# Online Social Networks and Media

## Diffusion:

Cascading Behavior in Networks

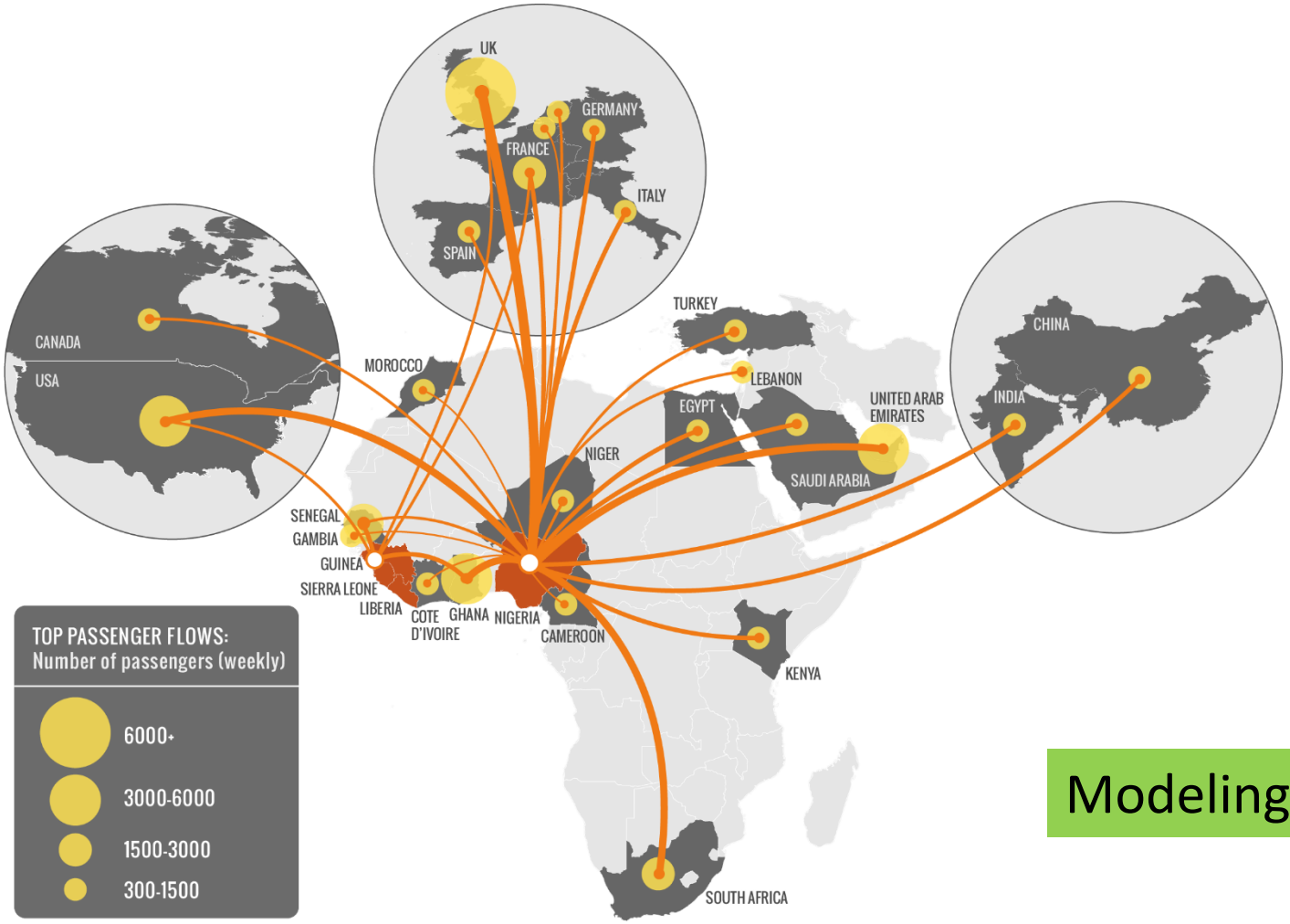
Epidemic Spread

Influence Maximization

# Introduction

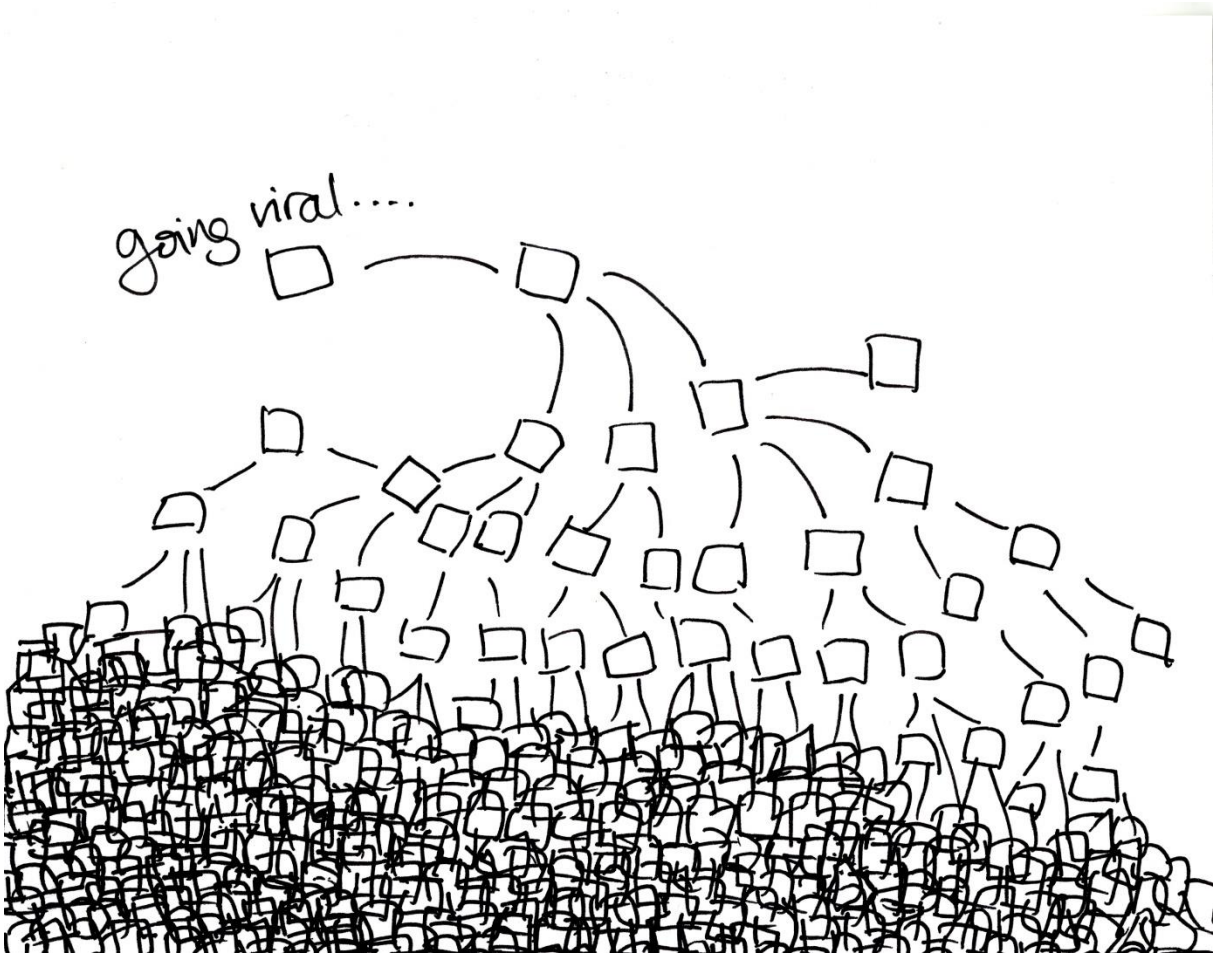
Diffusion: process by which a piece of information is spread and reaches individuals through interactions

# Why do we care?



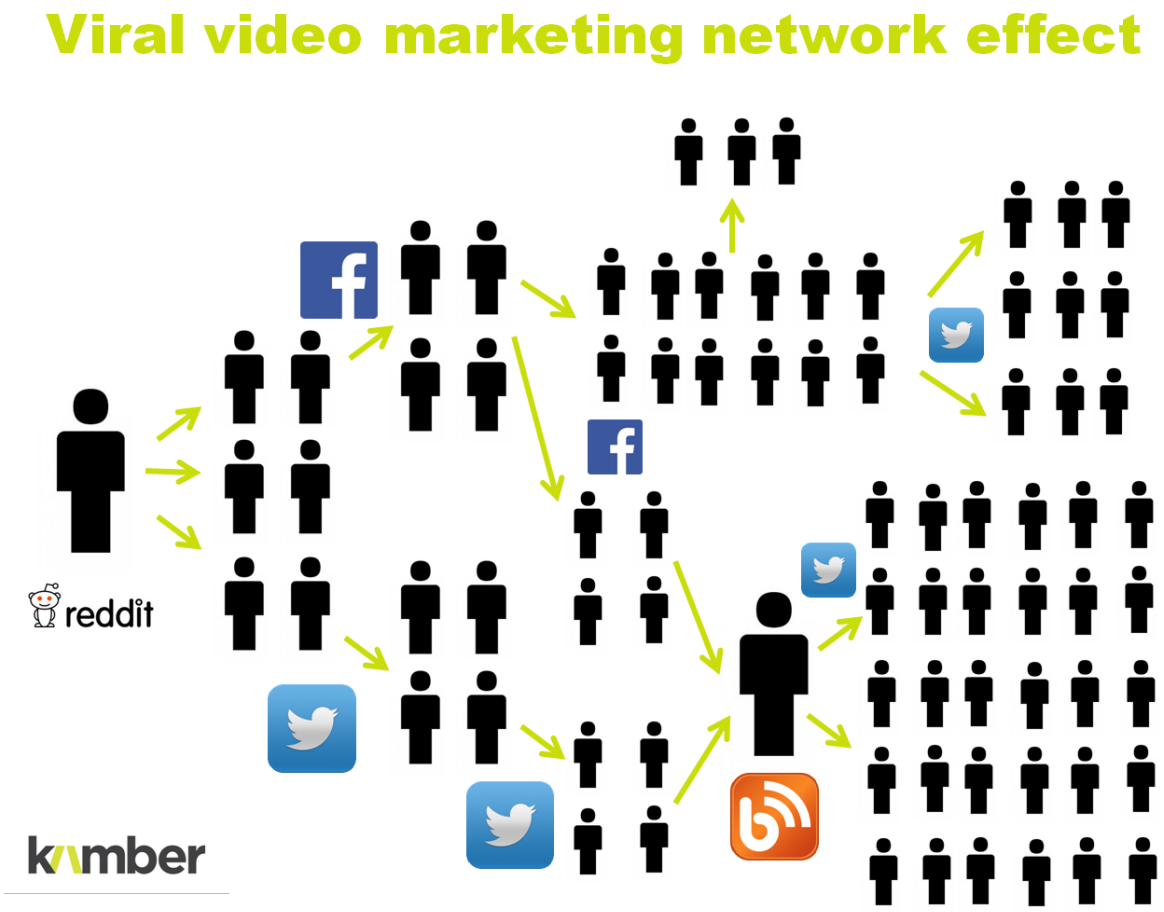

Modeling epidemics

# Why do we care?




Viral marketing

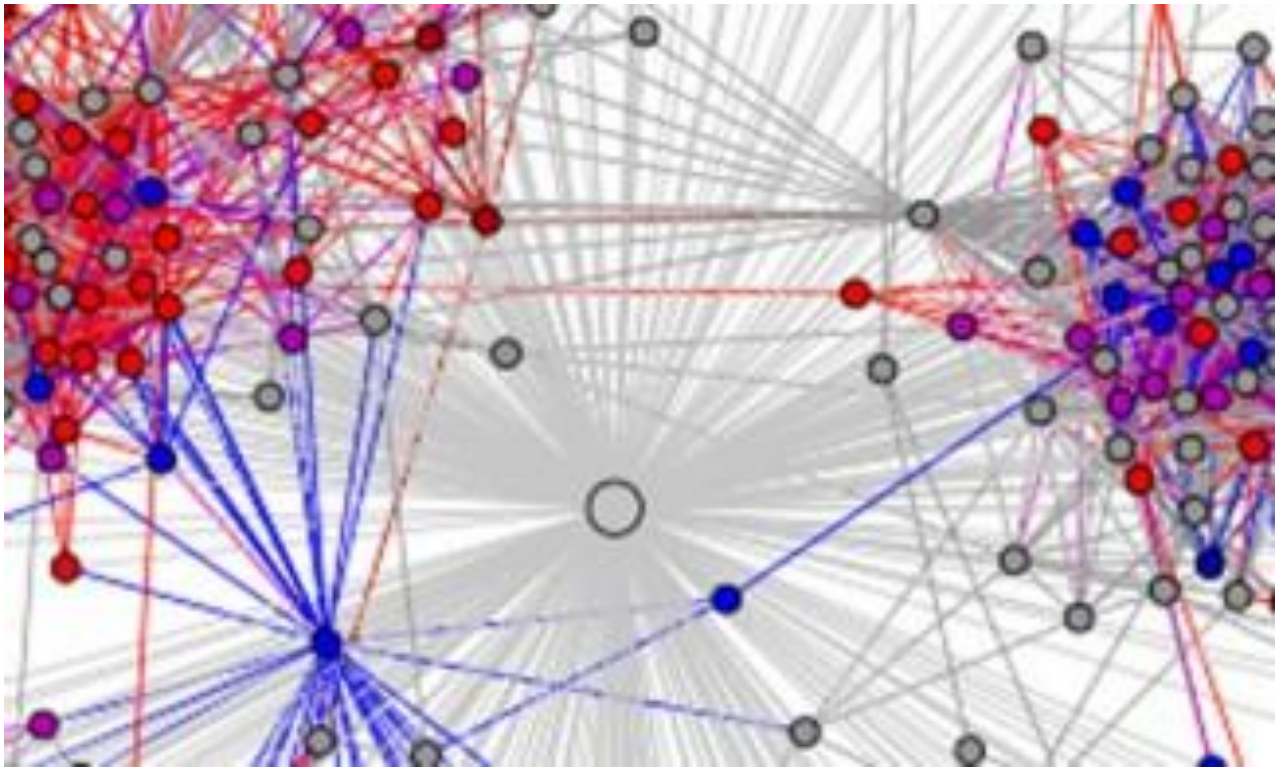
# Why do we care?




Viral marketing

# Why do we care?

Spread of innovation



# Outline

- Cascading behavior
- Epidemic models
- Influence maximization

# CASCADING BEHAVIOR IN NETWORKS

# Innovation Diffusion in Networks

How new behaviors, practices, opinions and technologies *spread* from person to person through a social network as people *influence* their friends to adopt new ideas

Why? Two classes of rational reasons:

- *Direct-Benefit Effect*: there are direct *payoffs* from copying the decisions of others (relative advantage)
  - E.g., Phone becomes more useful if more people use it
- *Informational effect*

# Informational Effect

## Informational effect:

choices made by others can provide indirect information about what they know (e.g., choosing restaurants)

## Informational social influence (social proof):

a psychological phenomenon where people *assume the actions of others* in an attempt to reflect *correct behavior* for a given situation

- prominent in *ambiguous social situations* where people are unable to determine the appropriate mode of behavior
- driven by the assumption that *surrounding people possess more knowledge* about the situation

# Spread of Innovation

Diffusion of innovation

Old studies

(mainly informational effect):

- Adoption of hybrid seed corn among farmers in Iowa
- Adoption of tetracycline by physicians in US

(mainly direct benefit)

- Technology (phone, email, etc)

Basic observations:

- High risk but high benefit
- Characteristics of *early adopters*
- Decisions made in the context of *social structure*

# Spread of Innovation

Common principles:

✓ *Complexity* of people to understand and implement

✓ *Observability*, so that people can become aware that others are using it

✓ *Trialability*, so that people can mitigate its risks by adopting it gradually and incrementally

✓ *Compatibility* with the social system that is entering (homophily as a barrier?)

# A Direct-Benefit Model

An *individual* level model of *direct-benefit effects* in networks due to S. Morris

The benefits of adopting a new behavior increase as more and more of the social network neighbors adopt it

## A Coordination Game

Two players (nodes), *u* and *w* linked by an edge
Two possible behaviors (strategies): A and B

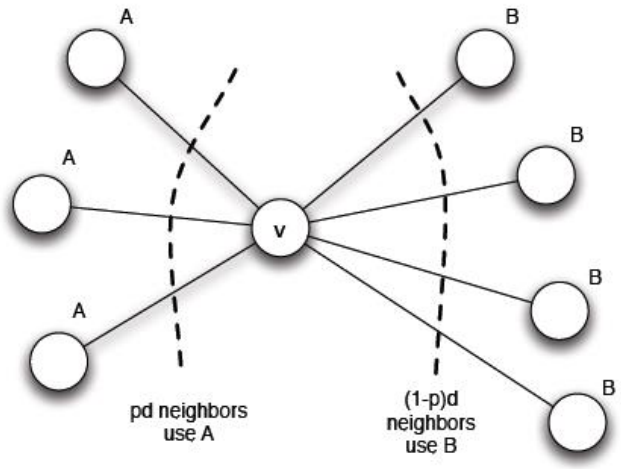
- If both *u* and *w* adapt A, get payoff $a > 0$
- If both *u* and *w* adapt B, get payoff $b > 0$
- If opposite behaviors, each gets a payoff 0

| | | $w$: $A$ | $w$: $B$ |
|---|---|---|---|
| $v$ | $A$ | $a, a$ | $0, 0$ |
| $v$ | $B$ | $0, 0$ | $b, b$ |

# Modeling Diffusion through a Network

*u* plays a copy of the game with each of its neighbors, its payoff is the *sum* of the payoffs in the games played on each edge

- Say *p* of the *d* neighbors of *u* neighbors adopt A and the other *(1- p)* adopt B, what should *u* do to maximize its payoff?




Threshold *q* *for preferring A* *(at least q of the neighbors follow A)*

$$q = b/(a+b)$$

Two obvious equilibria, which ones?

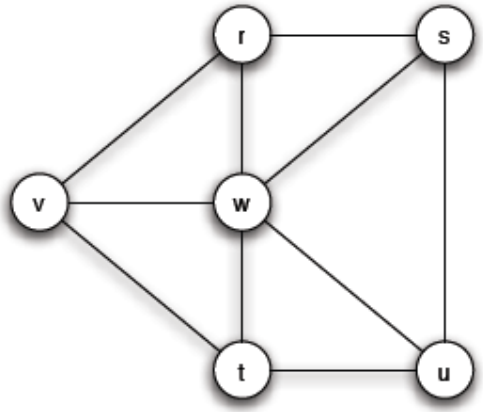
# Modeling Diffusion through a Network: Cascading Behavior

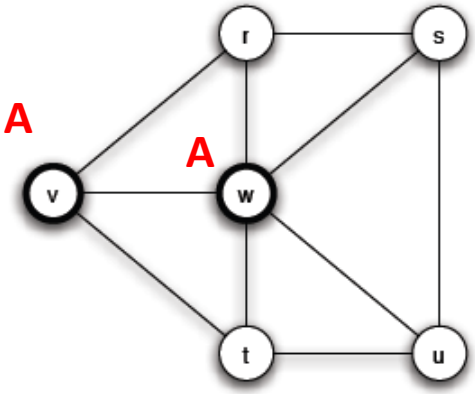
Suppose that initially everyone is using B as a default behavior
A small set of "initial adopters" decide to use A

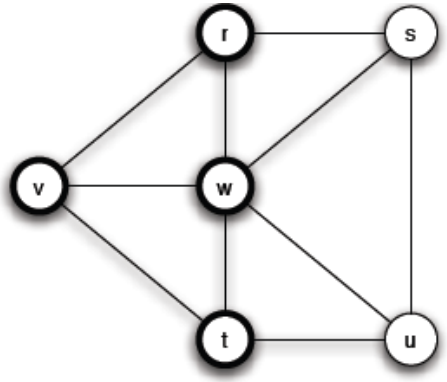
- When will this result in everyone eventually switching to A?
- If this does not happen, what causes the spread of A to stop?

# Modeling Diffusion through a Network: Cascading Behavior

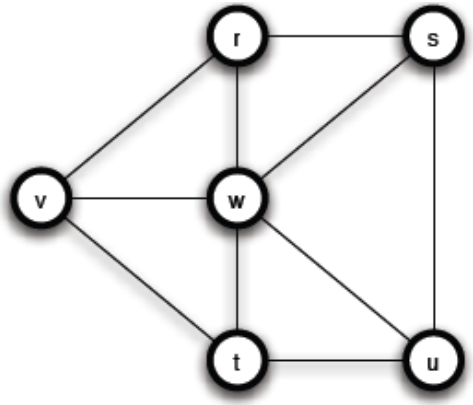
$a = 3, b = 2, q = 2/5$




**Step 1**



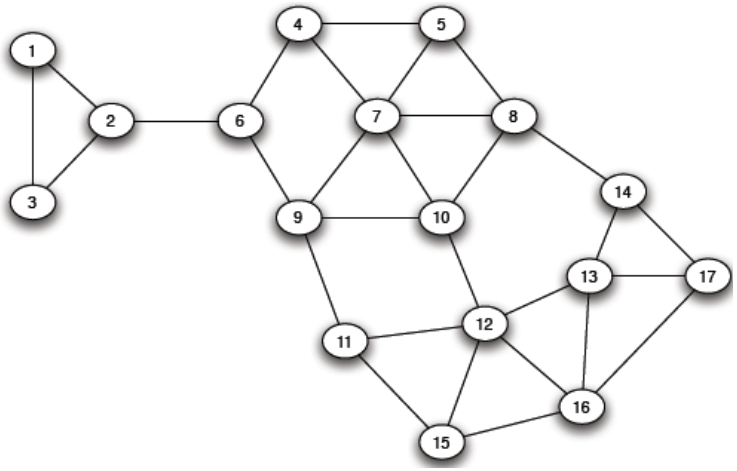

**Step 2**

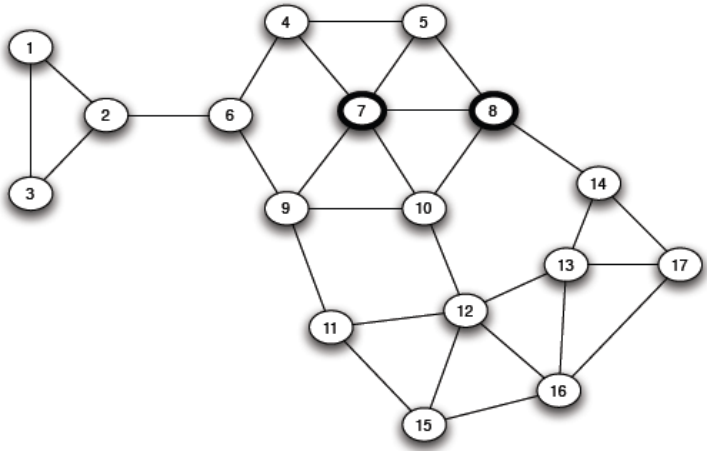
Chain reaction of switches to B -> A

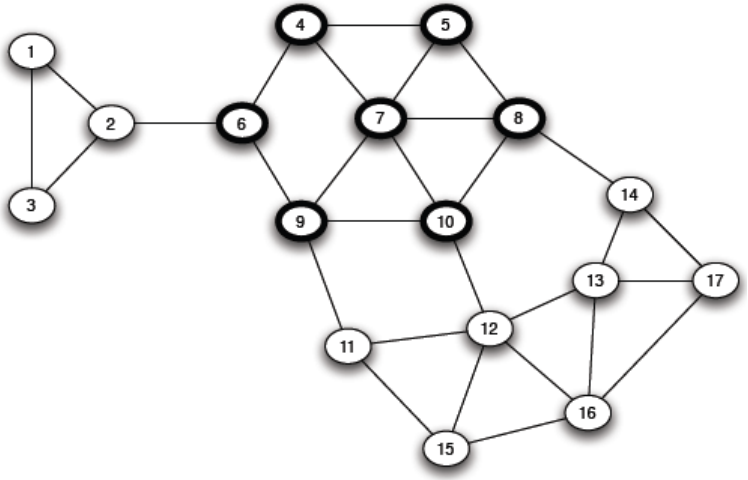
cascade of adoptions of A

# Modeling Diffusion through a Network: Cascading Behavior

$a = 3, b = 2, q = 2/5$




**Step 3**

# Modeling Diffusion through a Network: Cascading Behavior

- Observation: strictly progressive sequence of switches from B to A

- Depends on the choice of the *initial adapters* and threshold *q*

# Modeling Diffusion through a Network: Cascading Behavior

1. A set of initial adopters who start with a new behavior A, while every other node starts with behavior B.
2. Nodes repeatedly evaluate the decision to switch from B to A using a threshold of $q$.
3. If the resulting cascade of adoptions of A eventually causes every node to switch from B to A, then we say that the set of initial adopters causes a complete cascade at threshold $q$.

