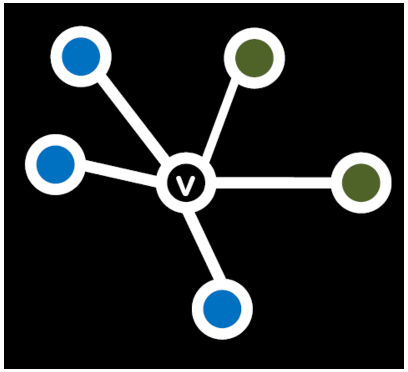
LECTURE 9:CONTAGION AND VIRAL MARKET

COMP4641: Social Information Network Analysis and Engineering

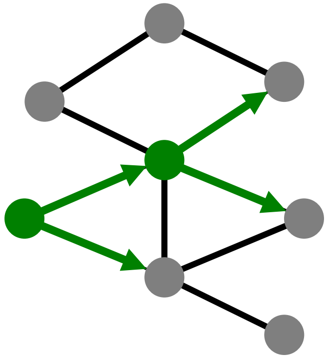
Friday March 13th 2015

Models of Cascading Behavior

- **Last time:**
  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



- **Today: Probabilistic Models**
  - Let's you do things by observing data
  - We loose "why people do things"

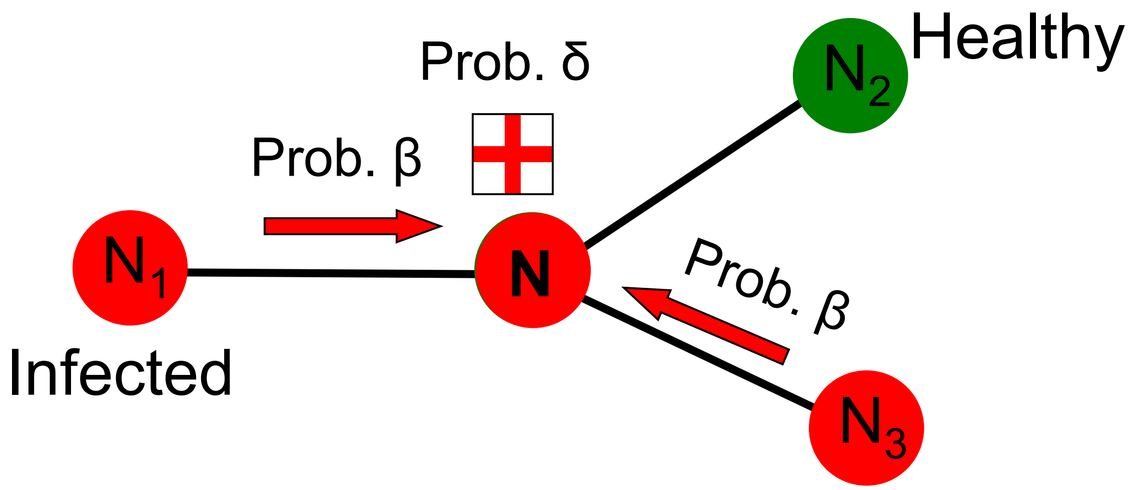


CLASSICAL MODELS OF DISEASE SPREADING

Spreading Models of Viruses

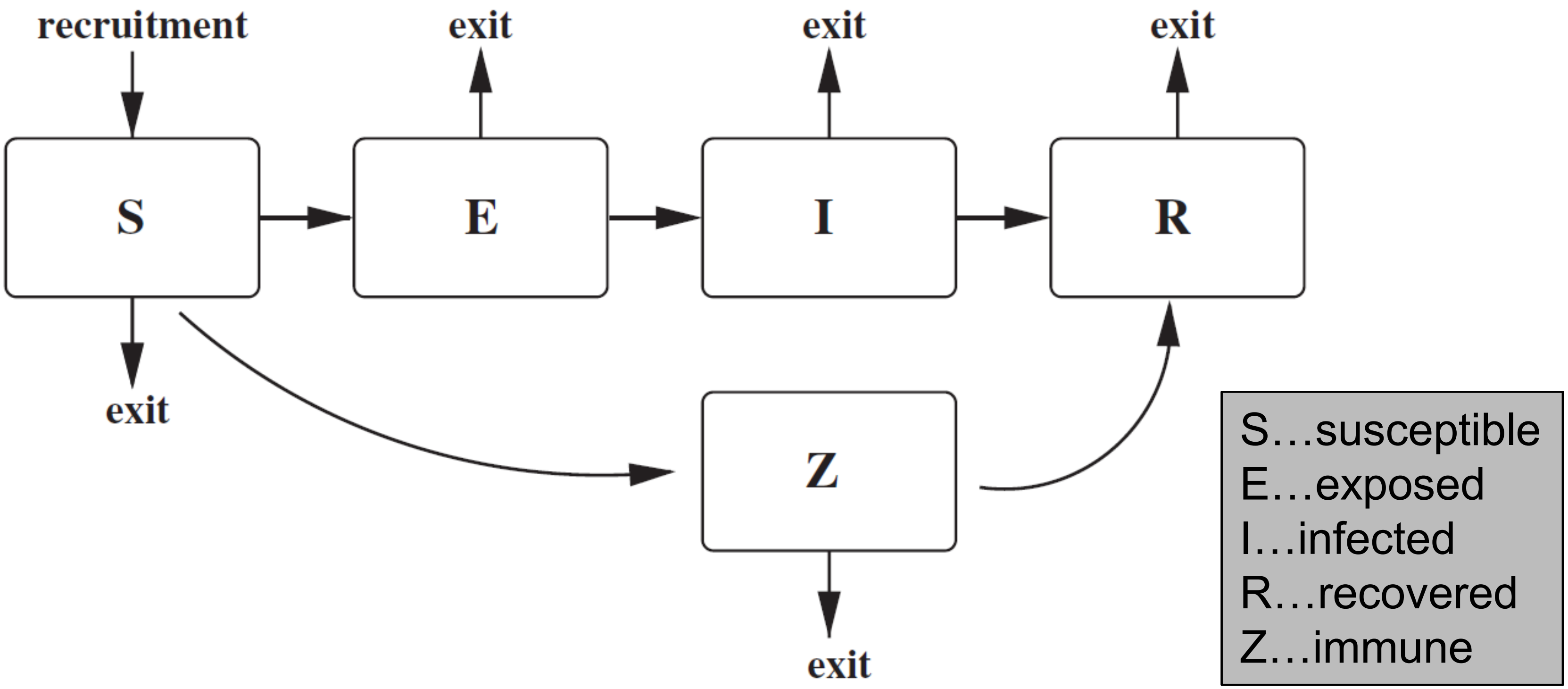
Virus Propagation: 2 Parameters:

- **(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



SIR Model

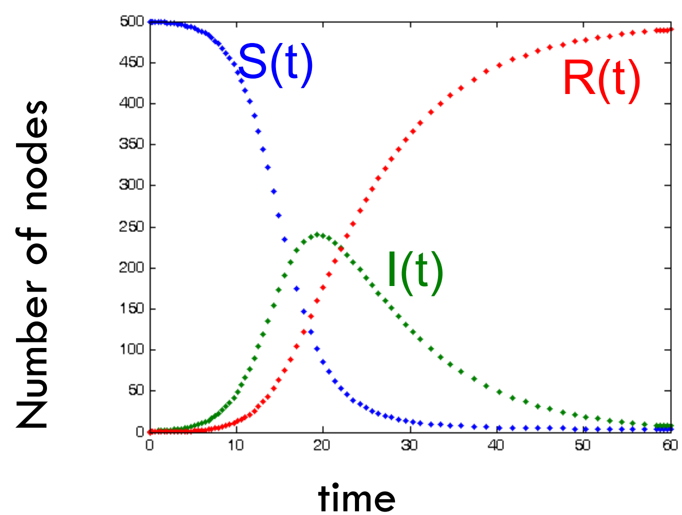
- **SIR model:** Node goes through phases

Susceptible $\xrightarrow{\beta}$ Infected $\xrightarrow{\delta}$ Recovered

  - 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:**

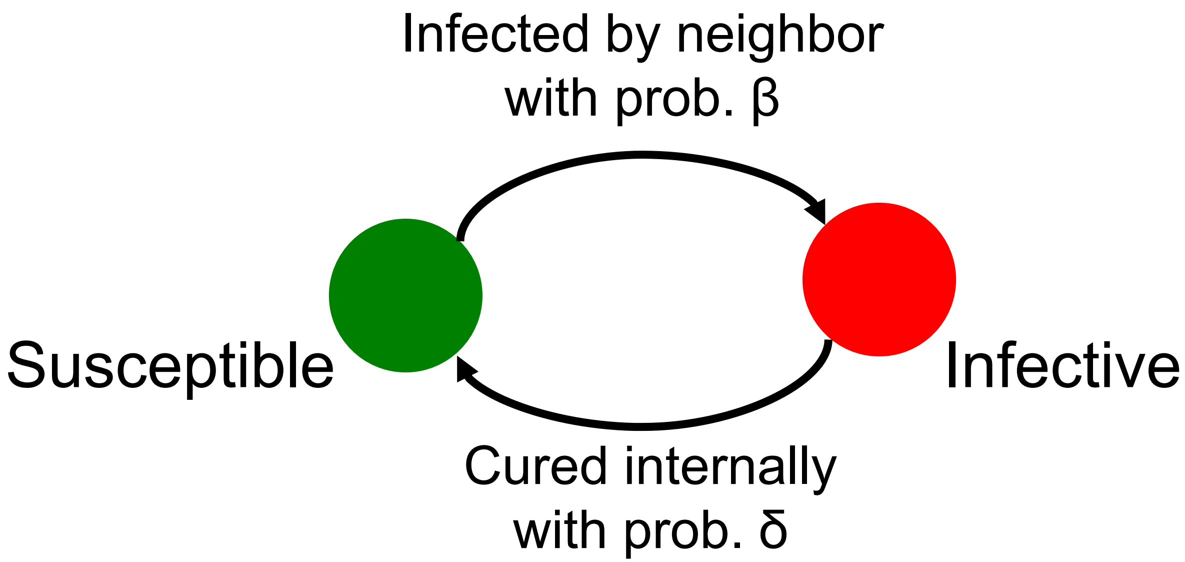
$$\frac{dS}{dt} = -\beta SI \qquad \frac{dR}{dt} = \delta I$$

