Online Social Networks and Media

Cascading Behavior in Networks
Epidemic Spread
Influence Maximization

CASCADING BEHAVIOR IN NETWORKS

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

Information effect: choices made by others can provide indirect information about what they know

Old studies:

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

Basic observations:

- Characteristics of early adopters
- Decisions made in the context of social structure

Diffusion in Networks

Direct-benefit Effect: there are direct payoffs from copying the decisions of others

Spread of technologies such as the phone, email, etc

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?)

Modeling Diffusion through a Network

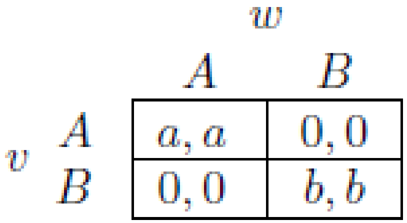
An *individual* level model of *direct-benefit effects* in networks due to S. Morris

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, than each get a payoff 0

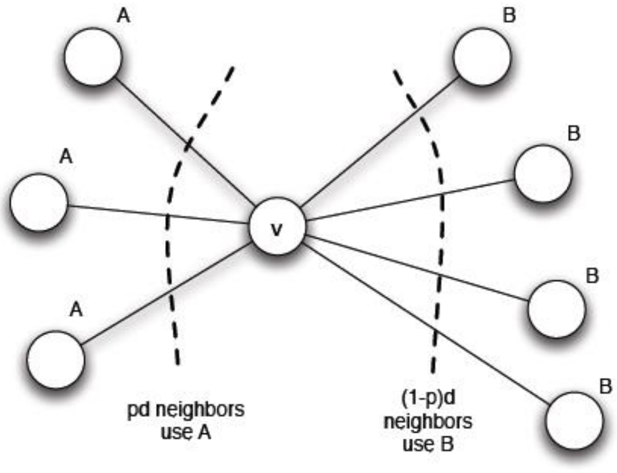


		w	
		A	B
v	A	a, a	$0, 0$
	B	$0, 0$	b, b

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 some of its neighbors adopt A and some B, what should *u* do to maximize its payoff?



Threshold $q = b/(a+b)$ for preferring A
(at least q of the neighbors follow A)

Modeling Diffusion through a Network: Cascading Behavior

Two obvious equilibria, which ones?

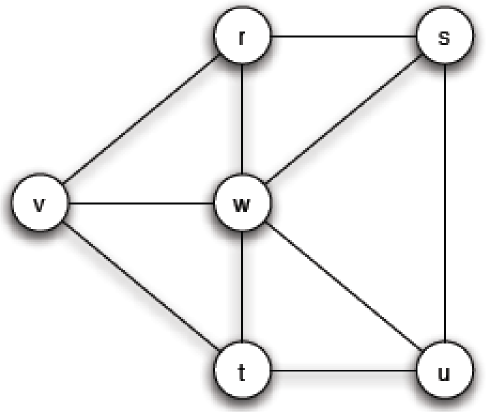
Suppose that initially everyone is using B as a default behavior
A small set of "initial adopters" decide to use A

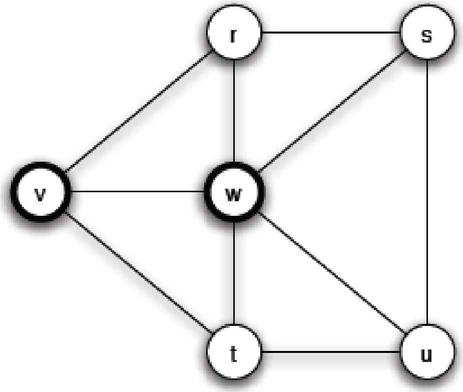
✓When will this result in everyone eventually switching to A?
✓If this does not happen, what causes the spread of A to stop?

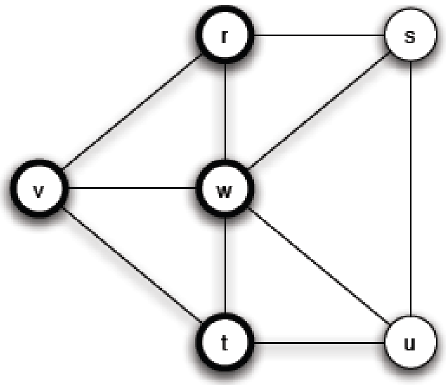
Observation: strictly progressive sequence of switches from A to B

