Online Social Networks and Media

Diffusion:

Epidemic Spread

Influence Maximization

Opinions

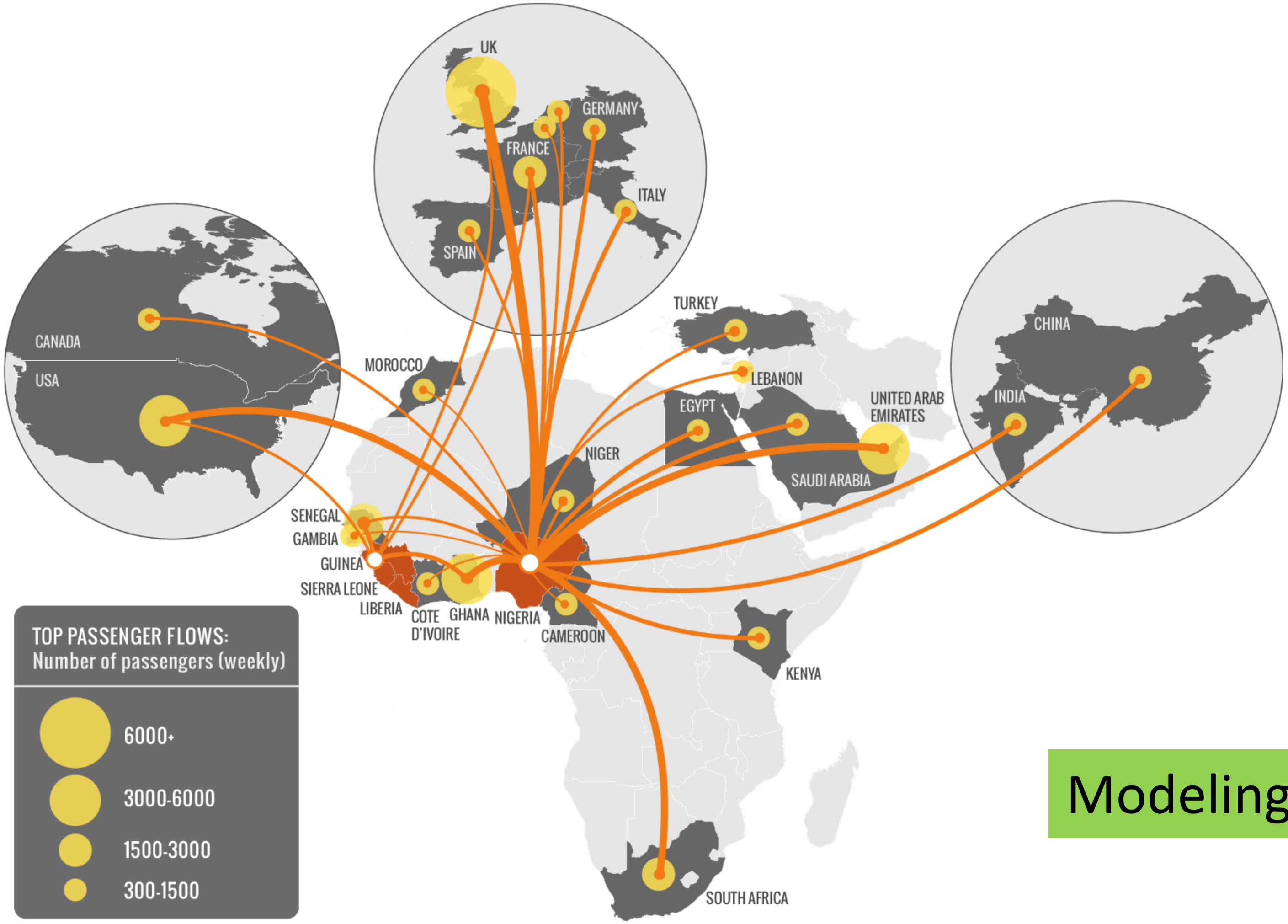
Introduction

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

Why do we care?

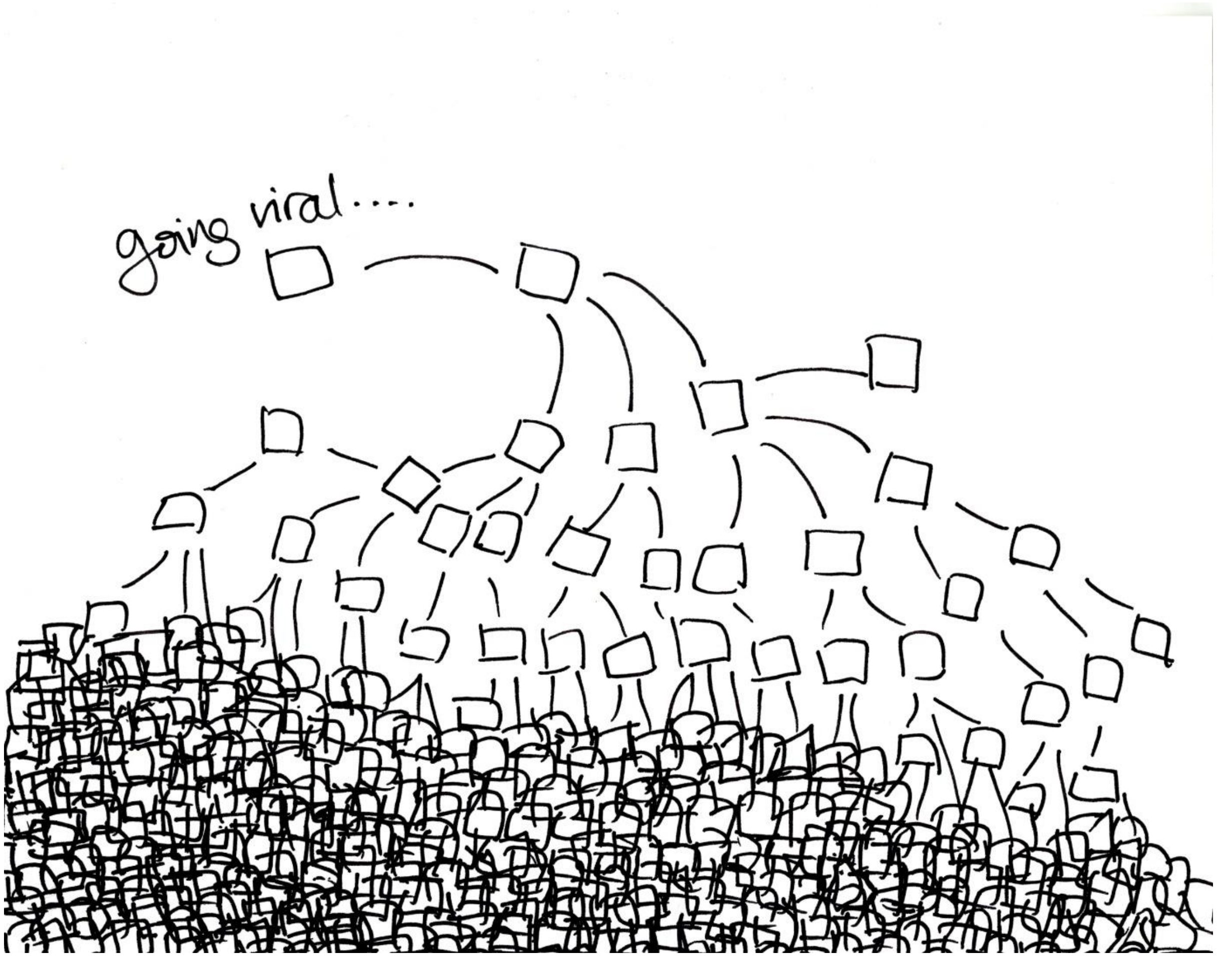
LHR
CDG
FRA
IST
DXB
ORD
JFK
LAX
DFW
ATL
PEK
PVG
HND
CAN
HKG
BKK
KUL
SIN
CGK
Africa
Americas
Asia
Europe
Oceania

Why do we care?



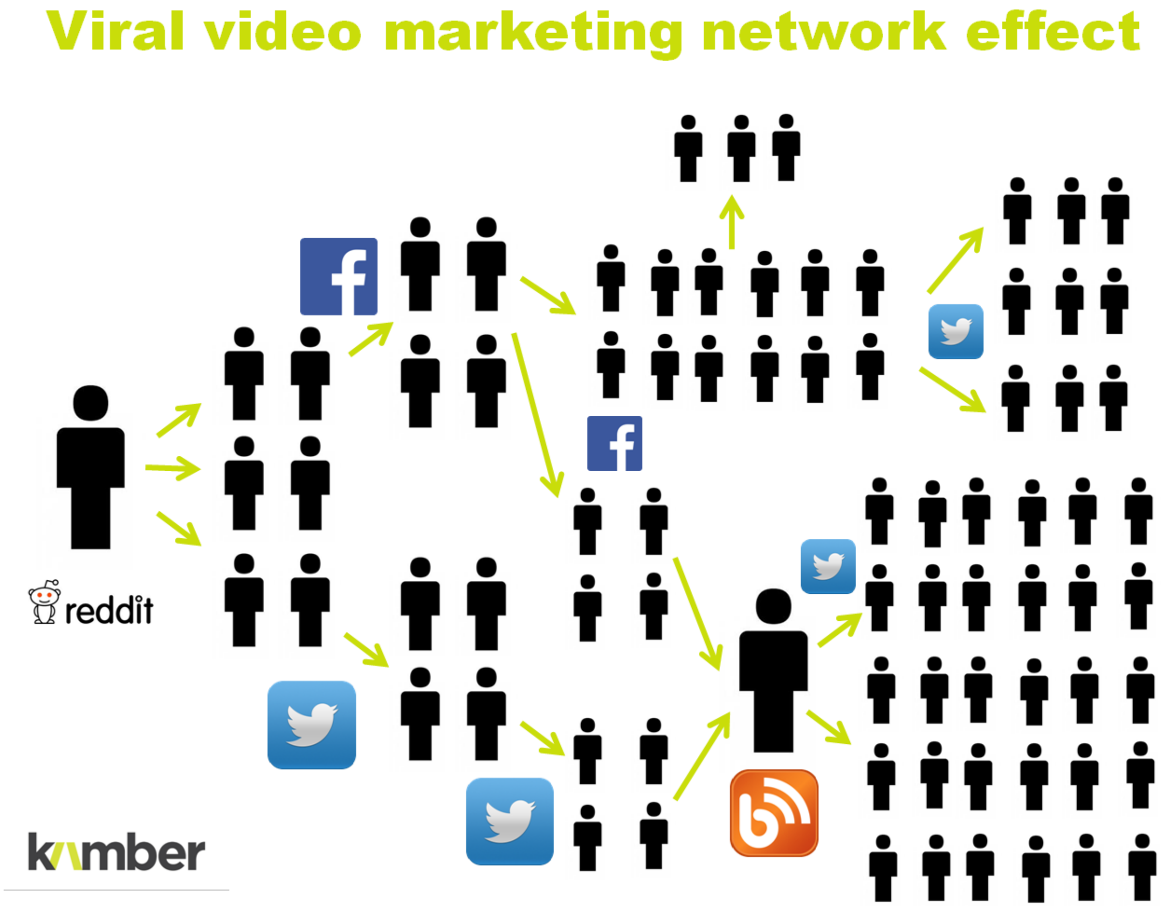
Modeling epidemics

Why do we care?



