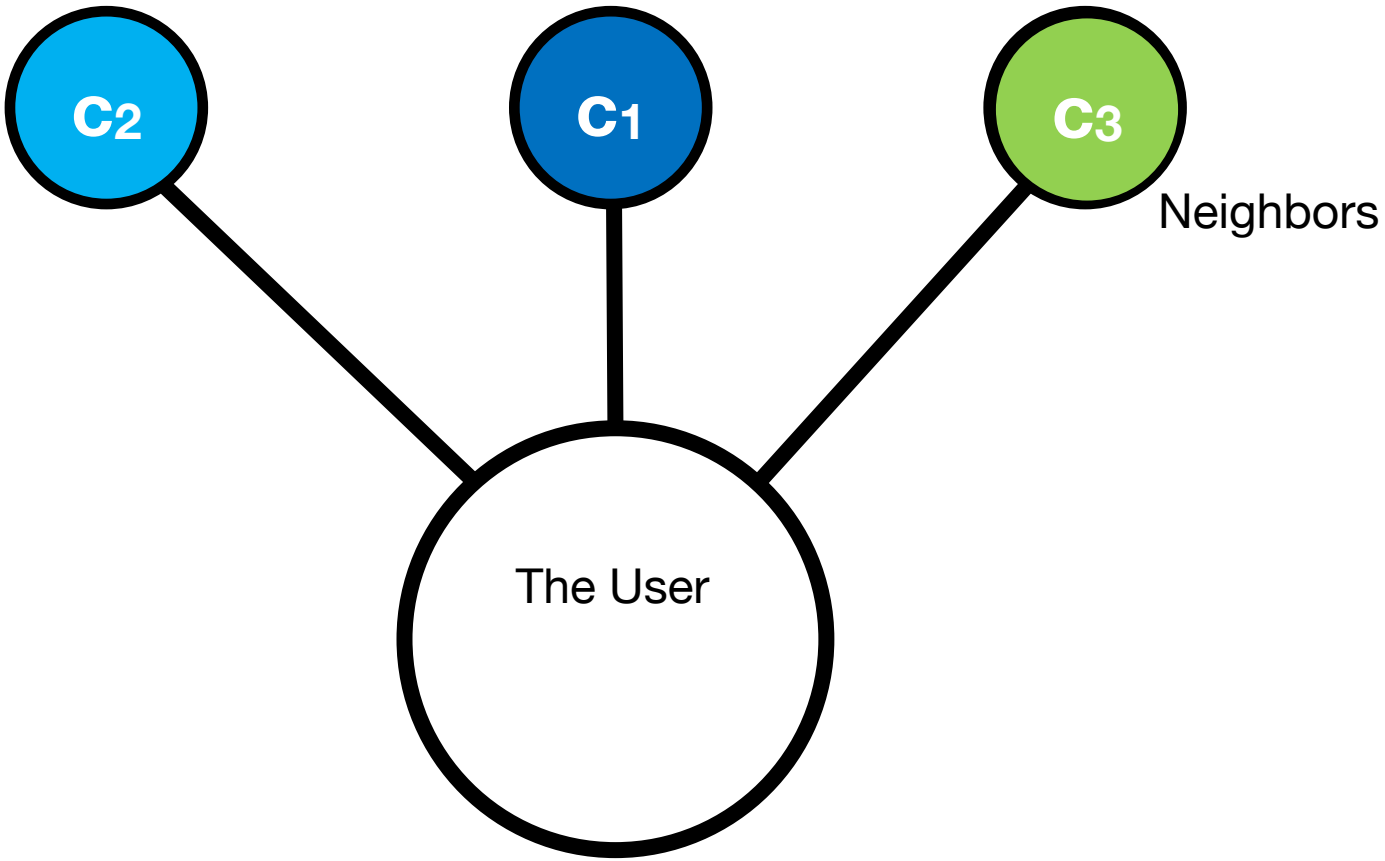




Modeling Interactions

- **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

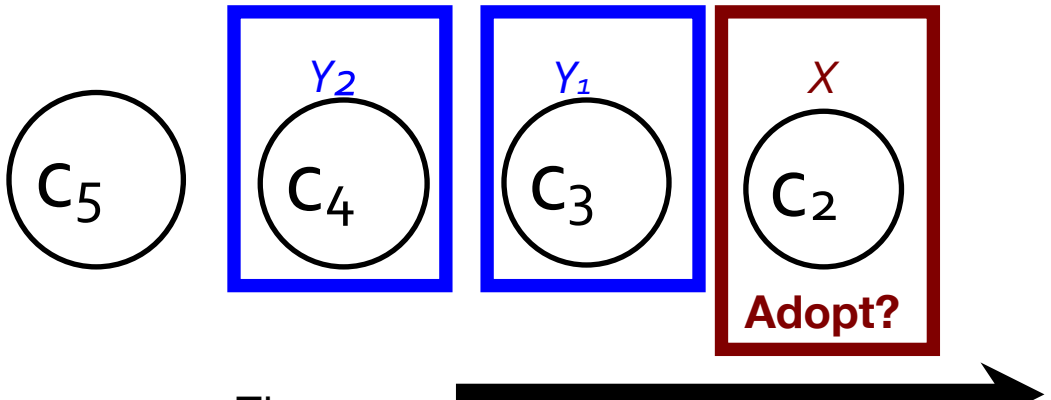


P(adopt c_3 | exposed *to* c_2 , c_1, c_0)

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 (X is exposed by them in order):



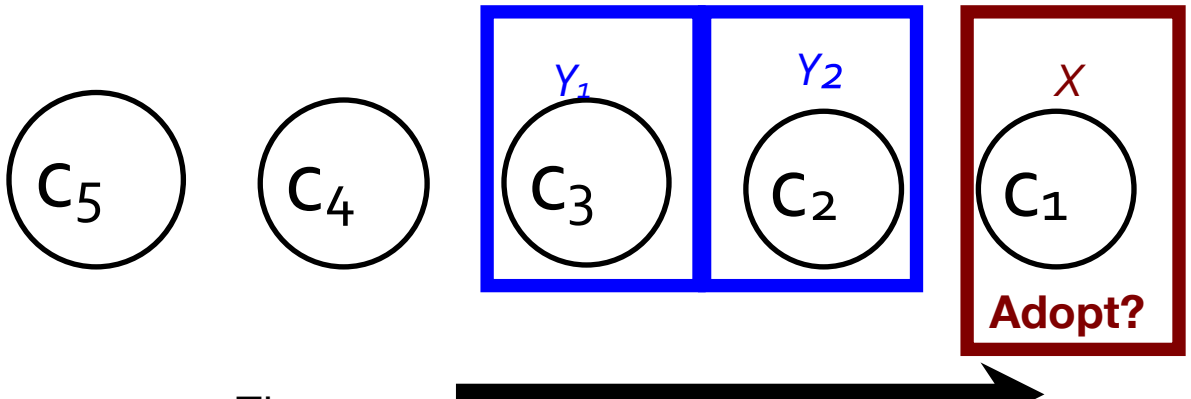