# Modeling Diffusion through a Network: Cascading Behavior and "Viral Marketing"

Tightly-knit communities in the network can work to hinder the spread of an innovation

(examples, age groups and life-styles in social networking sites, Mac users, political opinions)

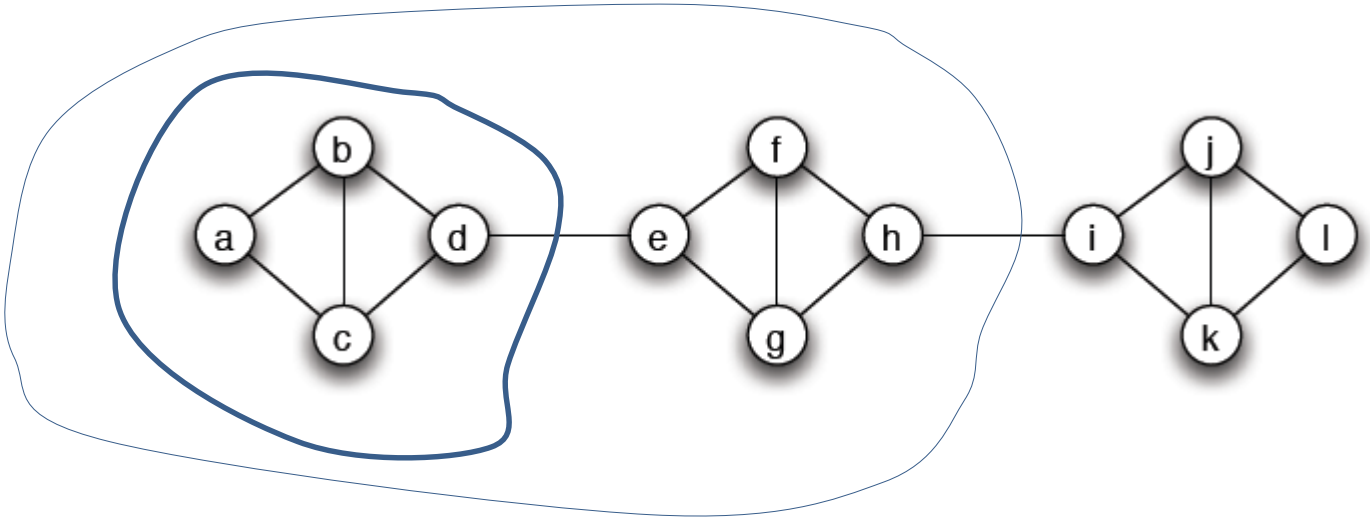
Strategies

- Improve the quality of A (increase the payoff *a*) (in the example, set a = 4)
- Convince a small number of *key people* to switch to A

Network-level cascade innovation adoption models vs population-level (decisions based on the entire population)

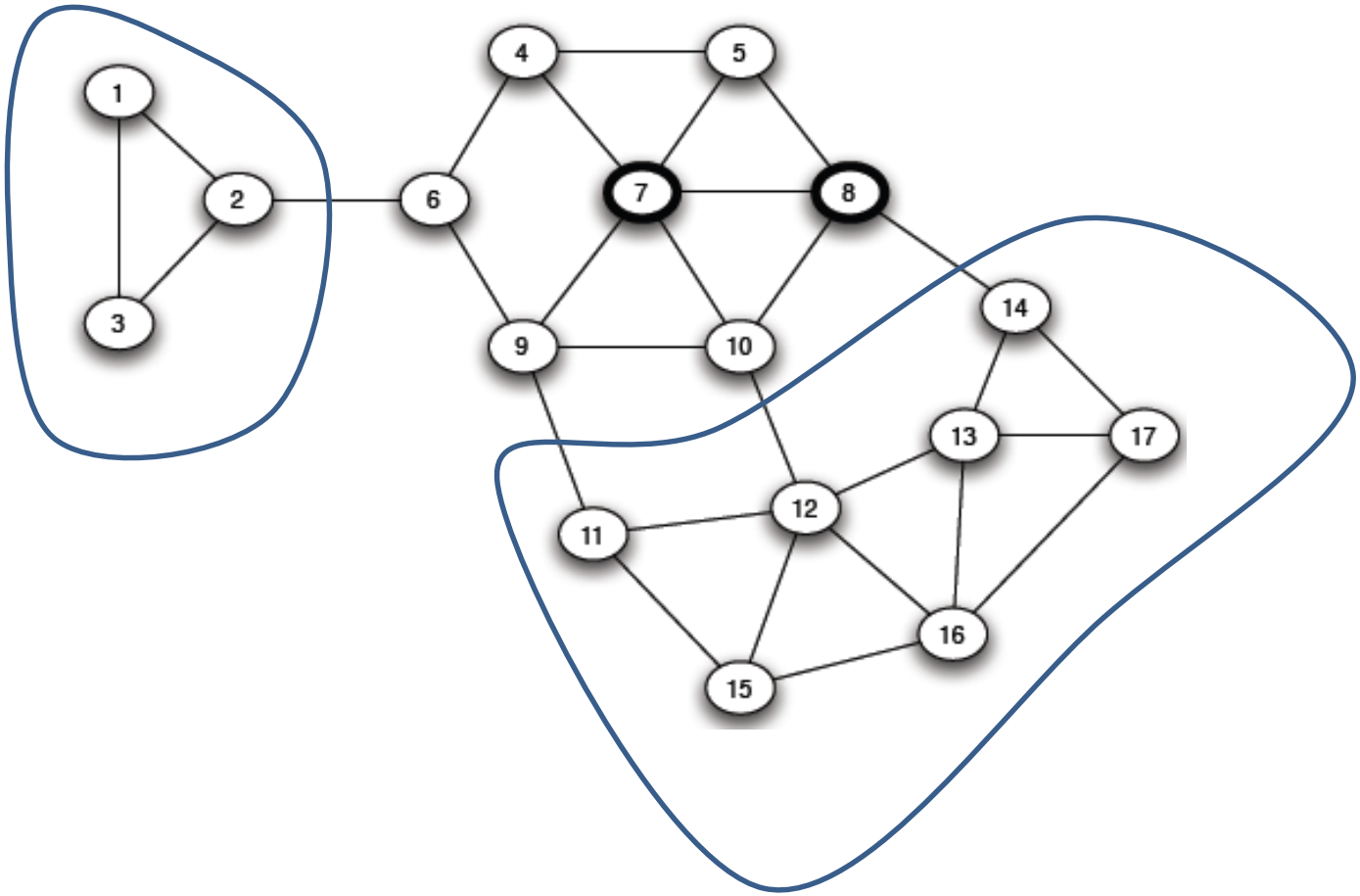
# Cascades and Clusters

A cluster of density *p* is a set of nodes such that each node in the set has *at least* a *p* fraction of its neighbors in the set




- Does not imply that any two nodes in the same cluster necessarily have much in common (what is the density of a cluster with all nodes?)
- The union of any two cluster of density *p* is also a cluster of density at least *p*

# Cascades and Clusters



# Cascades and Clusters

**Claim:** Consider a set of initial adopters of behavior A, with a threshold $q$ for nodes in the remaining network to adopt behavior A.

(i) (clusters as obstacles to cascades)

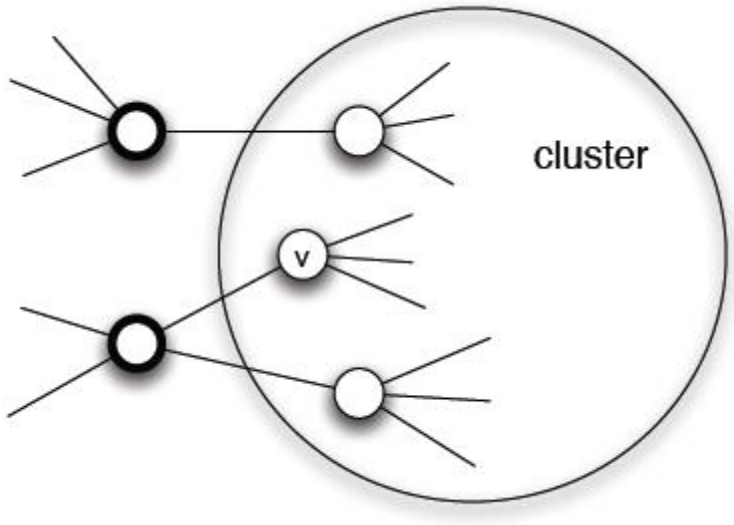
If the remaining network contains a cluster of density $> 1 - q$ =>
the set of initial adopters will not cause a complete cascade.

(ii) (clusters are the only obstacles to cascades)

Whenever a set of initial adopters does not cause a complete cascade =>
the remaining network contains a cluster of density $> 1 - q$.

# Cascades and Clusters

**Proof of** (i) (clusters as obstacles to cascades)

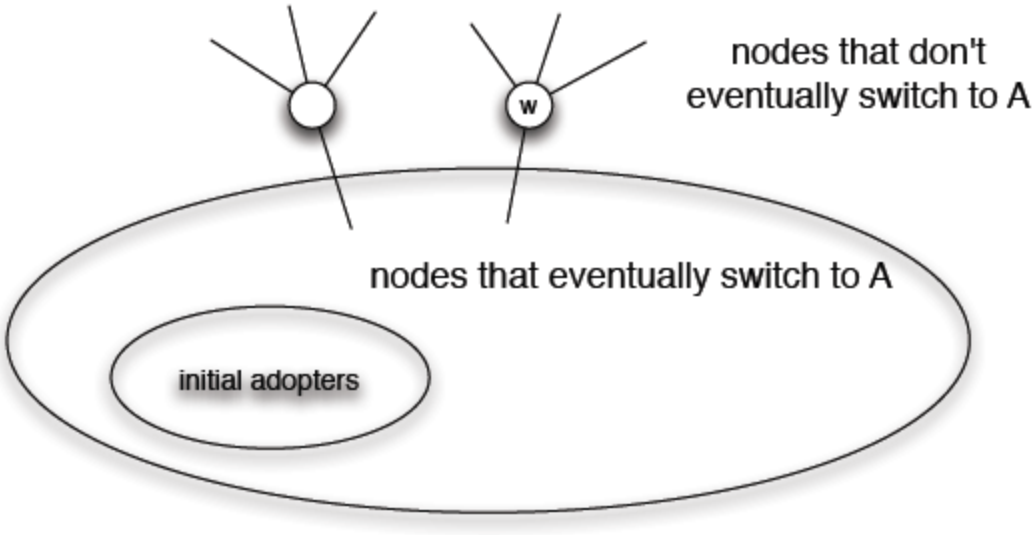



*Proof by contradiction*

Let v be the first node in the cluster that adopts A

# Cascades and Clusters

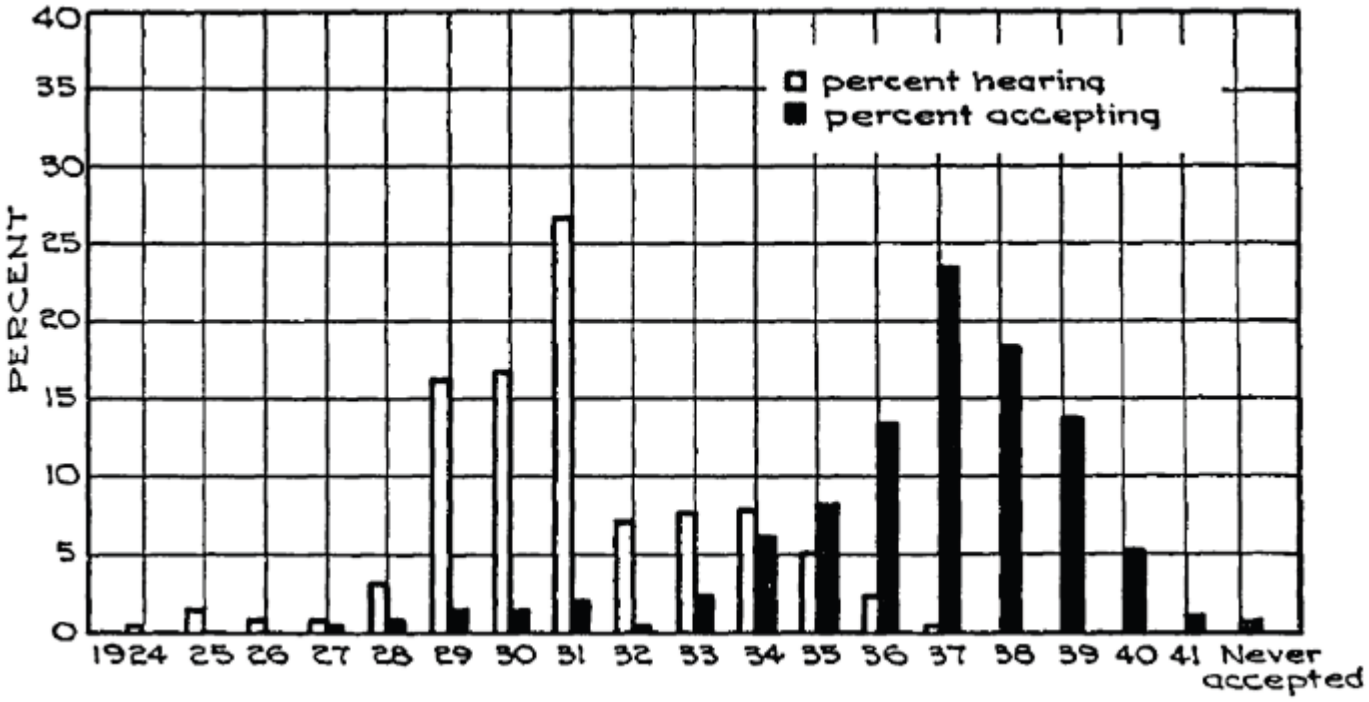
**Proof of** (ii) (clusters are the only obstacles to cascades)




Let *S* be the set of all nodes using B at the end of the process
Show that S is a cluster of density $> 1 - q$

# Innovation Adoption Characteristics

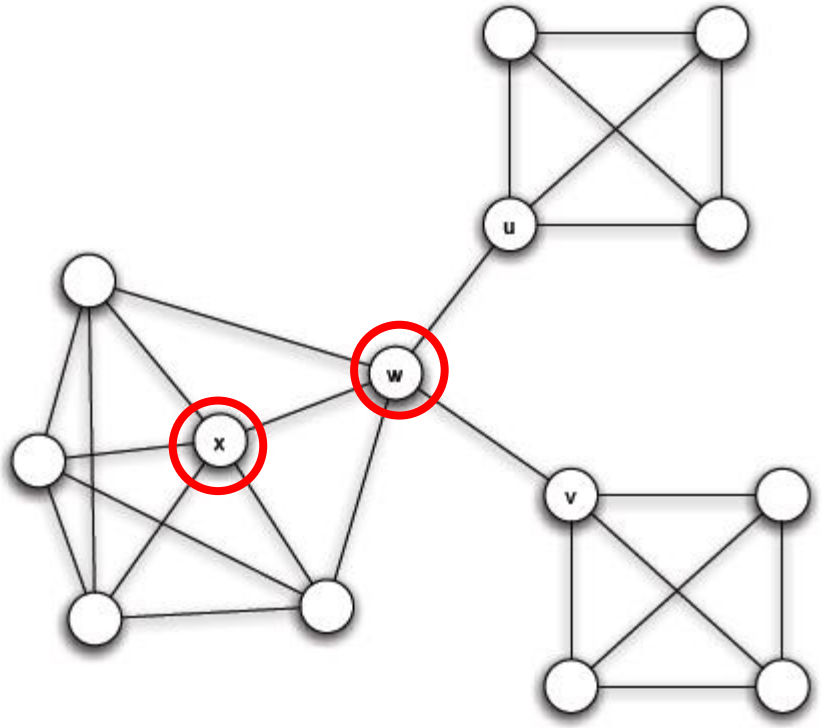
A crucial difference between learning a new idea and actually deciding to accept it (awareness vs adoption of an idea)



# Diffusion, Thresholds and the Role of Weak Ties

Relation to weak ties and local bridges

$q = 1/2$




Bridges convey awareness but are weak at transmitting costly to adopt behaviors

# Extensions of the Basic Cascade Model:
## Heterogeneous Thresholds

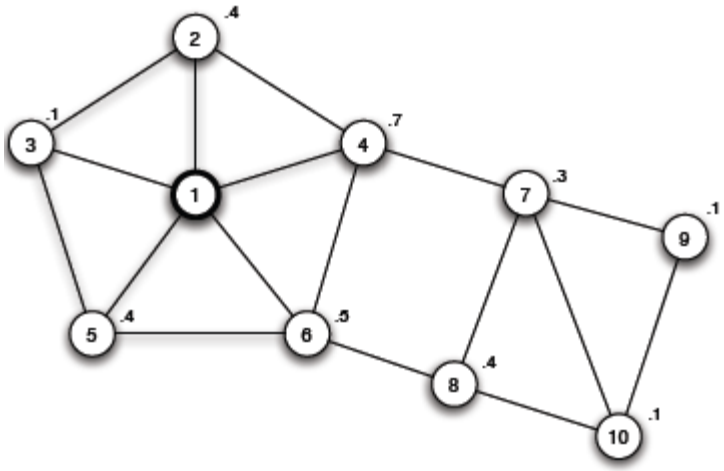
Each person values behaviors A and B differently:

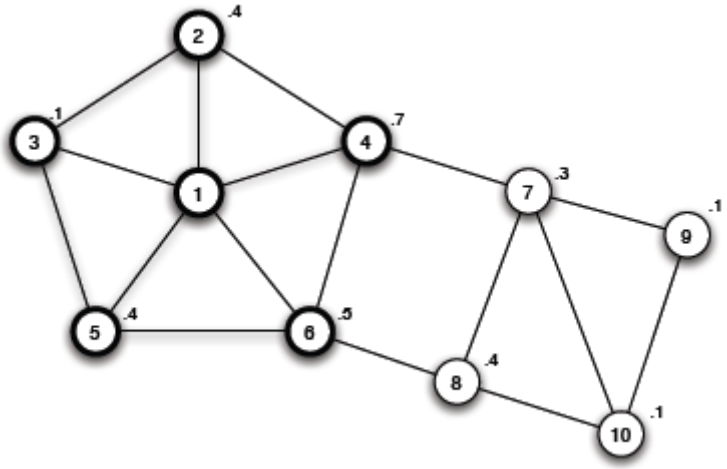
- If both *u* and *w* adapt A, *u* gets a payoff $a_u > 0$ and *w* a payoff $a_w > 0$
- If both *u* and *w* adapt B, u gets a payoff $b_u > 0$ and w a payoff $b_w > 0$
- If opposite behaviors, each gets a payoff 0

| | | $w$: $A$ | $w$: $B$ |
|---|---|---|---|
| $v$ | $A$ | $a_v, a_w$ | $0, 0$ |
| | $B$ | $0, 0$ | $b_v, b_w$ |

Each node u has its own personal threshold $q_u \geq b_u/(a_u + b_u)$

# Extensions of the Basic Cascade Model:
## Heterogeneous Thresholds




✓ Not just the power of influential people, but also the extent to which they have access to easily influenceable people

✓ What about the role of clusters?

A *blocking cluster* in the network is a set of nodes for which each node *u* has more that $1 - q_u$ fraction of its friends also in the set.

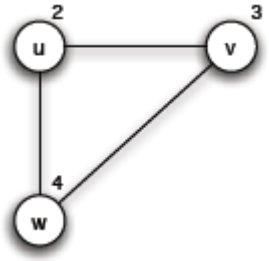
# Knowledge, Thresholds and Collective Action:
## Collective Action and Pluralistic Ignorance

A *collective action problem*: an activity produces benefits only if enough people participate (population level effect)

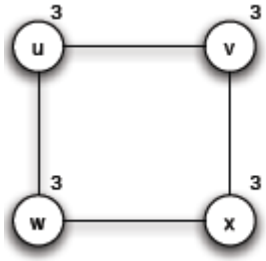
*Pluralistic ignorance*: a situation in which people have wildly erroneous estimates about the prevalence of certain opinions in the population at large (lack of knowledge)

# Knowledge, Thresholds and Collective Action:
## A model for the effect of knowledge on collective actions

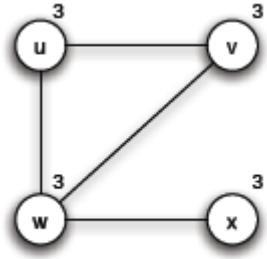
- Each person has a personal threshold which encodes her willingness to participate
- A threshold of *k* means that she will participate if at least *k* people in total (including herself) will participate
- Each person in the network knows the thresholds of her neighbors in the network




- w will never join, since there are only 3 people
- v
- u




- Is it safe for u to join?