Viral marketing

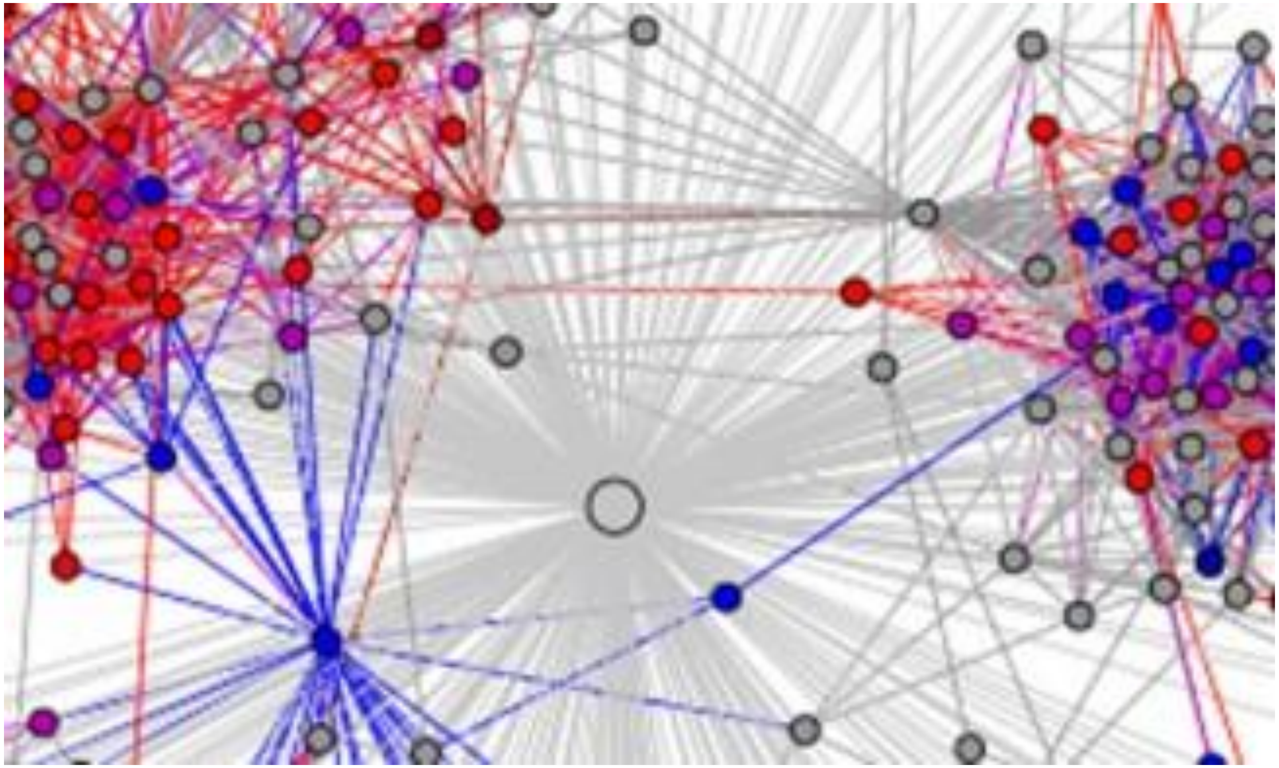
Why do we care?



Viral marketing

Why do we care?

Spread of innovation



Outline

- Epidemic models
- Influence maximization

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

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

Basic Reproductive Number R_0

- Basic Reproductive Number (R_0): the expected number of new cases of the disease caused by a single individual
- This is a dimensionless number (it does not have units) and it characterizes the spread of the virus.
- General computation:

$$R_0 \propto \left(\frac{infection}{contact}\right)\left(\frac{contact}{time}\right)\left(\frac{time}{infection}\right)$$

$$R_0 = \tau\,\bar{c}\,d$$

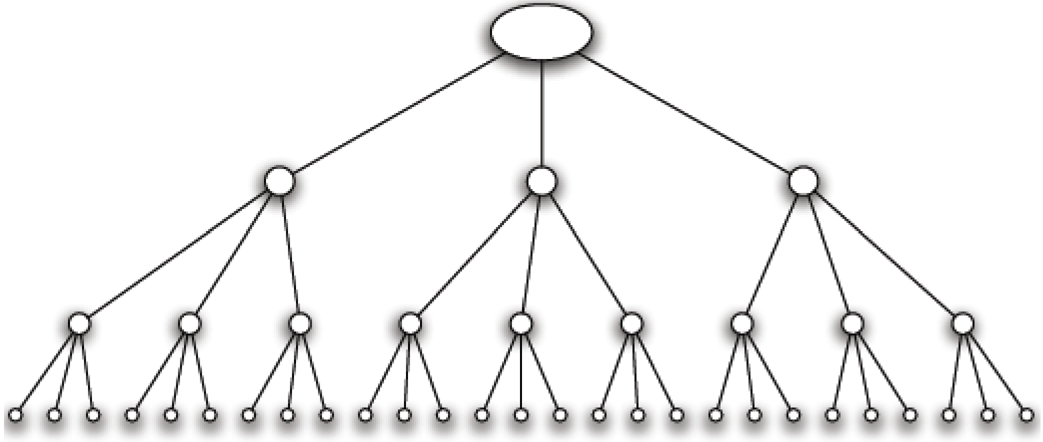
- In general, we want $R_0 < 1$ since this usually (but not always) implies that the infection will die out.

R_0 and R_t

- The computation of R_0assumes that everyone is susceptible to infection
- For monitoring the real-time development of an infection the real-time or effective R_t is used
- It takes into account the current state of the disease, who is sick, and who is immune
- We definitely want $R_t < 1$
- It is very hard to compute and depends on multiple factors.

A simple model

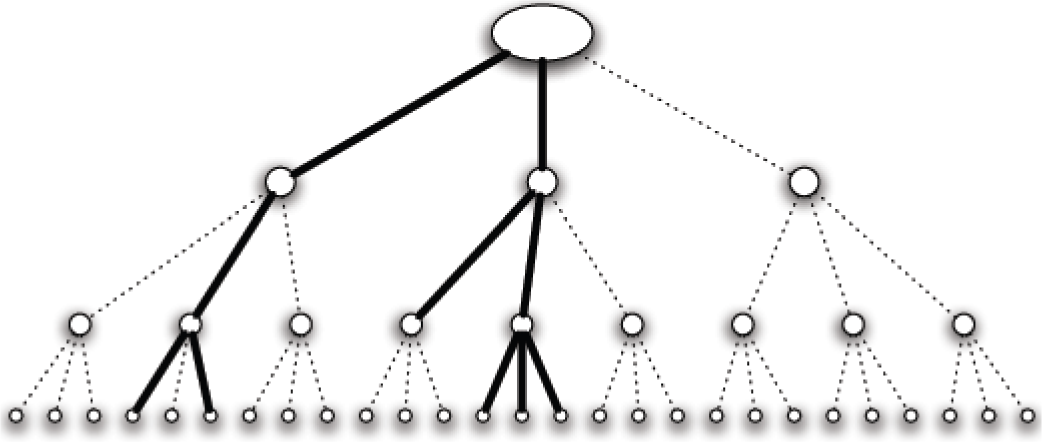
- Branching process: A person transmits the disease to each people she meets independently with a probability *p*
- An infected person meets *k* (new) people while she is contagious
- Infection proceeds in waves.



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*

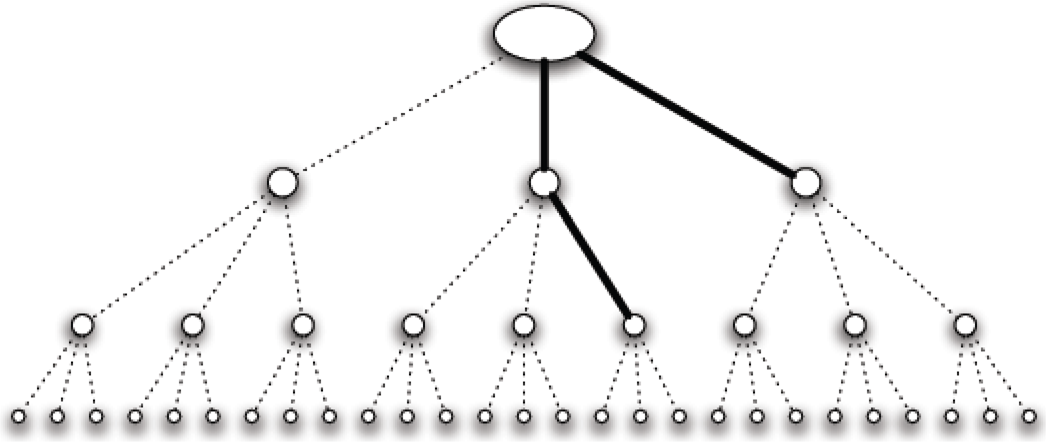


An aggressive epidemic with high infection probability

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.
In this case each person infects less than one person in expectation. The infection eventually *dies out*.

b) If $R_0 > 1$, then with probability greater than 0 the disease persists by infecting at least one person in each wave.
In this case each person infects more than one person in expectation. The infection *persists*.