Modeling Diffusion through a Network: Cascading Behavior

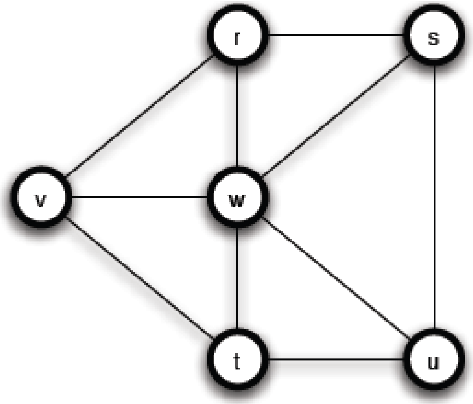
a = 3, *b* = 2, *q* = 2/5







Step 1

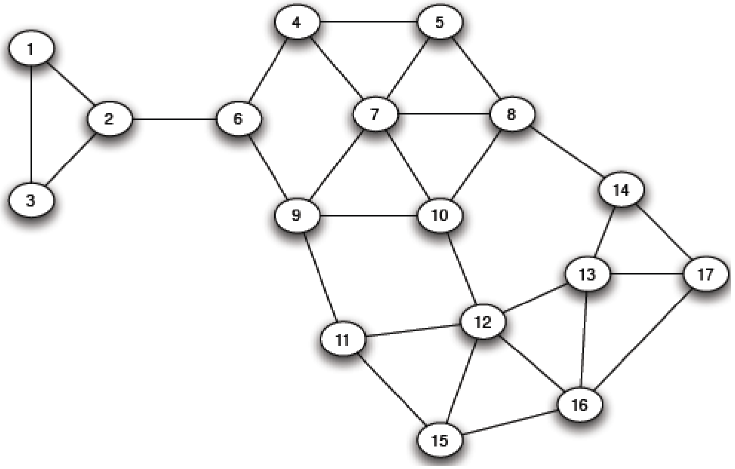


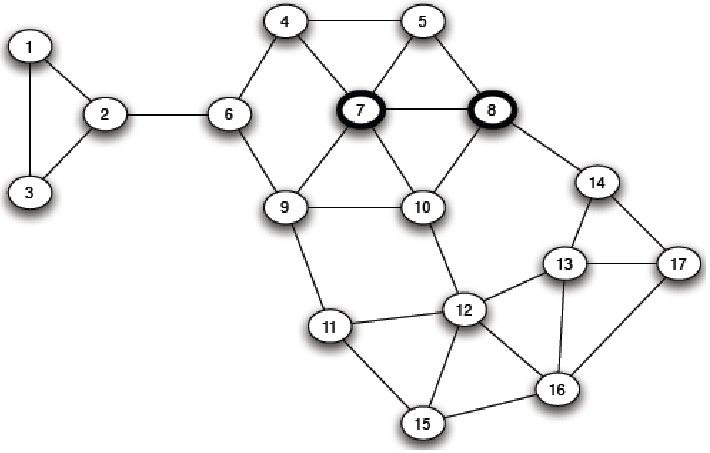
Step 2

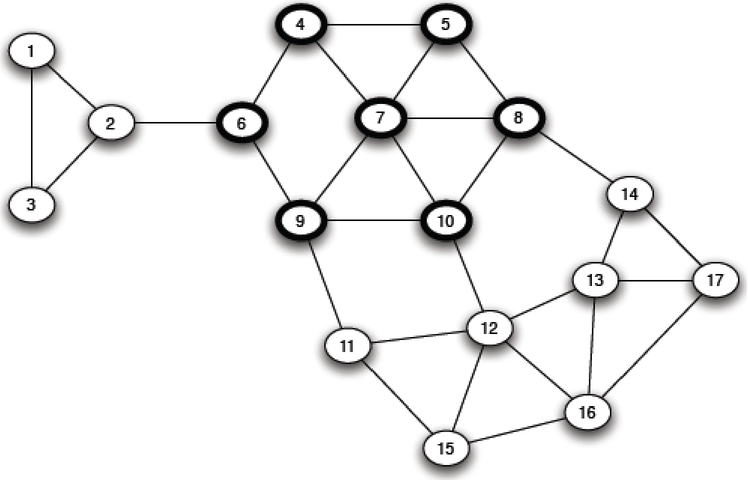
Chain reaction

Modeling Diffusion through a Network: Cascading Behavior

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







Step 3

Modeling Diffusion through a Network: Cascading Behavior

Chain reaction of switches to A -> a cascade of adoptions of A

1. Consider a set of initial adopters who start with a new behavior A, while every other node starts with behavior B.
2. Nodes then 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

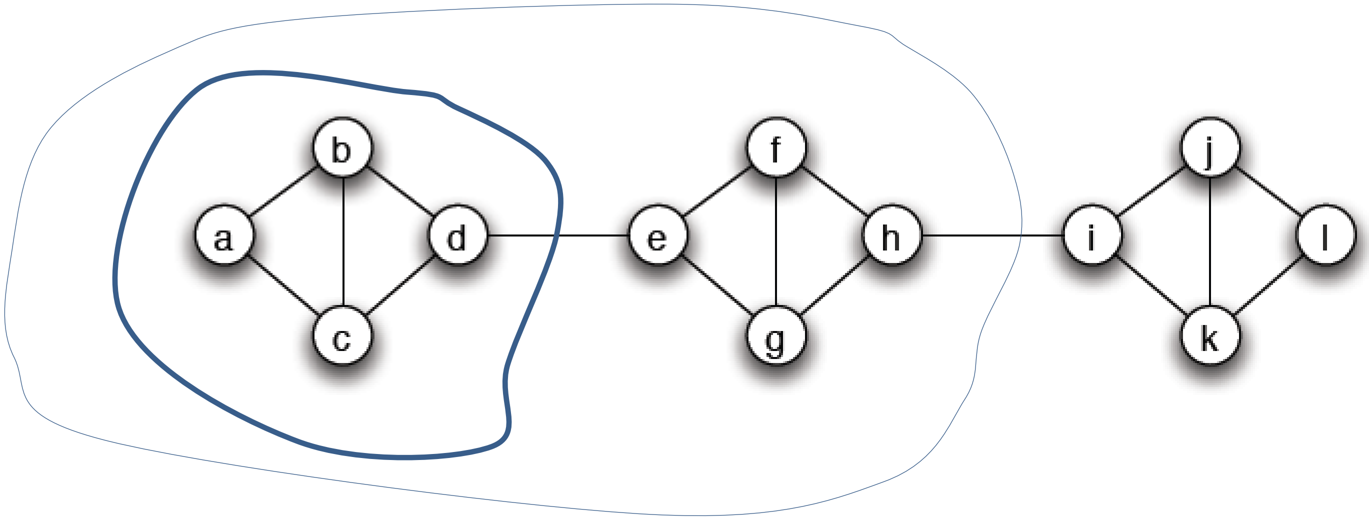
(examples, age groups and life-styles in social networking sites, Mac users, political opinions)

Strategies

- Improve the quality of A (increase the payoff *a*)
- Convince a small number of *key people* to switch to A

Cascades and Clusters

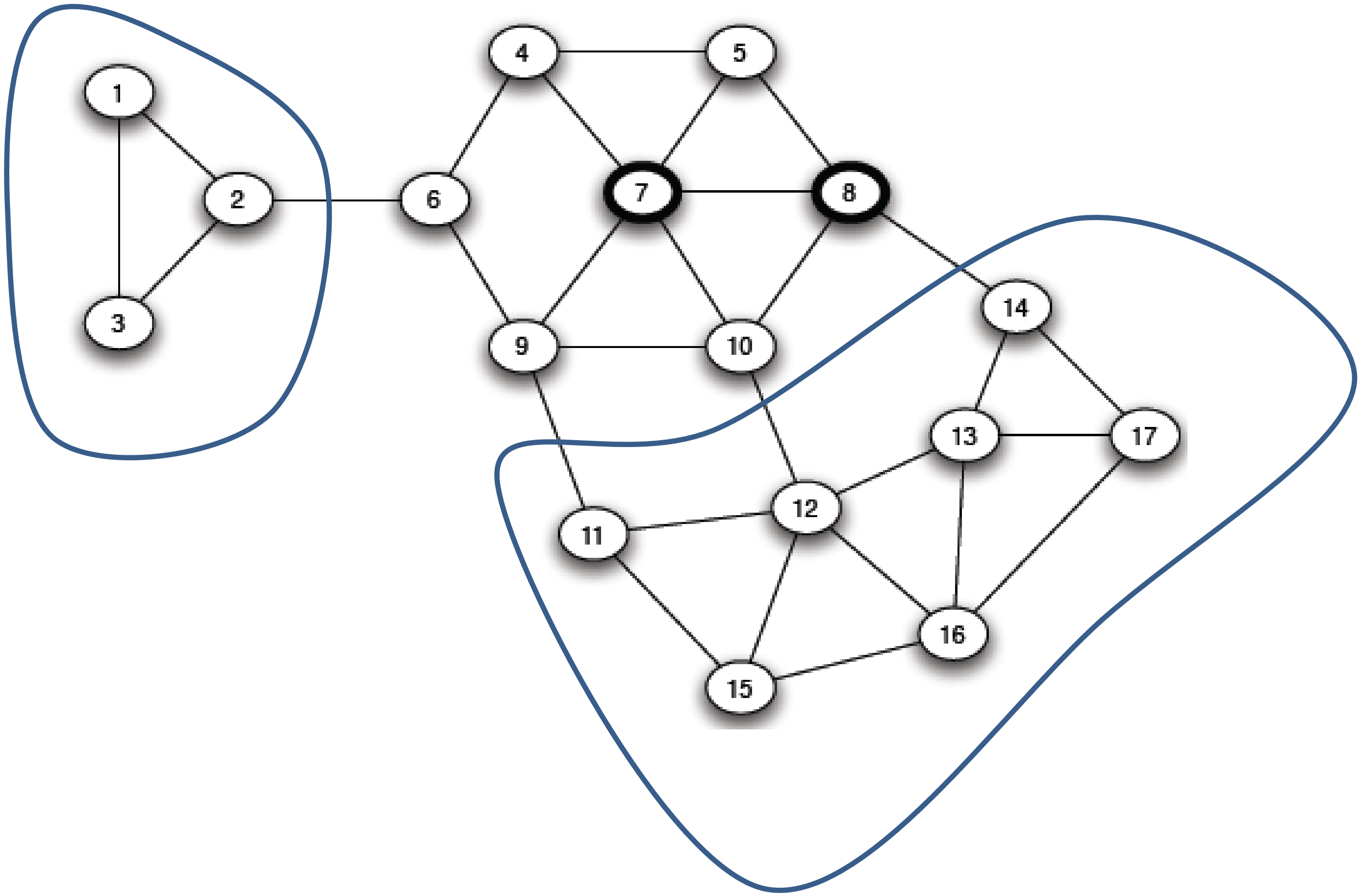
A cluster of density *p* is a set of nodes such that each node in the set has at least a *p* fraction of its neighbors in the set