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

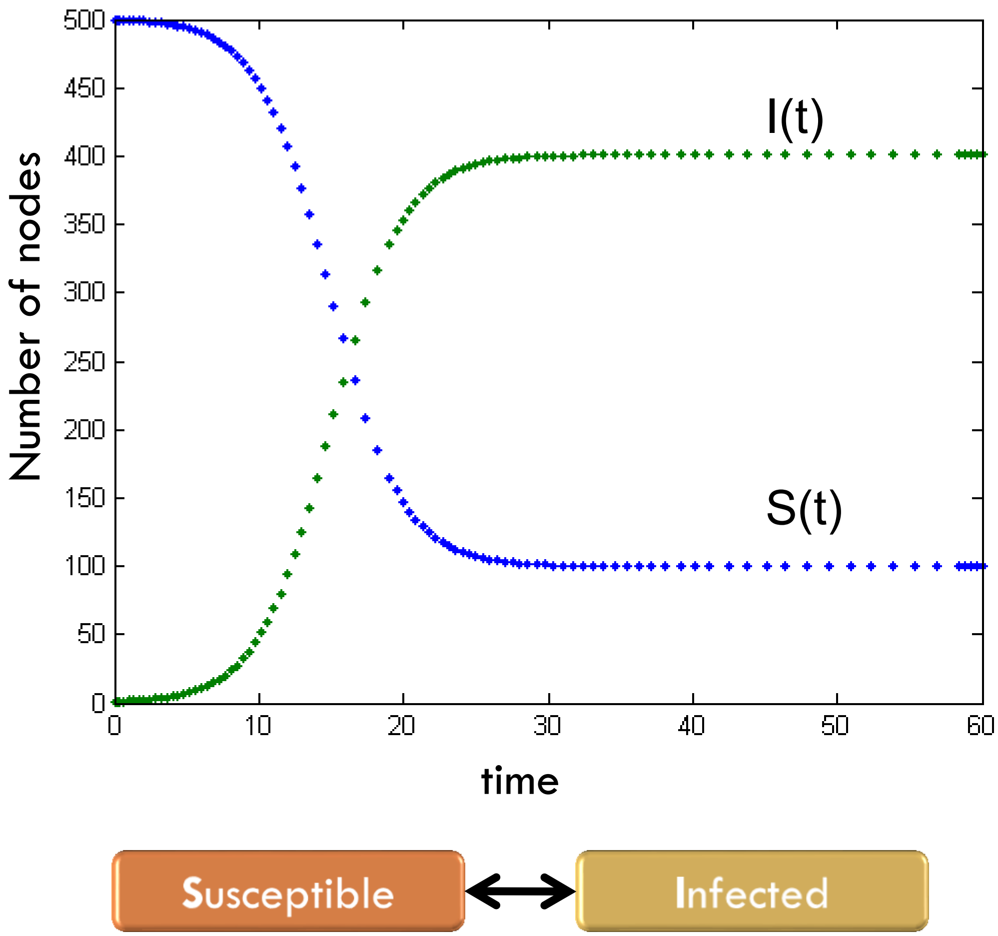


SIS Model

- **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?**

Epidemic Threshold in SIS Model

- **We have no epidemic if:**

Epidemic threshold

(Virus) Death rate

$$\boxed{\beta/\delta < \tau = 1/\lambda_{1,A}}$$

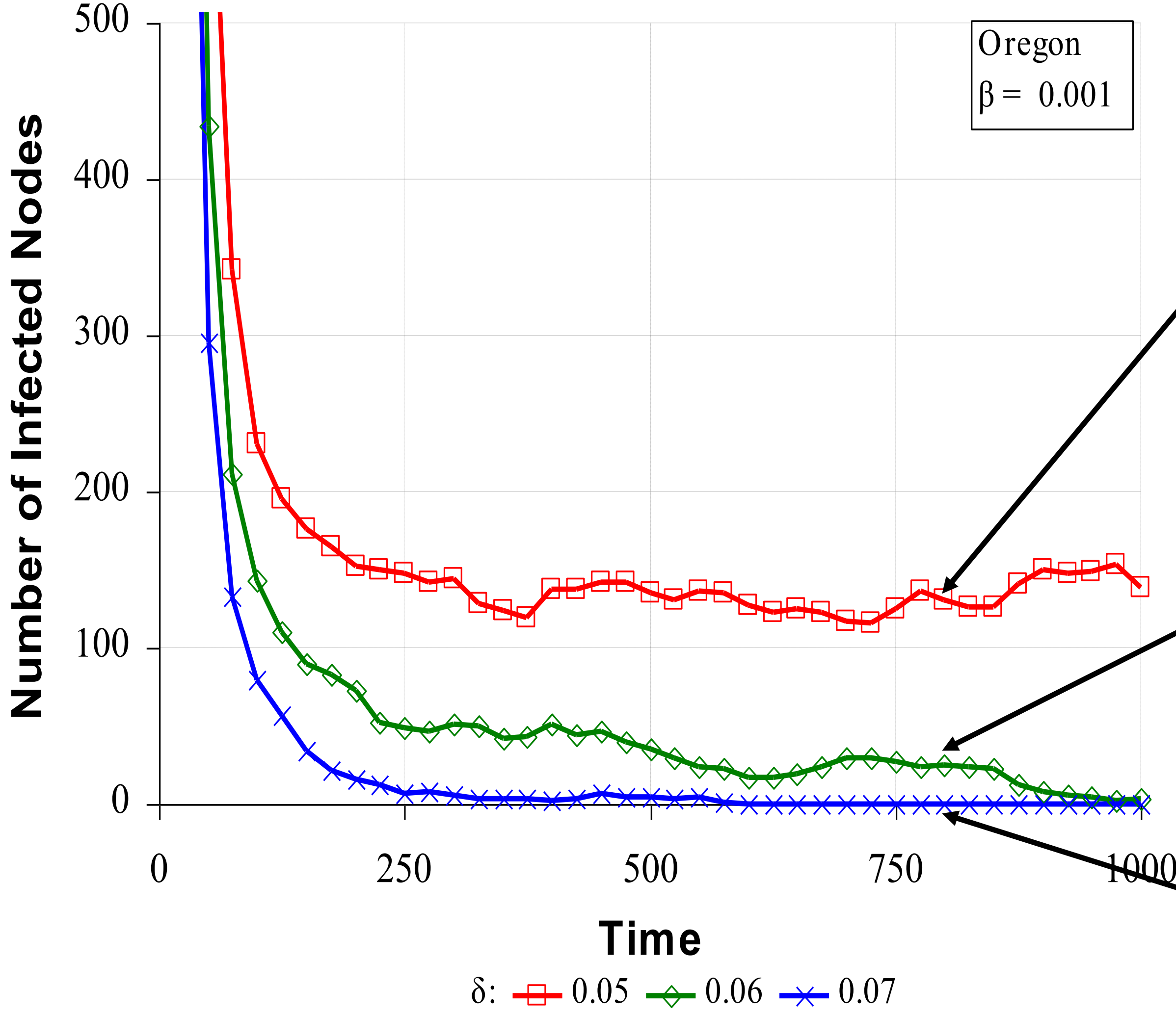
(Virus) Birth rate

largest eigenvalue of adj. matrix ***A***

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

Experiments (AS graph)

10,900 nodes and 31,180 edges



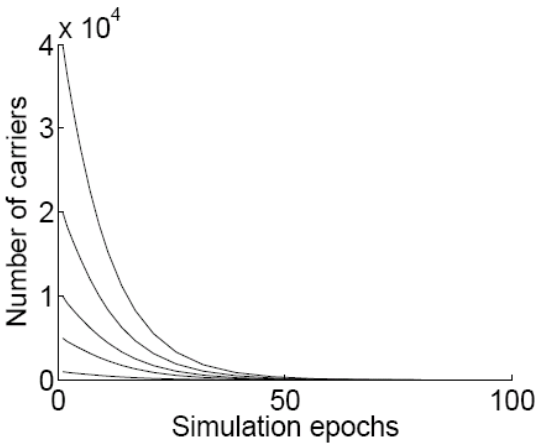
$\beta/\delta > \tau$
(above threshold)

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

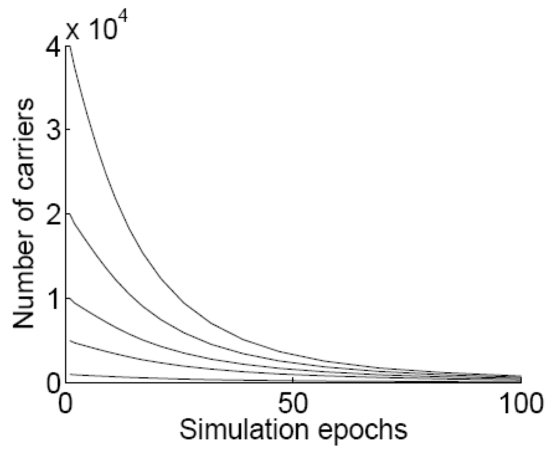
$\beta/\delta < \tau$
(below threshold)