Application: Reduce *k*, or *p* to combat an epidemic

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 $n \rightarrow \infty$

We want to show that

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

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

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

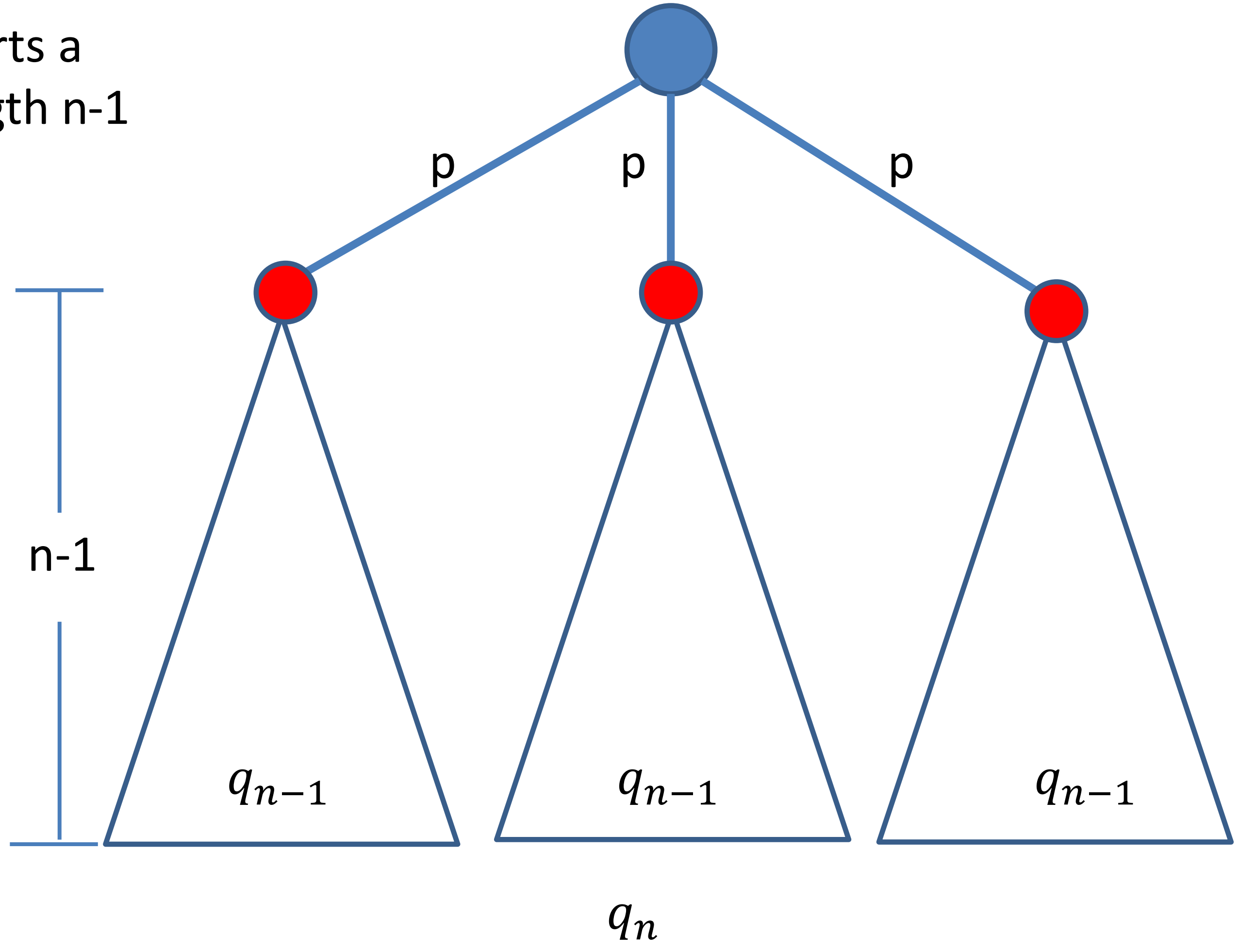
$$q_n = 1 - (1 - pq_{n-1})^k$$

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 $f(1) = 1 - (1-p)^k < 1$.

     passes through (0, 0); below y = x, once x = 1

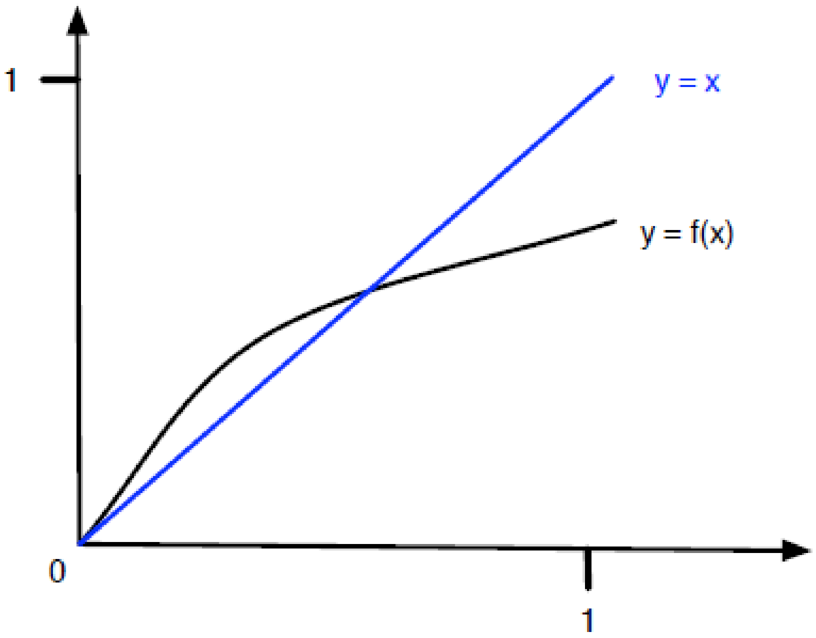
  2. $f'(x) = pk(1-px)^{k-1} > 0$, in the interval [0,1] but decreasing. Our function is increasing and concave.

  3. $f'(0) = pk = R_0$

     Slope at x = 0

Proof

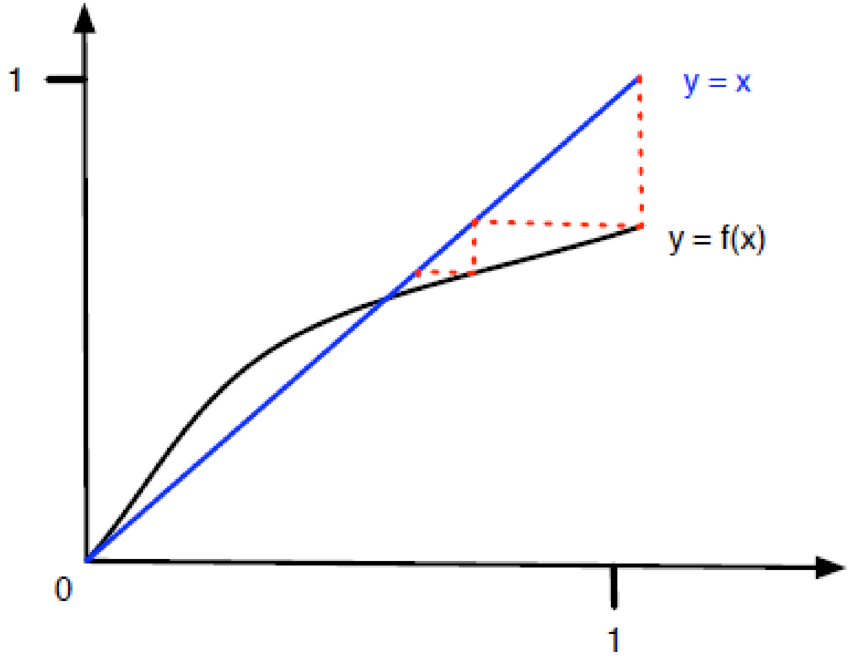
- Case 1: $R_0 = pk > 1$. The function starts above the line $y = x$ but then drops below the line.



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

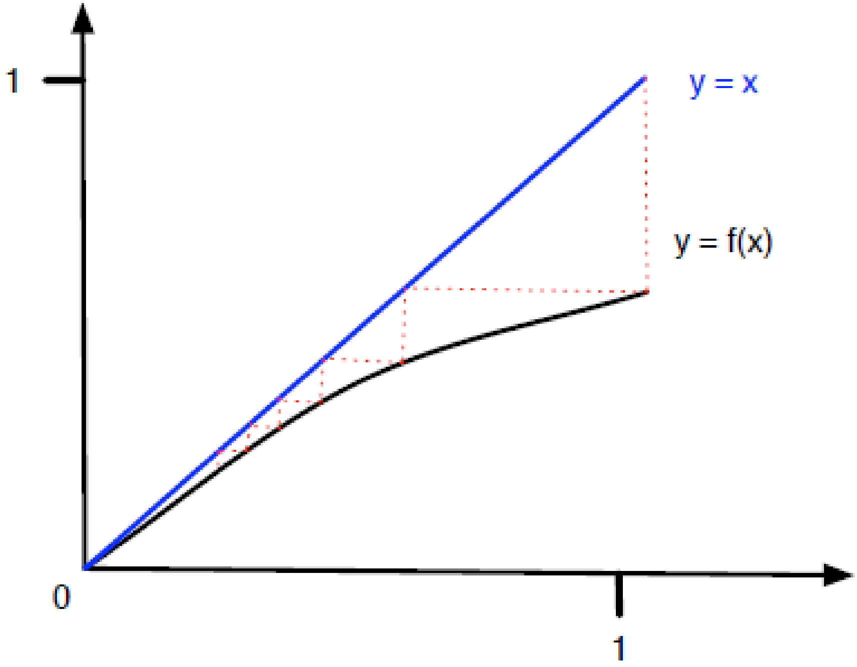
Proof

- 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

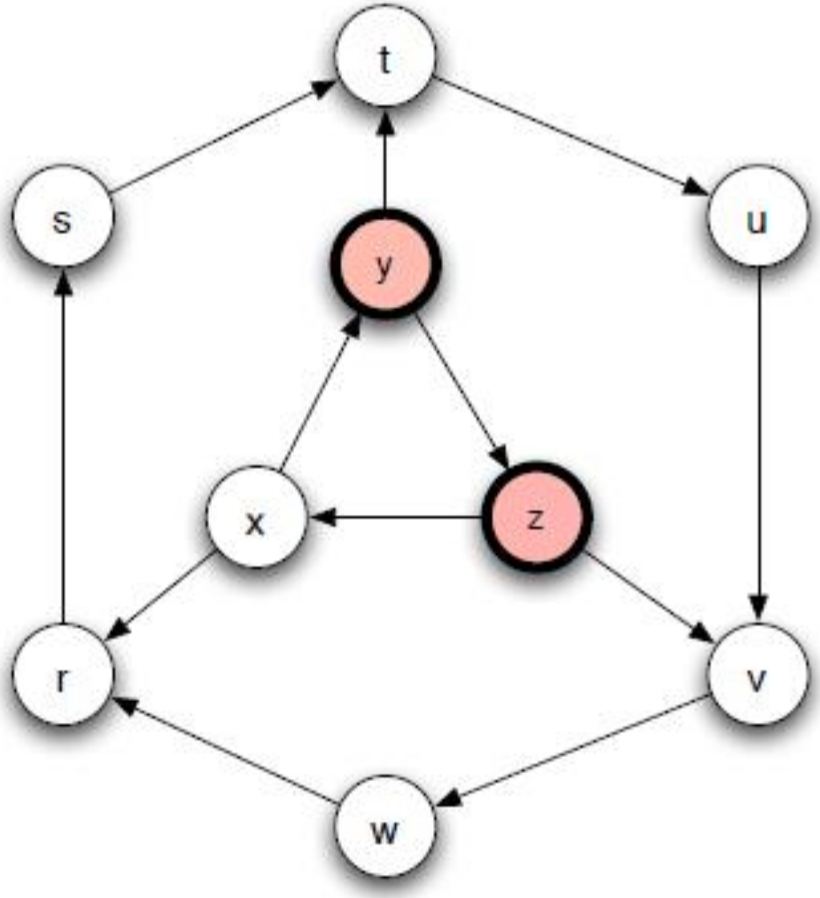
- Each node may be in the following states
  - Susceptible: healthy but not immune
  - Infected: has the virus and can actively propagate it
  - Removed: (Immune or Dead) had the virus but it is no longer active