The Model

- 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:



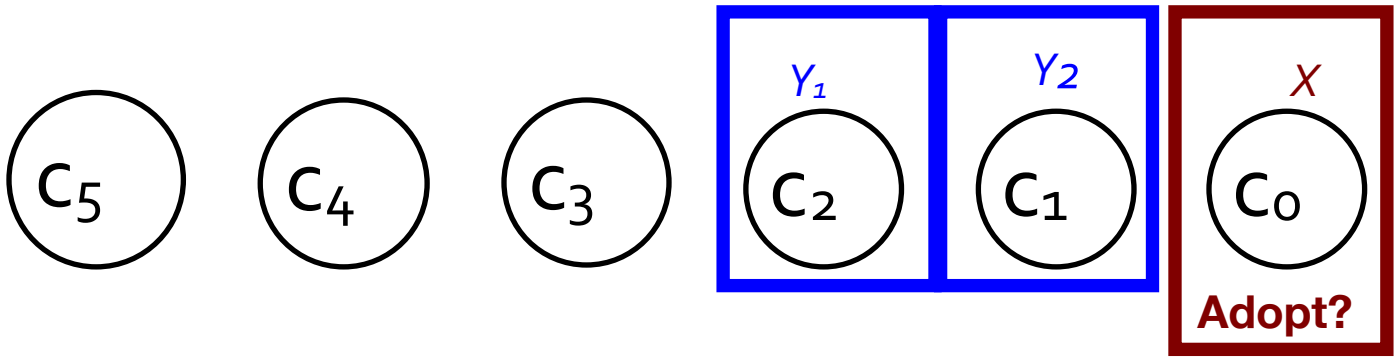
The Model

- 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

- We want to estimate: $P(X \mid Y_1, \ldots Y_5)$
- **What's the problem?**
  - **What's the size of probability table $P(X \mid Y_1, \ldots Y_5)$?**