Experiments

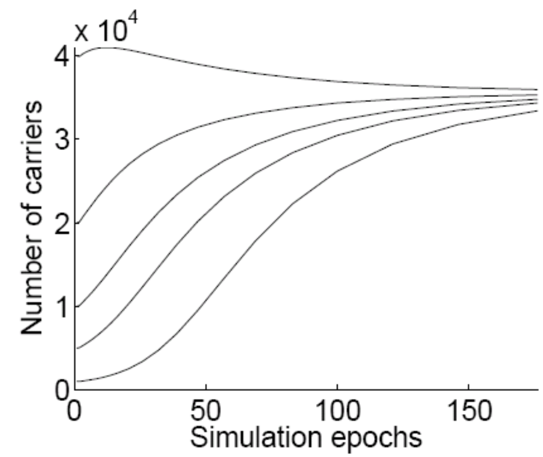
- **Does it matter how many people are initially infected?**



(a) Below the threshold, s=0.912



(b) At the threshold, s=1.003

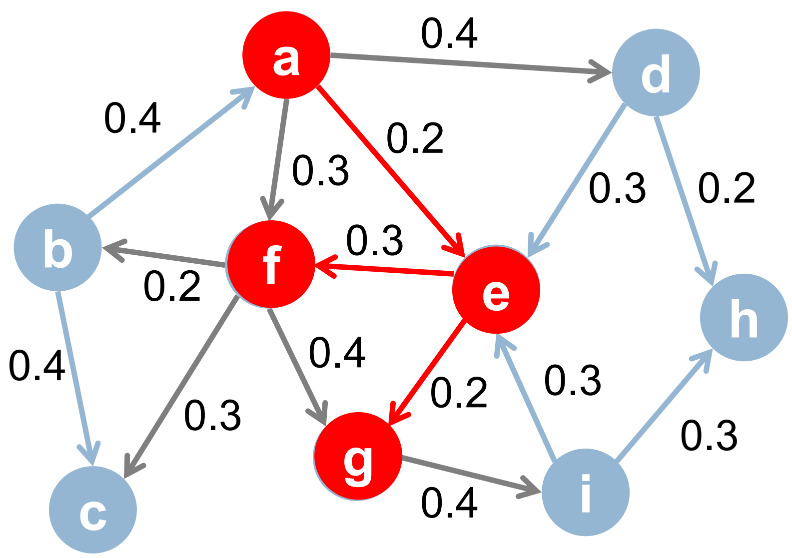


(c) Above the threshold, s=1.1

MODELS OF INFORMATION SPREAD

Independent Cascade Model

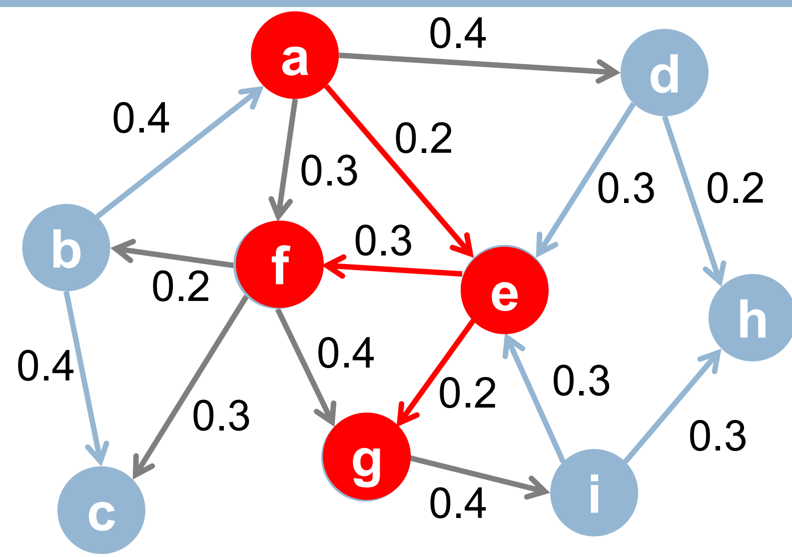
- **Initially some nodes *S* are active**
- Each edge (u,v) has probability (weight) p_{uv}



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

Independent Cascade Modal

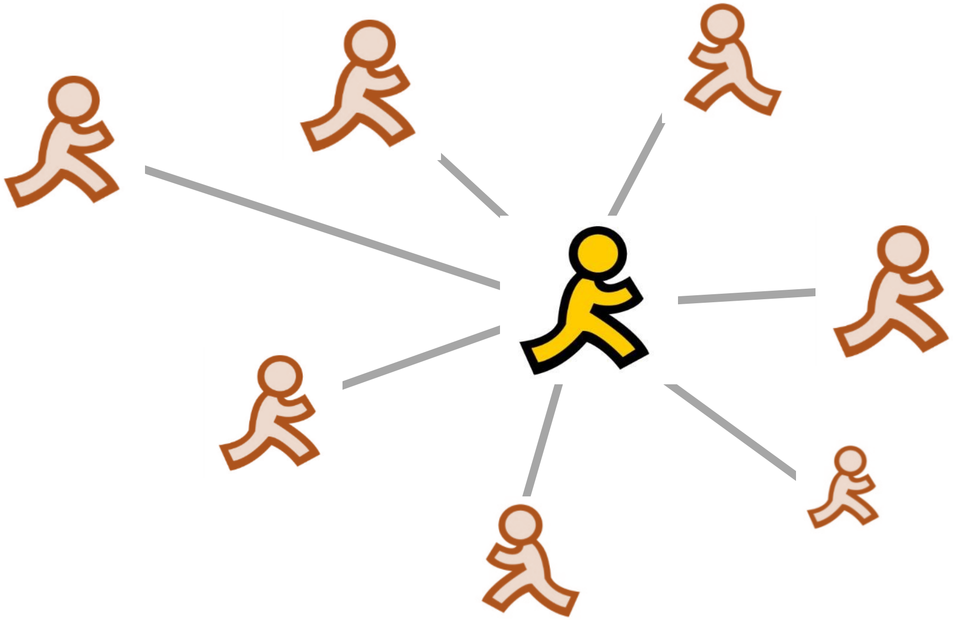
- **Independent cascade model is simple but requires many parameters!**
  - Estimating them from data is very hard [Goyal et al. 2010]



- **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?**

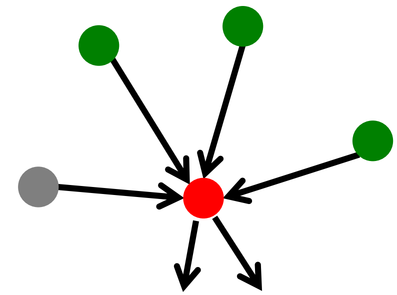
Exposures and Adoptions

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



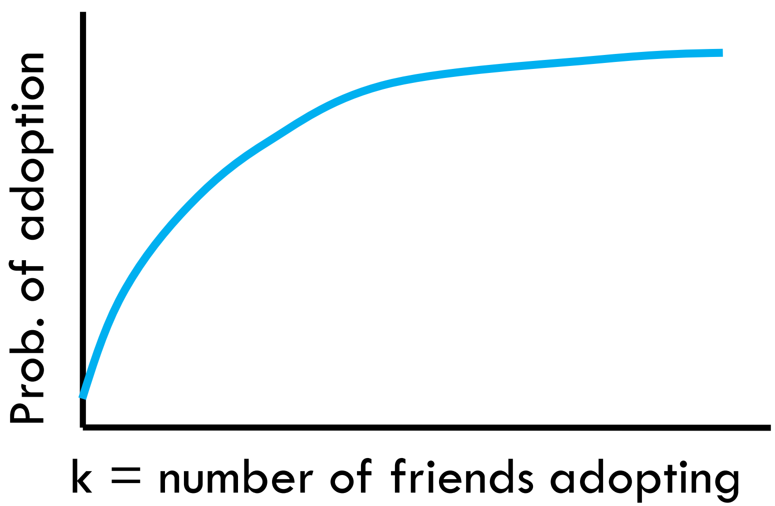
Exposure Curves

- <u>**Exposure curve:**</u>
  - **Probability of adopting new behavior depends on the number of friends who have already adopted**