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

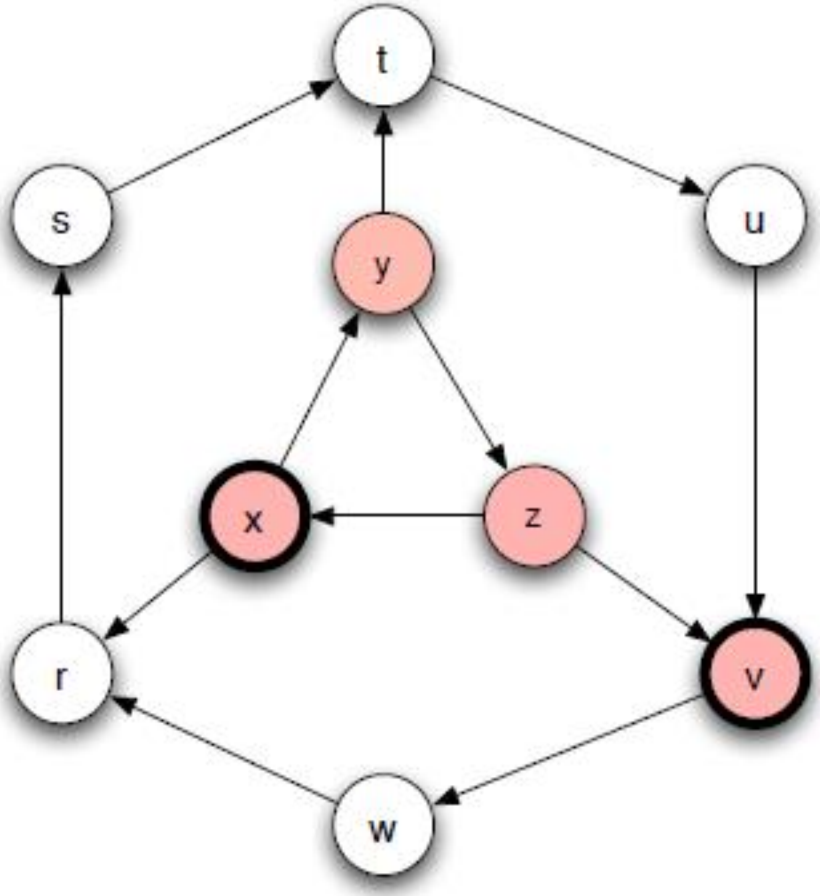
The SIR process

- 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

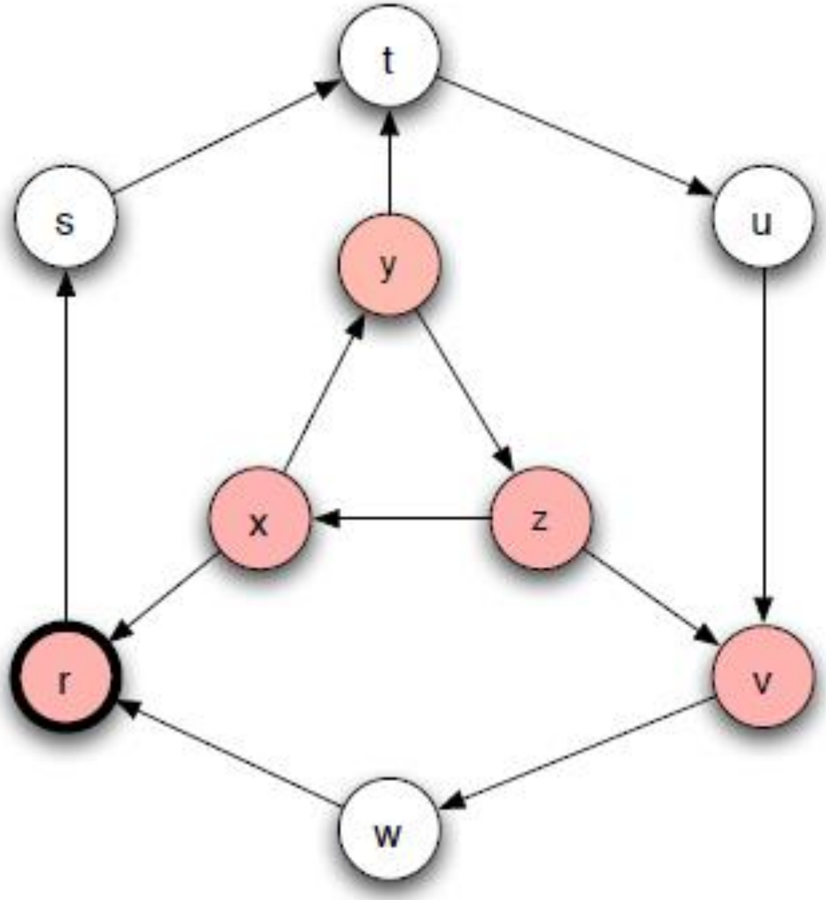
Example



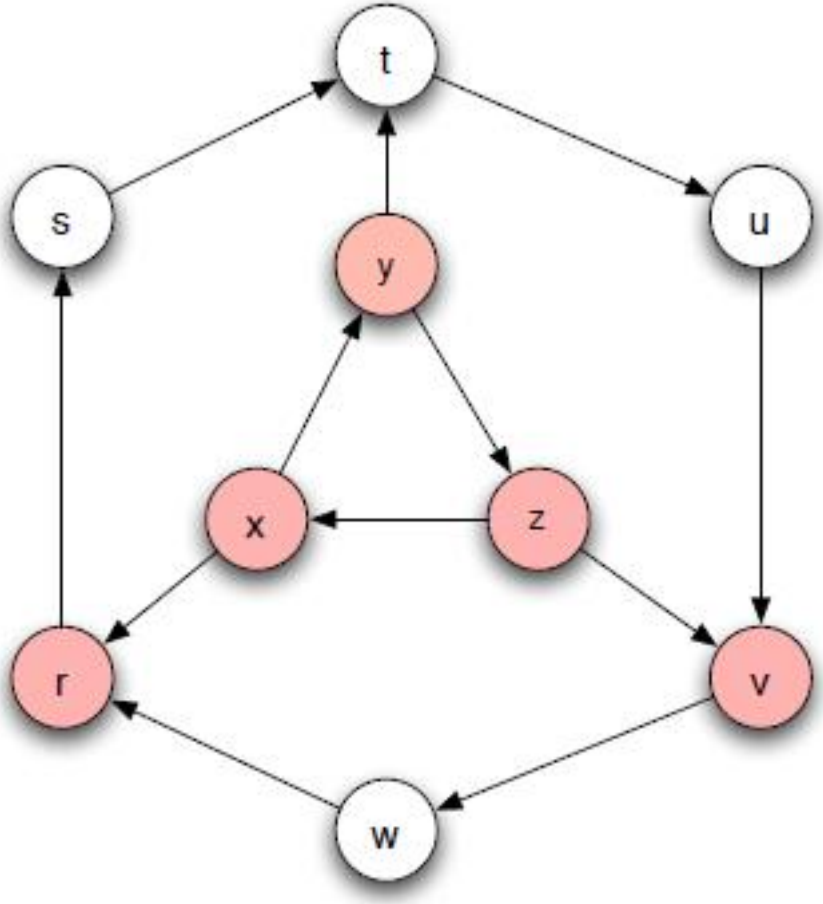
Example

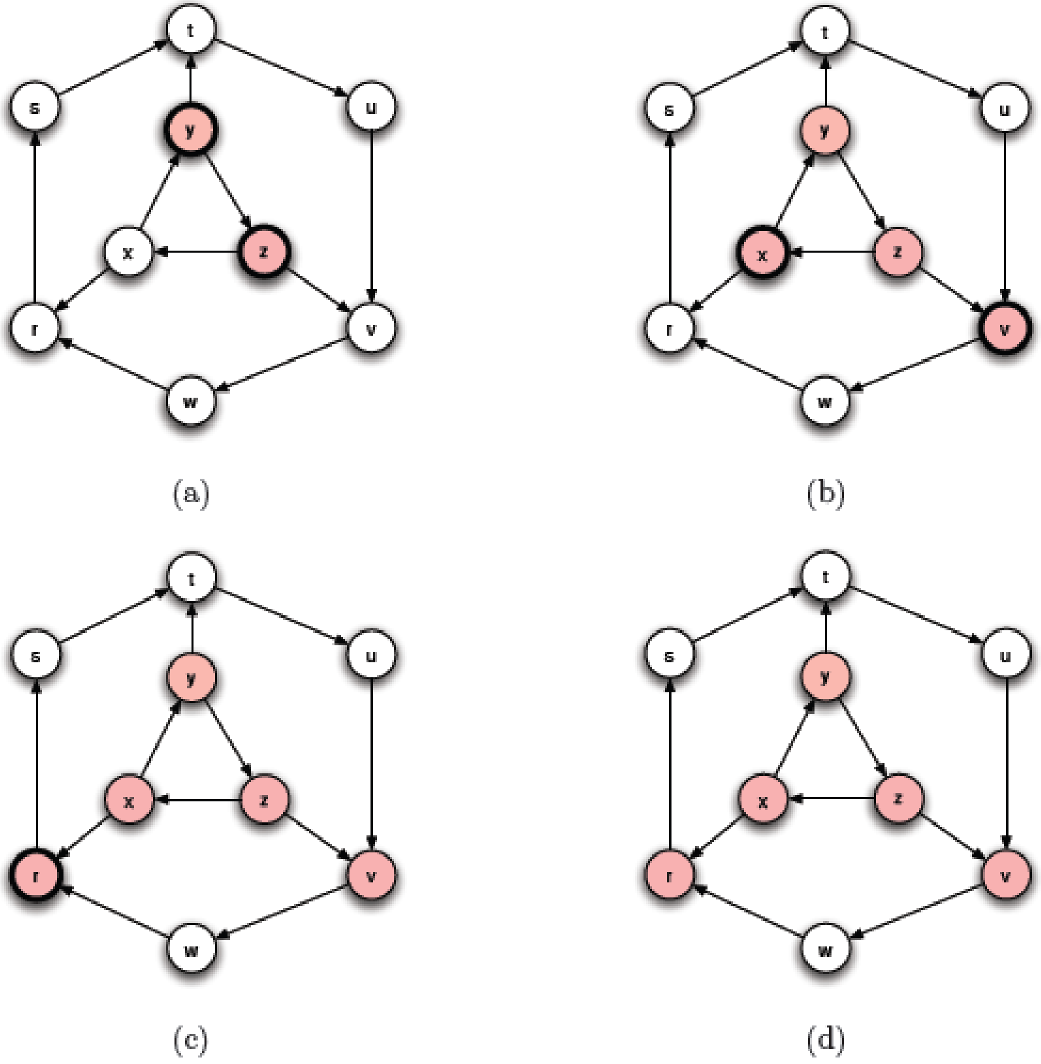


Example



Example





(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.

Extensions

- Probability per pair of nodes
- Sequence of several states (e.g. early, middle, and late periods of the infection), and allowing the contagion probabilities to vary across these states
- Mutating, change the characteristics

Continuous case

- We can analyze the SIR model assuming a continuous change in the number of Susceptible (S), Infected (I), and Removed (R) nodes.
- In the continuous model the infection probability is replaced by the infection rate β
- We also have the recovery (or removal) rate $\gamma = 1/t_I$ which is the rate by which nodes recover (or die)
- Let $s = \frac{S}{N}, i = \frac{I}{N}, r = \frac{R}{N}$, the fraction of S, I, R nodes, where N the size of the population
- We assumed that initially $s \approx 1$
- We assume that we have si contacts (random contacts)

Continuous case

- We can describe SIR with the following system of differential equations:

$$\frac{\partial s}{\partial t} = -\beta s i$$

$$\frac{\partial r}{\partial t} = \gamma i$$

$$\frac{\partial i}{\partial t} = \beta s i - \gamma i$$

- The epidemic persists if

$$\frac{\partial i}{\partial t} > 0 \Rightarrow \frac{\beta}{\gamma} > 1$$

$$\boxed{R_0 = \frac{\beta}{\gamma}}$$

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 $(1/3)^4 = 1/81$

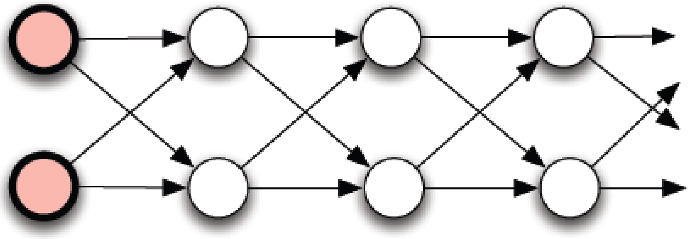


Figure 21.3: In this network, the epidemic is forced to pass through a narrow "channel" of nodes. In such a structure, even a highly contagious disease will tend to die out relatively quickly.