- Is it safe for u to join? (common knowledge)

# Knowledge, Thresholds and Collective Action:
## Common Knowledge and Social Institutions

- Not just transmit a message, but also make the listeners or readers *aware that many others have gotten the message* as well (*Apple Macintosh introduced in a Ridley-Scott-directed commercial during the 1984 Super Bowl)*

- Social networks do not simply allow for interaction and flow of information, but these processes in turn allow individuals to base decisions *on what other knows* and *on how they expect others to behave as a result*

# Cascade Capacity

Given a network, what is the *largest threshold* at which *any "small" set* of initial adopters can cause a *complete cascade*?

Called cascade capacity of the network

- Infinite network in which each node has a finite number of neighbors
- Small means finite set of nodes

# Cascade Capacity

Same model as before:

- Initially, *a finite set S* of nodes has behavior A and all others adopt B
- Time runs forwards in steps, $t = 1, 2, 3, \ldots$
- In each step $t$, each node other than those in S uses the decision rule with threshold $q$ to decide whether to adopt behavior A or B
- The set S causes a complete cascade if, starting from S as the early adopters of A, every node in the network eventually switched permanently to A.

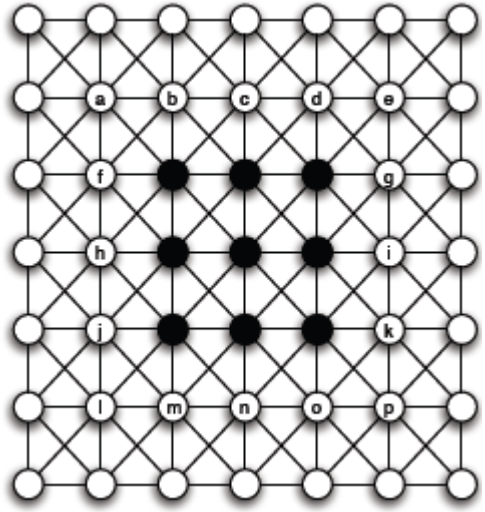
The cascade capacity of the network is the largest value of the threshold $q$ for which some finite set of early adopters can cause *a complete cascade*.

# Cascade Capacity

An infinite path


*Spreads if ≤ 1/2*

An infinite grid




*Spreads if ≤ 3/8*

- ✓ An intrinsic property of the network
- ✓ Even if A better than B, for *q* strictly between 3/8 and 1/2, A cannot win

# Cascade Capacity

How large can a cascade capacity be?

- At least 1/2

- *Is there any network with a higher cascade capacity*?

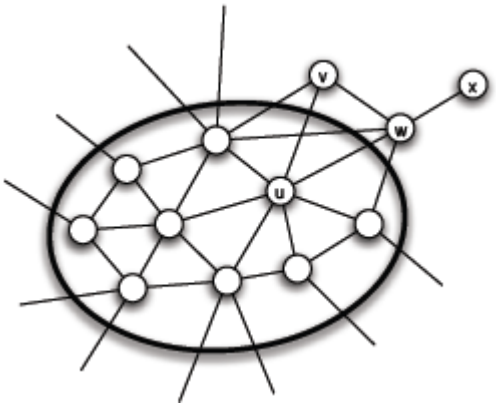
- This will mean that *an inferior technology* can displace a superior one, even when the inferior technology starts at only a small set of initial adopters.

# Cascade Capacity

**Claim**: There is no network in which the cascade capacity exceeds 1/2

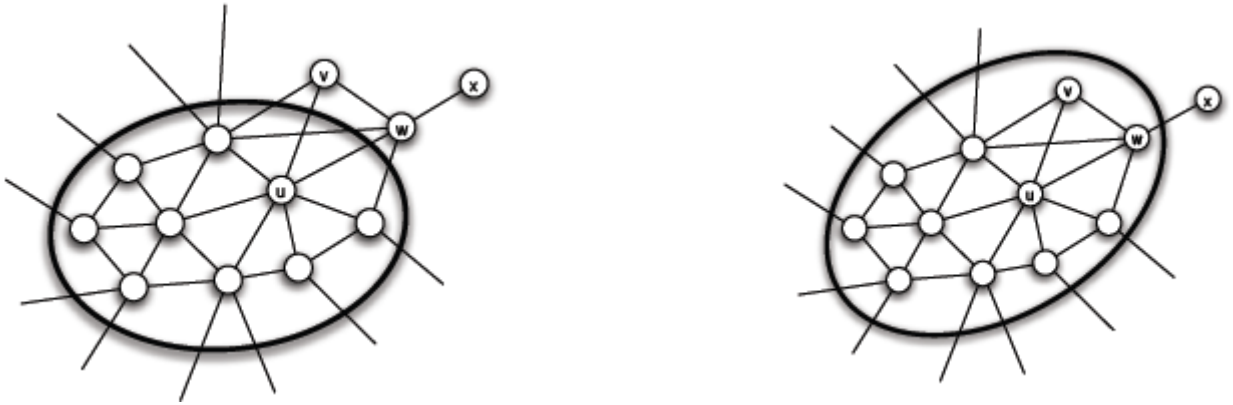
# Cascade Capacity

Interface: the set of A-B edges




In each step the size of the interface strictly decreases
Why is this enough?

# Cascade Capacity




At some step, a number of nodes decide to switch from B to A

*General Remark: In this simple model, a worse technology cannot displace a better and wide-spread one*

# Compatibility and Cascades

Extension: an individual can sometimes choose a combination of two available behaviors -> three strategies A, B and AB

## Coordination game with a bilingual option

- Two bilingual nodes can interact using the better of the two behaviors
- A bilingual and a monolingual node can only interact using the behavior of the monolingual node

| | | $w$: $A$ | $B$ | $AB$ |
|---|---|---|---|---|
| $v$: | $A$ | $a, a$ | $0, 0$ | $a, a$ |
| | $B$ | $0, 0$ | $b, b$ | $b, b$ |
| | $AB$ | $a, a$ | $b, b$ | $(a, b)^+, (a, b)^+$ |

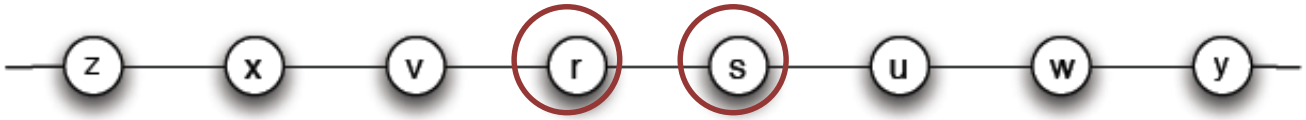
AB is a dominant strategy?

Cost **c** associated with the AB strategy

# Compatibility and Cascades

Example ($a$ = 2, $b$ =3, $c$ =1)

| $v$ \ $w$ | $A$ | $B$ | $AB$ |
|---|---|---|---|
| $A$ | $a, a$ | $0, 0$ | $a, a$ |
| $B$ | $0, 0$ | $b, b$ | $b, b$ |
| $AB$ | $a, a$ | $b, b$ | $(a, b)^+, (a, b)^+$ |




| Start | B | B | B | A | A | B | B | B |
|---|---|---|---|---|---|---|---|---|

B: 0+$b$ = 3
A: 0+$a$ = 2
AB: $b$+$a$-$c$ = 4 √

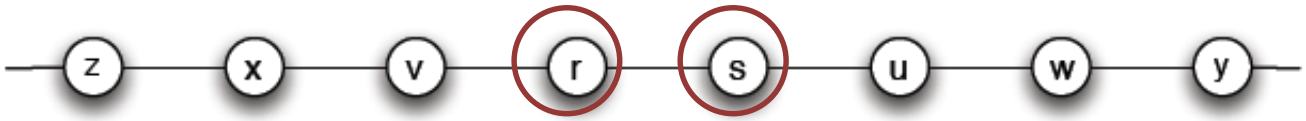
| Step 1 | B | B | AB | A | A | AB | B | B |
|---|---|---|---|---|---|---|---|---|

B: $b$+$b$ = 6 √
A: 0+$a$ = 2
AB: $b$+$b$-$c$ = 5

# Compatibility and Cascades

Example ($a$ = **5**, $b$ =3, $c$ =1)

| $v$ \ $w$ | $A$ | $B$ | $AB$ |
|---|---|---|---|
| $A$ | $a, a$ | $0, 0$ | $a, a$ |
| $B$ | $0, 0$ | $b, b$ | $b, b$ |
| $AB$ | $a, a$ | $b, b$ | $(a, b)^+, (a, b)^+$ |




Start B B B A A B B B

B: 0+$b$ = 3
A: 0+$a$ = 5
AB: $b$+$a$-$c$ = 7 √

Step 1 B B AB A A AB B B

B: 0+$b$ = 3
A: 0+$a$ = 5
AB: $b$+$a$-$c$ = 7 √

Step 2 B AB AB A A AB AB B

B: 0+$b$ = 3
A: $\alpha$+$a$ = 10 √
AB: $a$+$a$-$c$ = 9

# Compatibility and Cascades

Example ($a$ = 5, $b$ =3, $c$ =1)

| $v$ \ $w$ | $A$ | $B$ | $AB$ |
|---|---|---|---|
| $A$ | $a, a$ | $0, 0$ | $a, a$ |
| $B$ | $0, 0$ | $b, b$ | $b, b$ |
| $AB$ | $a, a$ | $b, b$ | $(a, b)^+, (a, b)^+$ |


| | z | x | v | r | s | u | w | y |
|---|---|---|---|---|---|---|---|---|
| Start | B | B | B | A | A | B | B | B |
| Step 1 | B | B | AB | A | A | AB | B | B |
| Step 2 | B | AB | AB | A | A | AB | AB | B |
| Step 3 | AB | AB | A | A | A | A | AB | AB |
| Step 4 | AB | A | A | A | A | A | A | AB |

- Strategy AB spreads, then behind it, nodes switch permanently from AB to A
- Strategy B becomes *vestigial*

# Compatibility and Cascades

- Given an infinite graph, for which payoff values of *a*, *b* and *c*, is it possible for a finite set of nodes to cause a complete cascade of *A*?

Set *b* = 1 (default technology)

- Given an infinite graph, for which payoff values of *a* *(how much better the new behavior A)* and *c* *(how compatible should it be with B),* is it possible for a finite set of nodes to cause a complete cascade of *A*?

*A does better when it has a higher payoff, but in general hard time cascading when the level of compatibility is "intermediate" (value of c neither too high nor too low)*

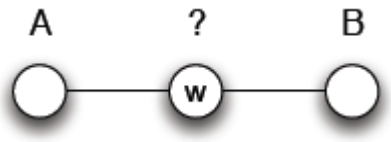
# Compatibility and Cascades

## Example: Infinite path

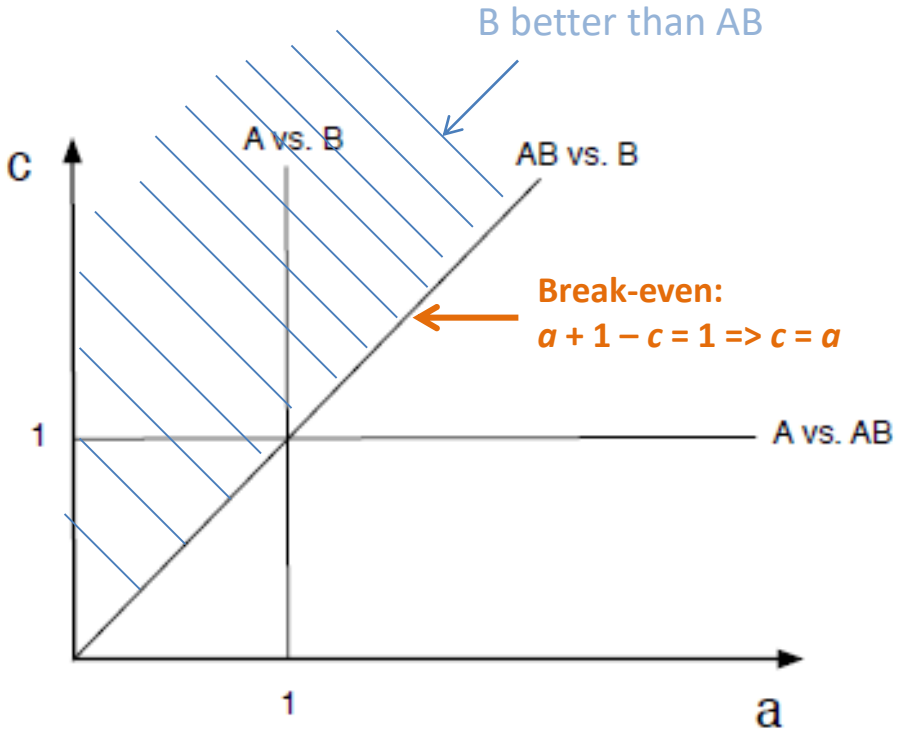
- (for two strategies) Spreads when $q \le 1/2$, $a \ge b$ (a better technology always spreads)

Assume that the set of initial adopters forms a contiguous interval of nodes on the path
Because of the symmetry, strategy changes to the right of the initial adopters

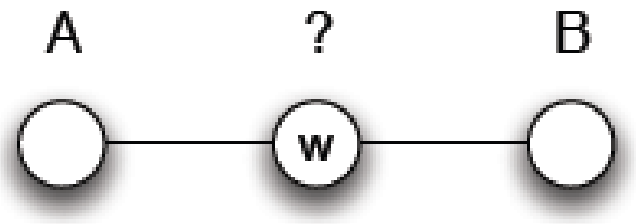
Initially,



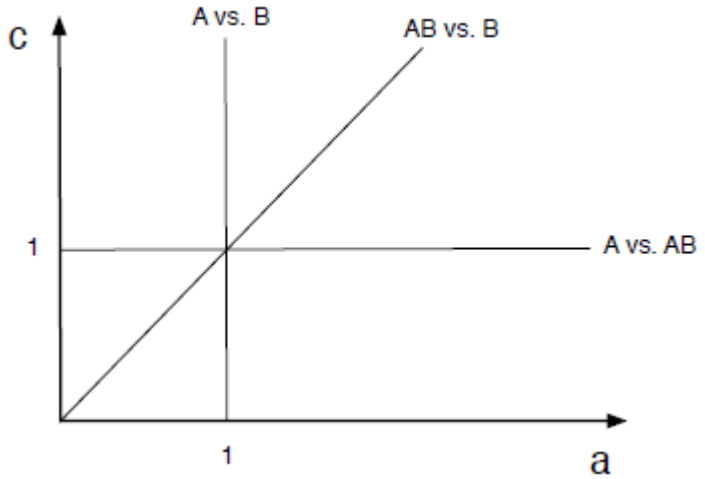

A: $0+a = a$
B: $0+b = 1$
AB: $a+b-c = a+1-c$

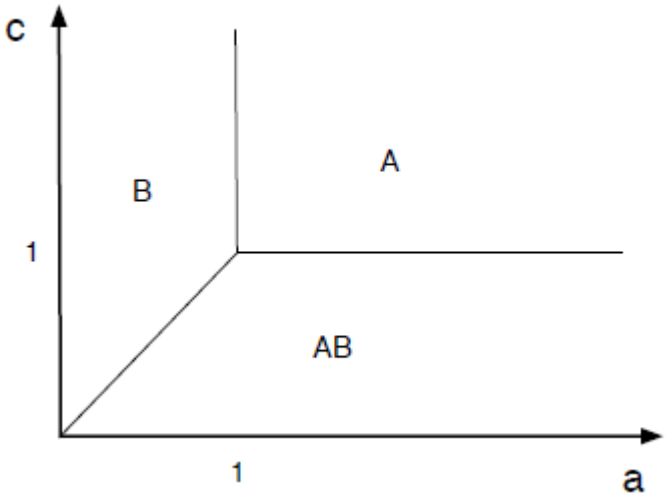


# Compatibility and Cascades

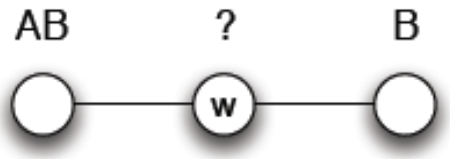



A: 0+$a$ = $a$
B: 0+$b$ = 1
AB: $a$+$b$-$c$ = $a$+1-$c$



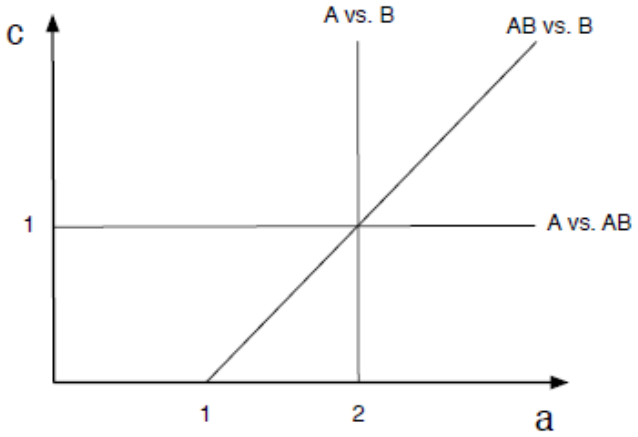



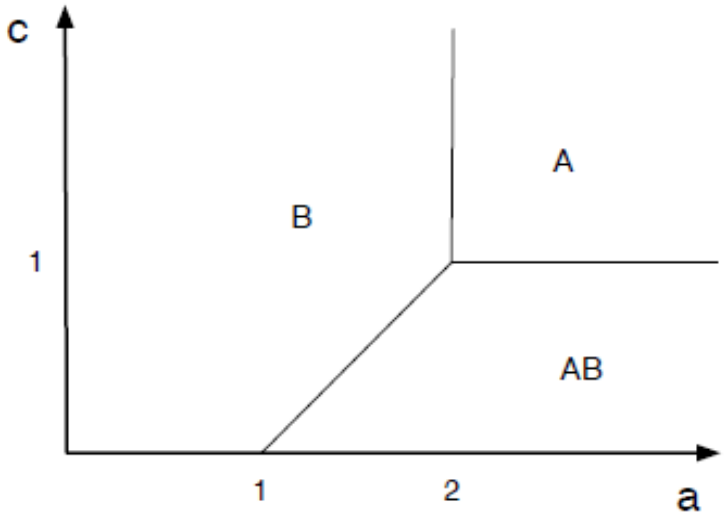
# Compatibility and Cascades



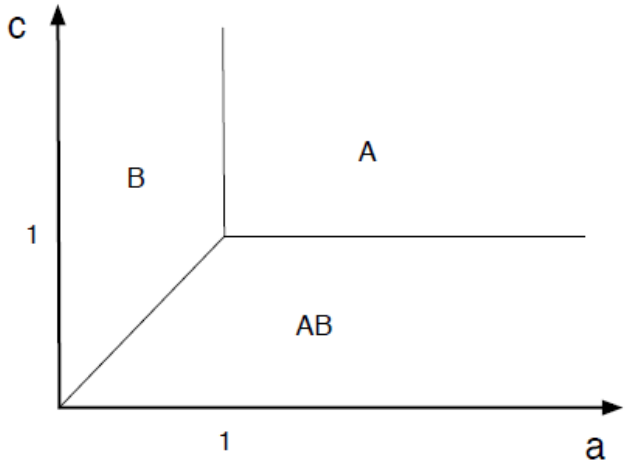

$a < 1$,
A: $0+a = a$
B: $b+b = 2$ √
AB: $b+b-c = 2-c$

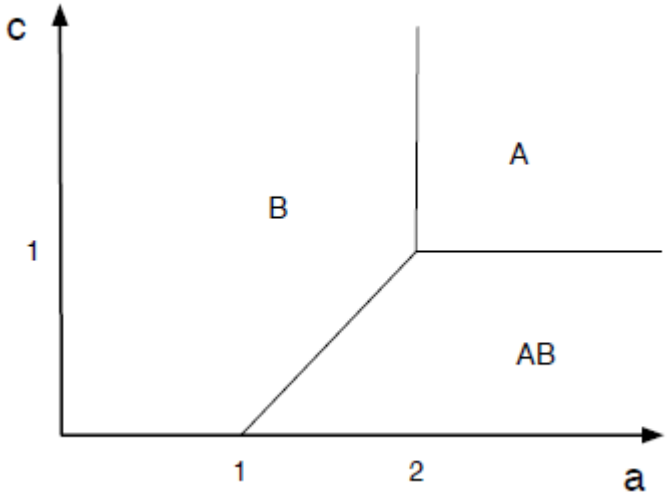
$a \geq 1$
A: $a$
B: 2
AB: $a+1-c$

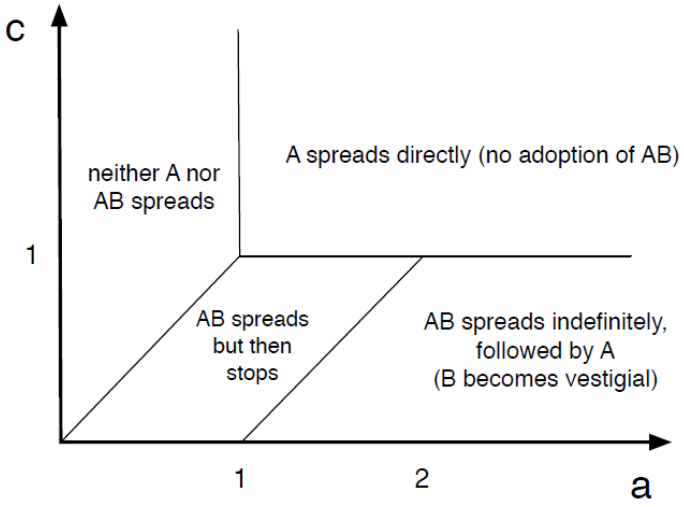


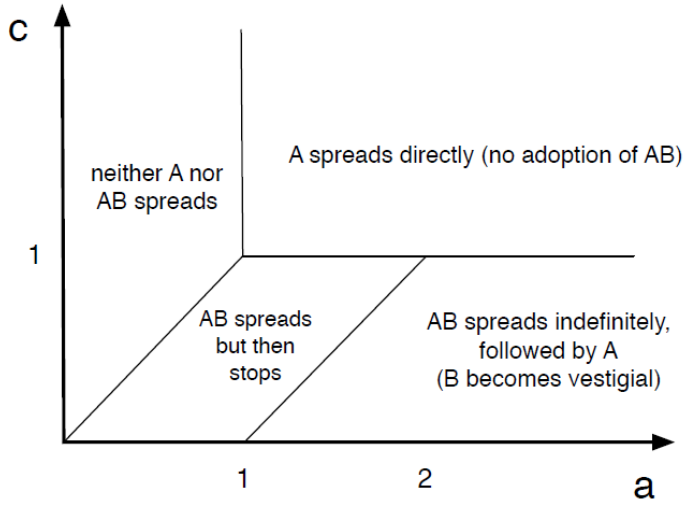
# Compatibility and Cascades

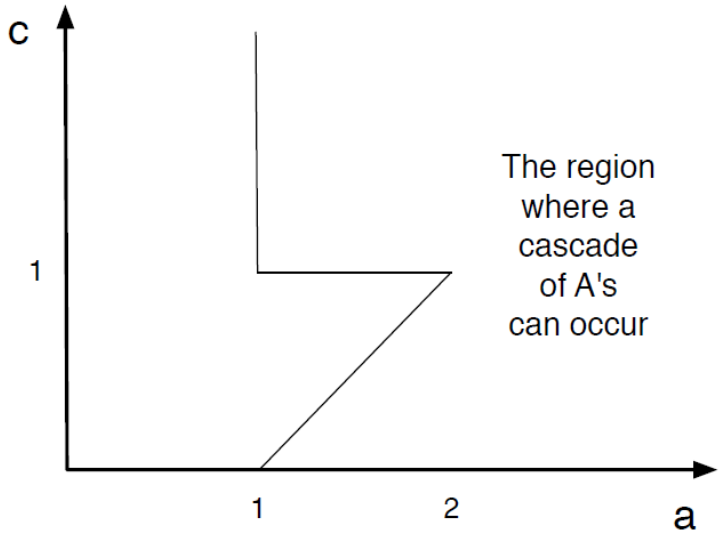


# Compatibility and Cascades




What does the triangular cut-out mean?