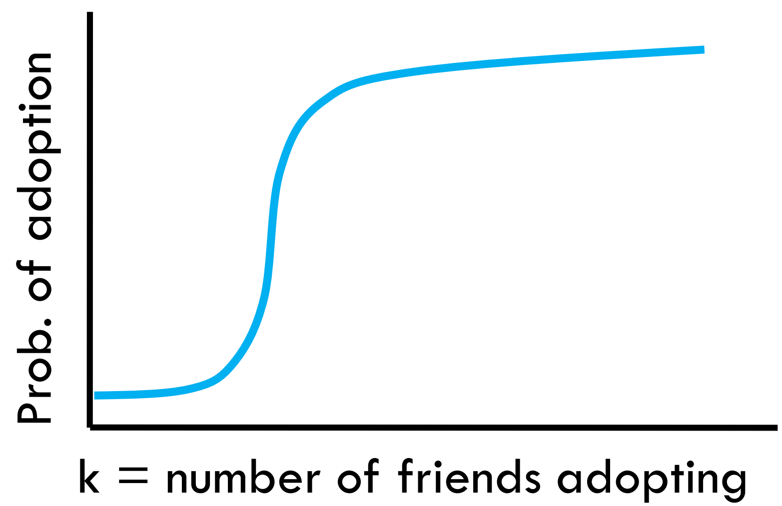


● … adopters

- **What's the dependence?**



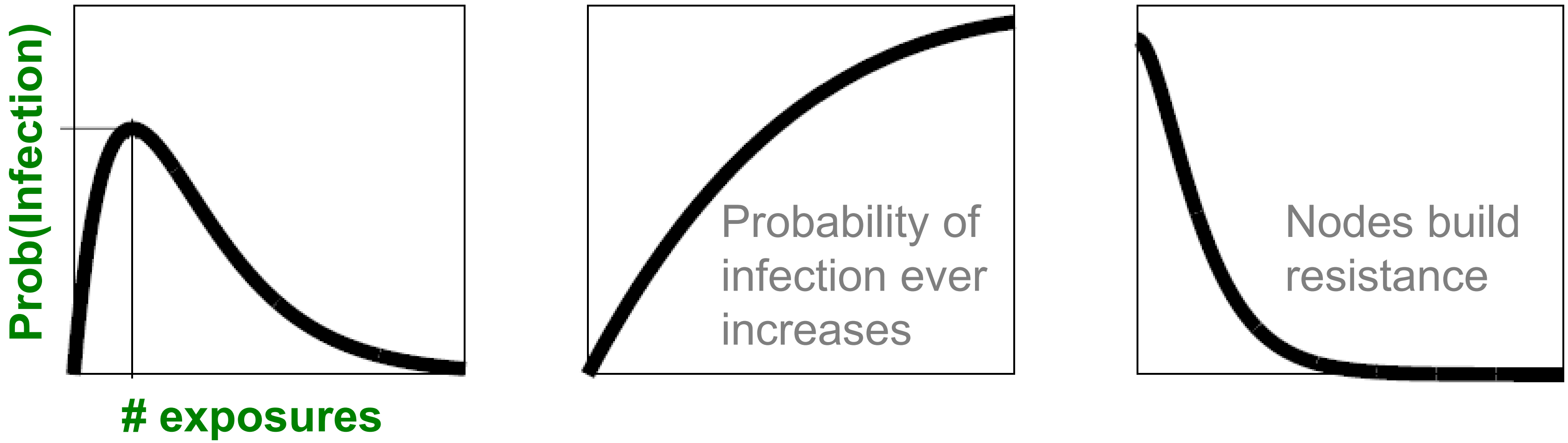
Diminishing returns:
Viruses, Information



Critical mass:
Decision making

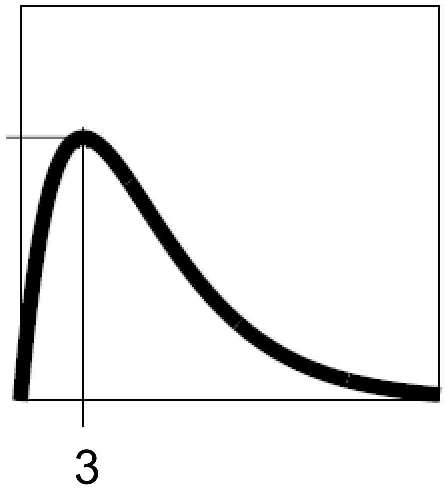
Exposure Curves

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



Example Application

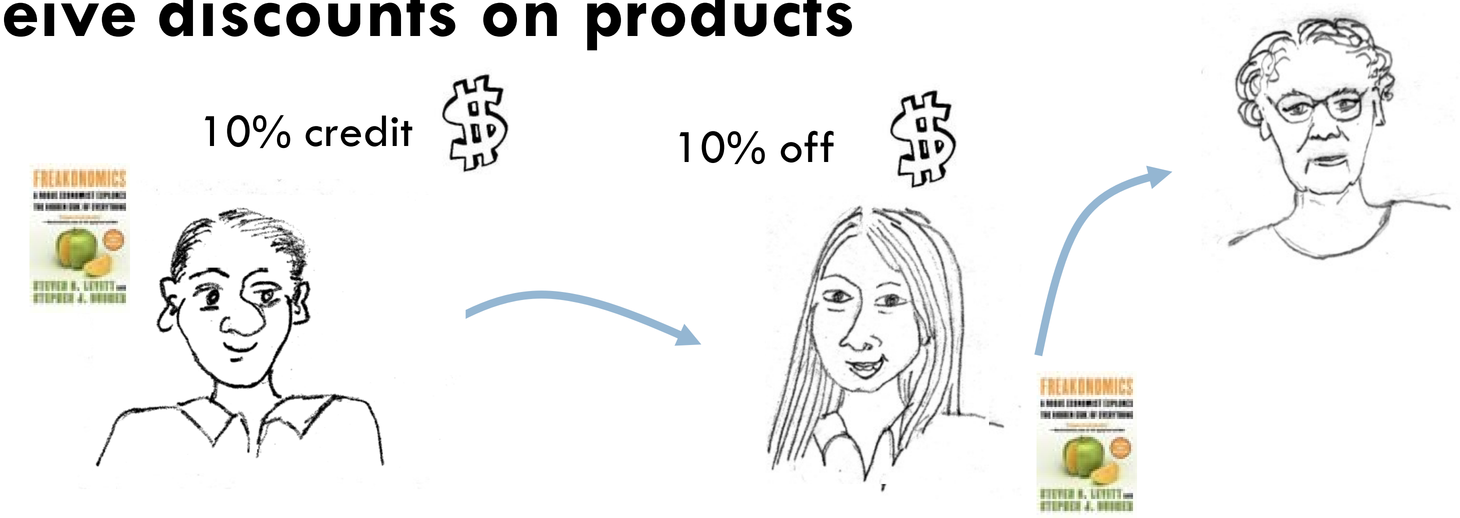
- **Marketing agency** would like you to adopt/buy product *X*
- They estimate the adoption curve



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

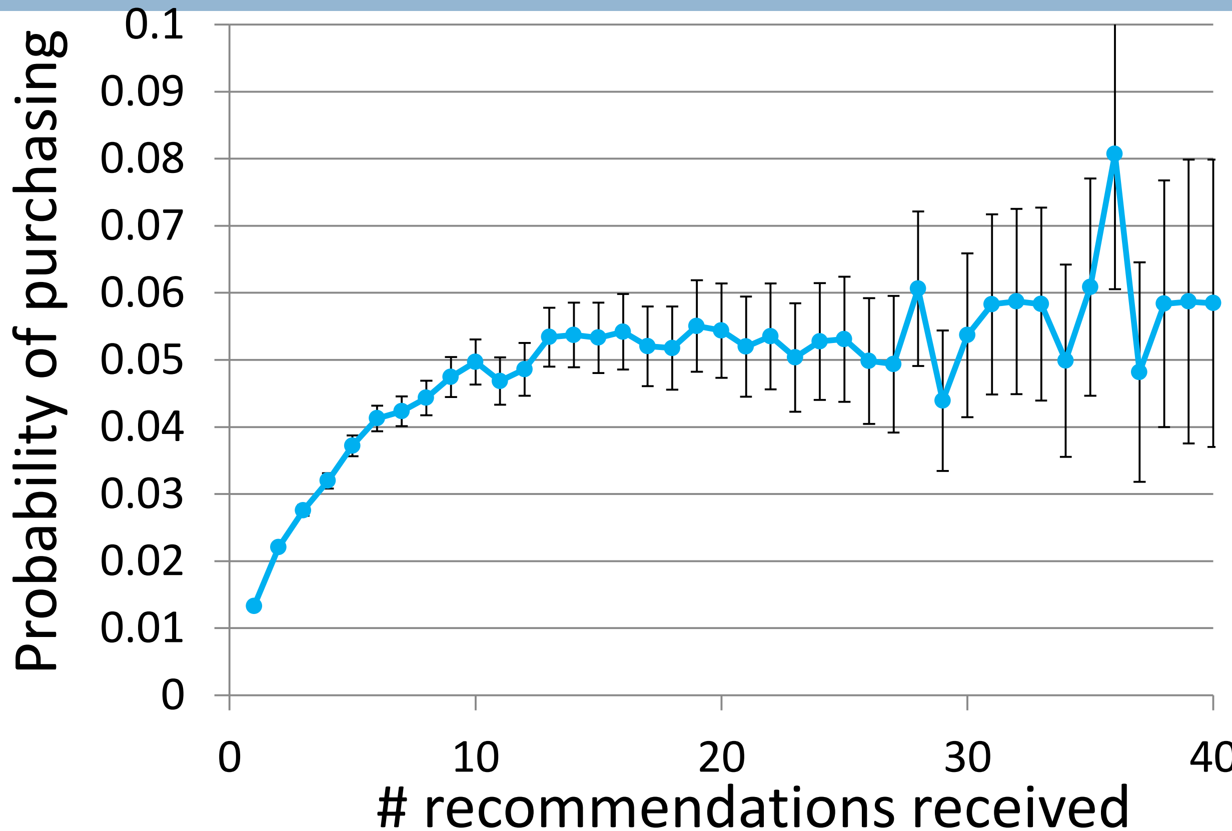
Diffusion in Viral Marketing

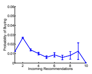
- Senders and followers of recommendations receive discounts on products