Percolation

- Percolation: we have a network of “pipes” which can carry liquids, and they can be either open, or closed with some probability
  - 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

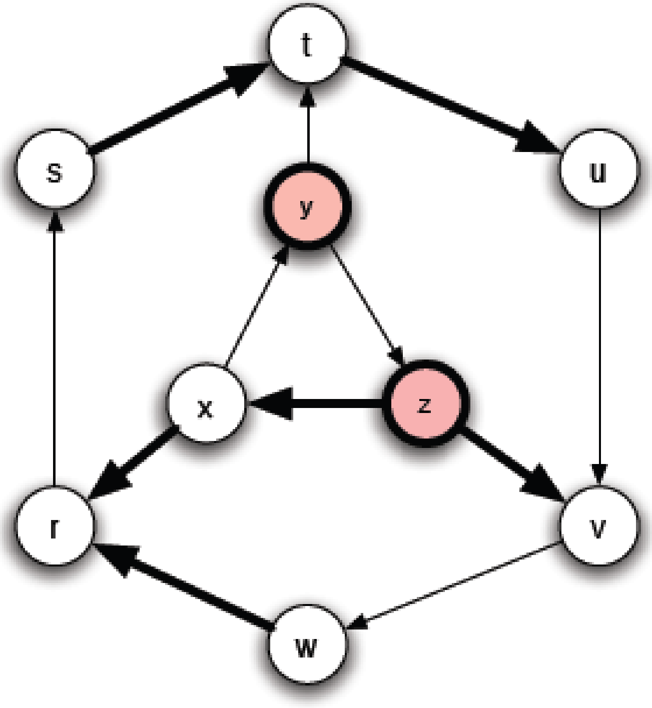


Figure 21.4: An equivalent way to view an SIR epidemic is in terms of *percolation*, where we decide in advance which edges will transmit infection (should the opportunity arise) and which will not.

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

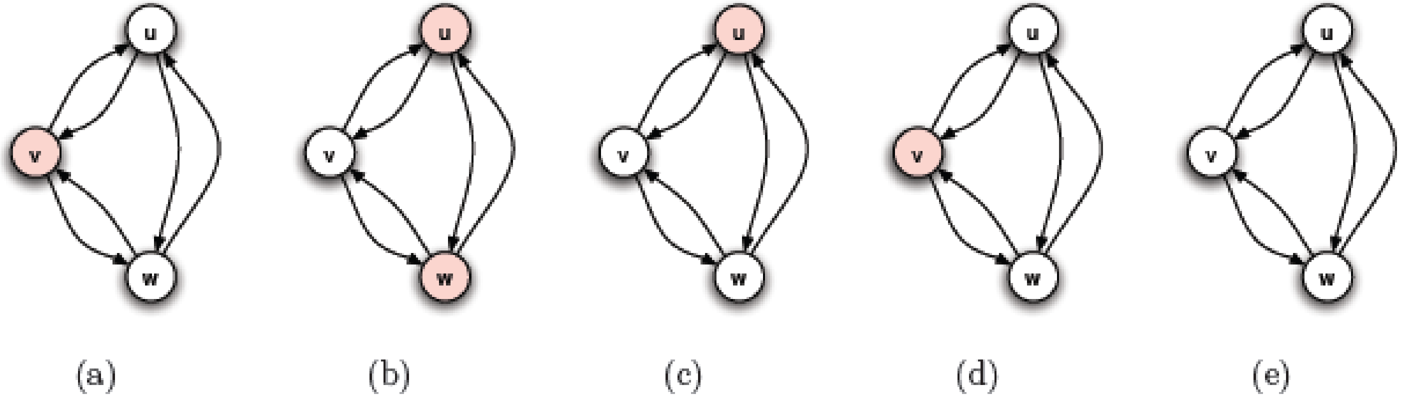


Figure 21.5: In an SIS epidemic, nodes can be infected, recover, and then be infected again. In each step, the nodes in the Infectious state are shaded.

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

An eigenvalue point of view

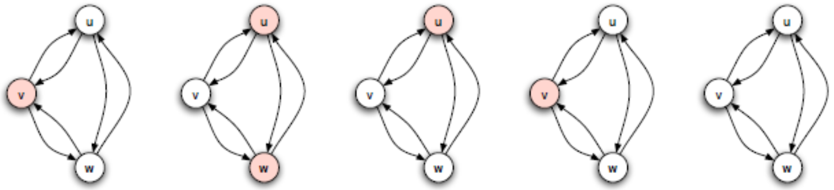
- If A is the adjacency matrix of the network, then the virus dies out if

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

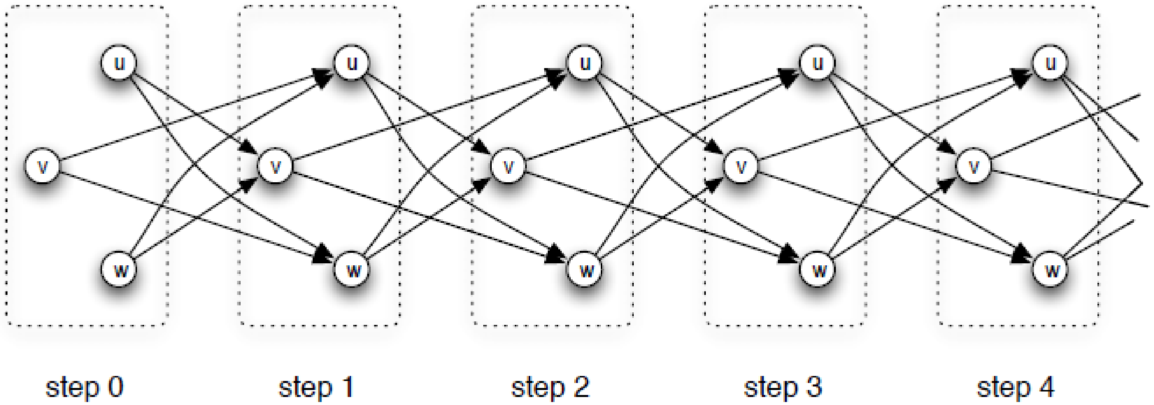
- Where $\lambda_1(A)$ is the first eigenvalue of A

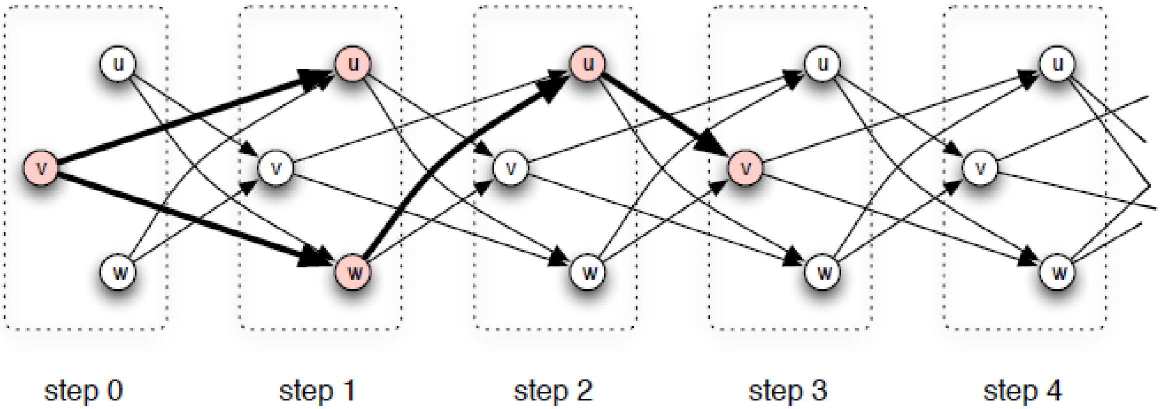
Y. Wang, D. Chakrabarti, C. Wang, C. Faloutsos. *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*. SRDS 2003

SIS and SIR



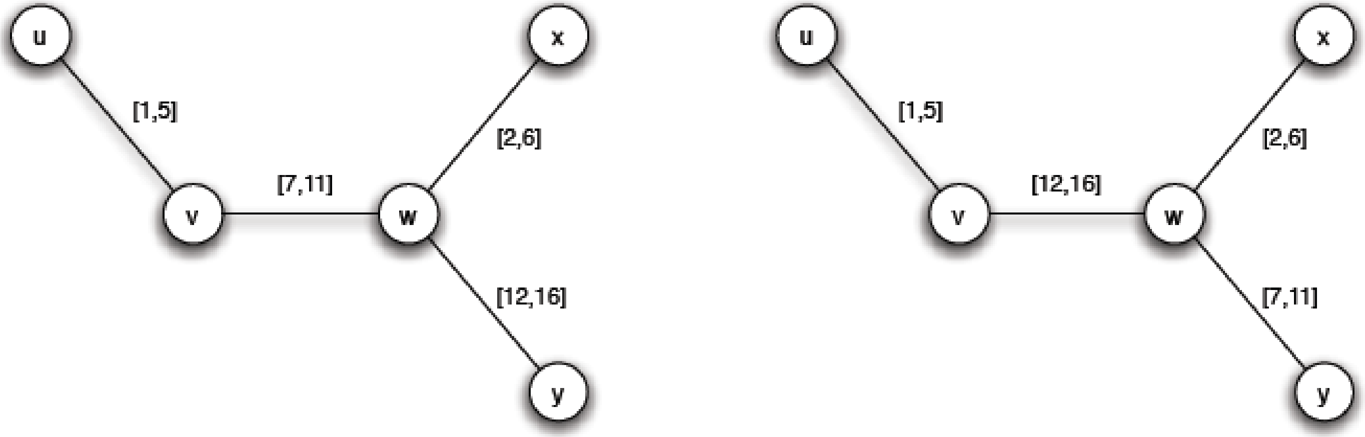
Time expanded network





Including time

- Infection can only happen within the active window



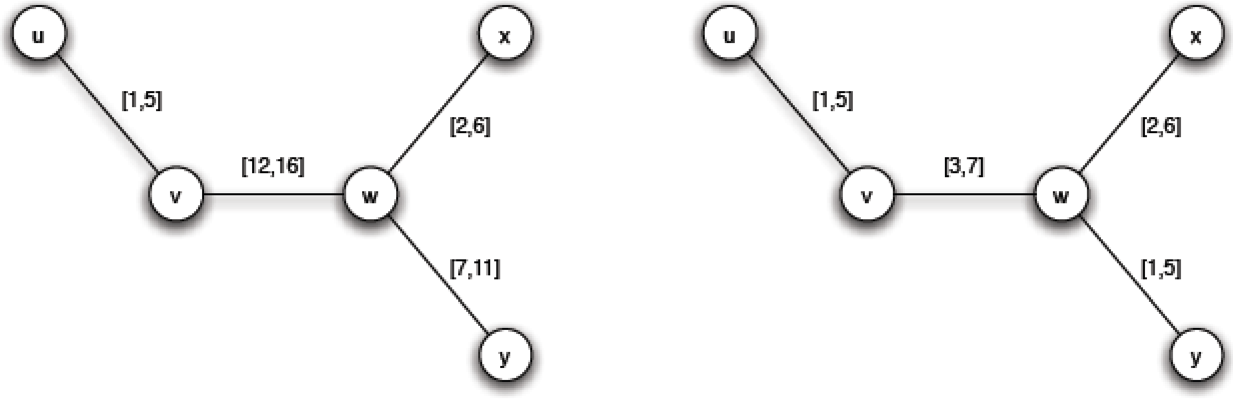
(a) *In a contact network, we can annotate the edges with time windows during which they existed.*