Ok, but it does not imply that any two nodes in the same cluster necessarily have much in common

The union of any two cluster of density *p* is also a cluster of density *p*

Cascades and Clusters



Cascades and Clusters

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

(i) (clusters as obstacles to cascades)

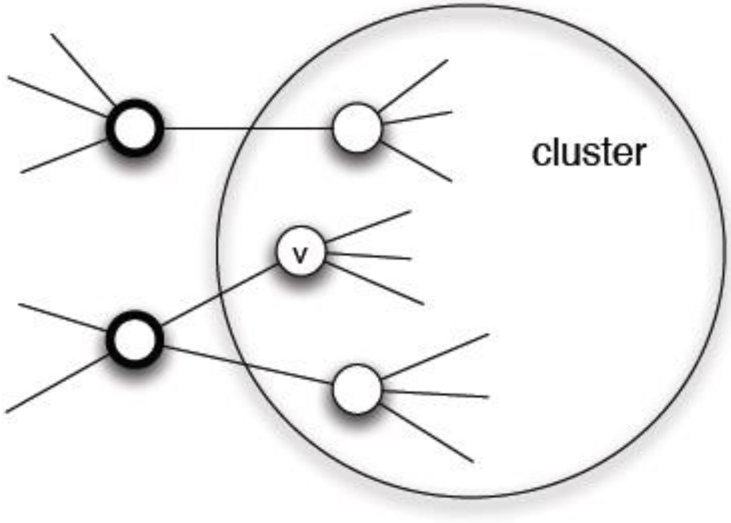
If the remaining network contains a cluster of density greater than $1 - q$, then the set of initial adopters will not cause a complete cascade.

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

Whenever a set of initial adopters does not cause a complete cascade with threshold q, the remaining network must contain a cluster of density greater than $1 - q$.

Cascades and Clusters

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

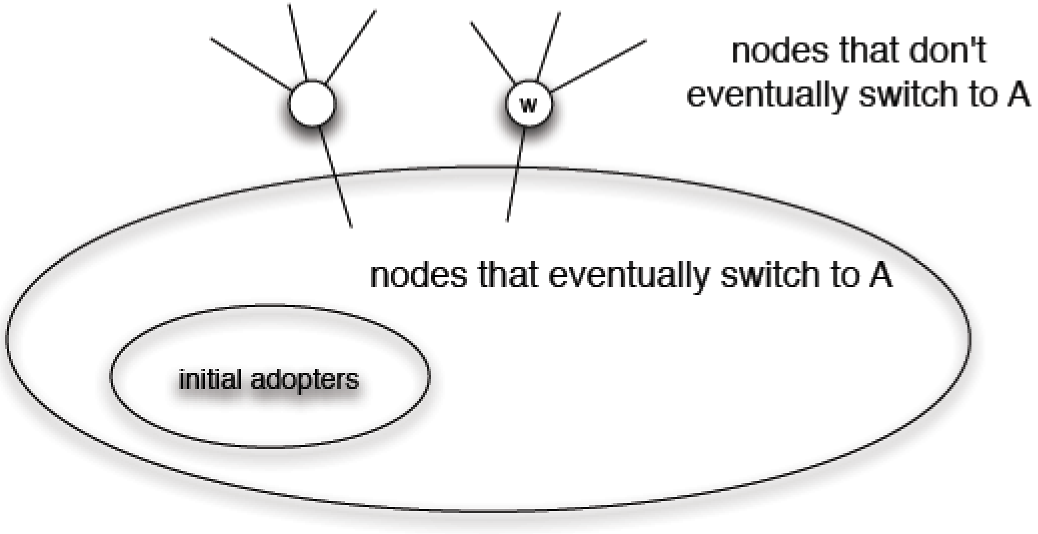


Proof by contradiction

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

Cascades and Clusters

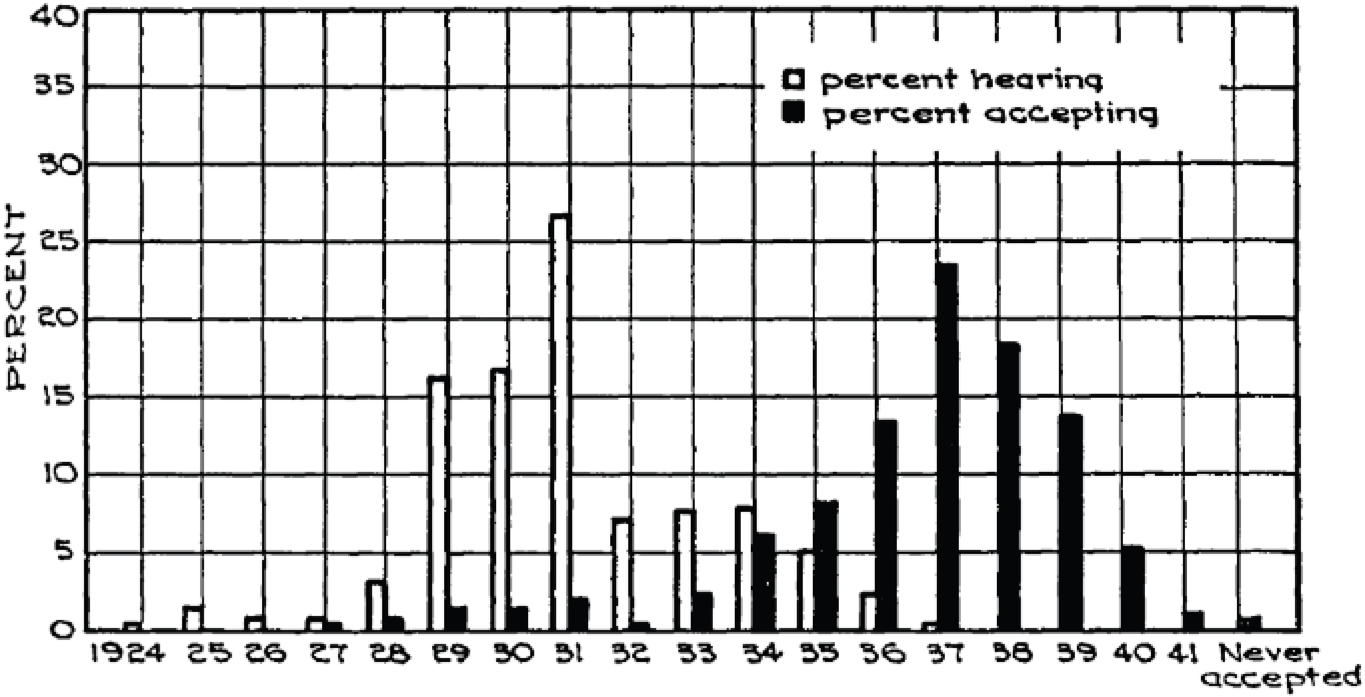
Proof of (ii) (clusters are the only obstacles to cascades)


w
nodes that don't eventually switch to A
nodes that eventually switch to A
initial adopters