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

Exposure Curve: Validation



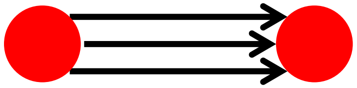


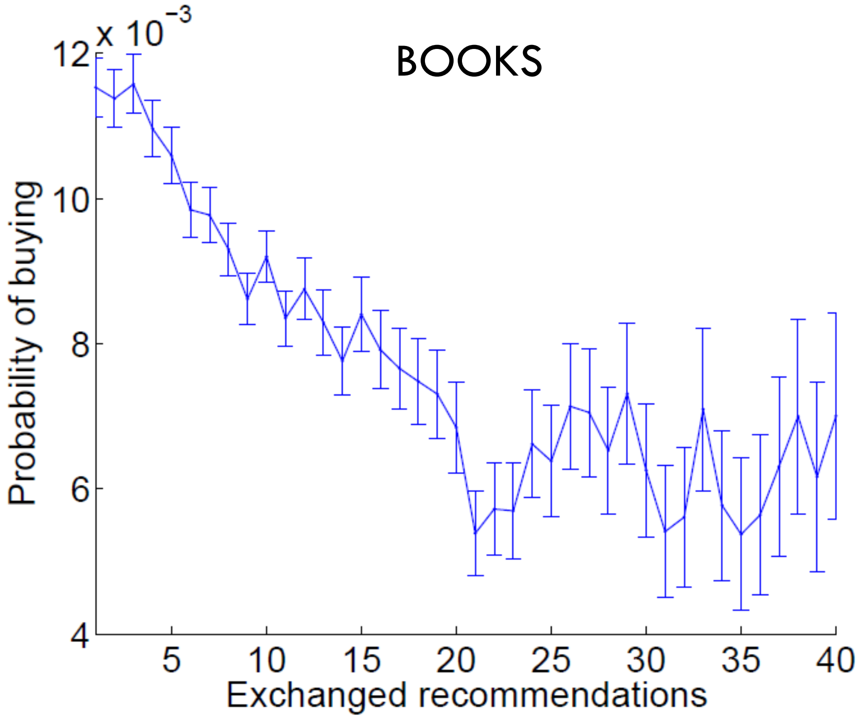
Books

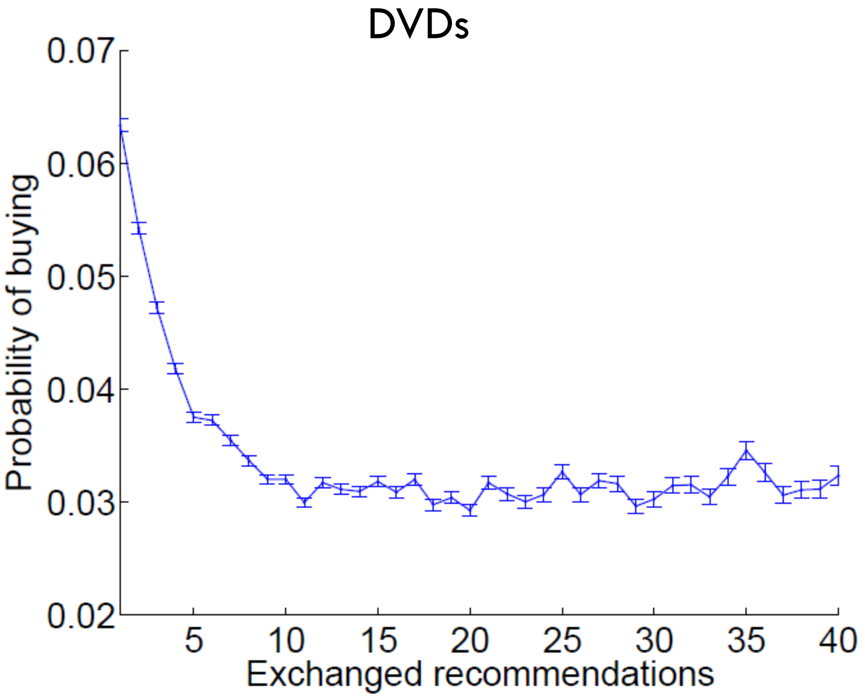
DVD recommendations
(8.2 million observations)

More Subtle Features

- What is the effectiveness of subsequent recommendations?

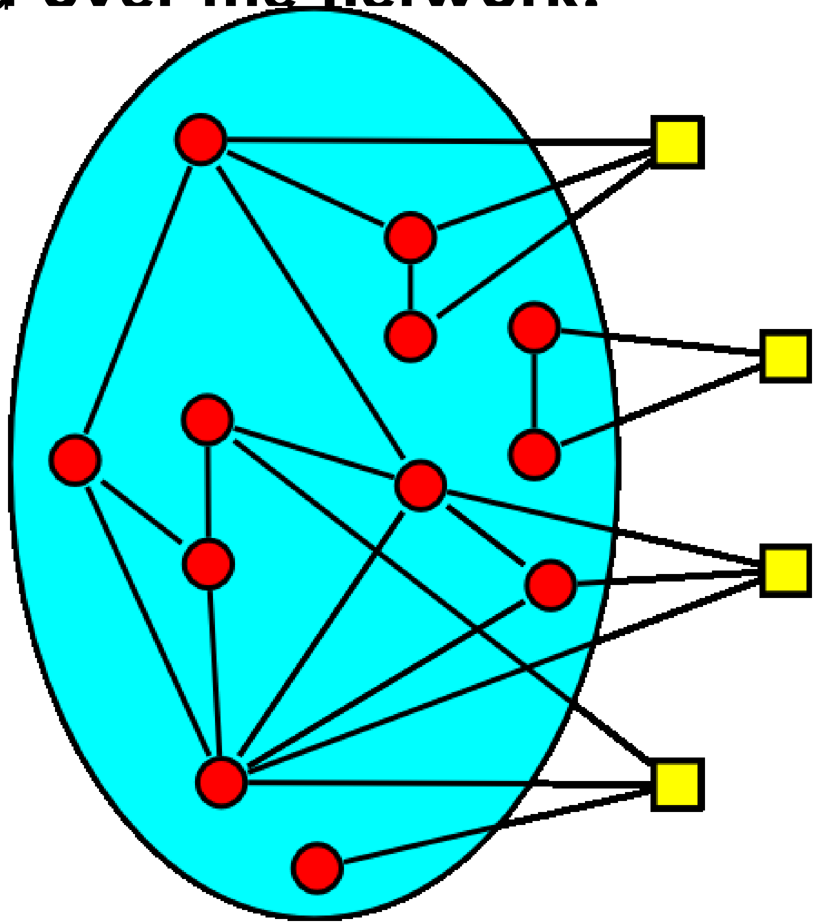






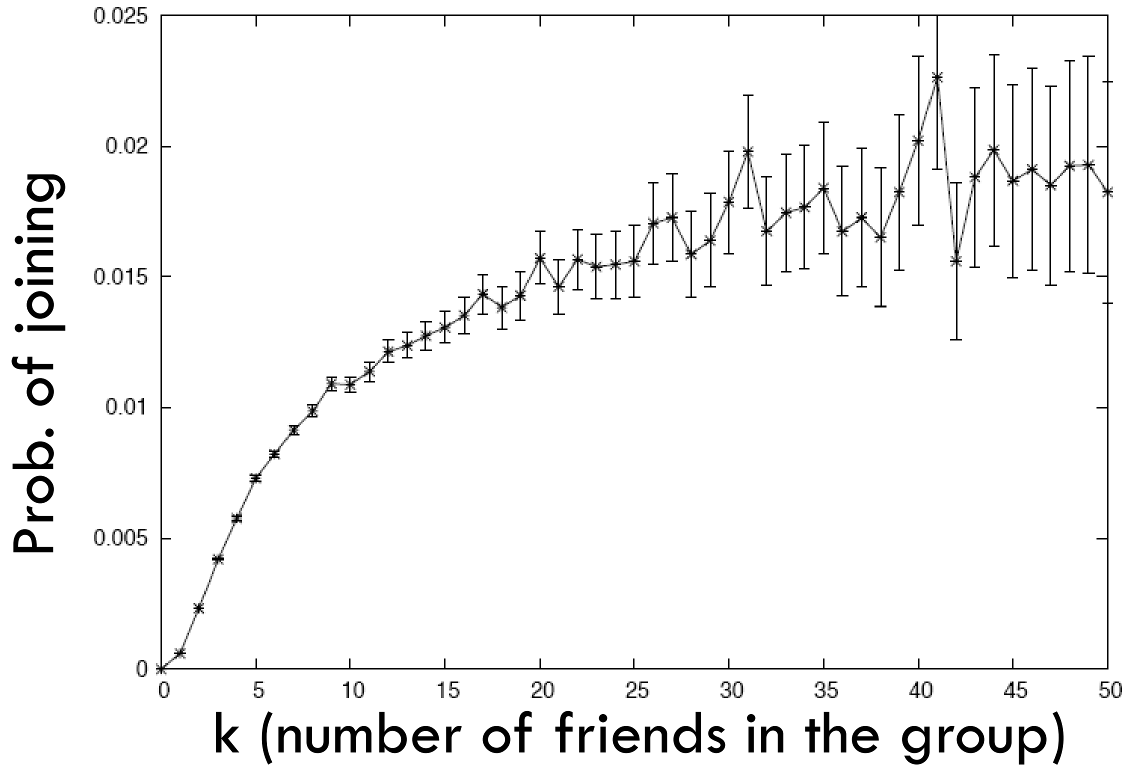
Exposure Curve: LiveJournal

- **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?



Exposure Curve: LiveJournal

- LiveJournal group membership

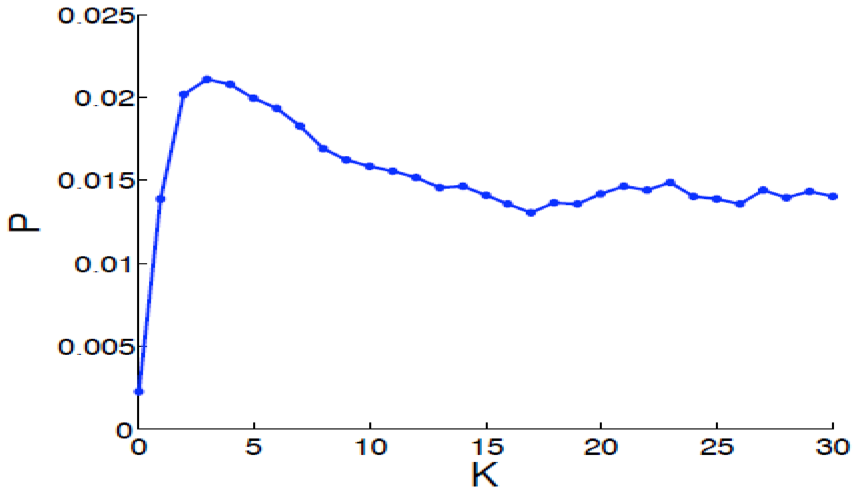


What are We Really Measuring?