- If too easy, infiltration
- If too hard, direct conquest
- In between, “buffer” of AB

# Reference

Networks, Crowds, and Markets (Chapter 19)

# EPIDEMIC SPREAD

# 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

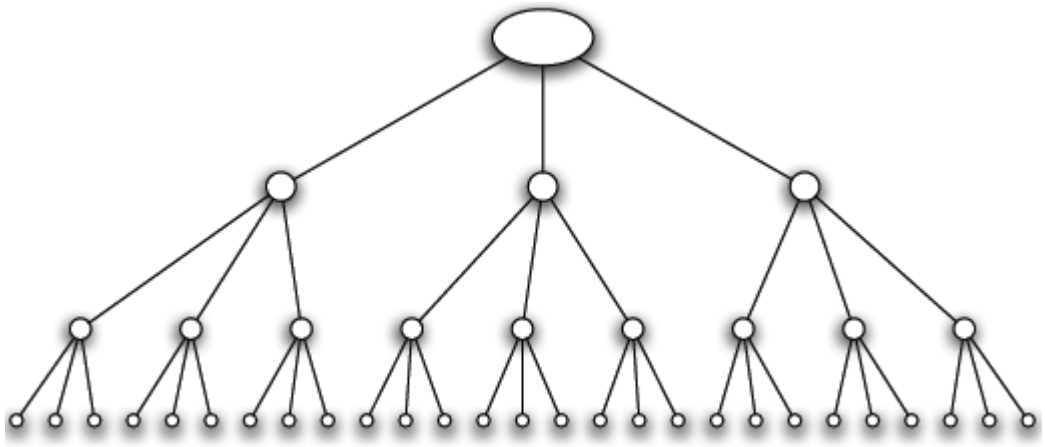
# Epidemics

- 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

# A simple model

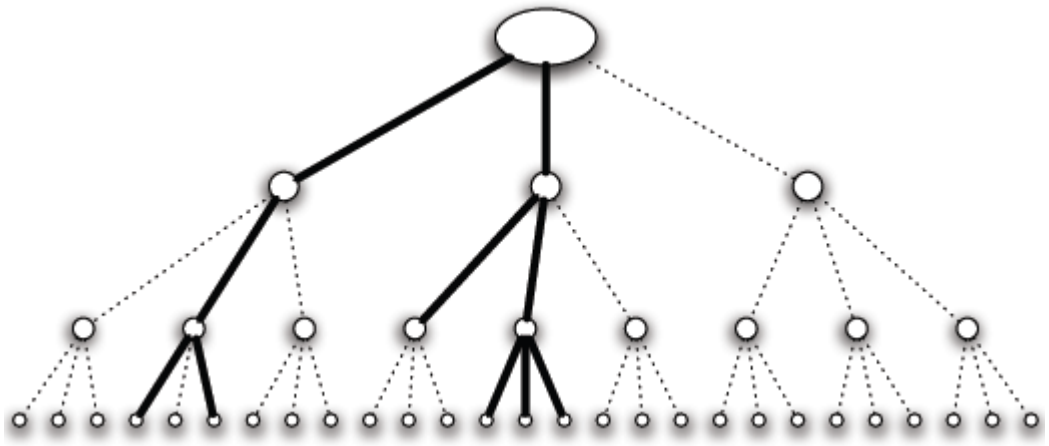
- Branching process: A person transmits the disease to each people she meets independently with a probability *p*
- An infected person meets *k* (new) people while she is contagious
- Infection proceeds in waves.



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*

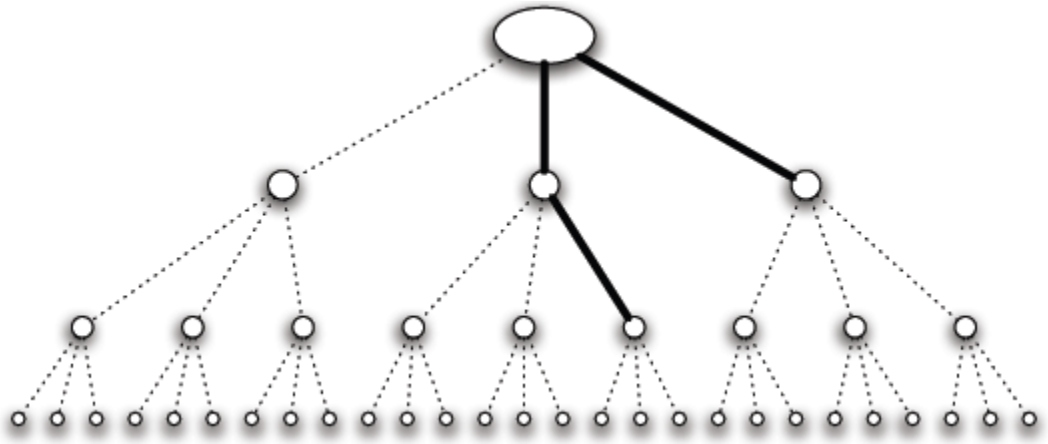


An aggressive epidemic with high infection probability

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 ($R_0$): the expected number of new cases of the disease caused by a single individual

$$R_0 = kp$$

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

Reduce *k*, or *p*

# Analysis

- $X_n$ : random variable indicating the number of infected nodes at level *n* (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 $n \to \infty$

We want to show that

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

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

# Proof

- At level n, $k^n$ nodes
- $Y_{nj}$: *1* if node *j* at level *n* is infected, 0 otherwise

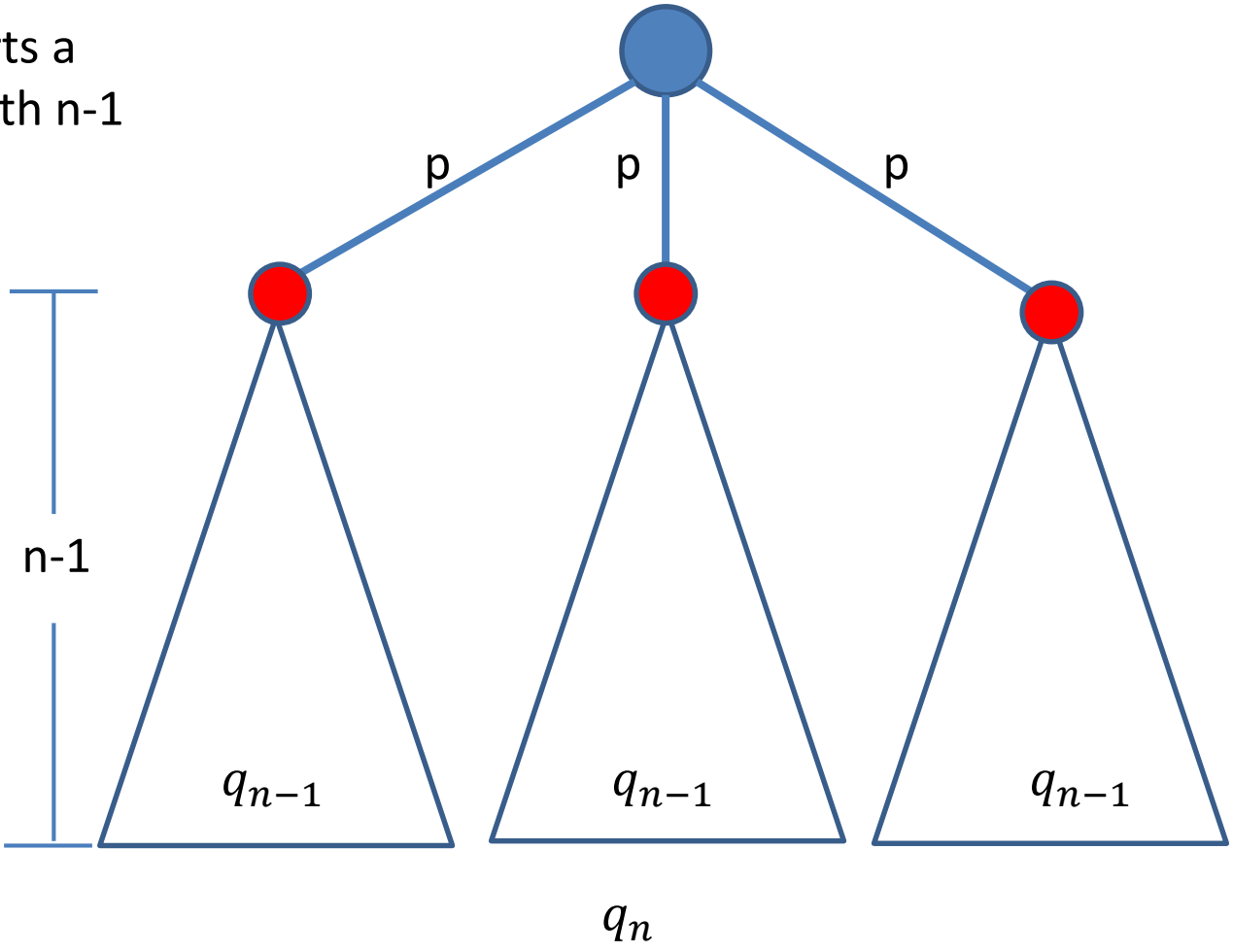
$$E[Y_{nj}] = p^n$$

- $E[X_n] = R_0^n$
- $E[X_n] \geq Pr[X_n \geq 1] \Rightarrow q_n \leq R_0^n$

This proves (a) but not (b)

# Proof

Each child of the root starts a branching process of length n-1




$$q_n = 1 - (1 - pq_{n-1})^k$$

if

$$f(x) = 1 - (1 - px)^k$$

then

$$q_n = f(q_{n-1})$$

We also have: $q_0 = 1$.

So we obtain a series of values: 1, $f(1)$, $f\big(f(1)\big)$, …

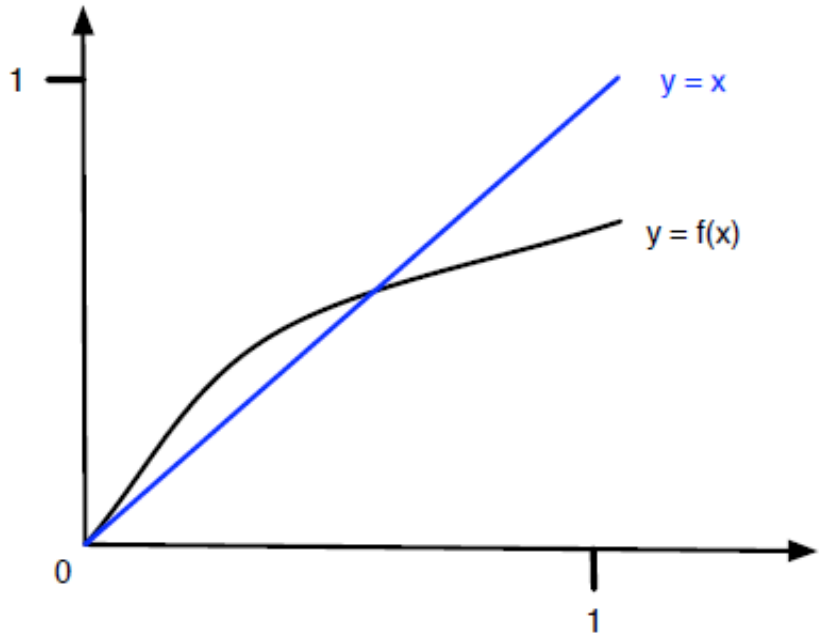
We want to find where this series converges

# Proof

- Properties of the function $f(x)$:
  1. $f(0) = 0$ and $f(1) = 1 - (1-p)^k < 1$.
  2. $f'(x) = pk(1-px)^{k-1} > 0$, in the interval [0,1] but decreasing. Our function is increasing and concave.
  3. $f'(0) = pk = R_0$

# Proof

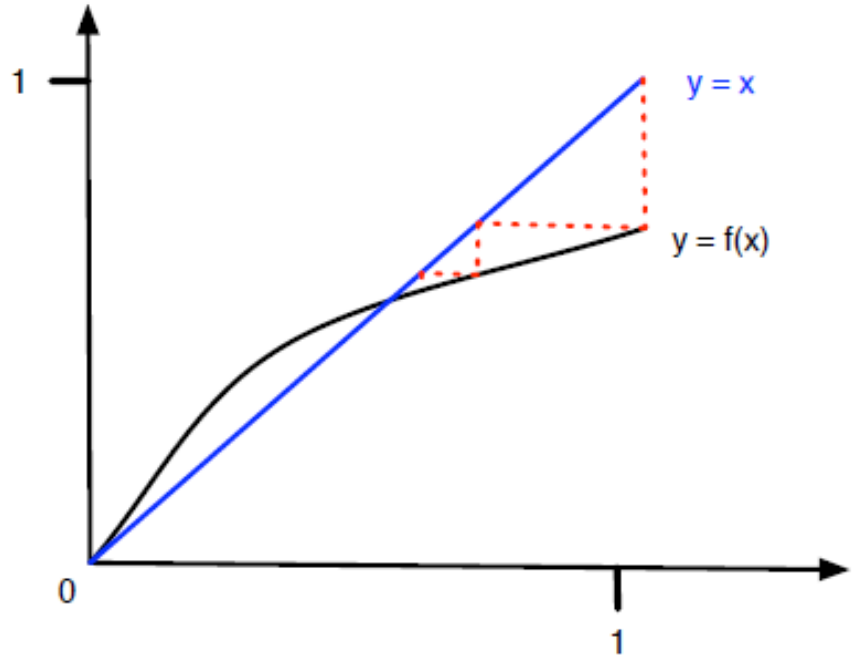
- Case 1: $R_0 = pk > 1$. The function starts with above the line $y = x$ but then drops below the line.




$f(x)$ crosses the line $y = x$ at some point

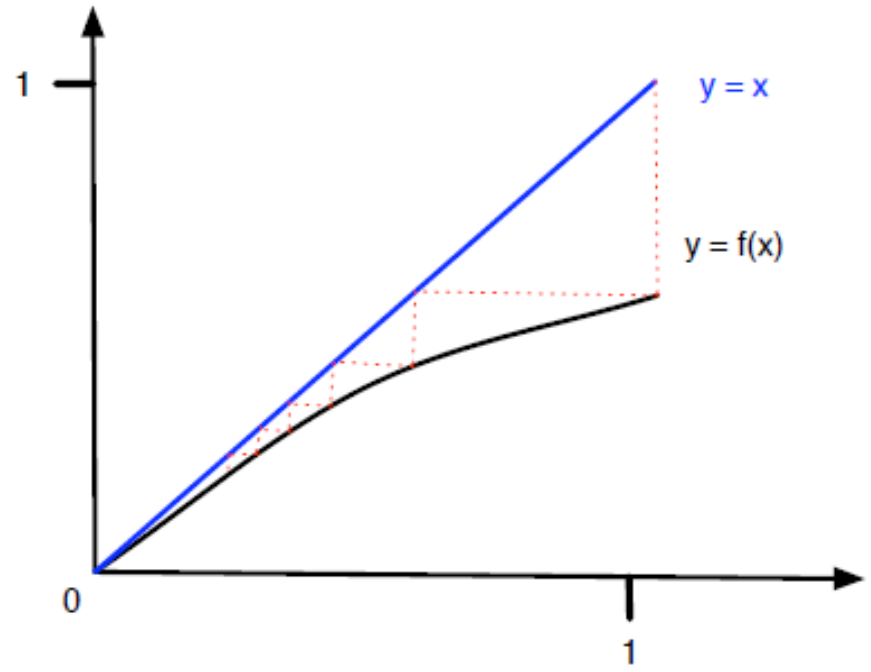
# Proof

- Starting from the value 1, repeated applications of the function $f(x)$ will converge to the value $q^* = q_n = f(q_n)$



# Proof

- Case 2: $R_0 = pk < 1$. The function starts with below the line $y = x$. Repeated applications of $f(x)$ converge to zero.



# Branching process

- Assumes no network structure, no triangles or shared neighbors

# The SIR model

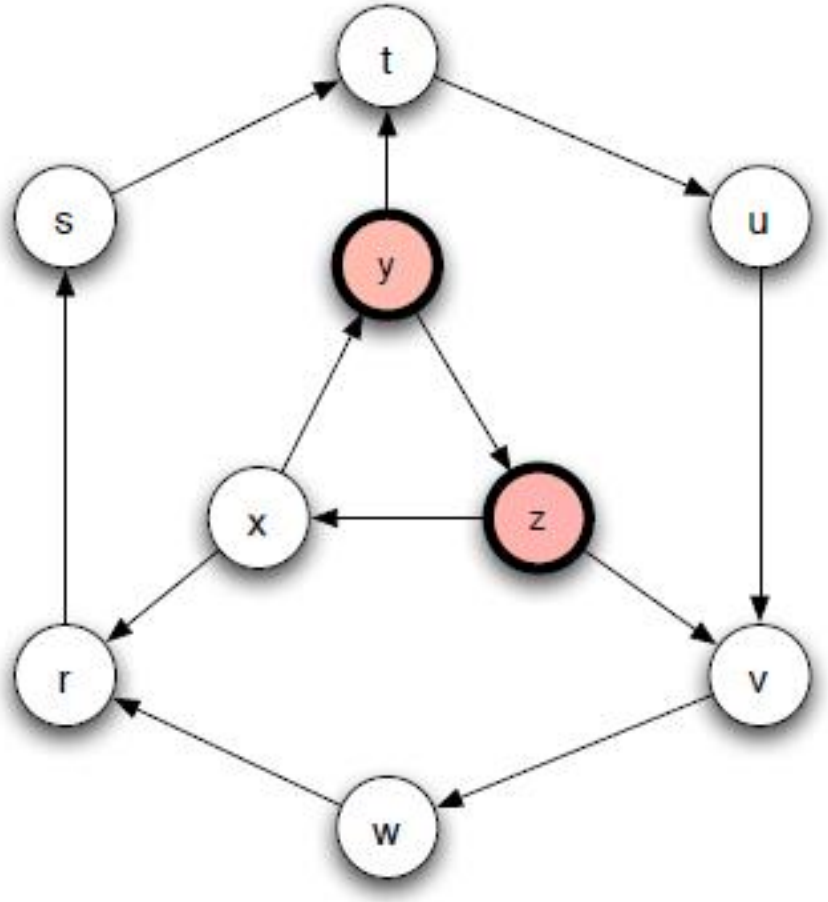
- Each node may be in the following states
  - Susceptible: healthy but not immune
  - Infected: has the virus and can actively propagate it
  - Removed: (Immune or Dead) had the virus but it is no longer active