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

Diffusion, Thresholds and the Role of Weak Ties

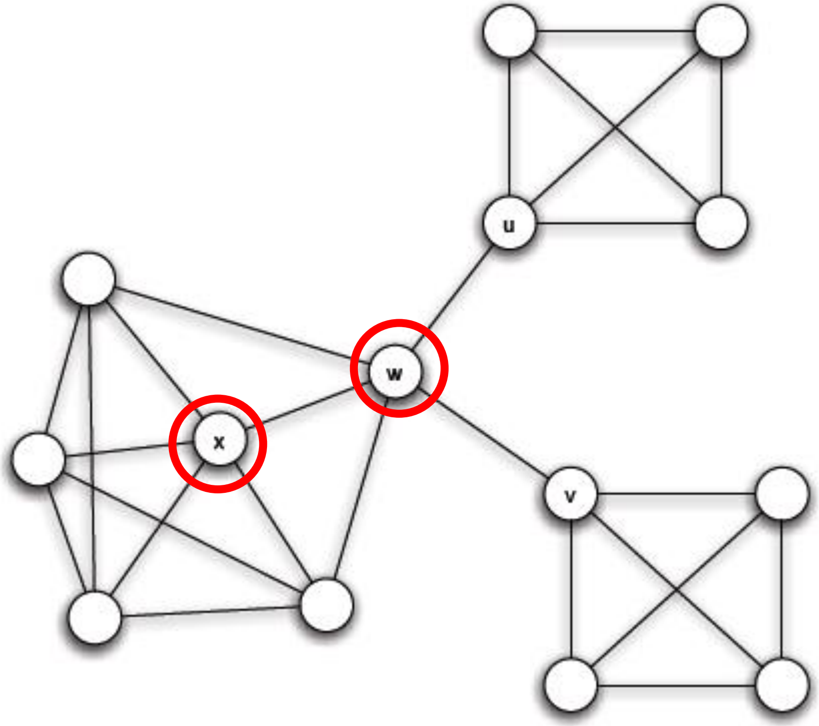
A crucial difference between learning a new idea and actually deciding to accept it



Diffusion, Thresholds and the Role of Weak Ties

Relation to weak ties and local bridges

$q = 1/2$

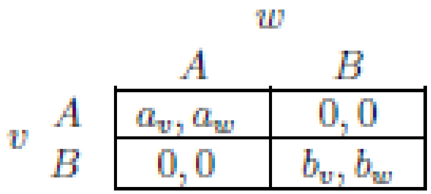


Bridges convey awareness but weak at transmitting costly to adopt behaviors

Extensions of the Basic Cascade Model:
Heterogeneous Thresholds

Each person values behaviors A and B differently:

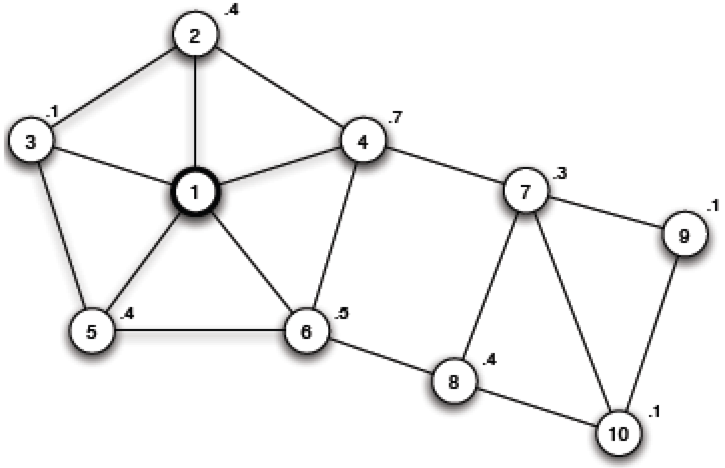
- If both *u* and *w* adapt A, *u* gets a payoff $a_u > 0$ and *w* a payoff $a_w > 0$
- If both *u* and *w* adapt B, u gets a payoff $b_u > 0$ and w a payoff $b_w > 0$
- If opposite behaviors, than each gets a payoff 0

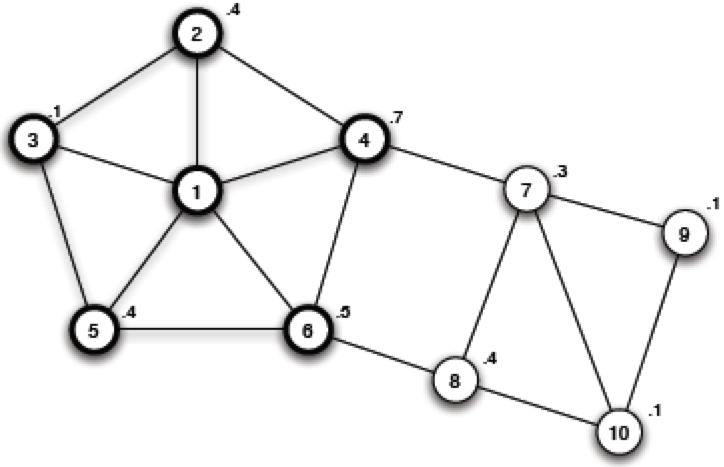


$$\begin{array}{cc|c|c|} & & \multicolumn{2}{c}{w} \\ & & A & B \\ \hline \multirow{2}{*}{v} & A & a_v, a_w & 0, 0 \\ \hline & B & 0, 0 & b_v, b_w \\ \hline \end{array}$$

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:

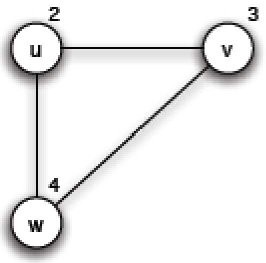
Collective Action and Pluralistic Ignorance

A collective action problem: an activity produces benefits only if enough people participate

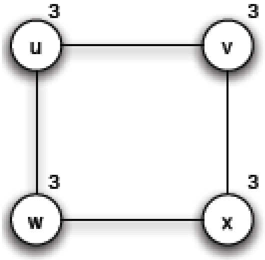
Pluralistic ignorance: a situation in which people have wildly erroneous estimates about the prevalence of certain opinions in the population at large

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

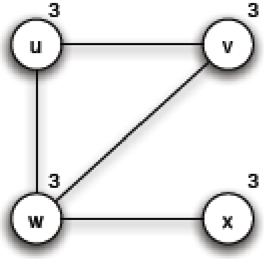
- Each person has a personal threshold which encodes her willingness to participate
- A threshold of *k* means that she will participate if at least *k* people in total (including herself) will participate
- Each person in the network knows the thresholds of her neighbors in the network



- 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

- Social networks do not simply allow or 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*

The Cascade Capacity

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

Cascade capacity of the network

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

The Cascade Capacity: Cascades on Infinite Networks

- Initially, **a finite set S** of nodes has behavior A and all others adopt B

- Time runs forwards in steps, t = 1, 2, 3, …

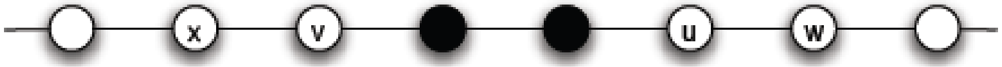
- In each step t, each node other than those in S uses the decision rule with threshold q to decide whether to adopt behavior A or B

- 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.