- **For viral marketing:**
  - We see that node v receiving the *i*-th recommendation and then purchased the product
- **For groups:**
  - At time t we see the behavior of node v's friends
- **Good questions:**
  - When did v become aware of recommendations or friends' behavior?
  - When did it translate into a decision by v to act?
  - How long after this decision did v act?

Exposure Curve: Information

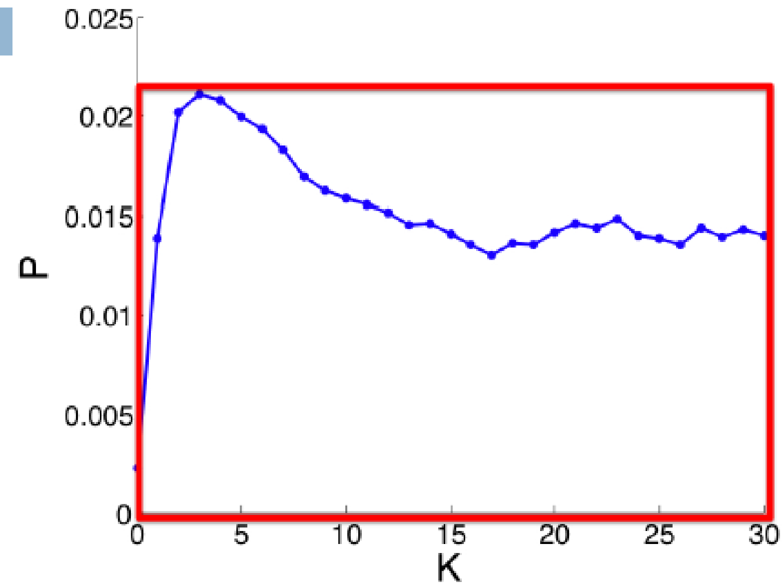
- **Twitter** [Romero et al. '11]
  - Aug '09 to Jan '10, 3B tweets, 60M users

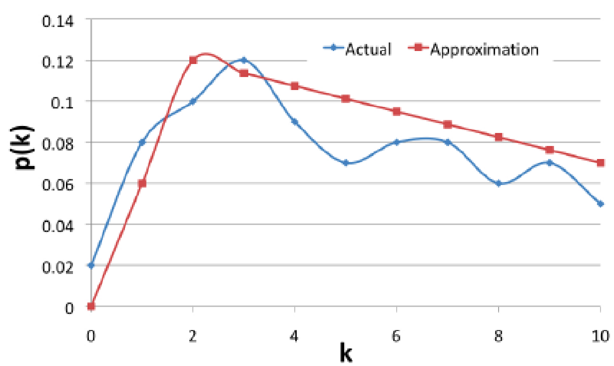


  - Avg. exposure curve for the top 500 hashtags
  - What are the most important aspects of the shape of exposure curves?
  - Curve reaches peak fast, decreases after!

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 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**

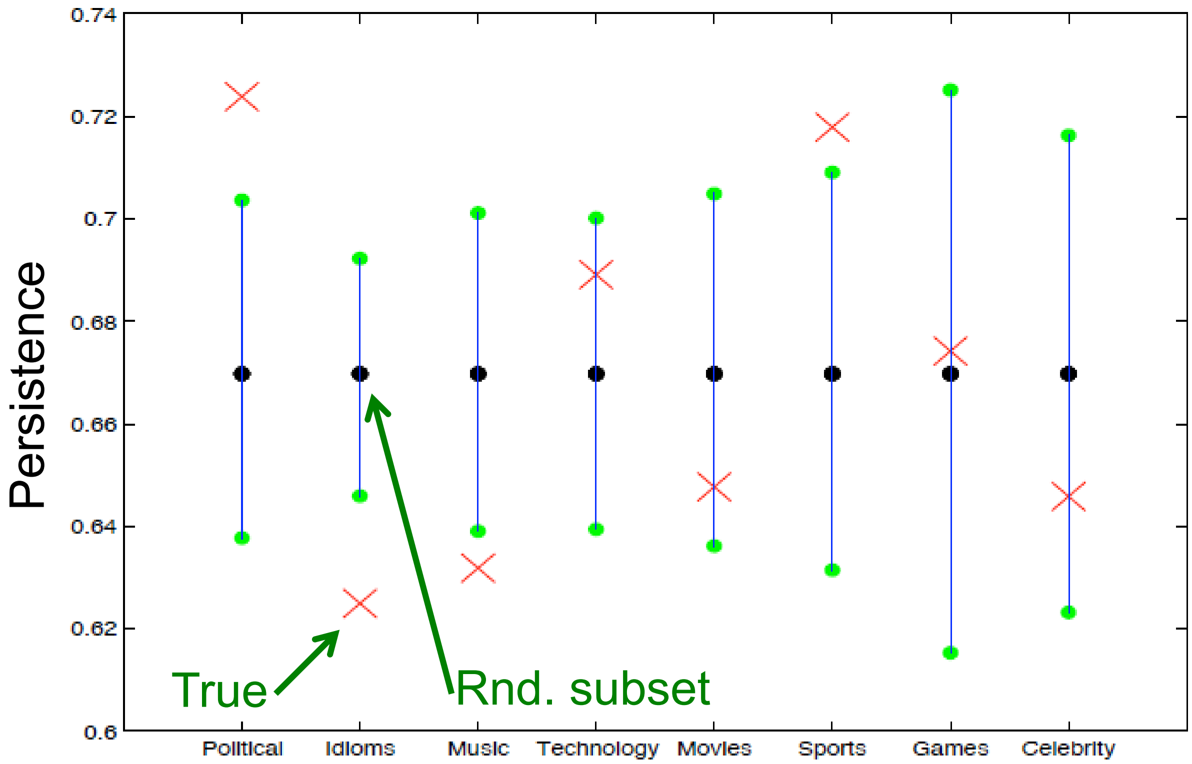




Exposure Curve: Persistence

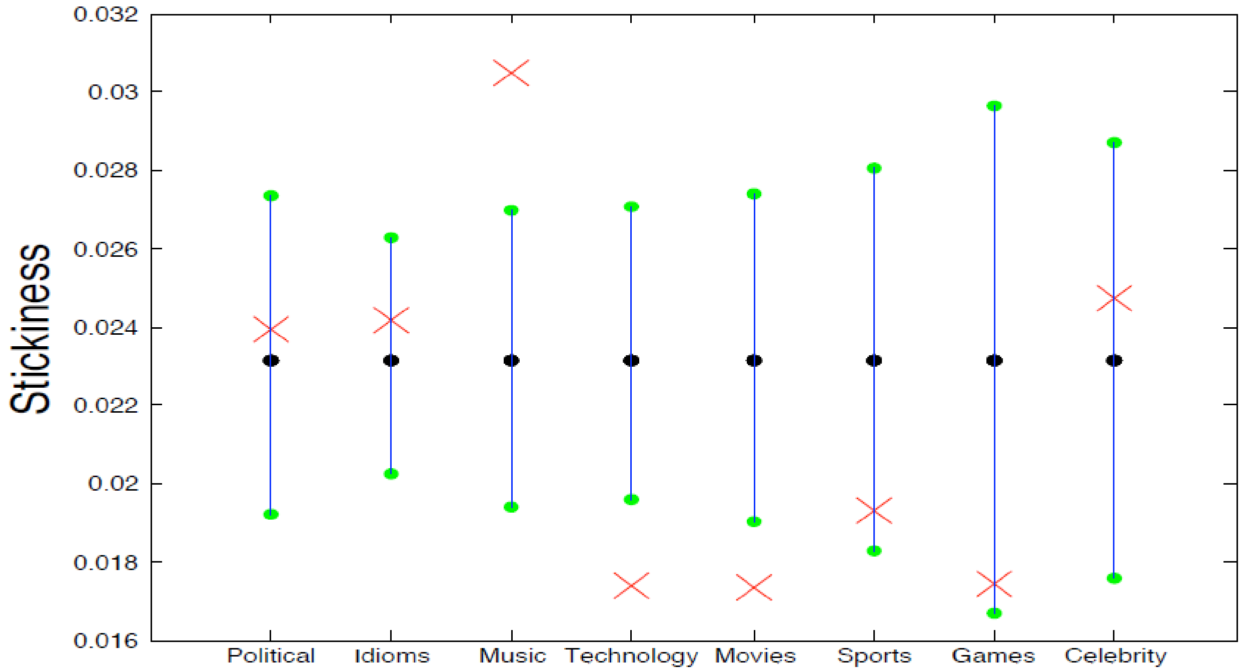
- Manually identify 8 broad categories with at least 20 HTs in each