    $= (\text{Num. Contagions})^5 \approx 1.9 \times 10^{21}$

- **Simplification: Assume Y_i is independent of Y_j**

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

We apply Bayes theorem twice and use the independence assumption

The Model

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

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

- **Next, assume "topics":**
  - $\boldsymbol{\Delta^{(k)}_{cont}(u_i, u_j)}$ models the change in infection prob. of $\boldsymbol{u_j}$ given that exposure $\boldsymbol{k}$-steps ago was $\boldsymbol{u_j}$
  - **We estimate P(X) and $\Delta^{(k)}(u_i, u_j)$ by simply counting**
    - **P(X) … fraction of people exposed to X that got infected by X**
    - **$\Delta^{(k)}(u_i, u_j)$ … P(X) – fraction of people first exposed to u_i and then to u_j and then got infected by u_j.**

Dataset: Twitter

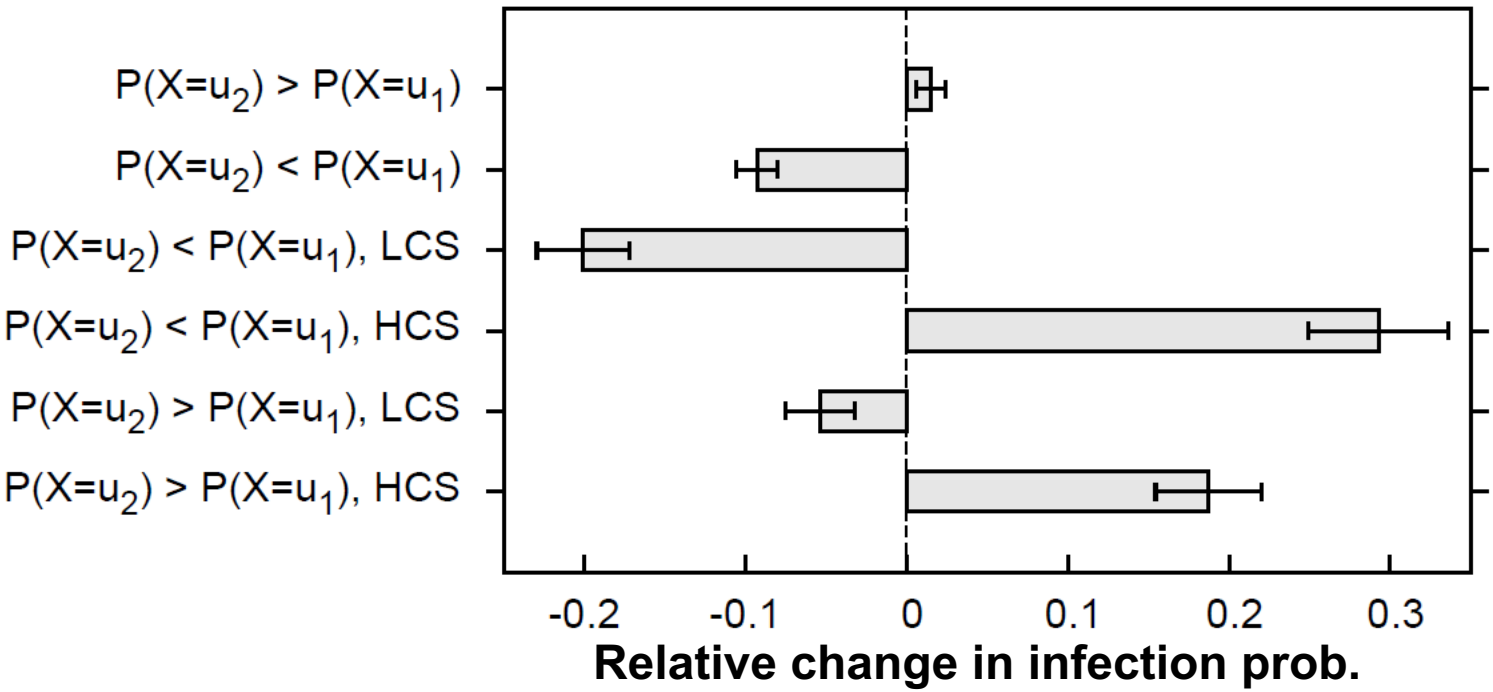
- **Data from Twitter**
  - *Complete* data from Jan 2011: 3 billion tweets
  - All URLs tweeted by at least 50 users: 191k
- **Task:**

  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?
    - LCS (low content similarity), HCS (high content similarity)
  - **u_1 is highly viral? Prob. of infection P(u):**



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!

Conclusion

- **R_0: Epidemics die out if R_0<1**
  - R_0: reproductive number
- **Epidemic Threshold:** Virus "strength" $s = \beta / \delta < \tau$ the epidemic can not happen (it eventually dies out)
- **Shape of the adoption curve:**
- **Modeling interactions between contagions**