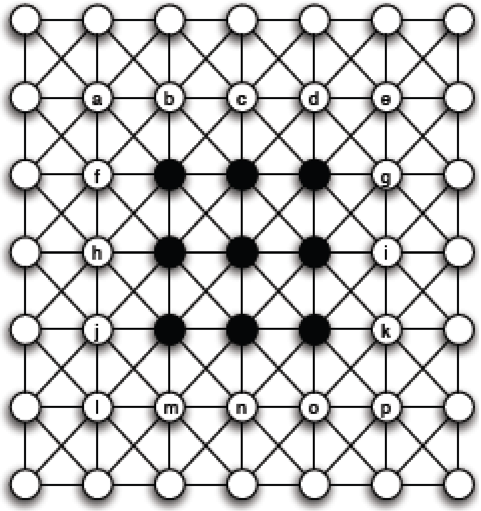
The cascade capacity of the network is the largest value of the threshold q for which some finite set of early adopters can cause a complete cascade.

The Cascade Capacity: Cascades on Infinite Networks

An infinite path



Spreads if ≤ 1/2



An infinite grid

Spreads if ≤ 3/8

✓An intrinsic property of the network

✓Even if A better, for *q* strictly between 3/8 and ½, A cannot win

The Cascade Capacity: Cascades on Infinite Networks

How large can a cascade capacity be?

At least 1/2, but is there any network with a higher cascade capacity?

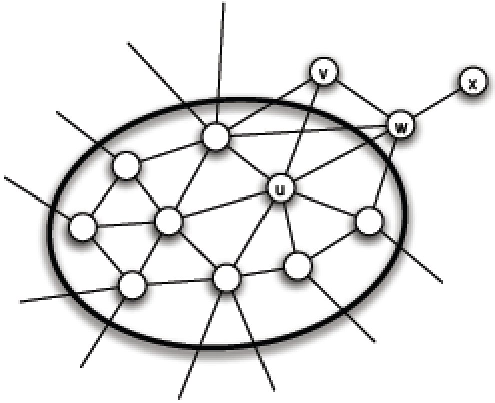
Will mean that an inferior technology can displace a superior one, even when the inferior technology starts at only a small set of initial adopters.

The Cascade Capacity: Cascades on Infinite Networks

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

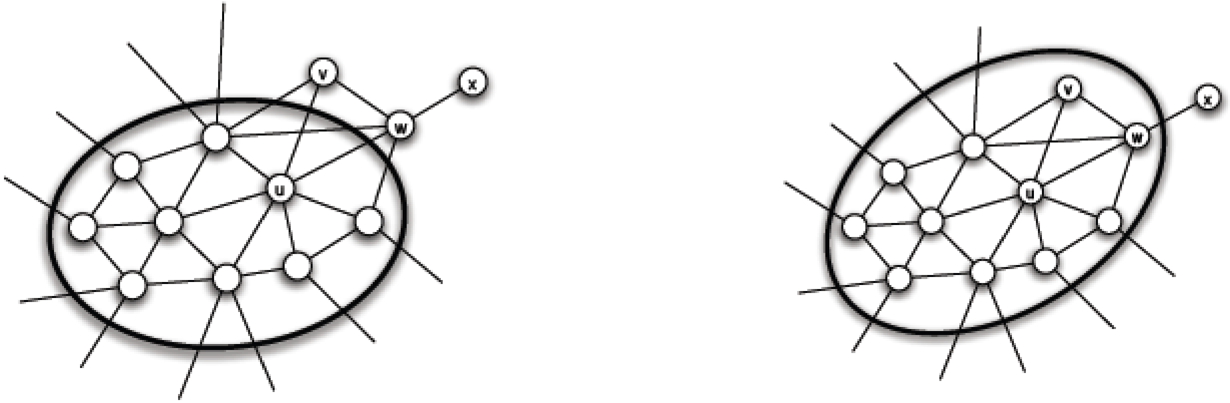
The Cascade Capacity: Cascades on Infinite Networks

Interface: the set of A-B edges



Prove that in each step the size of the interface strictly decreases
Why is this enough?

The Cascade Capacity: Cascades on Infinite Networks



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

EPIDEMIC SPREAD

Epidemics

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

- Health officials
- Sociologists
- Mathematicians
- Hollywood

COTILLARD DAMON FISHBURNE LAW PALTROW WINSLET
NOTHING SPREADS LIKE FEAR
CONTAGION
SEPTEMBER 9

The underlying contact network clearly affects the spread of an epidemic

Model epidemic spread as a random process on the graph and study its properties

- Main question: will the epidemic take over most of the network?

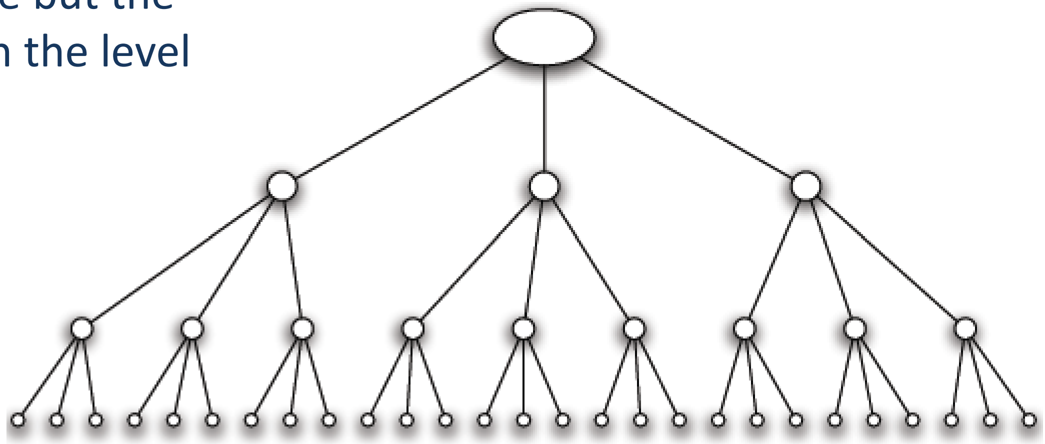
Diffusion of ideas and the spread of influence can also be modeled as epidemics

Branching Processes

- A person transmits the disease to each people she meets *independently with a probability* ***p***
- Meets ***k*** *people* while she is contagious