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



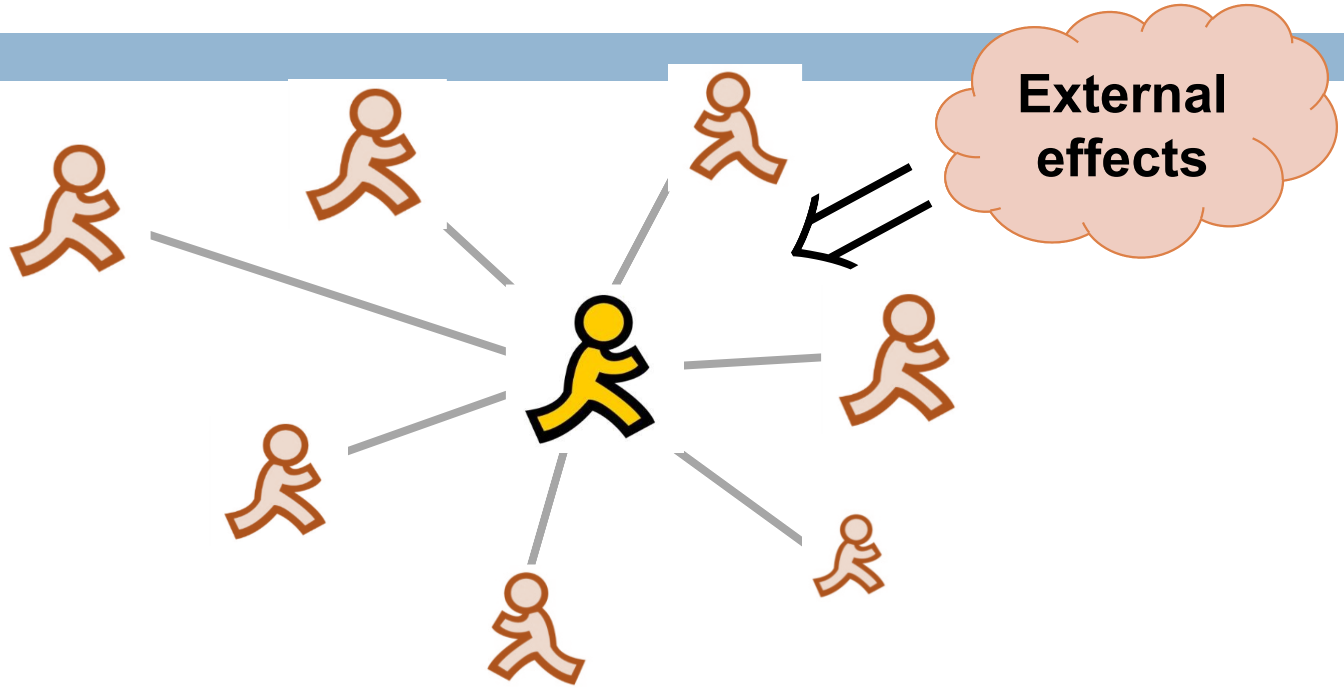
- Idioms and Music have lower persistence than that of a random subset of hashtags of the same size
- Politics and Sports have higher persistence than that of a random subset of hashtags of the same size

Exposure Curve: Stickiness



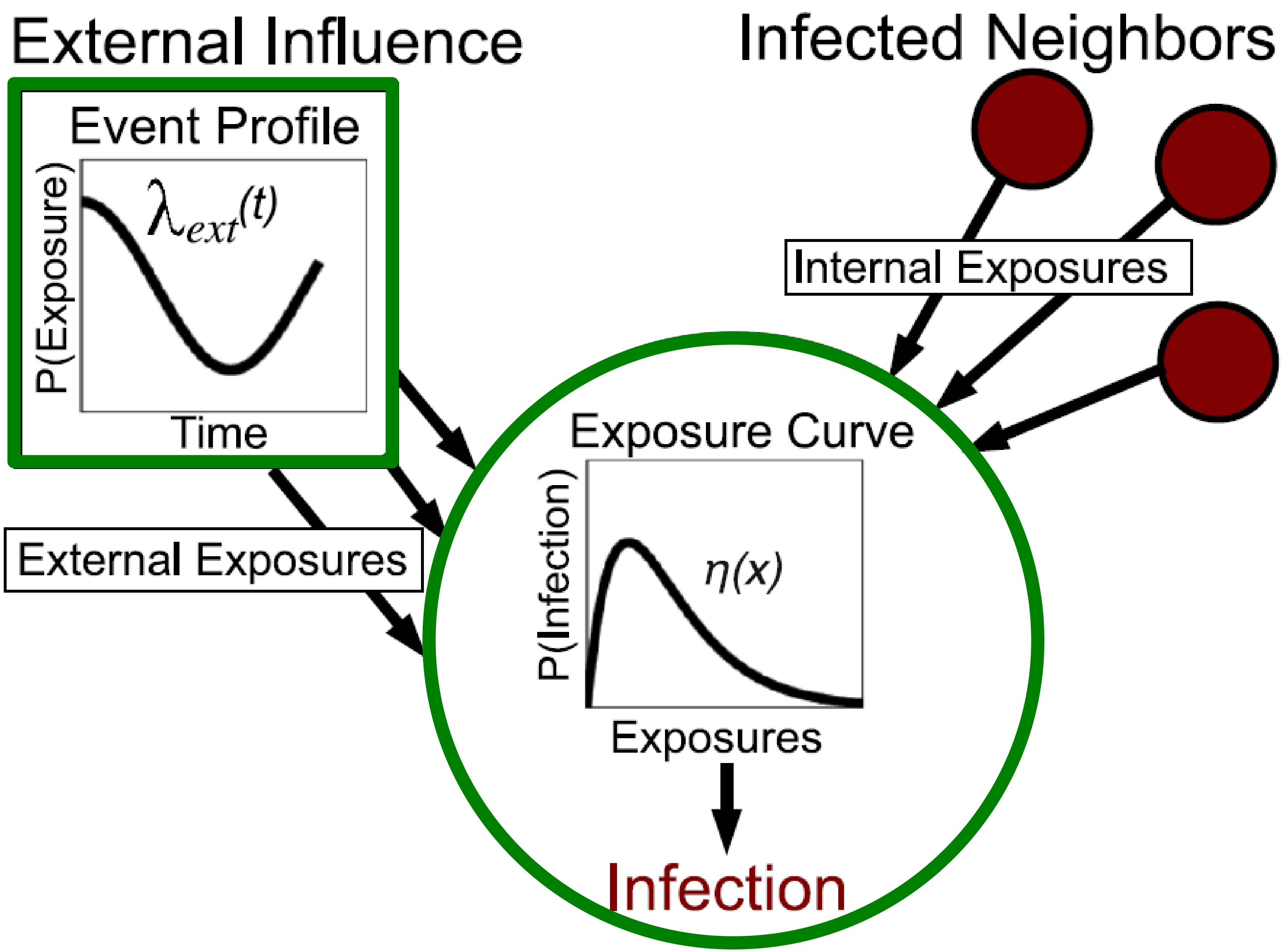
- Technology and Movies have lower stickiness than that of a random subset of hashtags
- Music has higher stickiness than that of a random subset of hashtags (of the same size)

Network & External Exposures

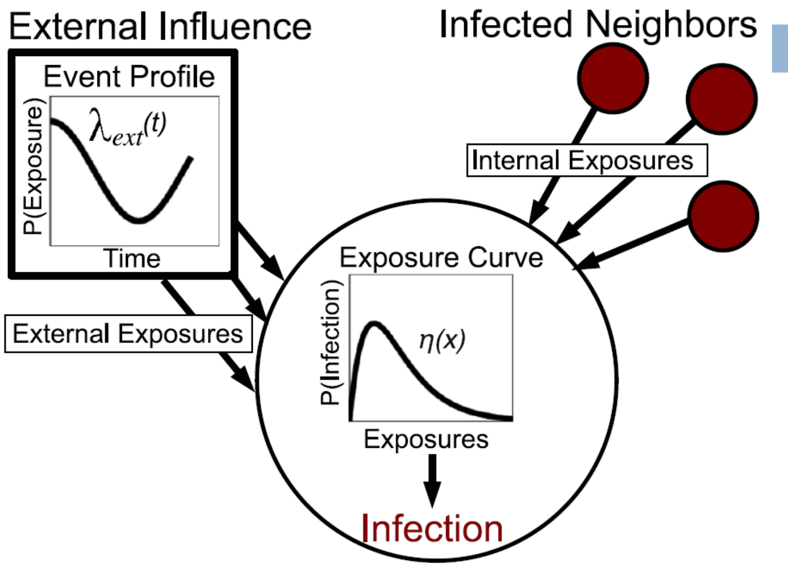


- **Two sources of exposures** *[Myers et al., KDD, 2012]*
  - Exposures from the network
  - External exposures

Putting it all together

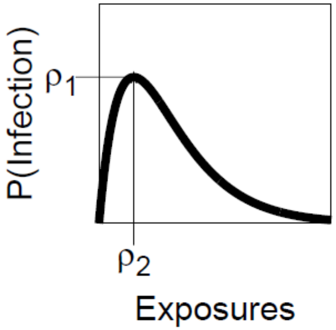


Model Inference Task



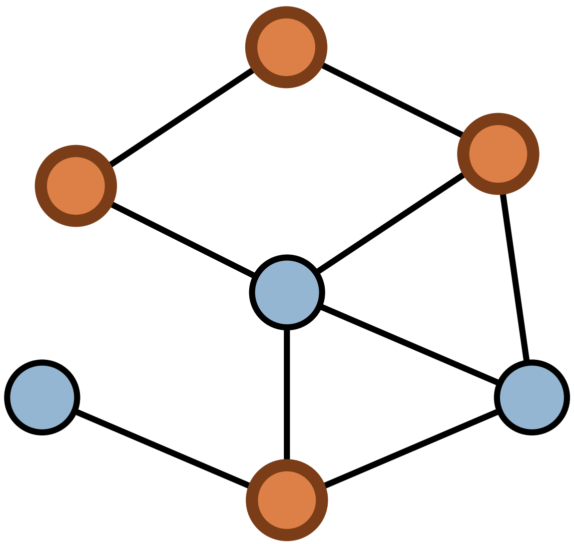
- **Given:**
  - Network G
  - A set of node adoption times *(u, t)* single piece of info
- **Goal: Infer**
  - **External event profile:**

    $\lambda_{ext}(t)$ … # external exposures over time
  - **Adoption curve:**



Experiment

- **In social networks people post links to interesting articles**
  - You hear about an article from a friend
  - You read the article and then post it
- **Data from Twitter**
  - *Complete* data from Jan 2011: 3 billion tweets
  - Trace the emergence of URLs
    - Label each URL by its topic