(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 are 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
- James Holland Jones, Notes on R_0
- Y. Wang, D. Chakrabarti, C. Wang, C. Faloutsos. *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*. SRDS 2003

INFLUENCE MAXIMIZATION

Maximizing spread

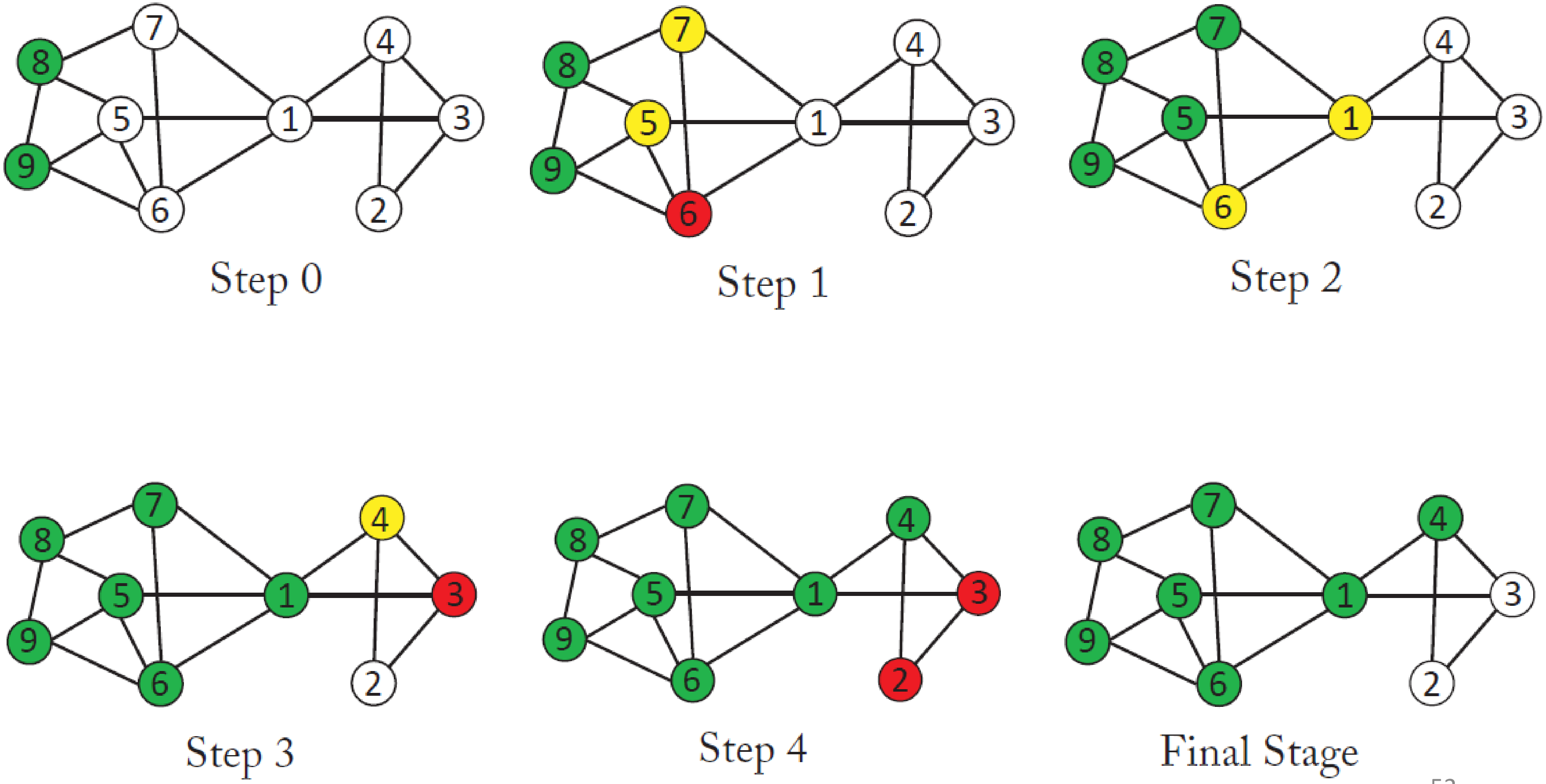
- Suppose that instead of a virus we have an item (product, idea, video) that propagates through contact
  - Word of mouth propagation.
- An advertiser is interested in maximizing the spread of the item in the network
  - The holy grail of "viral marketing"
- Question: which nodes should we "infect" 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 a 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 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?

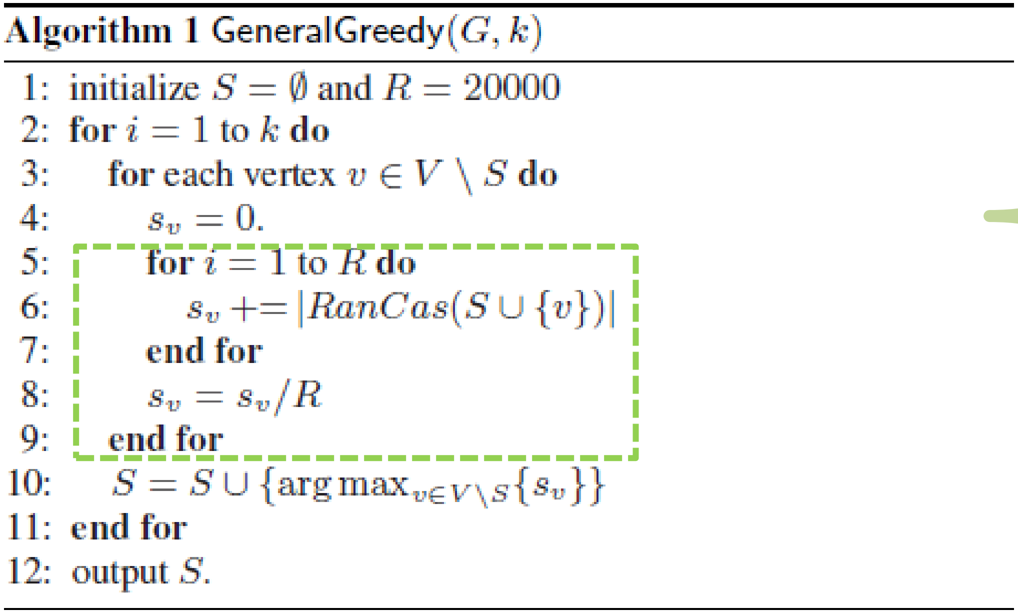
- We will use the fact that probabilistic propagation on a fixed graph can be viewed as deterministic propagation over a randomized graph
  - Express s(A) as an expectation over the input graph rather than the choices of the algorithm

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

Computation of Expected Spread

Computing s(A): perform multiple Monte-Carlo simulations of the process and take the average.



Algorithm 1 GeneralGreedy(G, k)

```
initialize $S = \emptyset$ and $R = 20000$
for $i = 1$ to $k$ do
    for each vertex $v \in V \setminus S$ do
        $s_v = 0$.
        for $i = 1$ to $R$ do
            $s_v \mathrel{+}= |RanCas(S \cup \{v\})|$
        end for
        $s_v = s_v / R$
    end for
    $S = S \cup \{\arg\max_{v \in V \setminus S}\{s_v\}\}$
end for
output $S$.
```

To estimate the influence spread of $S \cup \{u\}$, R repeated simulations of $RanCas(S \cup \{u\})$ are used

Each run takes O(m)

Complexity for computing the marginal gain of adding u: $O(Rm)$

For each k, all n nodes are tested, thus

$O(knRm)$

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

Degree discount

General idea

- Select seed nodes based on their degree
- If node v is selected, decrease the degree of all its neighbors

Wei Chen, Yajun Wang, Siyu Yang: Efficient influence maximization in social networks. KDD 2009: 199-208

Maximum influence path

General idea

- For each node, use the maximum influence paths (paths with the largest probability) to all other nodes
  - Shortest weighted path
- Assumption: influence propagates through these paths
- Given this assumption, estimate the probability that a node is activated

Wei Chen, Chi Wang, Yajun Wang: Scalable influence maximization for prevalent viral marketing in large-scale social networks. KDD 2010: 1029-1038

Reverse Reachable Sets

Construct graph X from G by activating edges with probability $p(e)$.

Let *v* be a node in G, the reverse reachable (RR) set for v in X is the set of nodes in X *that can reach v*.

That is, for each node u in the RR set, there is a directed path from u to v in X.

Youze Tang, Xiaokui Xiao, Yanchen Shi: Influence maximization: near-optimal time complexity meets practical efficiency. SIGMOD Conference 2014: 75-86

Reverse Reachable Sets

Let p be the probability for an RR set generated for v to overlap with a node set A, then when we use A as the seed set to run an influence propagation process on G, we have probability p to activate v

A random RR set is an RR set generated on an instance of X randomly sampled from G, for a node selected uniformly at random from X.