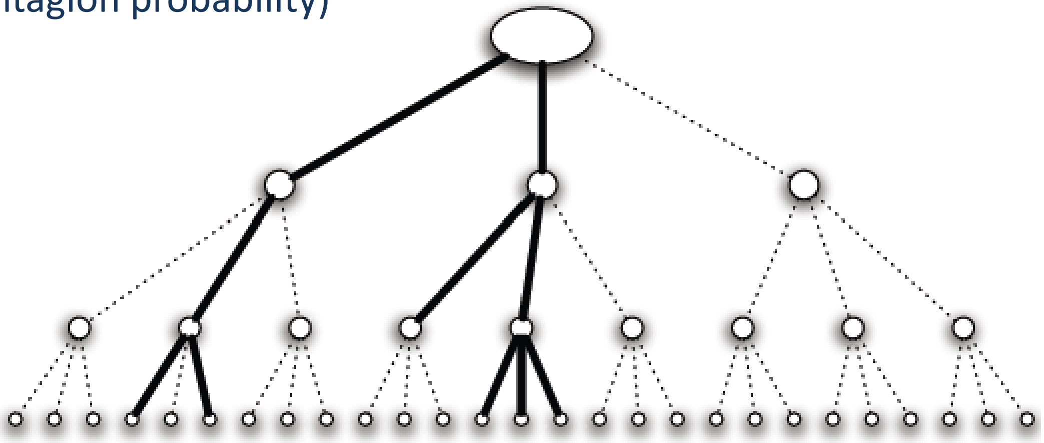
1. A person carrying a new disease enters a population, first *wave* of k people
2. Second wave of k^2 people
3. Subsequent waves

A contact network with *k* =3
Tree (root, each node but the root, a single node in the level above it)



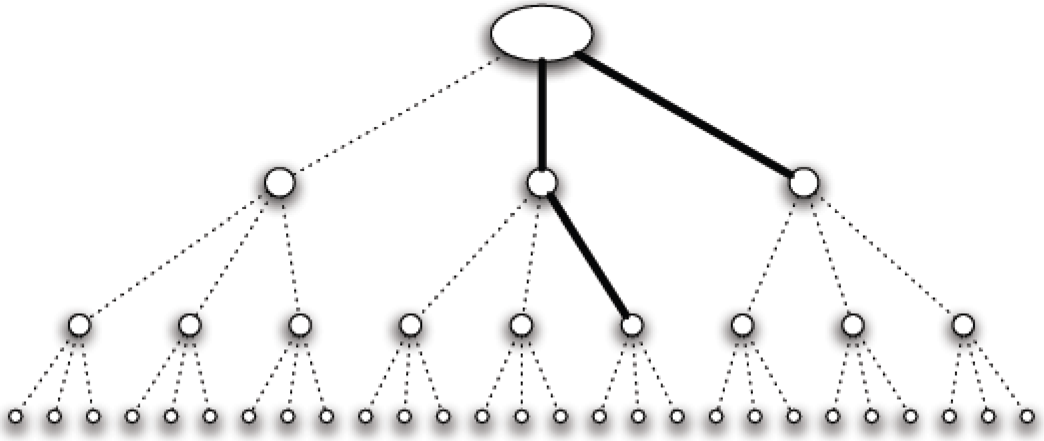
Branching Processes

Aggressive epidemic (high contagion probability)



Mild epidemic (low contagion probability)

- If it ever reaches a wave where it infects no one, then it dies out
- Or, it continues to infect people in every wave infinitely



Branching Processes: Basic Reproductive Number

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

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.

$R_0 = pk$

(a) $R_0 < 1$ -- Each infected person produces less than one new case in expectation
Outbreak constantly trends downwards

(b) $R_0 > 1$ – trends upwards, and the disease persists with positive probability
(when $p < 1$, the disease can get unlucky!)

A “knife-edge” quality around the critical value of $R_0 = 1$

Branching process

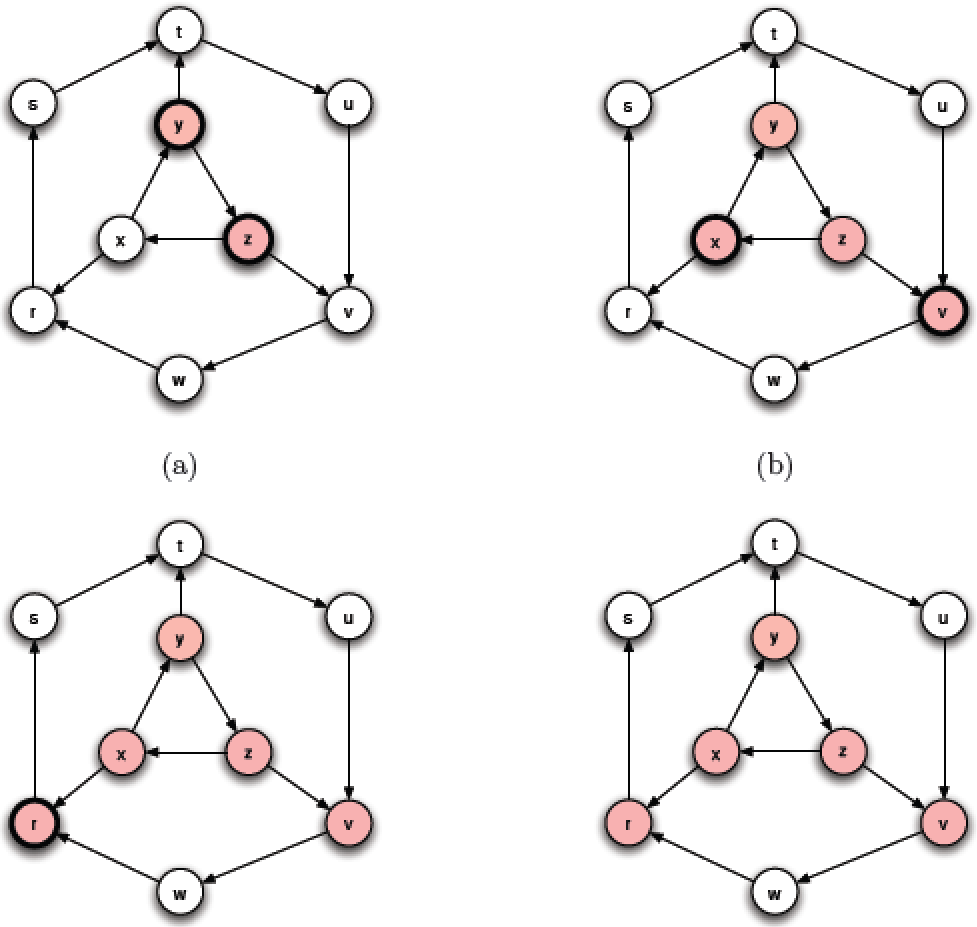
- Assumes no network structure, no triangles or shared neihgbors

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
- 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



(a)

(b)

(c)

(d)

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 basic reproductive number is not necessarily informative in the general case

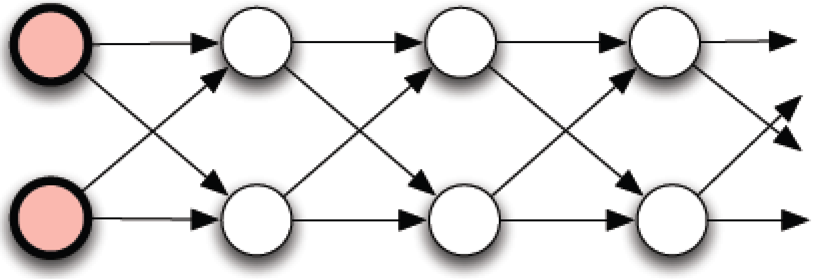


Figure 21.3: In this network, the epidemic is forced to pass through a narrow "channel" of nodes. In such a structure, even a highly contagious disease will tend to die out relatively quickly.