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

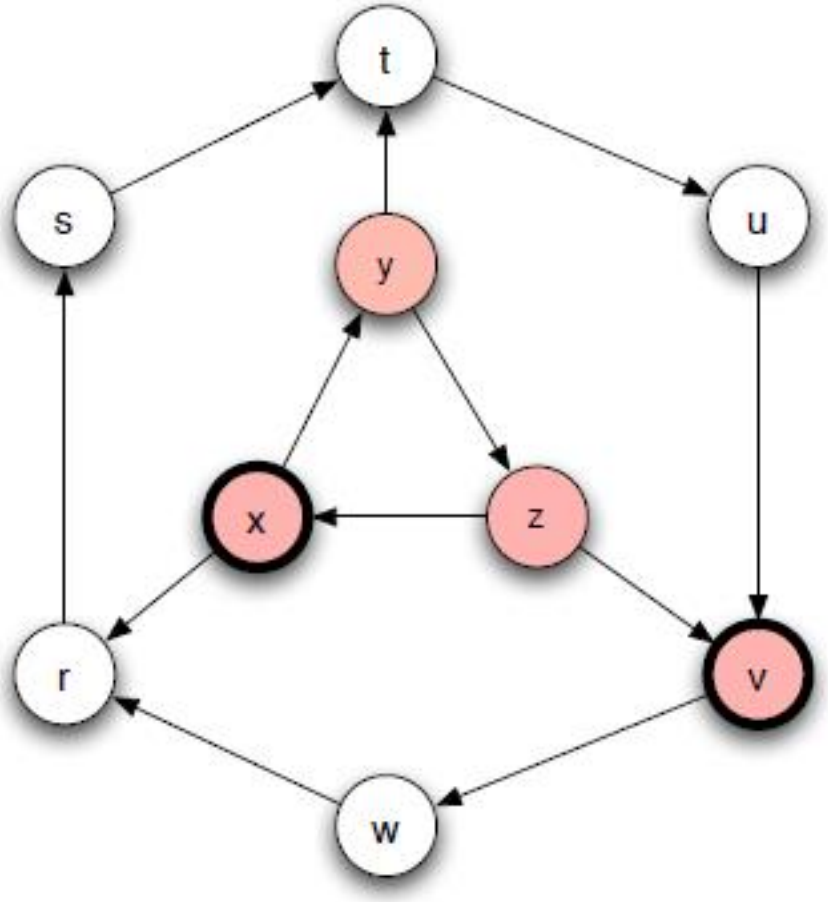
# The SIR process

- 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

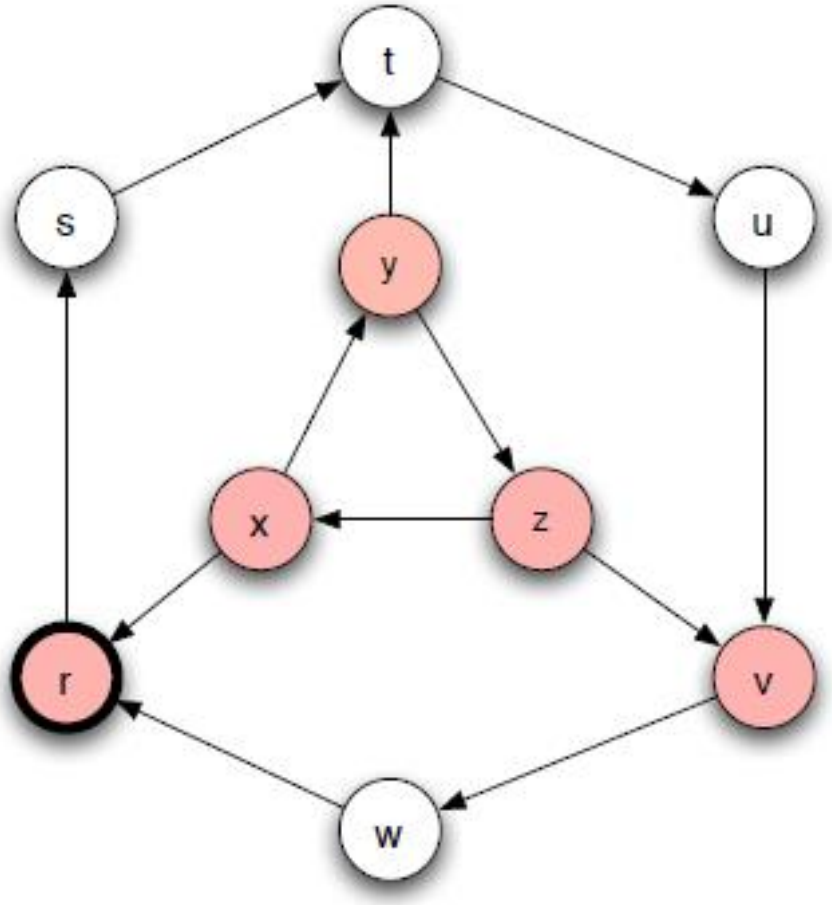
# Example



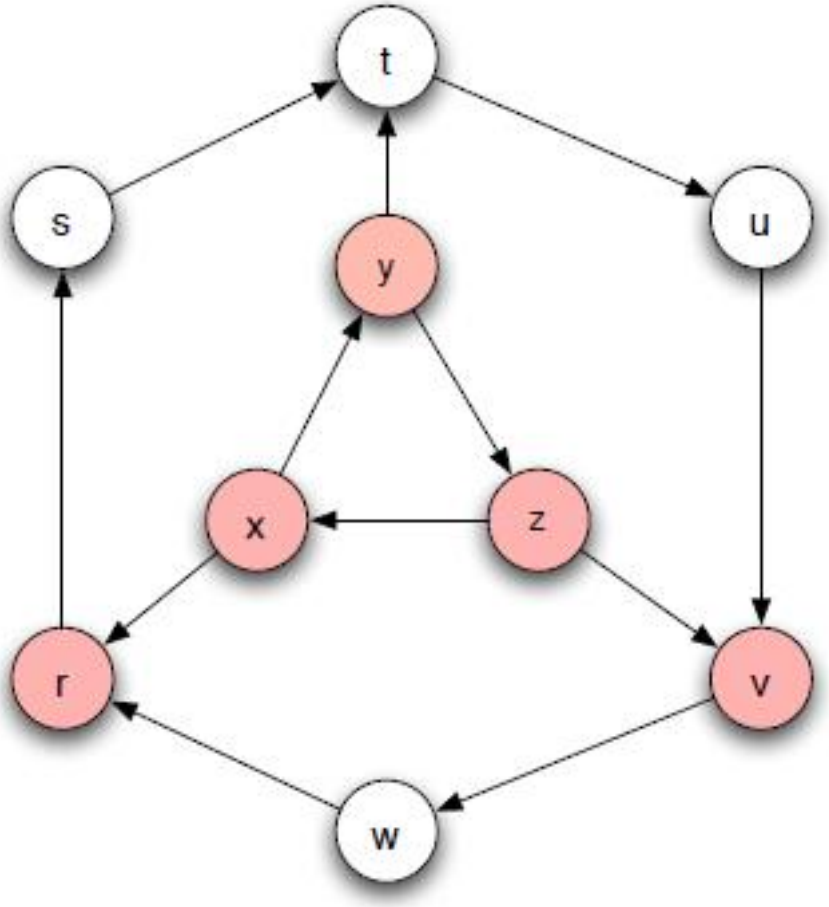
# Example



# Example



# Example



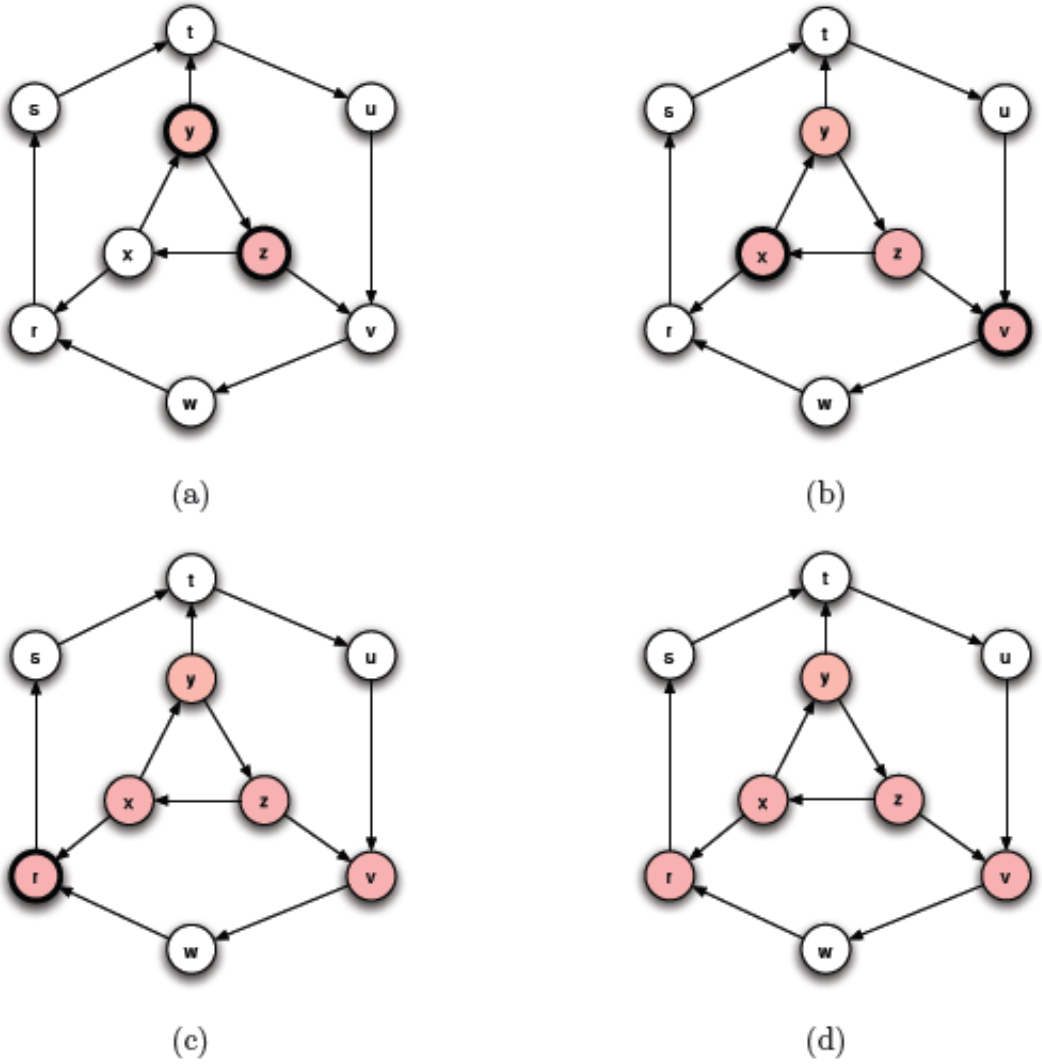



Figure 21.2: The course of an SIR epidemic in which each node remains infectious for a number of steps equal to $t_I = 1$. Starting with nodes $y$ and $z$ initially infected, the epidemic spreads to some but not all of the remaining nodes. In each step, shaded nodes with dark borders are in the Infectious ($I$) 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 $(1/3)^4 = 1/81$

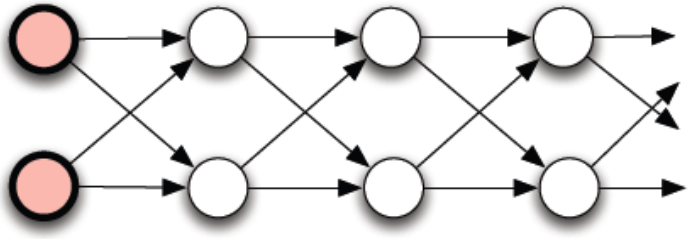


Figure 21.3: In this network, the epidemic is forced to pass through a narrow "channel" of nodes. In such a structure, even a highly contagious disease will tend to die out relatively quickly.

# 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

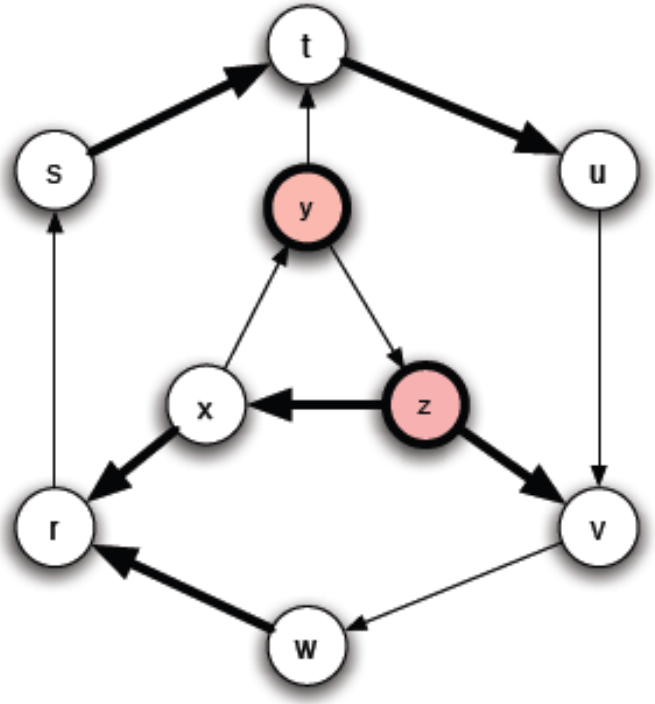



Figure 21.4: An equivalent way to view an SIR epidemic is in terms of *percolation*, where we decide in advance which edges will transmit infection (should the opportunity arise) and which will not.

# 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

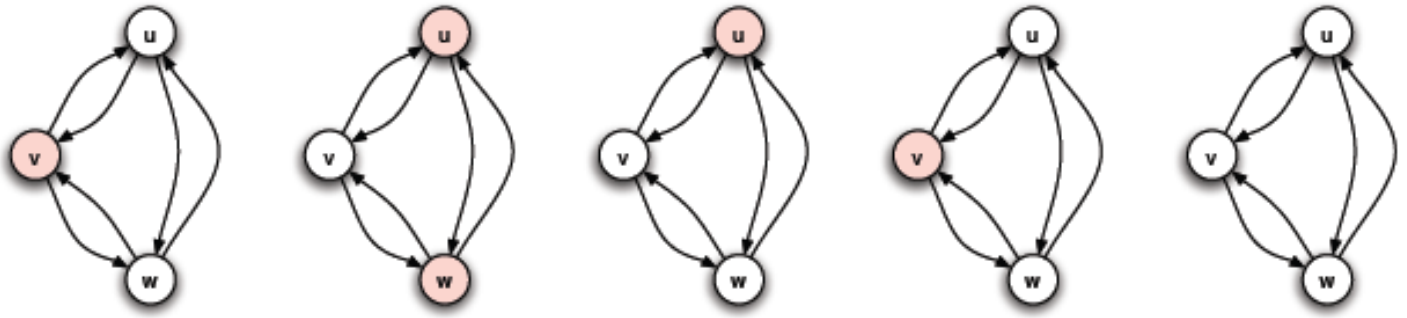



Figure 21.5: In an SIS epidemic, nodes can be infected, recover, and then be infected again. In each step, the nodes in the Infectious state are shaded.

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

# An eigenvalue point of view

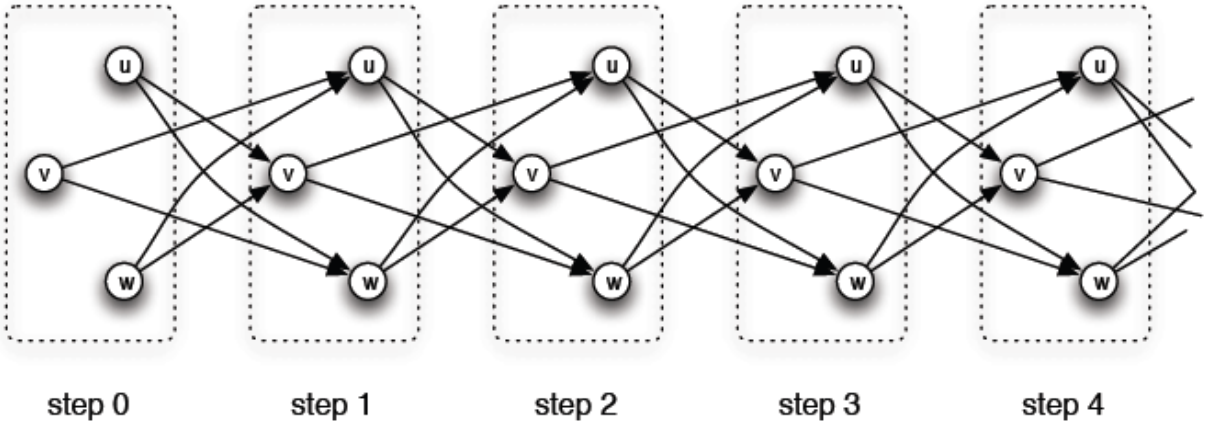
- If A is the adjacency matrix of the network, then the virus dies out if

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

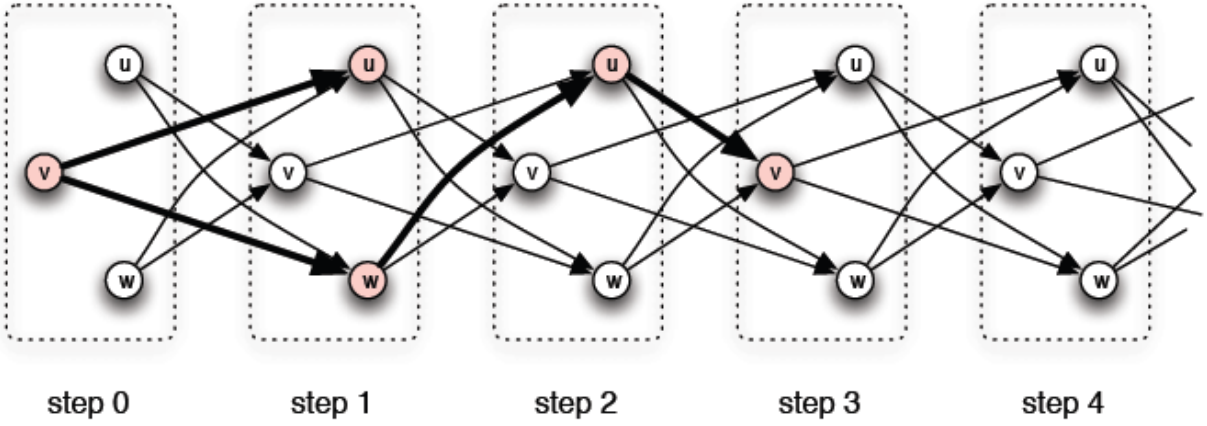
- Where $\lambda_1(A)$ is the first eigenvalue of A

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*

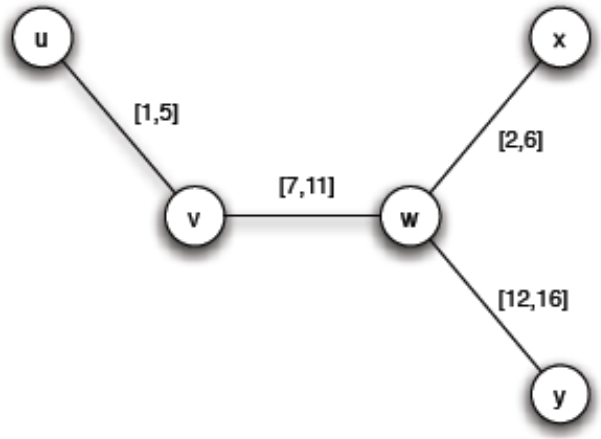



(b) *The SIS epidemic can then be represented as an SIR epidemic on this time-expanded network.*

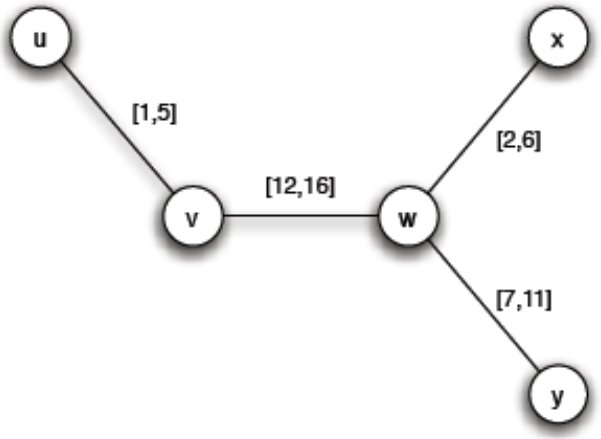
Figure 21.6: An SIS epidemic can be represented in the SIR model by creating a separate copy of the contact network for each time step: a node at time $t$ can infect its contact neighbors at time $t + 1$.

# 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.*

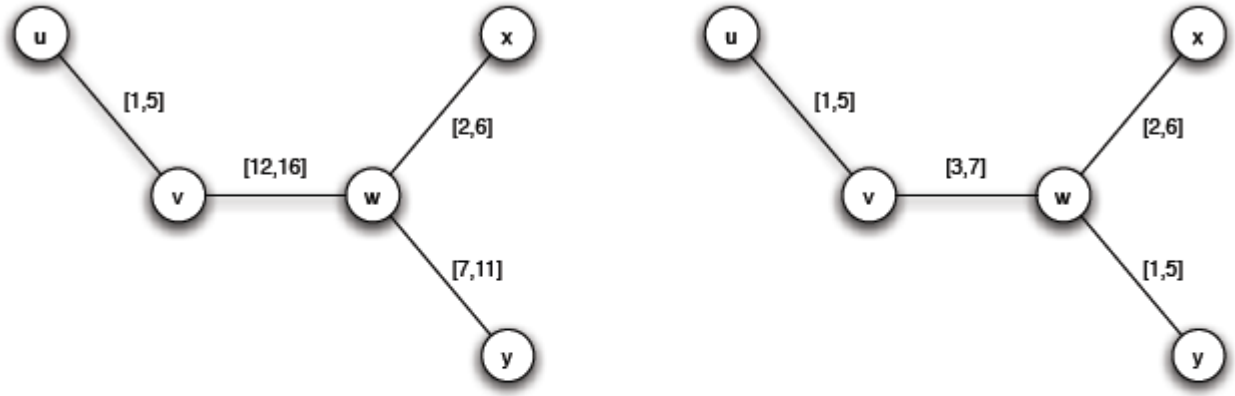



(b) *The same network as in (a), except that the timing of the w-v and w-y partnerships have been reversed.*

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 infected 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 $t_R$ steps in the ***R*** state, node *u* returns to the ***S*** state.

# References


- D. Easley, J. Kleinberg. *Networks, Crowds and Markets: Reasoning about a highly connected world*. Cambridge University Press, 2010 – Chapter 21
- Y. Wang, D. Chakrabarti, C. Wang, C. Faloutsos. *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*. SRDS 2003

# INFLUENCE MAXIMIZATION

# Maximizing spread

- Suppose that instead of a virus we have an item (product, idea, video) that propagates through contact
  - Word of mouth propagation.

- An advertiser is interested in maximizing the spread of the item in the network
  - The holy grail of "viral marketing"

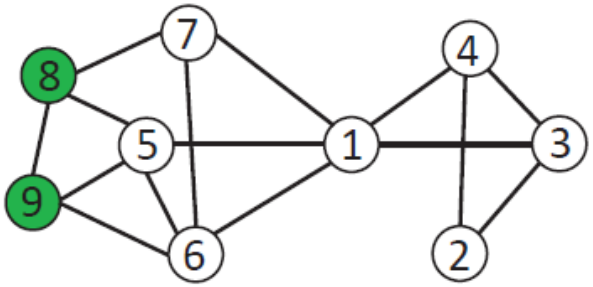
- Question: which nodes should we "infect" so that we maximize the spread? [KKT2003]

# Independent cascade model

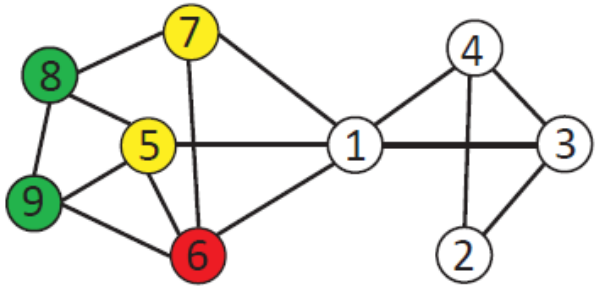
- Each node may be active (has the item) or inactive (does not have the item)
- Time proceeds at discrete time-steps.
- At time t, every node v that became active in time t-1 activates a non-active neighbor w with probability $p_{uw}$. If it fails, it does not try again

- The same as the simple SIR model

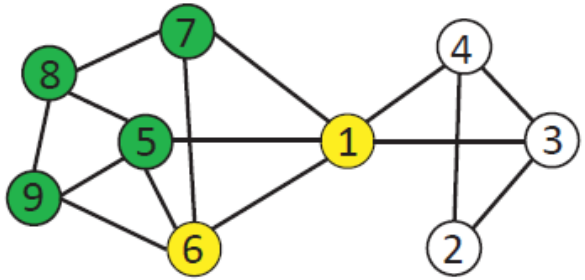
# Independent cascade



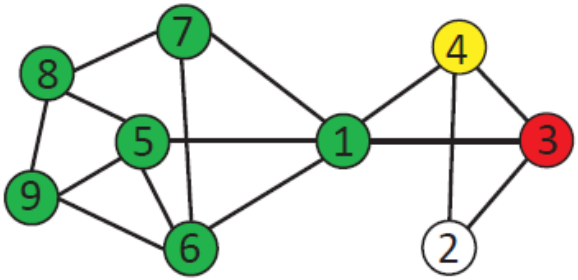

Step 0



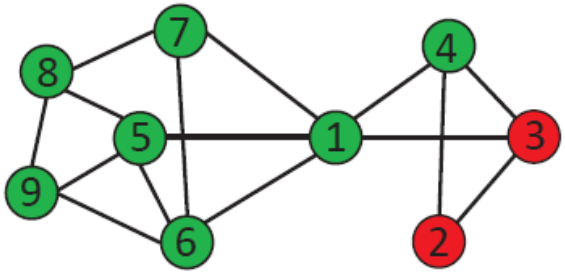

Step 1



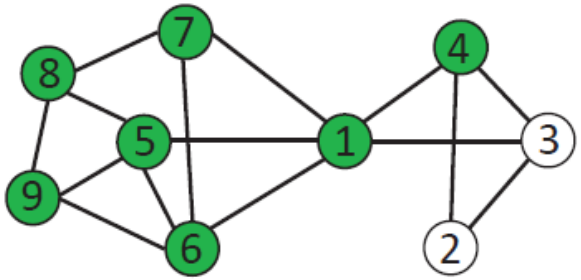

Step 2




Step 3




Step 4




Final Stage

# Influence maximization

- Influence function: for a set of nodes A (target set) the influence s(A) (spread) is the expected number of active nodes at the end of the diffusion process if the item is originally placed in the nodes in A.