Reverse Reachable Sets

1. Generate a certain number of random RR sets from G.

2. Select k nodes to cover the maximum number of RR sets generated. (maximum coverage)

3. Return the k nodes as seed

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

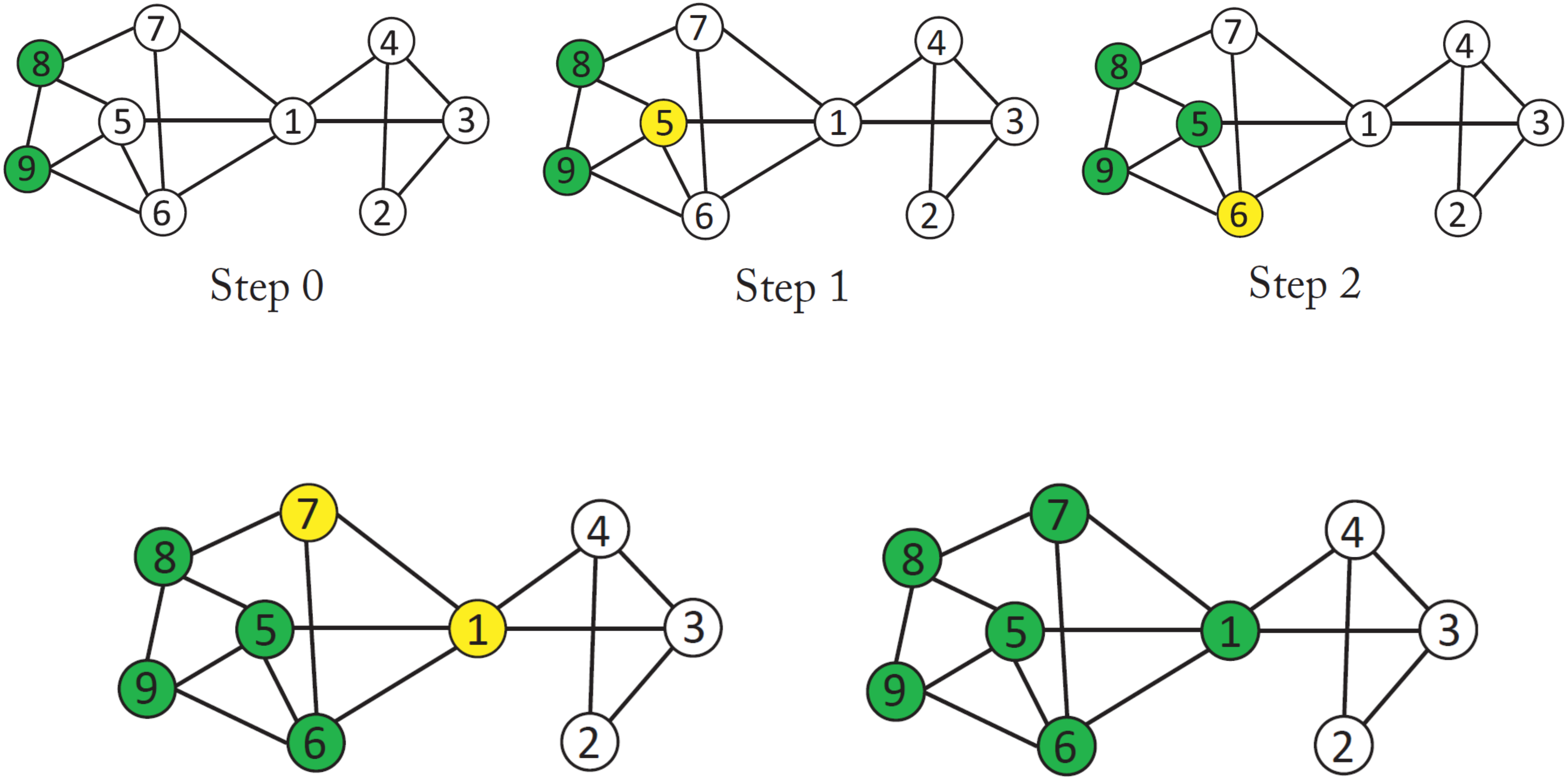
$$\sum_{v \text{ is a neighbor of } u} b_{vu} \leq 1$$

- 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

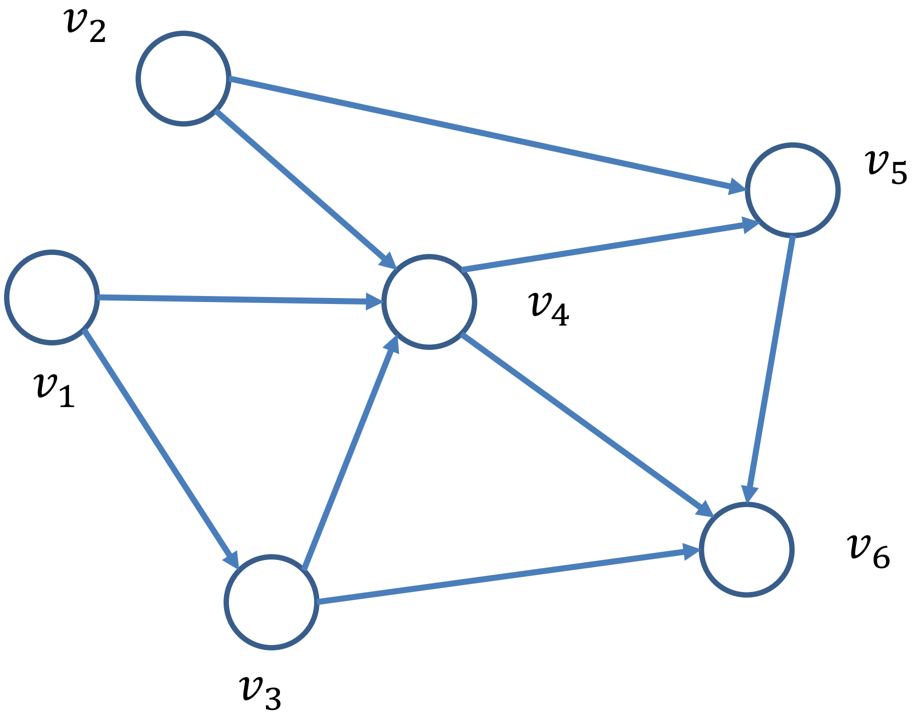
Proof idea

- 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 (submodularity LT model)

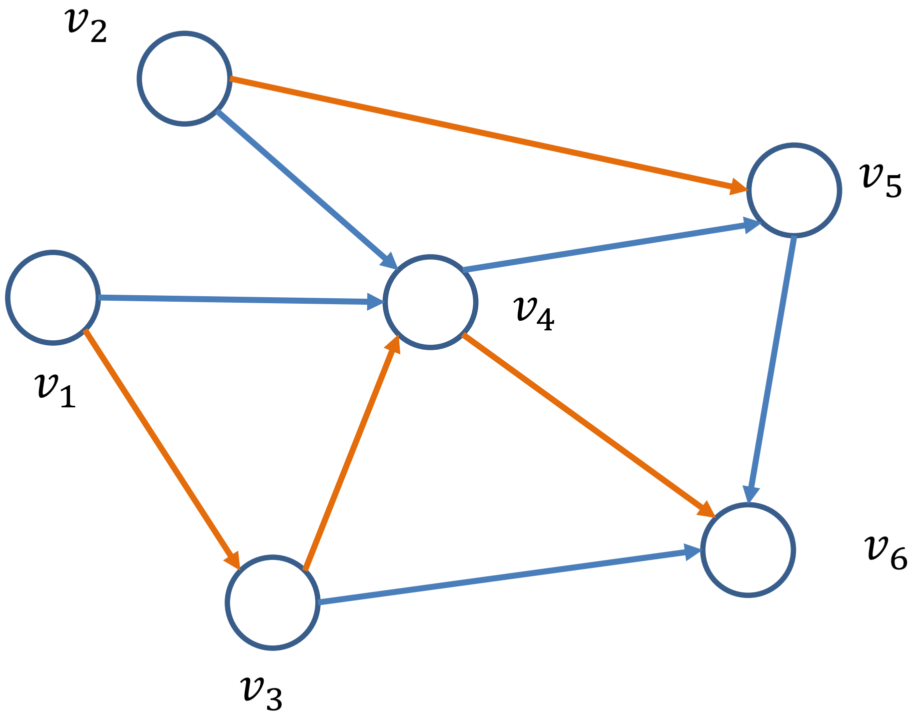
- Consider the special case of a DAG (Directed Acyclic Graph)
  - There is a topological ordering of the nodes $v_0, v_1, \dots, v_n$ such that edges go from left to right
- Consider node v_i in this ordering and assume that S_i is the set of incoming neighbors of v_i that are active.
- What is the probability that node v_i becomes active in either of the two models?
  - In the Linear Threshold model the random threshold θ_i must be $\sum_{u \in S_i} b_{ui} \geq \theta_i$
  - 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



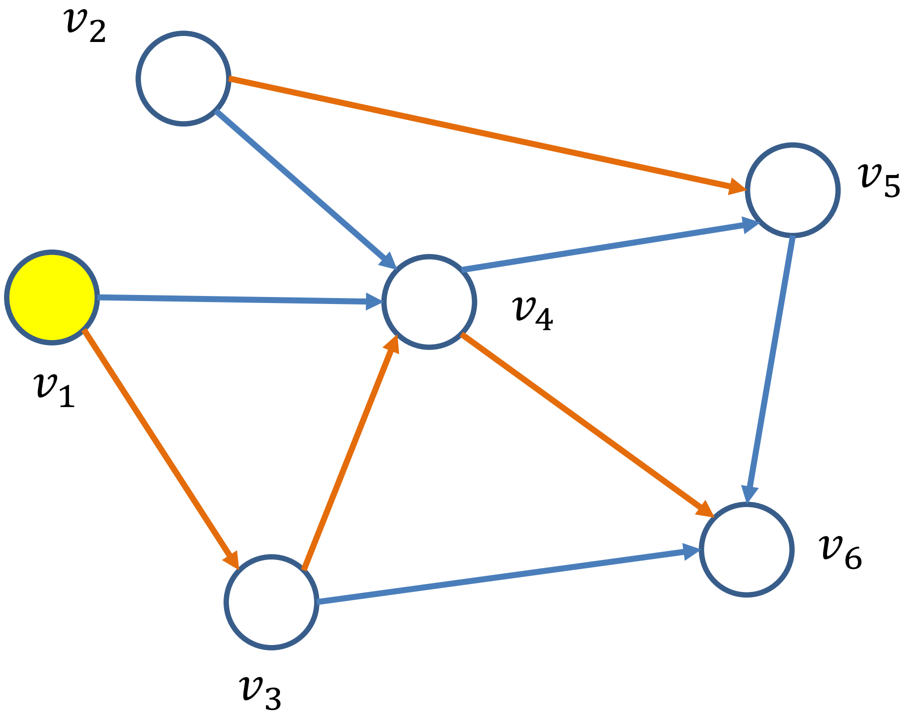
Assume that all edge weights incoming to any node sum to 1

Example



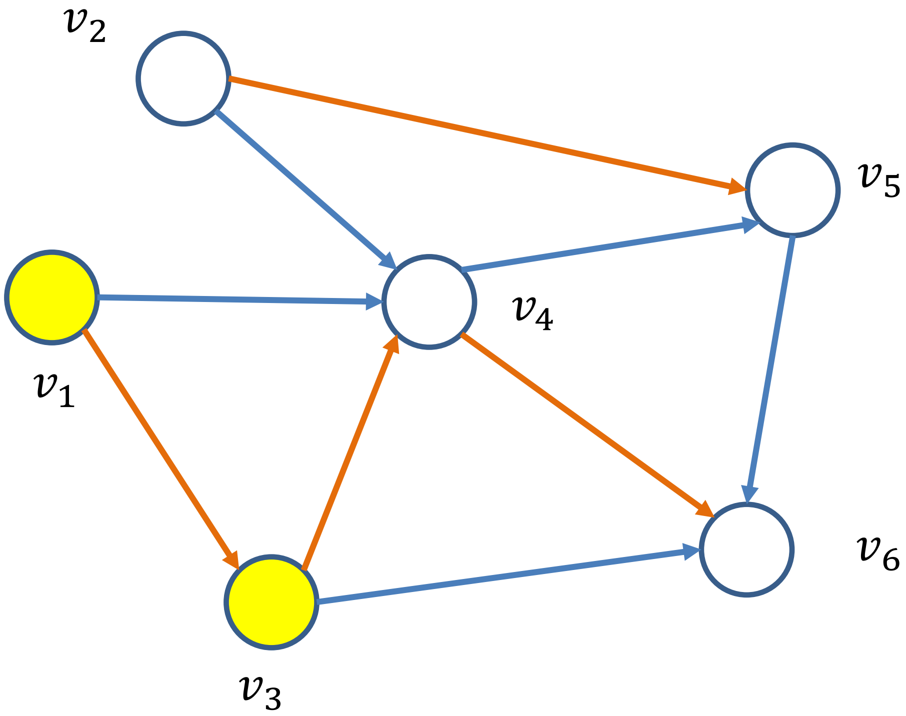
The nodes select a single incoming edge with probability equal to the weight (uniformly at random in this case)