Percolation

- Percolation: we have a network of “pipes” which can curry liquids, and they can be either open with probability p, or close with probability (1-p)
  - 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.

Example

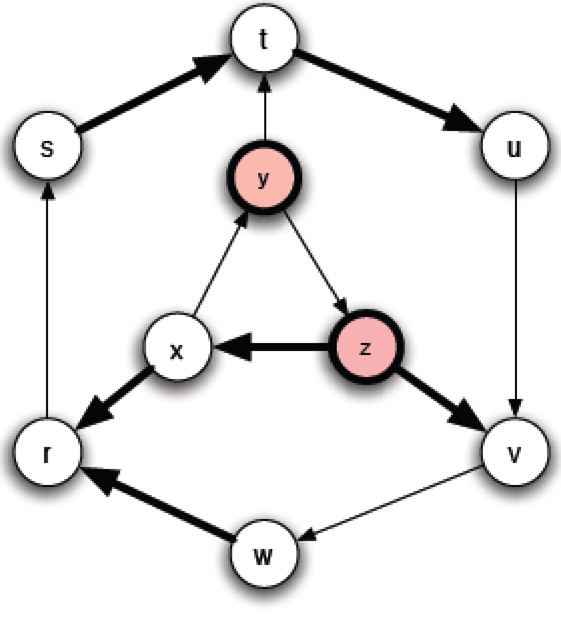


Figure 21.4: An equivalent way to view an SIR epidemic is in terms of *percolation*, where we decide in advance which edges will transmit infection (should the opportunity arise) and which will not.

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)
- Nodes alternate between Susceptible and Infected status

Exampe

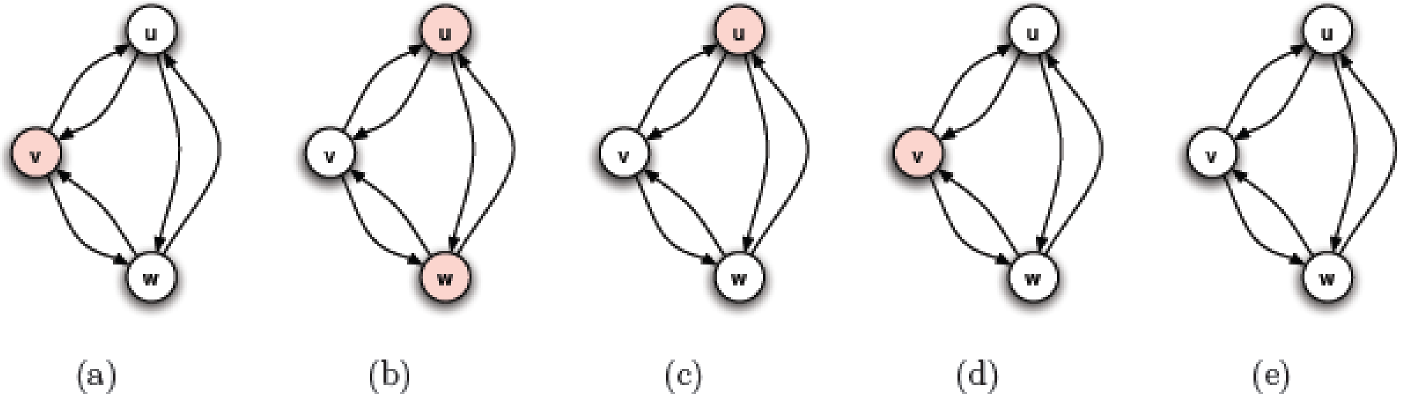


Figure 21.5: In an SIS epidemic, nodes can be infected, recover, and then be infected again. In each step, the nodes in the Infectious state are shaded.

- 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 λ_1 is the first eigenvalue of A

Multiple copies model

- Each node may have multiple copies of the same virus
  - **v**: state vector : v_i : number of virus copies at node i

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

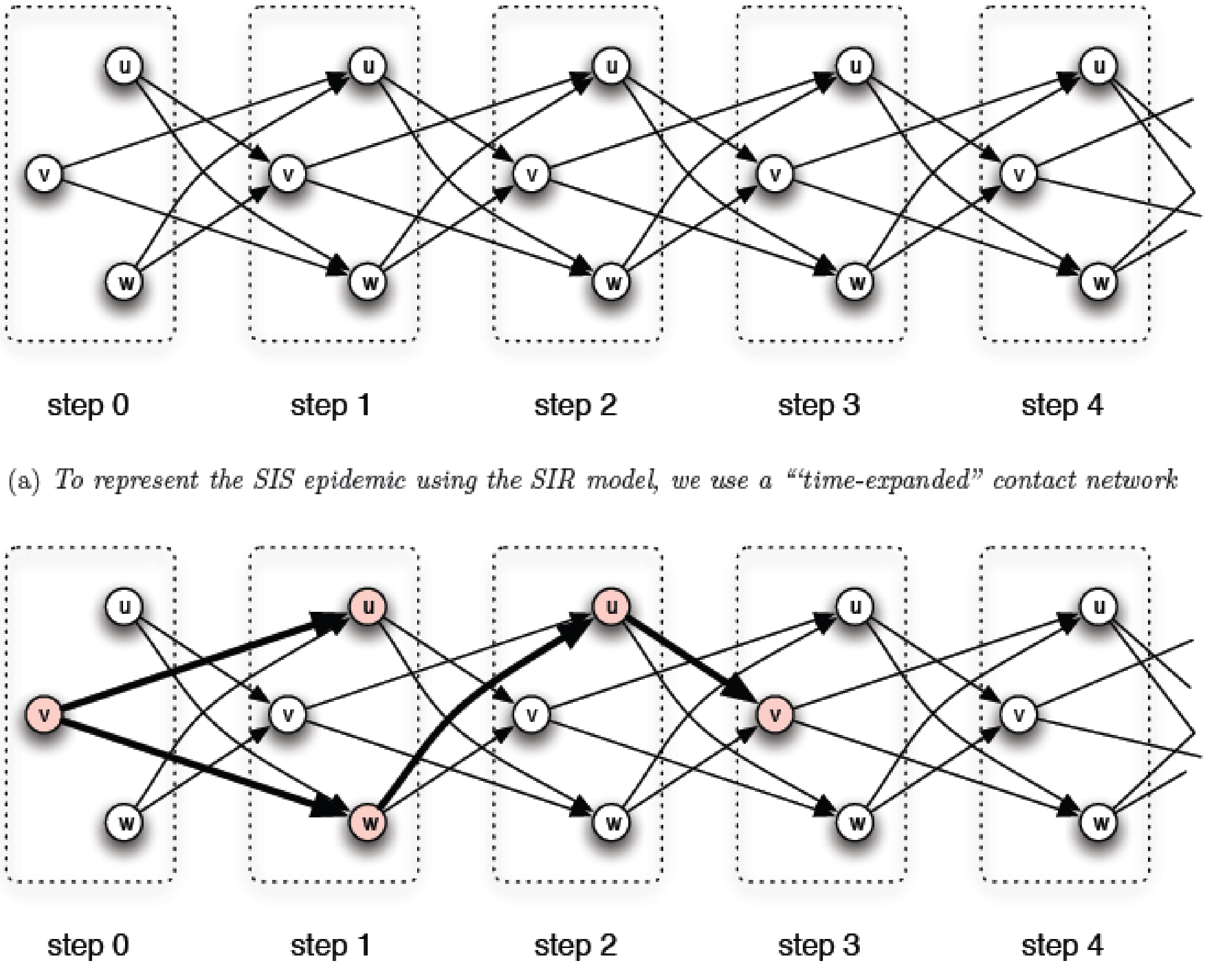
Analysis

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

$$\overline{\mathbf{v}^{t}} = \left(p\mathbf{A} + (1-q)\mathbf{I}\right)\overline{\mathbf{v}^{t-1}}$$