- Influence maximization problem [KKT03]: Given an network, a diffusion model, and a value k, identify a set A of k nodes in the network that maximizes s(A).

- The problem is NP-hard

# A Greedy algorithm

- What is a simple algorithm for selecting the set A?

Greedy algorithm

Start with an empty set A

Proceed in k steps

At each step add the node u to the set A the maximizes the increase in function s(A)

- The node that activates the most additional nodes

- Computing s(A): perform multiple Monte-Carlo simulations of the process and take the average.
- How good is the solution of this algorithm compared to the optimal solution?

# Approximation Algorithms

- Suppose we have a (combinatorial) optimization problem, and X is an instance of the problem, OPT(X) is the value of the optimal solution for X, and ALG(X) is the value of the solution of an algorithm ALG for X
  - In our case: X = (G, k) is the input instance, OPT(X) is the spread S(A*) of the optimal solution, GREEDY(X) is the spread S(A) of the solution of the Greedy algorithm
- ALG is a good approximation algorithm if the ratio of OPT and ALG is bounded.

# Approximation Ratio

- For a maximization problem, the algorithm ALG is an $\alpha$-approximation algorithm, for $\alpha < 1$, if for all input instances X,

$$ALG(X) \geq \alpha OPT(X)$$

- The solution of ALG(X) has value at least α% that of the optimal
- α is the approximation ratio of the algorithm
  - Ideally we would like α to be a constant close to 1

# Approximation Ratio for Influence Maximization

- The GREEDY algorithm has approximation ratio $\alpha = 1 - \frac{1}{e}$

$$GREEDY(X) \geq \left(1 - \frac{1}{e}\right) OPT(X), \text{ for all X}$$

# Proof of approximation ratio

- The spread function s has two properties:
- S is monotone:

$$S(A) \leq S(B) \text{ if } A \subseteq B$$

- S is submodular:

$$S(A \cup \{x\}) - S(A) \geq S(B \cup \{x\}) - S(B) \;\; if \;\; A \subseteq B$$

- The addition of node x to a set of nodes has greater effect (more activations) for a smaller set.
  - The diminishing returns property

# Optimizing submodular functions

- Theorem: A greedy algorithm that optimizes a monotone and submodular function S, each time adding to the solution A, the node x that maximizes the gain $S(A \cup \{x\}) - s(A)$has approximation ratio $\alpha = \left(1 - \frac{1}{e}\right)$

- The spread of the Greedy solution is at least 63% that of the optimal

# Submodularity of influence

- Why is S(A) submodular?
  - How do we deal with the fact that influence is defined as an expectation?

- We will use the fact that probabilistic propagation on a fixed graph can be viewed as deterministic propagation over a randomized graph
  - Express S(A) as an expectation over the input graph rather than the choices of the algorithm

# Independent cascade model

- Each edge (u,v) is considered only once, and it is "activated" with probability $p_{uv}$.
- We can assume that all random choices have been made in advance
  - generate a sample subgraph of the input graph where edge (u, v) is included with probability $p_{uv}$
  - propagate the item deterministically on the input graph
  - the active nodes at the end of the process are the nodes reachable from the target set A
- The influence function is obviously(?) submodular when propagation is deterministic
- The linear combination of submodular functions is also a submodular function

# Linear threshold model

- Again, each node may be active or inactive
- Every directed edge (v,u) in the graph has a weight $b_{vu}$, such that

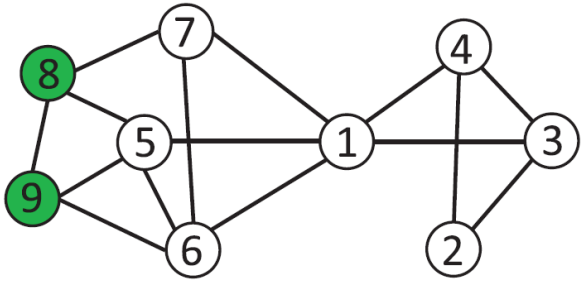
$$\sum_{v \text{ is a neighbor of } u} b_{vu} \leq 1$$

- Each node u has a randomly generated threshold value $T_u$
- Time proceeds in discrete time-steps. At time t an inactive node u becomes active if

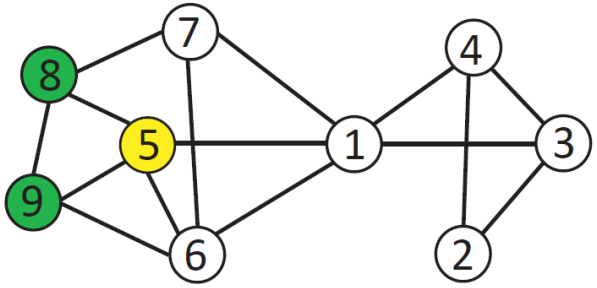
$$\sum_{v \text{ is an active neighbor of } u} b_{vu} \geq T_u$$

- Related to the game-theoretic model of adoption.

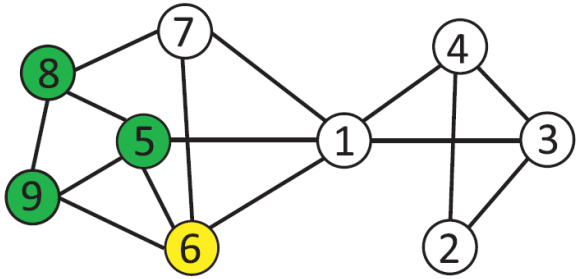
# Linear threshold model



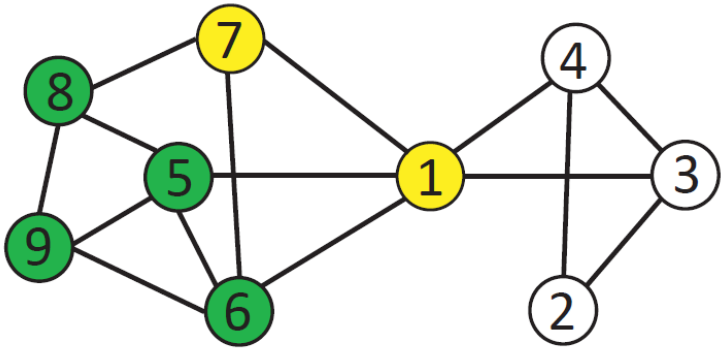

Step 0



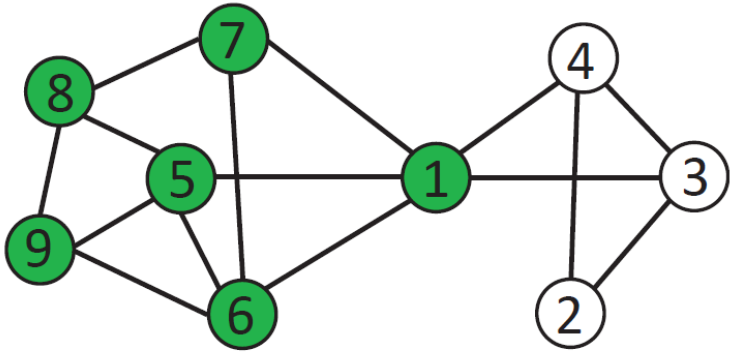

Step 1




Step 2




Step 3




Final Stage

# Influence Maximization

- KKT03 showed that in this case the influence S(A) is still a submodular function, using a similar technique
  - Assumes uniform random thresholds
- The Greedy algorithm achieves a (1-1/e) approximation

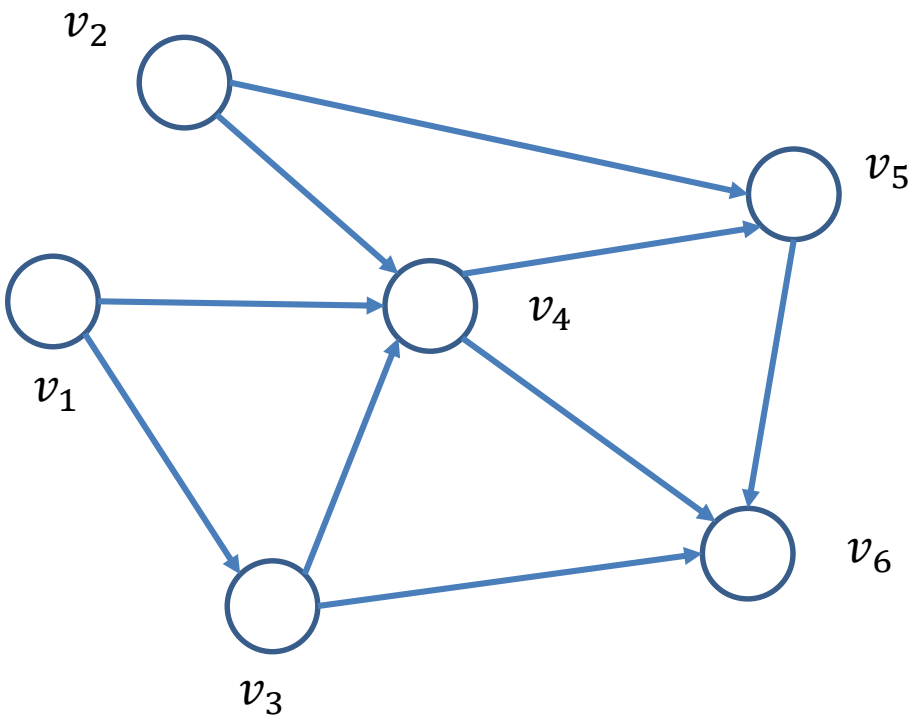
# Proof idea

- For each node $u$, pick one of the edges $(v, u)$ incoming to $u$ with probability $b_{vu}$ and make it live. With probability $1 - \sum b_{vu}$ it picks no edge to make live
- Claim: Given a set of seed nodes A, the following two distributions are the same:
  - The distribution over the set of activated nodes using the Linear Threshold model and seed set A
  - The distribution over the set of reachable nodes from A using live edges.

# Proof idea

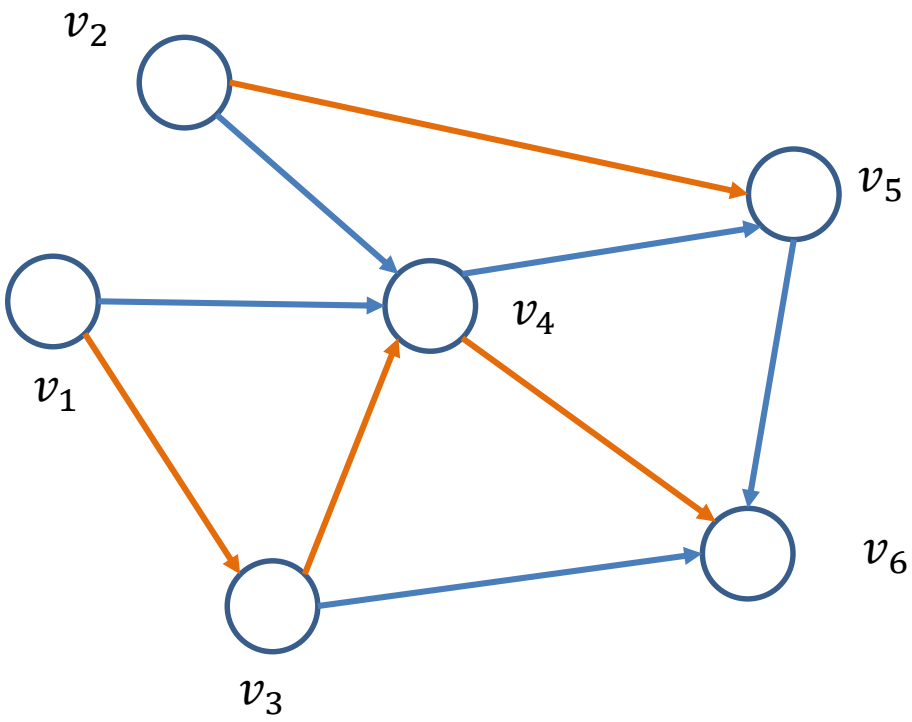
- Consider the special case of a DAG (Directed Acyclic Graph)
  - There is a topological ordering of the nodes $v_0, v_1, \dots, v_n$ such that edges go from left to right
- Consider node $v_i$ in this ordering and assume that $S_i$ is the set of neighbors of $v_i$ that are active.
- What is the probability that node $v_i$ becomes active in either of the two models?
  - In the Linear Threshold model the random threshold $\theta_i$ must be greater than $\sum_{u \in S_i} b_{ui} \geq \theta_i$
  - In the live-edge model we should pick one of the edges in $S_i$
- This proof idea generalizes to general graphs
  - Note: if we know the thresholds in advance submodularity does not hold!

# Example



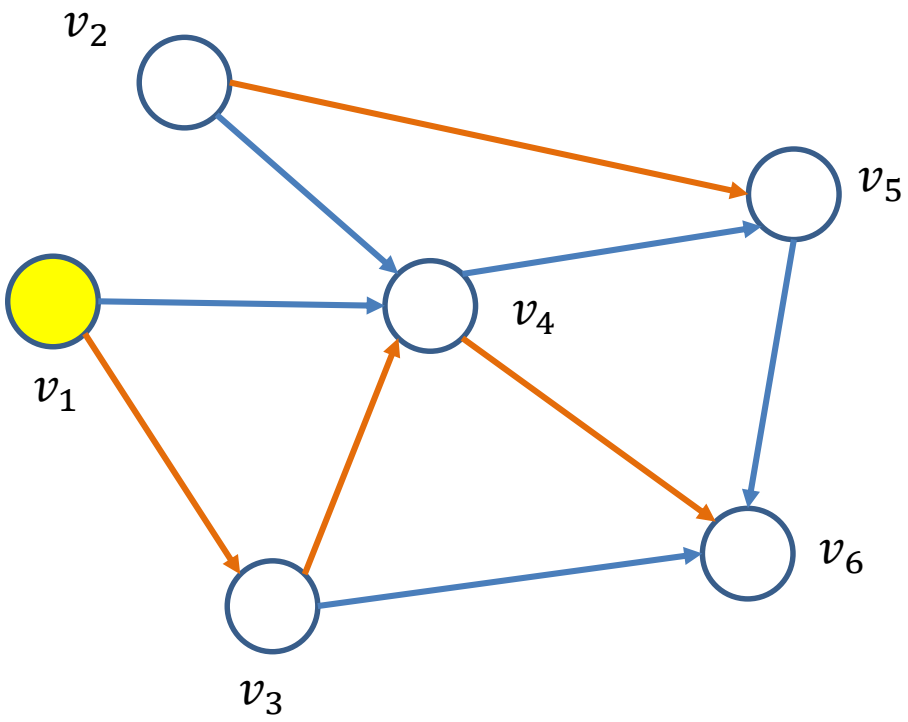

Assume that all edge weights incoming to any node sum to 1

# Example



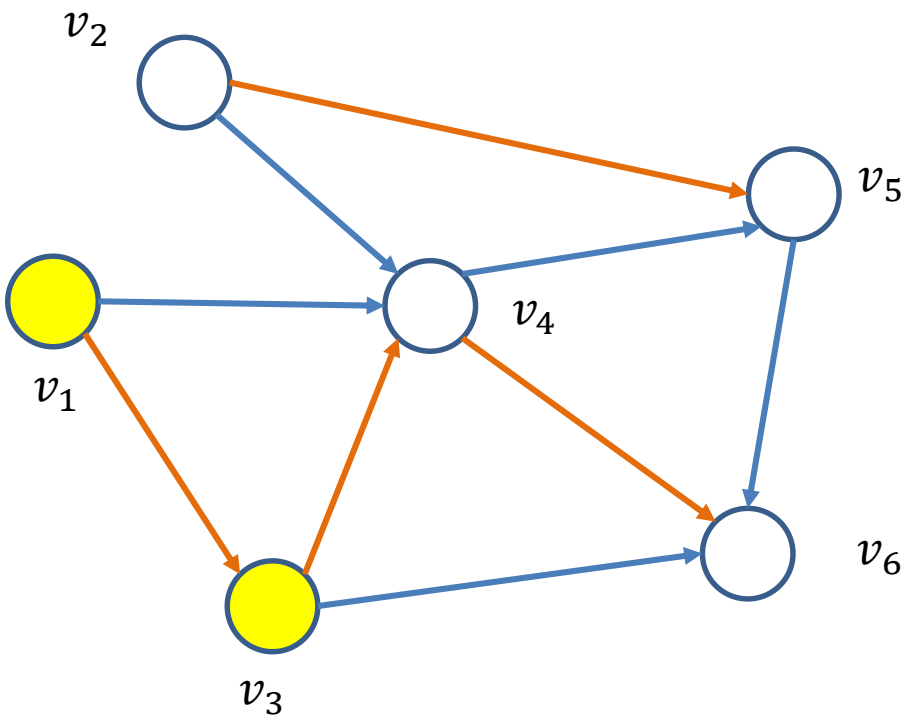

The nodes select a single incoming edge with probability equal to the weight (uniformly at random in this case)

# Example



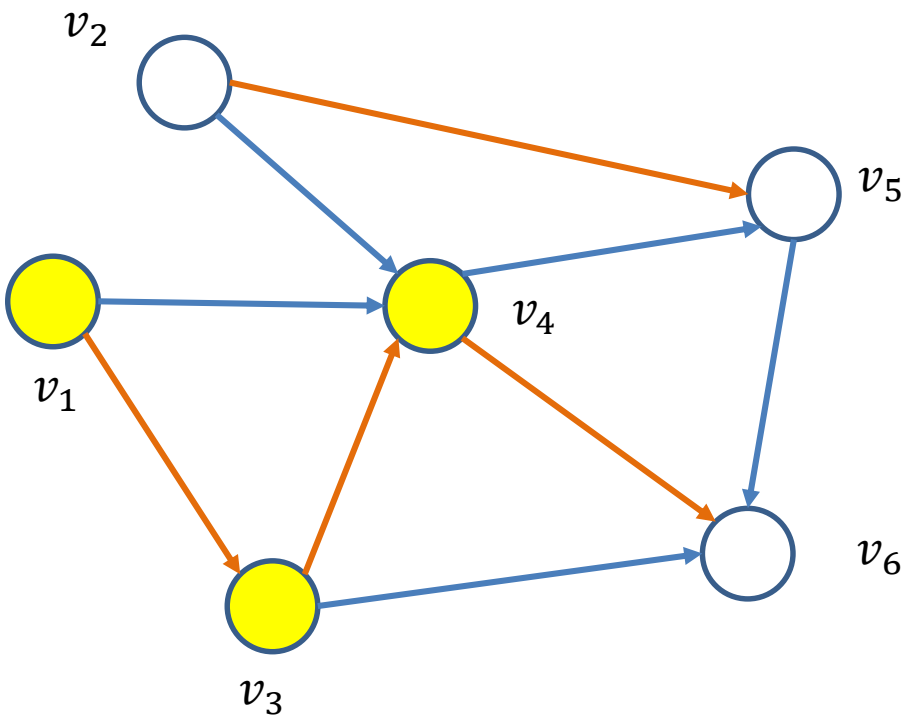

Node $v_1$ is the seed

# Example



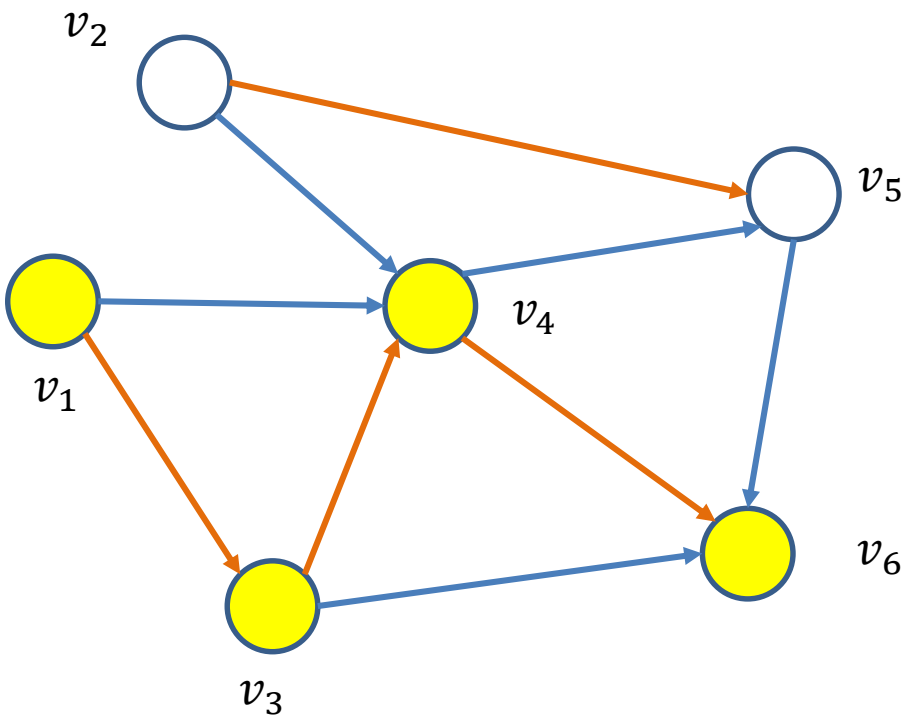

Node $v_3$ has a single incoming neighbor, therefore for any threshold it will be activated

# Example



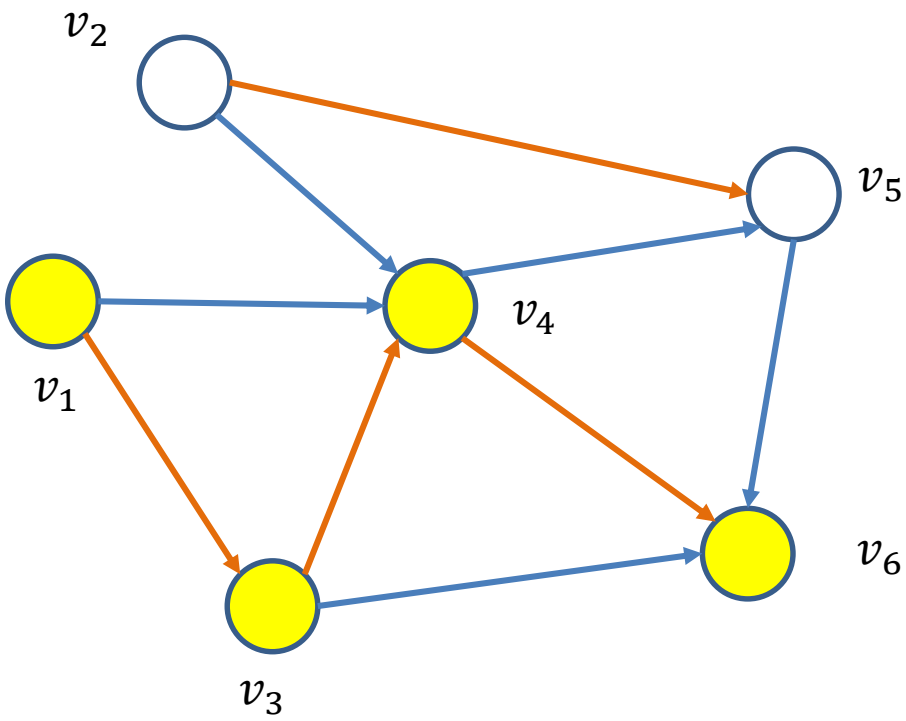

The probability that node $v_4$ gets activated is 2/3 since it has incoming edges from two active nodes.
The probability that node $v_4$ picks one of the two edges to these nodes is also 2/3

# Example




Similarly the probability that node $v_6$ gets activated is 2/3 since it has incoming edges from two active nodes.
The probability that node $v_6$ picks one of the two edges to these nodes is also 2/3

# Example




The set of active nodes is the set of nodes reachable from $v_1$ with live edges (orange).