Results: Different Topics

- **Adoption of URLs across Twitter:**

	max P(k)	k at max P(k)	Duration (hours)	% Ext. Exposures
Politics (25)	0.0007 +/- 0.0001	4.59 +/- 0.76	51.24 +/- 16.66	47.38 +/- 6.12
World (824)	0.0013 +/- 0.0000	2.97 +/- 0.10	43.54 +/- 2.94	26.07 +/- 1.19
Entertain. (117)	0.0015 +/- 0.0002	3.52 +/- 0.28	89.89 +/- 16.13	17.87 +/- 2.51
Sports (24)	0.0010 +/- 0.0003	4.76 +/- 0.83	87.85 +/- 38.03	43.88 +/- 6.97
Health (81)	0.0016 +/- 0.0002	3.25 +/- 0.30	100.09 +/- 17.57	18.81 +/- 3.33
Tech. (226)	0.0013 +/- 0.0001	3.00 +/- 0.16	83.05 +/- 8.73	18.36 +/- 1.80
Business (298)	0.0015 +/- 0.0001	3.18 +/- 0.16	49.61 +/- 5.14	22.27 +/- 1.79
Science (106)	0.0012 +/- 0.0002	4.06 +/- 0.30	135.28 +/- 16.19	20.53 +/- 2.78
Travel (16)	0.0005 +/- 0.0001	2.33 +/- 0.29	151.73 +/- 39.70	39.99 +/- 6.60
Art (32)	0.0006 +/- 0.0001	5.26 +/- 0.66	188.55 +/- 48.17	27.54 +/- 5.30
Edu. (31)	0.0009 +/- 0.0001	3.77 +/- 0.51	130.53 +/- 38.63	21.45 +/- 6.40

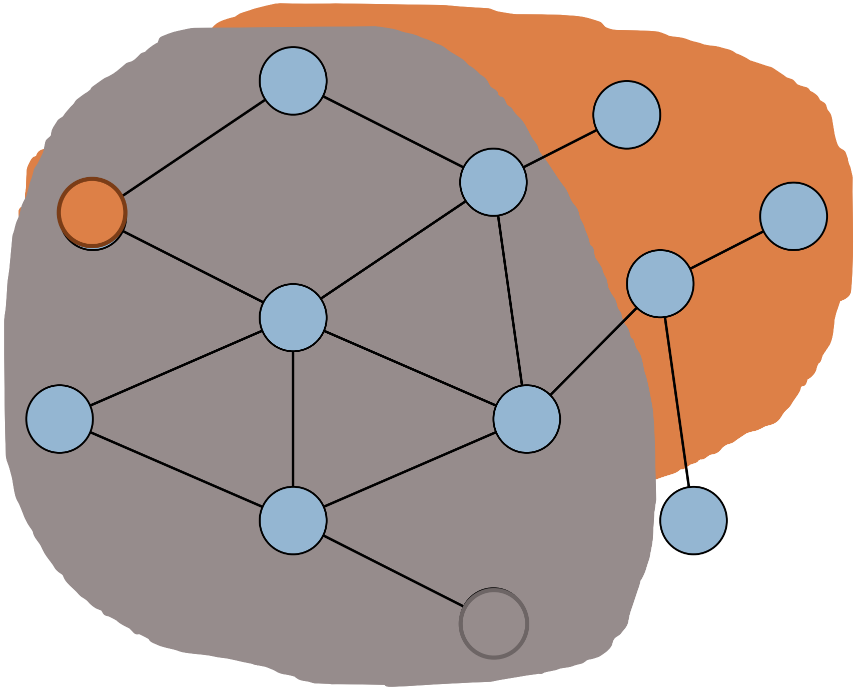
- More in *Myers et al., KDD, 2012*

MODELING INTERACTIONS BETWEEN CONTAGIONS

Interactions

- So far we considered pieces of information as **independently** propagating

- **Do pieces of information interact?**
  - Does being exposed to **blue** change the probability of talking about **red**?

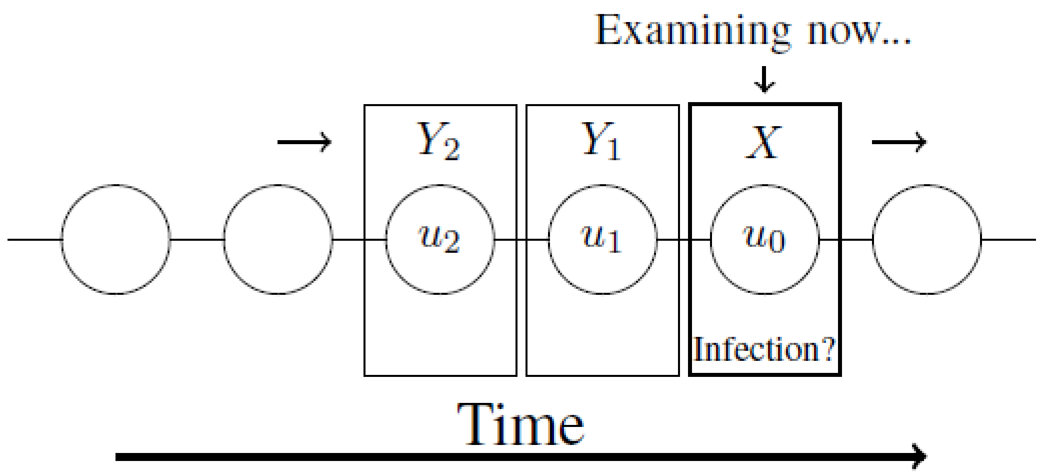


Modeling Interactions

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

Modeling Interactions

- **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?



$$P(\text{post } X \mid \text{exposed to } X, Y_1, Y_2, Y_3) = ?$$

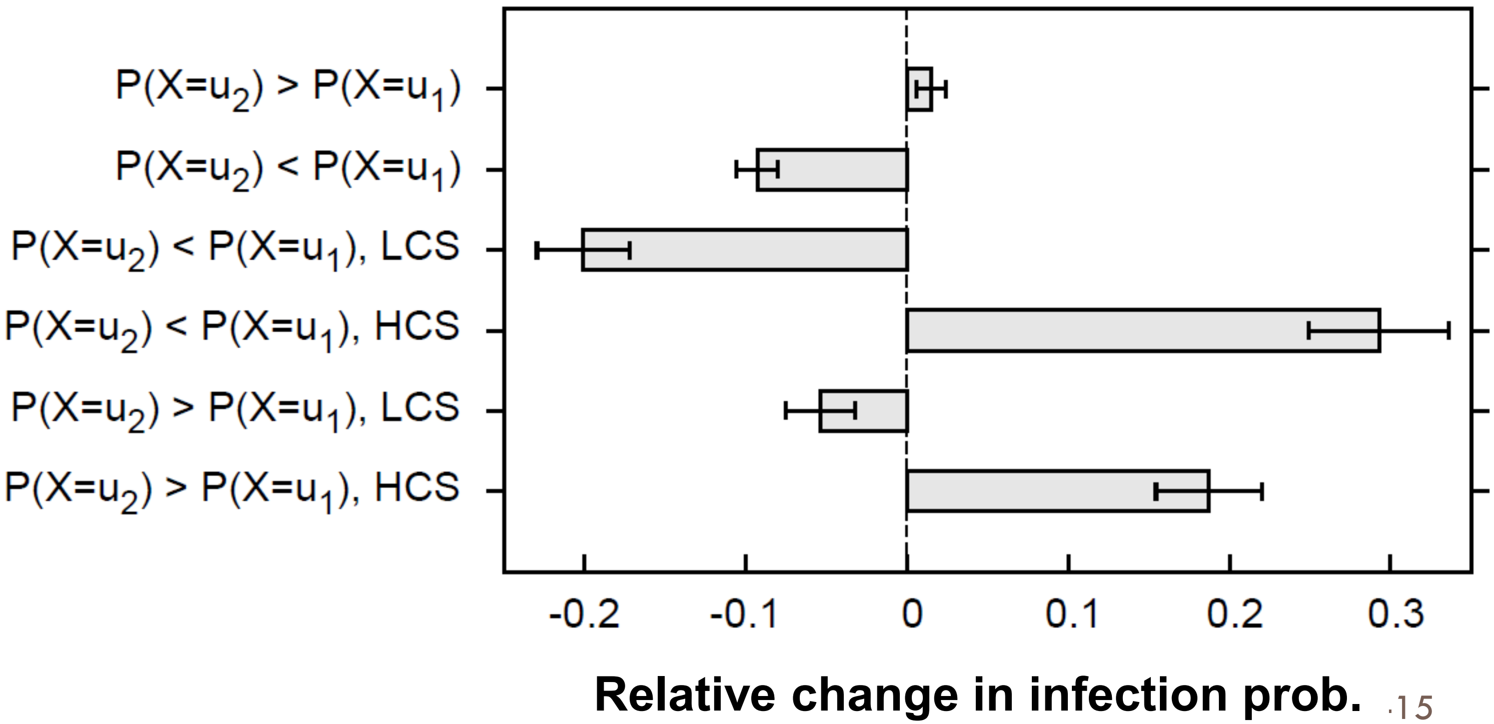
Dataset: Twitter

- **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
  - Train on 90% of the data, test on 10%
- **Baselines:**
  - **Infection Probability (IP):** $P(X = u_i | Y_k = u_j) = \; = P(X = u_i)$
  - **IP + Node bias (NB):** $= P(X = u_i) + \gamma_n$
  - **Exposure curve (EC):** $= P(X \mid \#\, times\ exposed\ to\ X)$

How to Tweets Interact?

- **How $P(\text{post } u_2 \mid \text{exp. } u_1)$ 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

.15