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

SIS and SIR



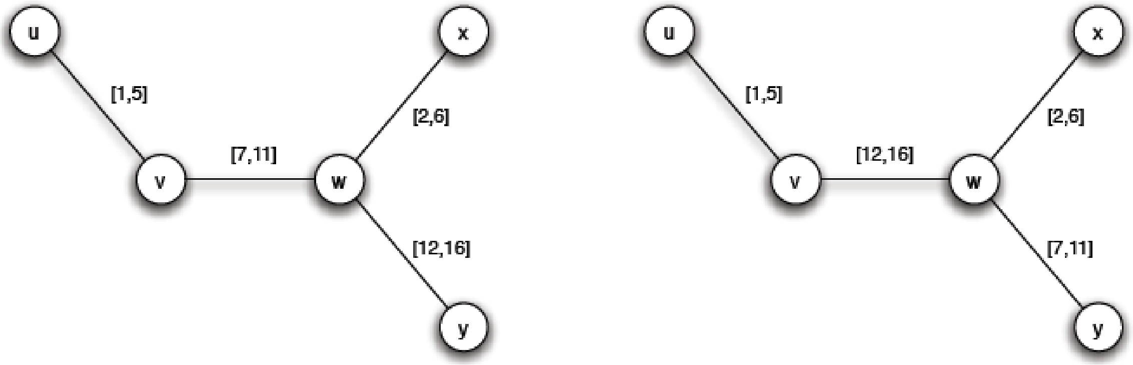
(a) *To represent the SIS epidemic using the SIR model, we use a "“time-expanded” contact network*

(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



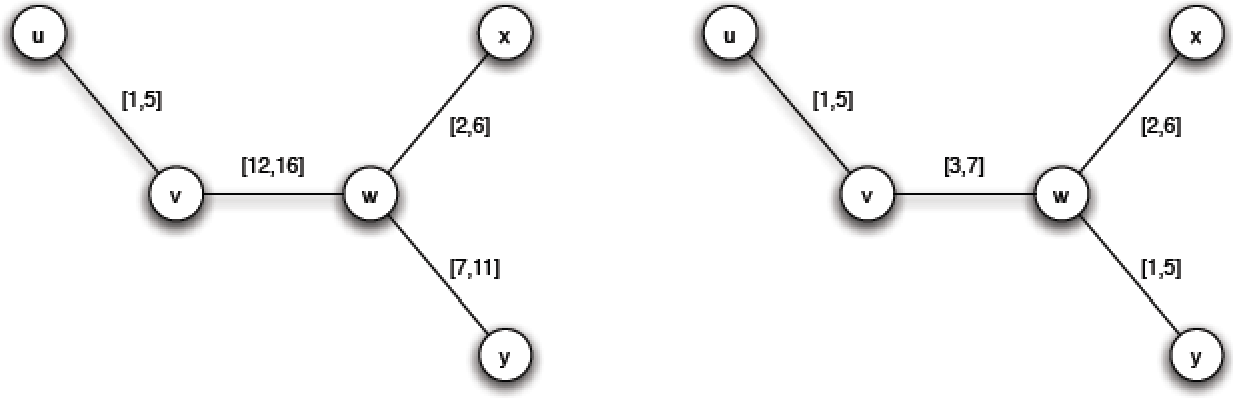
(a) *In a contact network, we can annotate the edges with time windows during which they existed.*

(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.

Concurency

- 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.

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 $t-1$ actives a non-active neighbor w with probability p_{uw}. If it fails, it does not try again

- The same as the simple SIR model

Influence maximization

- Influence function: for a set of nodes A (target set) the influence s(A) 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 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 α-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 $\alpha = 1 - \frac{1}{e}$

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

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 nodes 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 θ_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!

Experiments

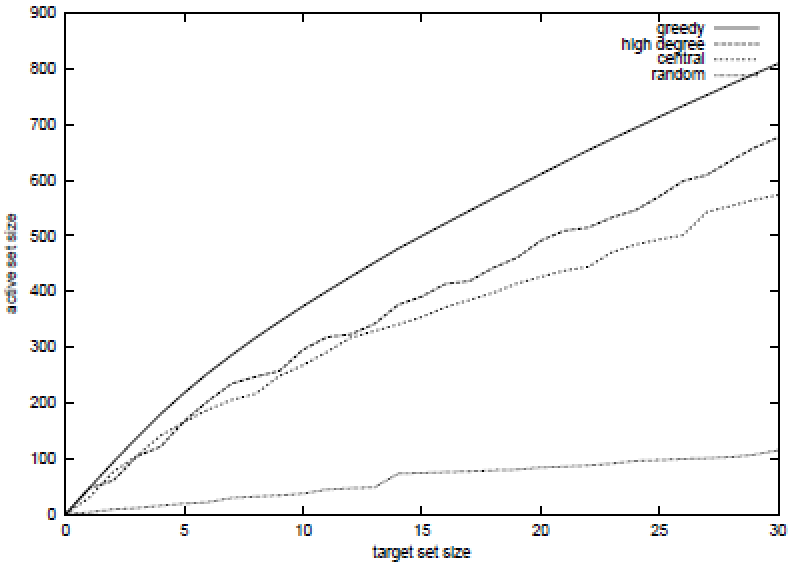


Figure 2: Results for the weighted cascade model

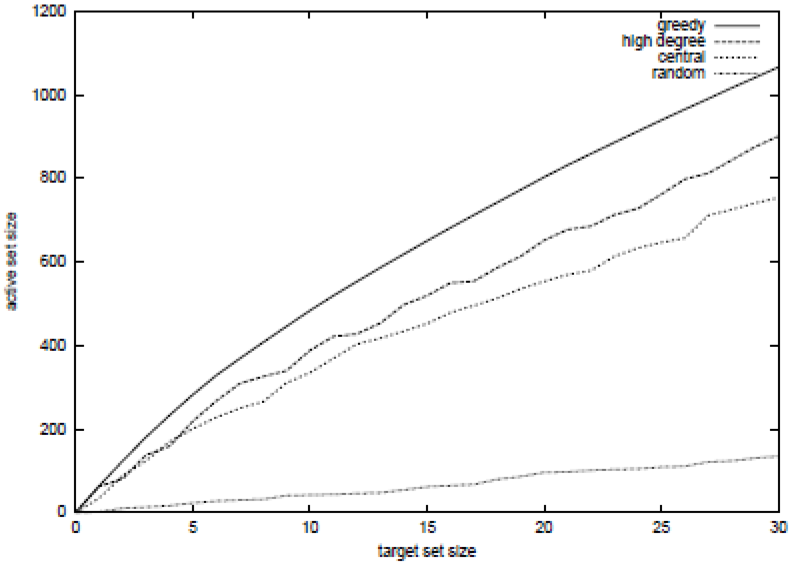


Figure 1: Results for the linear threshold model