# Improvements

## Computation of Expected Spread

- Performing simulations for estimating the spread on multiple instances is very slow. Several techniques have been developed for speeding up the process.
  - CELF: exploiting the submodularity property

    (the marginal gain of a node in the current iteration cannot be better than its marginal gain in the previous iteration) J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, N. S. Glance. *Cost-effective outbreak detection in networks*. KDD 2007

  - Maximum Influence Paths: store paths for computation

    W. Chen, C. Wang, and Y. Wang. *Scalable influence maximization for prevalent viral marketing in large-scale social networks*. KDD 2010.

  - Sketches: compute sketches for each node for approximate estimation of spread

    Edith Cohen, Daniel Delling, Thomas Pajor, Renato F. Werneck. *Sketch-based Influence Maximization and Computation: Scaling up with Guarantees*. CIKM 2014

# Experiments

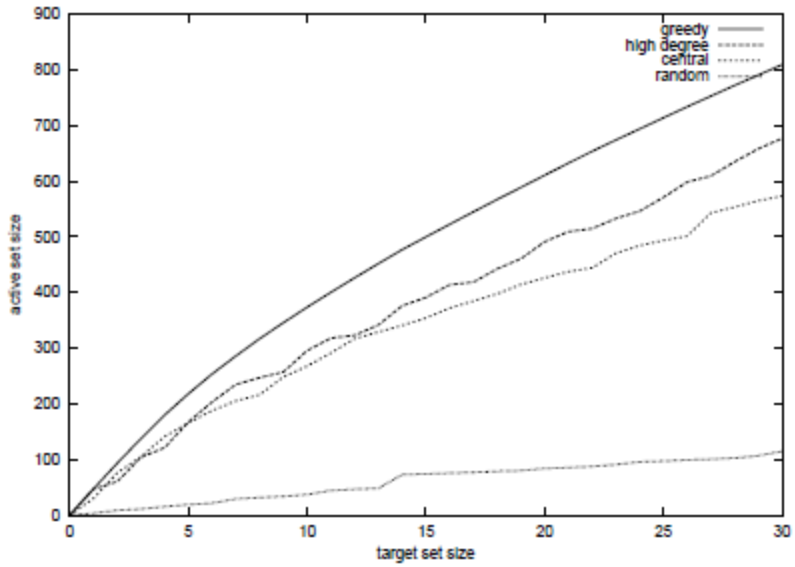



**Figure 2: Results for the weighted cascade model**

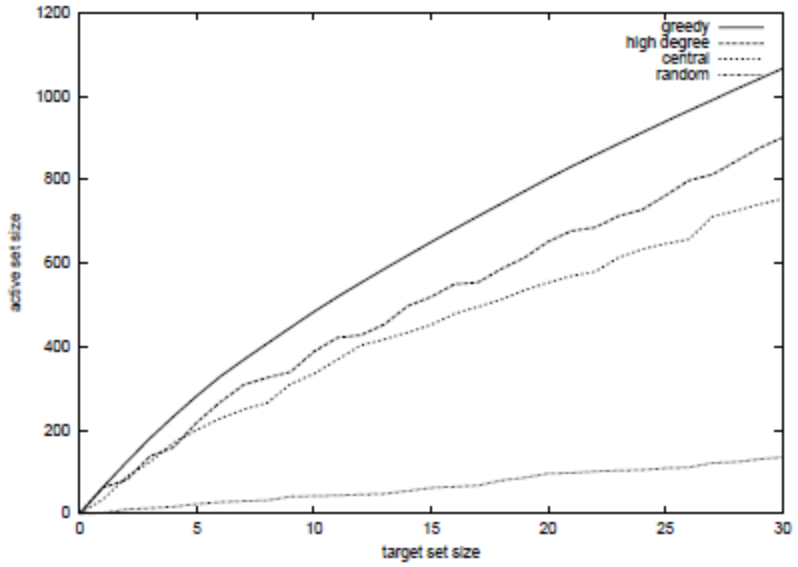



**Figure 1: Results for the linear threshold model**

# One-slide summary

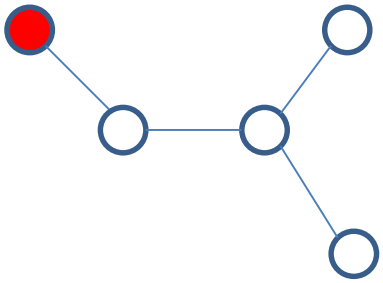
- Influence maximization: Given a graph $G$ and a budget $k$, for some diffusion model, find a subset of $k$ nodes $A$, such that when activating these nodes, the spread of the diffusion $s(A)$ in the network is maximized.
- Diffusion models:
  - Independent Cascade model
  - Linear Threshold model
- Algorithm: Greedy algorithm that adds to the set each time the node with the maximum marginal gain, i.e., the node that causes the maximum increase in the diffusion spread.
- The Greedy algorithm gives a $\left(1-\frac{1}{e}\right)$ approximation of the optimal solution
  - Follows from the fact that the spread function $s(A)$ is
    - Monotone
    - Submodular

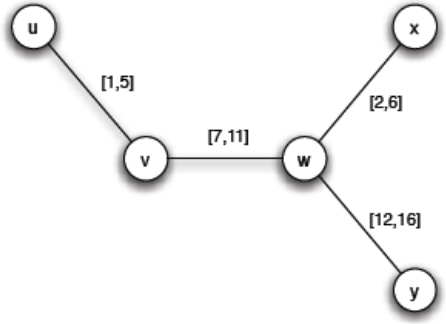
$$s(A) \leq s(B), \text{if } A \subseteq B$$

$$s(A \cup \{x\}) - s(A) \geq s(B \cup \{x\}) - s(B), \forall x \text{ if } A \subseteq B$$

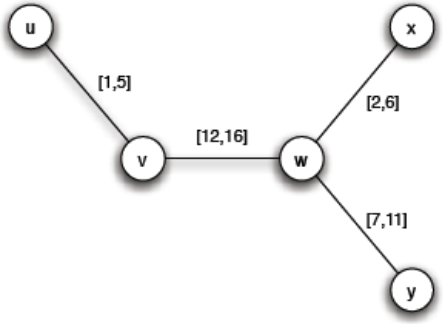
# Another example

- What is the spread from the red node?






(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.*

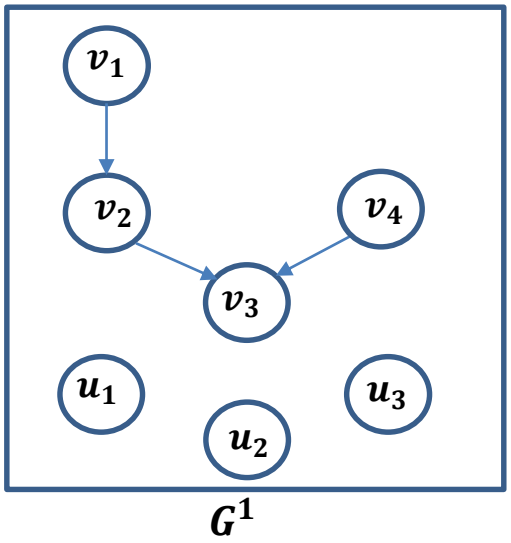
- Inclusion of time changes the problem of influence maximization
  - N. Gayraud, E. Pitoura, P. Tsaparas, Diffusion Maximization on Evolving networks

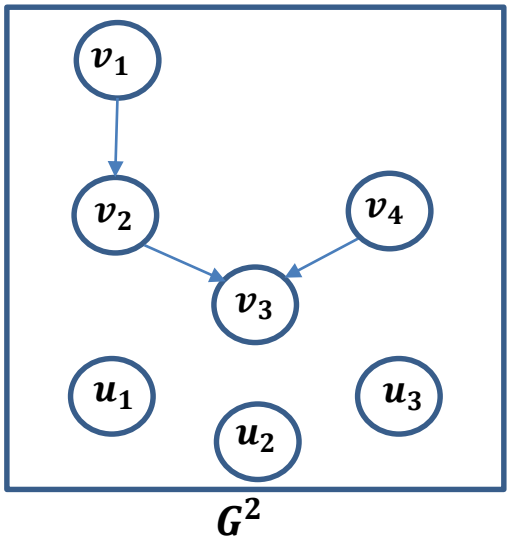
# Evolving network

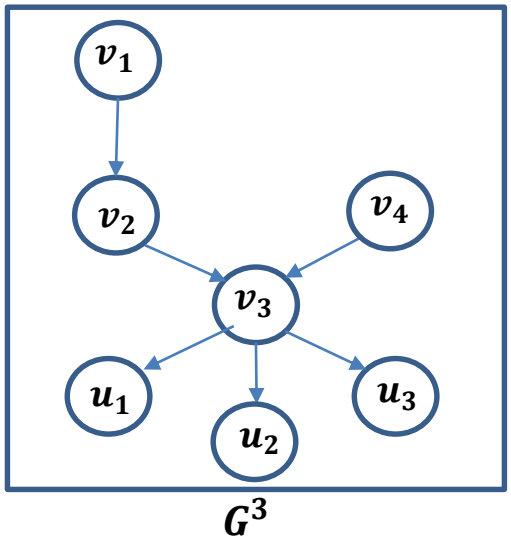
- Consider a network that changes over time
  - Edges and nodes can appear and disappear at discrete time steps
- Model:
  - The evolving network is a sequence of graphs $\{G_1, G_2, \dots, G_n\}$ defined over the same set of vertices $V$, with different edge sets $E_1, E_2, \dots, E_n$
    - Graph snapshot $G_i$ is the graph at time-step $i$ .

N. Gayraud, E. Pitoura, P. Tsaparas. *Maximizing Diffusion in Evolving Networks*. COSN 2015

# Example



# Time

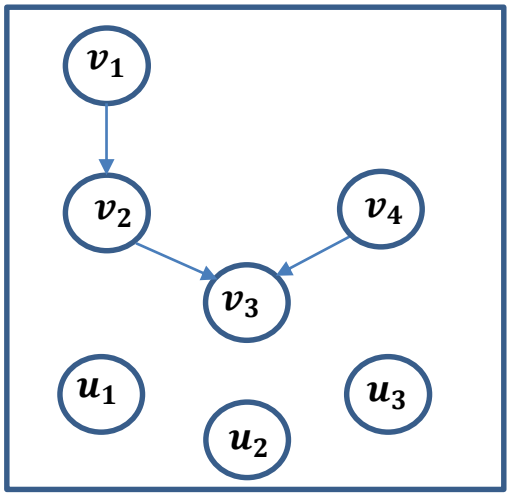
- How does the evolution of the network relates to the evolution of the diffusion?
  - How much physical time does a diffusion step last?
- Assumption: The two processes are in sync. One diffusion step happens in on one graph snapshot
- Evolving IC model: at time-step $t$, the infectious nodes try to infect their neighbors in the graph $G_t$.
- Evolving LT model: at time-step $t$ if the weight of the active neighbors of node $v$ in graph $G_t$ is greater than the threshold the nodes gets activated.

# Submodularity

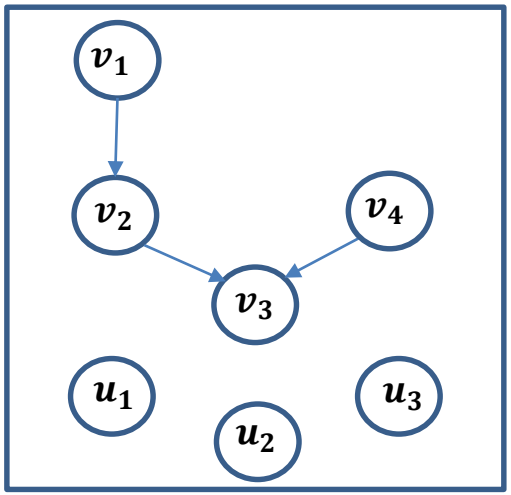
- Will the spread function remain monotone and submodular?

- No!

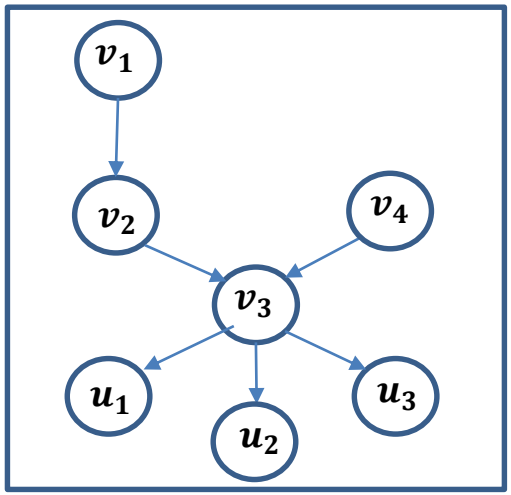
# Monotonicity for the EIC model




$G^1$

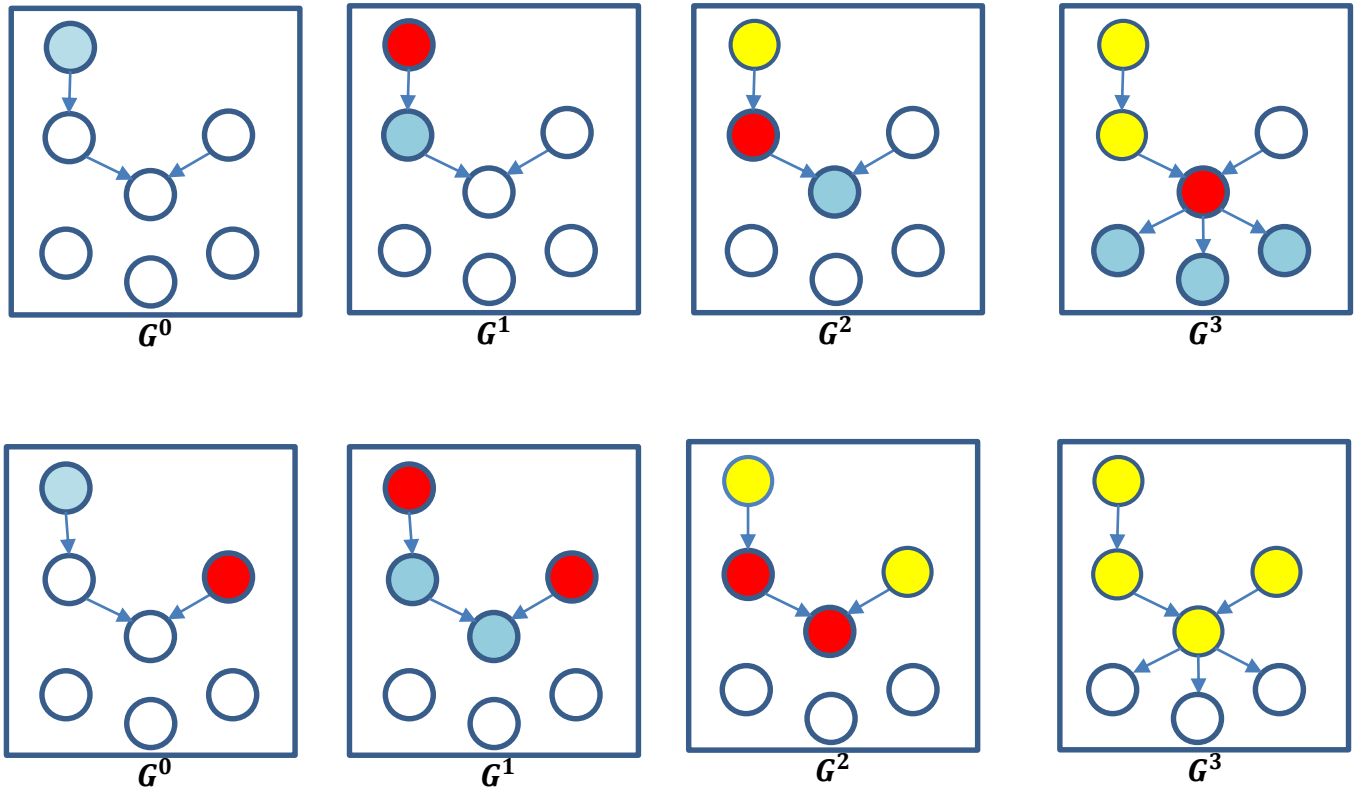



$G^2$



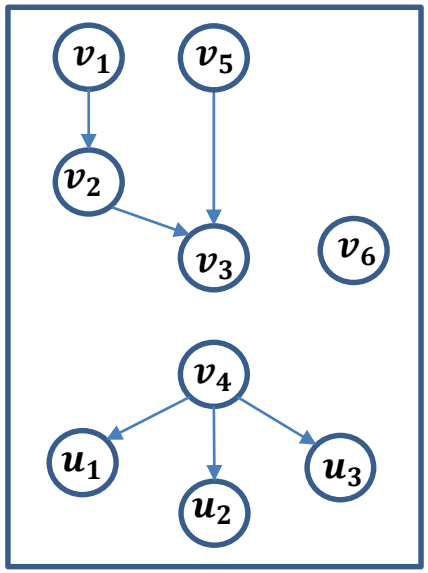

$G^3$

# Monotonicity for the EIC model

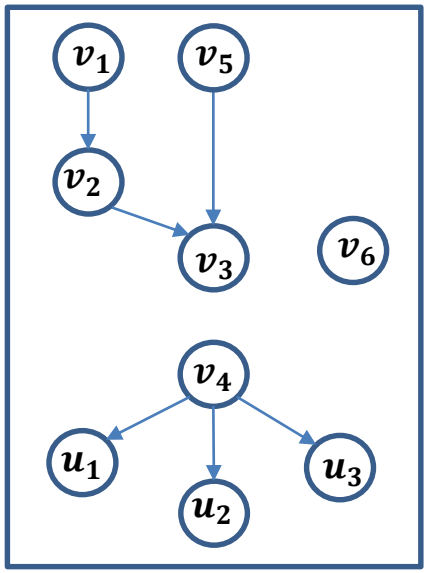



The spread is not monotone in the case of the Evolving IC model

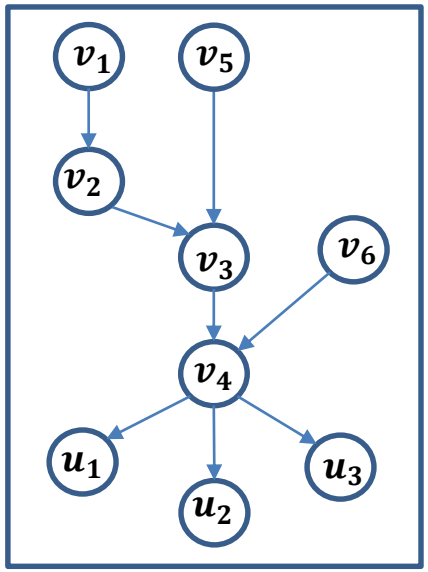
# Submodularity for the EIC model



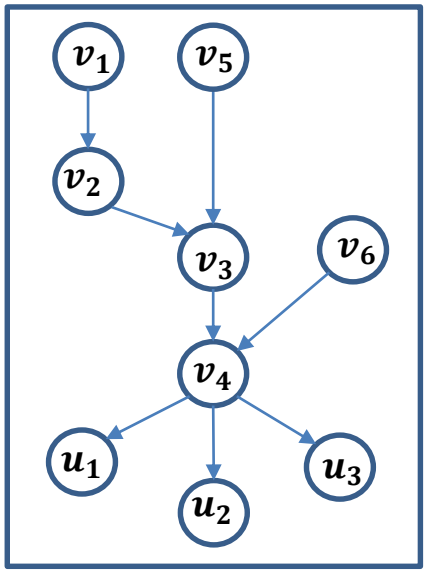

$G^1$




$G^2$

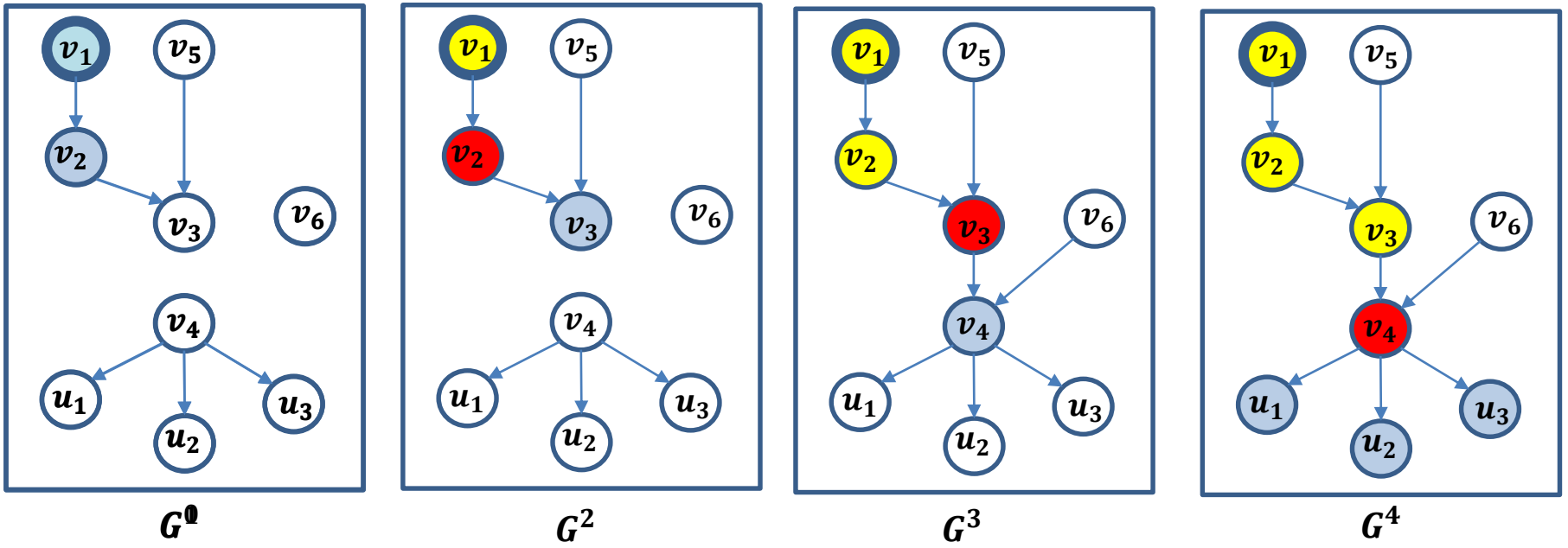



$G^3$



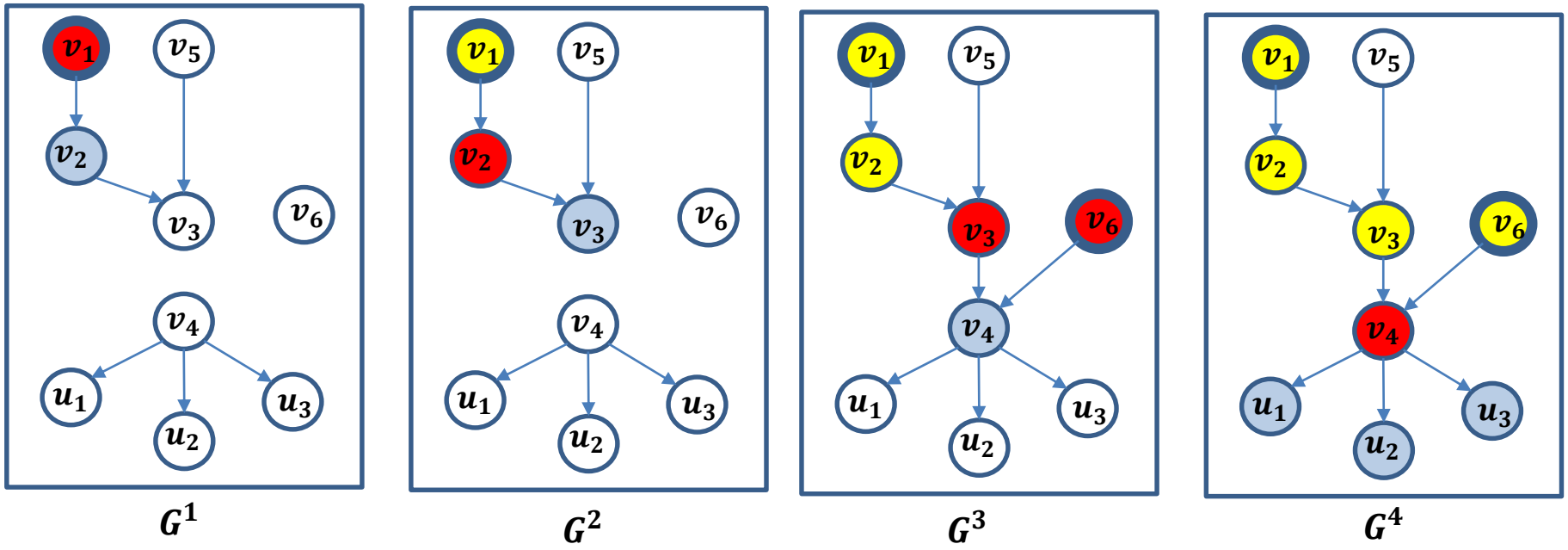

$G^4$

# Submodularity for the EIC model




Activating node $v_1$ at time $t = 0$ has spread 7

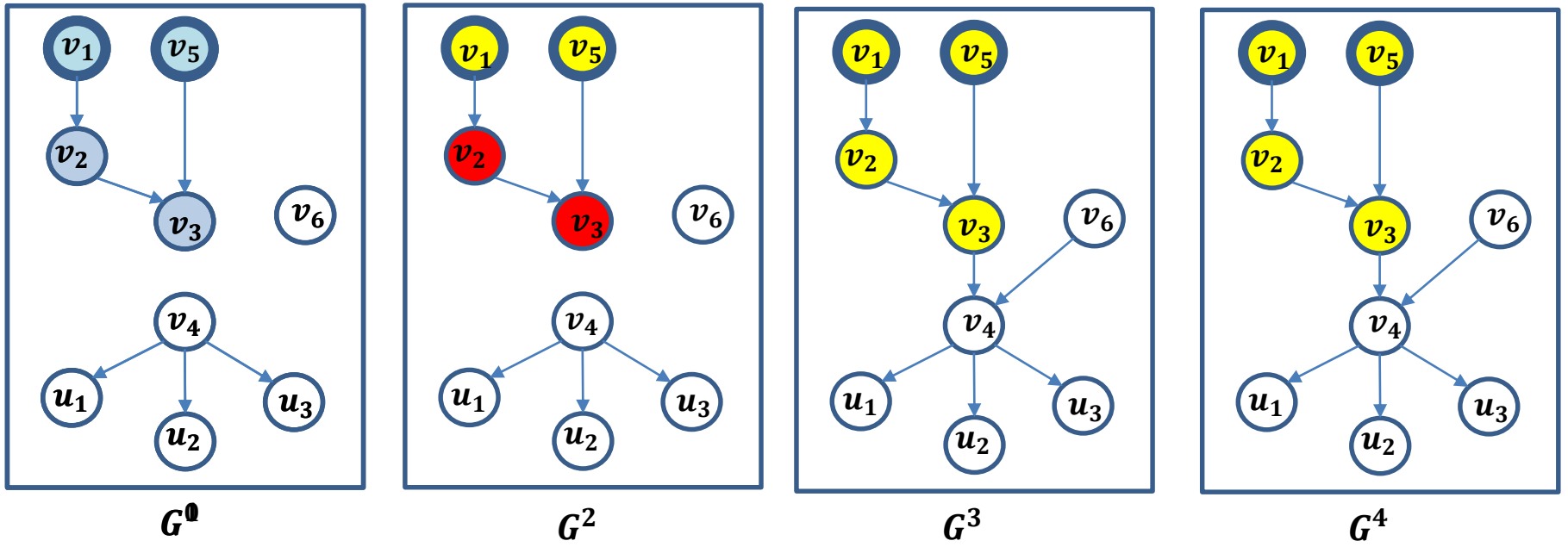
# Submodularity for the EIC model




Activating node $v_1$ at time $t = 0$ has spread 7

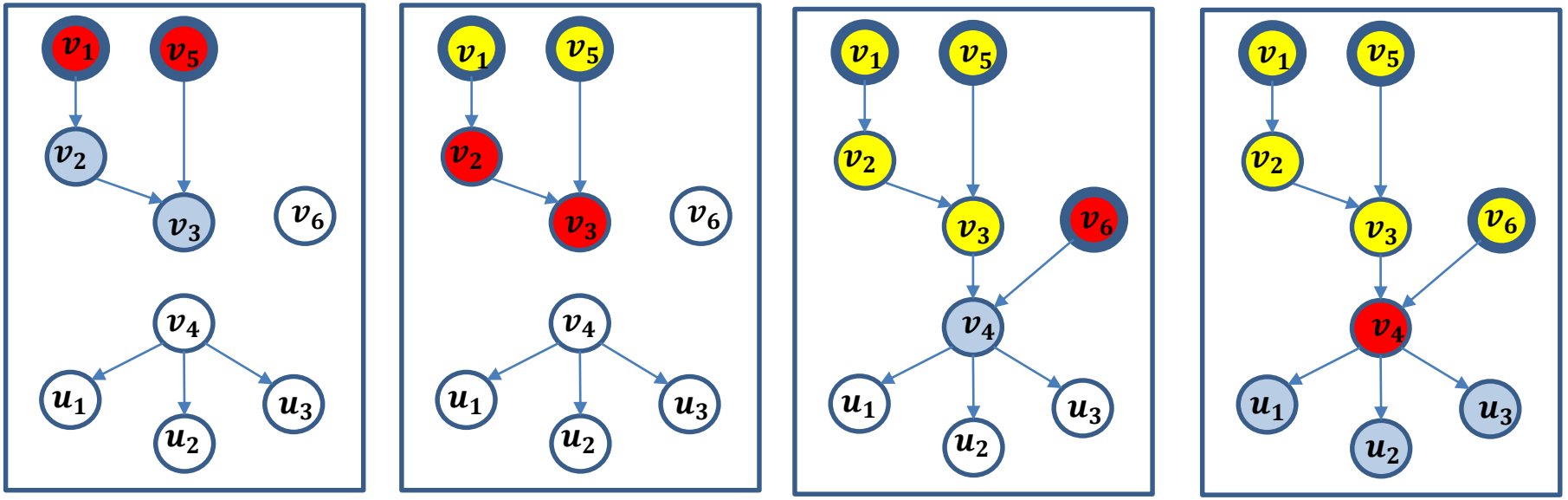
Adding node $v_6$ at time $t = 3$ does not increase the spread

# Submodularity for the EIC model




Activating nodes $v_1$ and $v_5$ at time $t = 0$ has spread 4

# Submodularity for the EIC model

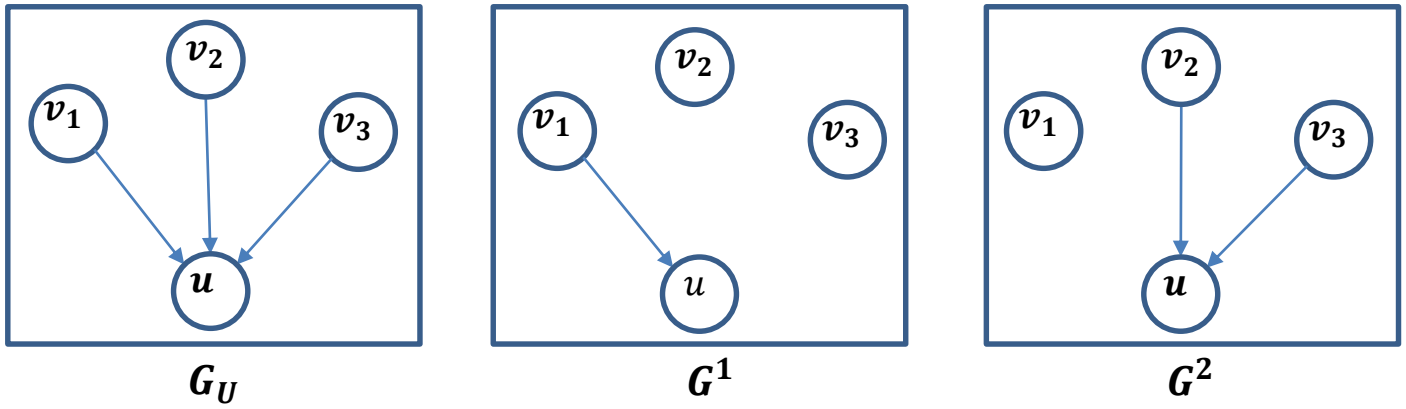



Activating nodes $v_1$ and $v_5$ at time $t = 0$ has spread 4

Adding node $v_6$ at time $t = 3$ increases the spread to 9

# Evolving LT model

- The evolving LT model is monotone but it is not submodular




- Expected Spread: the probability that $u$ gets infected
  - Adding node $v_3$ has a larger effect if added to the set $\{v_1, v_2\}$ than to set $\{v_1\}$.

# Extensions

- Other models for diffusion
  - Deadline model: There is a deadline by which a node can be infected

W. Chen, W. Lu, N. Zhang. *Time-critical influence maximization in social networks with time-delayed diffusion process*. AAAI, 2012.

  - Time-decay model: The probability of an infected node to infect its neighbors decays over time

B. Liu, G. Cong, D. Xu, and Y. Zeng. *Time constrained influence maximization in social networks.* ICDM 2012.

  - Timed influence: Each edge has a speed of infection, and you want to maximize the speed by which nodes are infected.

N. Du, L. Song, M. Gomez-Rodriguez, H. Zha. *Scalable influence estimation in continuous-time diffusion networks*. NIPS 2013.

- Competing diffusions
  - Maximize the spread while competing with other products that are being diffused.

A. Borodin, Y. Filmus, and J. Oren. *Threshold models for competitive influence in social networks*. WINE, 2010.
M. Draief and H. Heidari. M. Kearns. *New Models for Competitive Contagion.* AAAI 2014.

# Extensions

- Reverse problems:
  - Initiator discovery: Given the state of the diffusion, find the nodes most likely to have initiated the diffusion

    H. Mannila, E. Terzi. *Finding Links and Initiators: A Graph-Reconstruction Problem*. SDM 2009

  - Diffusion trees: Identify the most likely tree of diffusion tree given the output

    M. Gomez Rodriguez, J. Leskovec, A. Krause. *Inferring networks of diffusion and influence*. KDD 2010

  - Infection probabilities: estimate the true infection probabilities

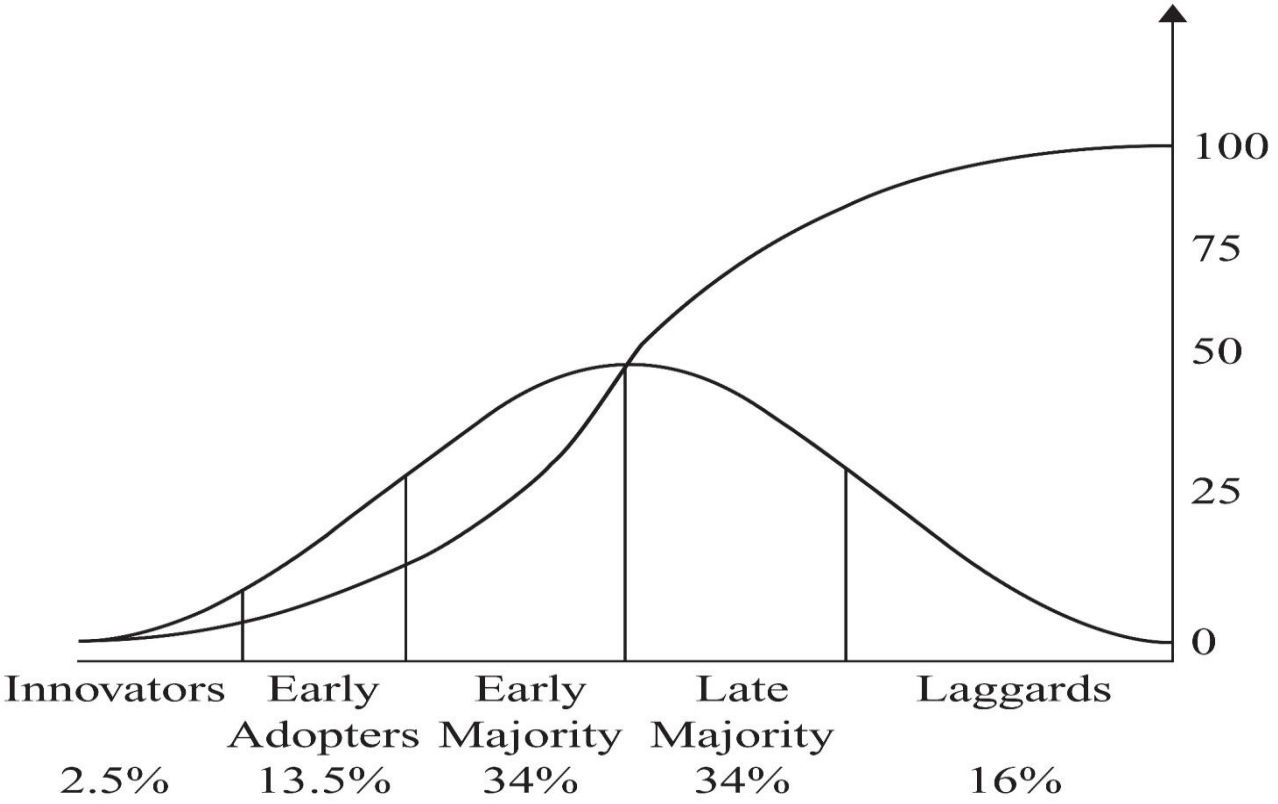
    M. Gomez-Rodriguez, D. Balduzzi, B. Scholkopf. *Uncovering the temporal dynamics of diffusion networks*. ICML, 2011.

# References


- D. Kempe, J. Kleinberg, E. Tardos. *Maximizing the Spread of Influence through a Social Network*. Proc. 9th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2003.
- N. Gayraud, E. Pitoura, P. Tsaparas. *Maximizing Diffusion in Evolving Networks*. ICCSS 2015
- J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. M. VanBriesen, Natalie S. Glance. *Cost-effective outbreak detection in networks*. KDD 2007
- W. Chen, C.Wang, and Y.Wang. *Scalable influence maximization for prevalent viral marketing in large-scale social networks*. In 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD 2010.
- B. Liu, G. Cong, D. Xu, and Y. Zeng. *Time constrained influence maximization in social networks.* ICDM 2012.
- Edith Cohen, Daniel Delling, Thomas Pajor, Renato F. Werneck. *Sketch-based Influence Maximization and Computation: Scaling up with Guarantees*. CIKM 2014
- W. Chen, W. Lu, N. Zhang. *Time-critical influence maximization in social networks with time-delayed diffusion process*. AAAI, 2012.
- N. Du, L. Song, M. Gomez-Rodriguez, H. Zha. *Scalable influence estimation in continuous-time diffusion networks*. NIPS 2013.
- A. Borodin, Y. Filmus, and J. Oren. *Threshold models for competitive influence in social networks*. In Proceedings of the 6th international conference on Internet and network economics, WINE'10, 2010.
- M. Draief and H. Heidari. M. Kearns. *New Models for Competitive Contagion.* AAAI 2014.
- H. Mannila, E. Terzi. *Finding Links and Initiators: A Graph-Reconstruction Problem*. SDM 2009
- Manuel Gomez Rodriguez, Jure Leskovec, Andreas Krause. *Inferring networks of diffusion and influence*. KDD 2010
- M. Gomez-Rodriguez, D. Balduzzi, B. Scholkopf. *Uncovering the temporal dynamics of diffusion networks*. ICML, 2011.

# EXTRA SLIDES

# Innovation Adoption Characteristics




Category of Adopters in the corn study

# Multiple copies model

- Each node may have multiple copies of the same virus
  - $\boldsymbol{v}$: state vector : $v_i$ : number of virus copies at node $i$

- At time $t = 0$, the state vector is initialized to $\boldsymbol{v}^0$
- At time t,

  For each node i

    For each of the $v_i^t$ virus copies at node $i$

      the copy is copied to a neighbor $j$ with prob $p$

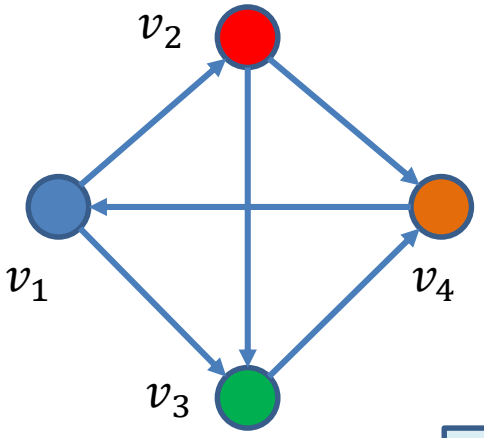
      the copy dies with probability $q$

G. Giakkoupis, A. Gionis, E. Terzi, P. T. Models and algorithms for network immunization. Technical Report C-2005-75, Department of Computer Science, University of Helsinki, 2005

# Analysis

- The expected state of the system at time t is given by

$$\overline{\boldsymbol{v^t}} = (p\boldsymbol{A} + (1-q)\boldsymbol{I})\overline{\boldsymbol{v^{t-1}}} = \boldsymbol{M}\overline{\boldsymbol{v^{t-1}}}$$




$$M = \begin{bmatrix} 1-q & p & p & 0 \\ 0 & 1-q & p & p \\ 0 & 0 & 1-q & p \\ p & 0 & 0 & 1-q \end{bmatrix}$$

Probability that the copy from node $v_4$is copied to node $v_1$

Probability that the copy from node $v_4$ survives at $v_4$

# Analysis

- As $t \to \infty$
  - if $\lambda_1(M) < 1 \Leftrightarrow \lambda_1(A) < q/p$ then $\overline{v^t} \to 0$
    - the probability that all copies die converges to 1
  - if $\lambda_1(M) = 1 \Leftrightarrow \lambda_1(A) = q/p$ then $\overline{v^t} \to c$
    - the probability that all copies die converges to 1
  - if $\lambda_1(M) > 1 \Leftrightarrow \lambda_1(A) > q/p$ then $\overline{v^t} \to \infty$
    - the probability that all copies die converges to a constant < 1