Example



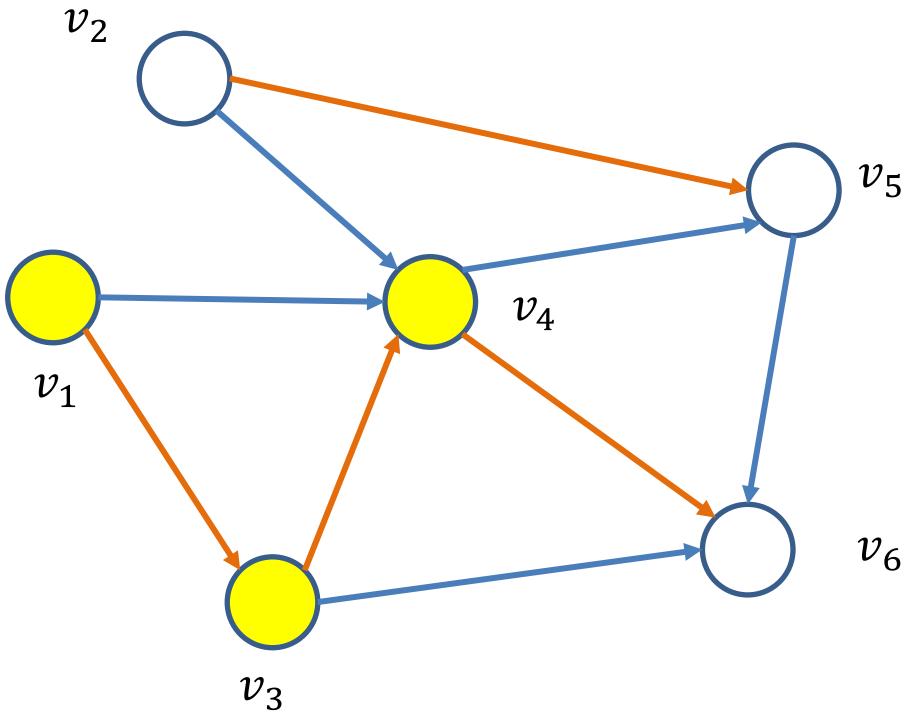
Node v_1 is the seed

Example



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

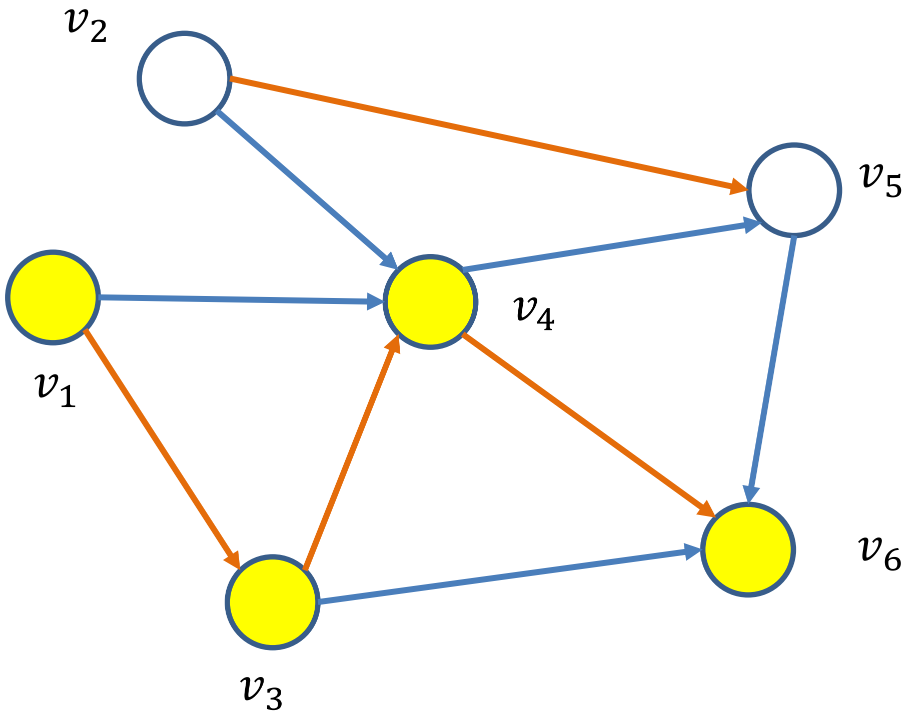
Example



The probability that node v_4 gets activated is 2/3 since it has incoming edges from two active nodes.

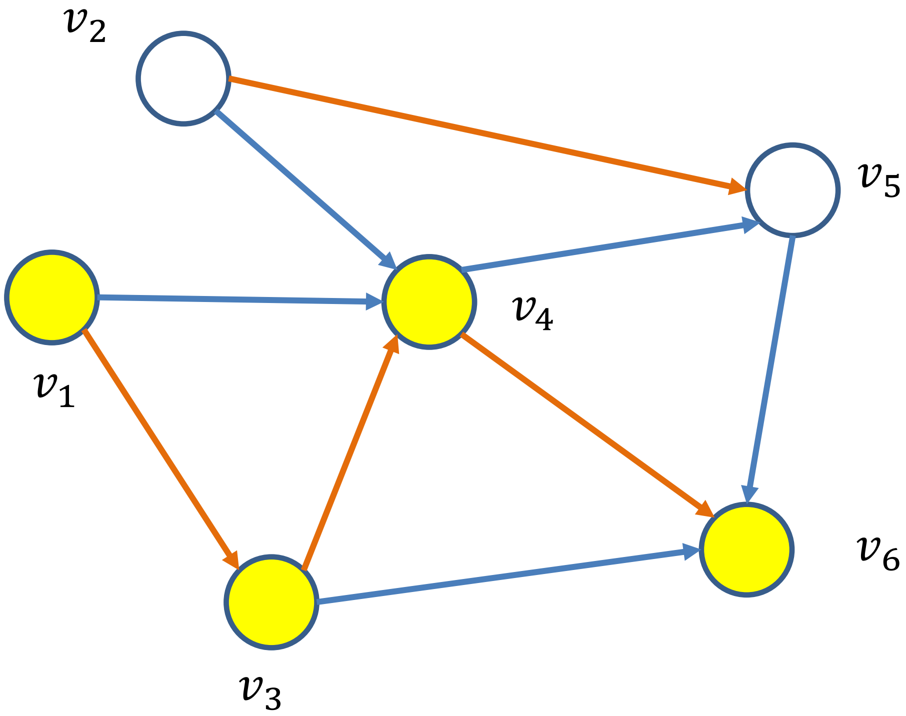
The probability that node v_4 picks one of the two edges to these nodes is also 2/3

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).

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

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

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

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.

OPINION FORMATION IN SOCIAL NETWORKS

Diffusion of items

- So far we have assumed that what is being diffused in the network is some discrete item:
  - E.g., a virus, a product, a video, an image, a link etc.
- For each network user a binary decision is being made about the item being diffused
  - Being infected by the virus, adopt the product, watch the video, save the image, retweet the link, etc.
  - This decision may happen with some probability, but the probability is over the discrete values {0,1} and the decisions usually do not change

Diffusion of opinions

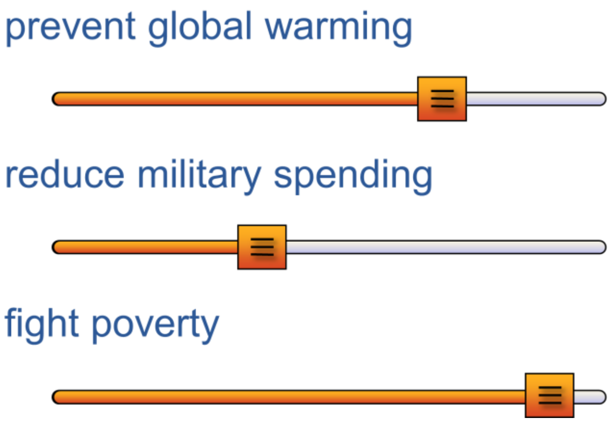
- The network can also diffuse opinions.
  - What people believe about an issue, a person, an item, is shaped by their social network
- People hold opinions that may change due to social influence
- Opinions may assume a continuous range of values, from completely negative to completely positive.
  - Opinion diffusion is different from item diffusion
  - It is often referred to as opinion formation.

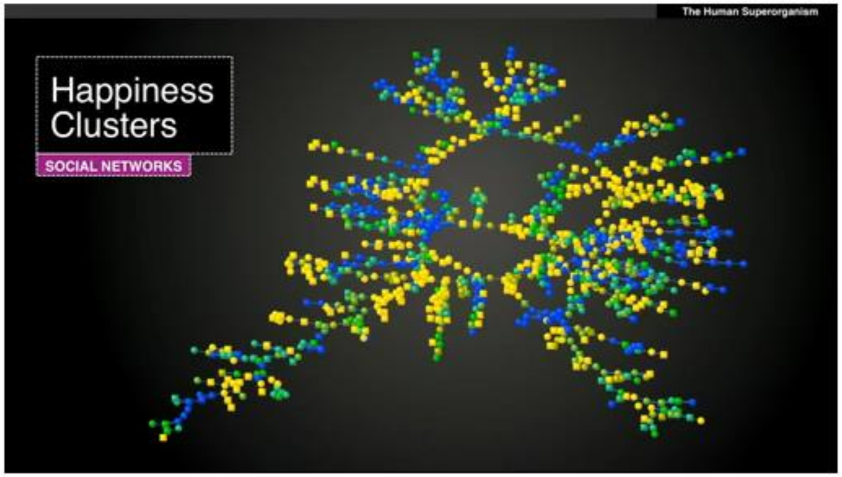
Modeling opinion formation

- There is a lot of work from different perspectives:
  - Psychologists/Sociologists: field experiments and decades of observations
  - Statistical Physicists: model humans as particles and predict their behavior
  - Mathematicians/Economists: Use game theory to model human behavior
  - Computer Scientists: build algorithms on top of the models
- Questions asked:
  - How do societies reach consensus?
    - Not always the case, but necessary for many issues in order for society to function
  - When do we get polarization or opinion clusters?
    - More realistic in the real world where consensus tends to be local

Opinion formation models

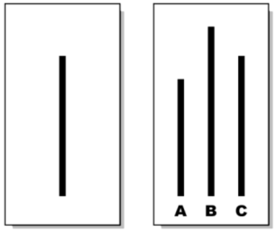
- An opinion is a real value
  - E.g., a value in the interval [0,1], or [-1,1]
- Opinions are shaped through our interactions with our social network





Social Influence

- There are two main types of social influence:
  - Normative Influence: Users influenced by opinion of neighbors due to social norms, conformity, group acceptance, avoiding ridicule, etc
  - Informational Influence: Users lacking necessary information, or not trusting their information, use opinion of neighbors to form their opinions
- Asch's conformity experiment:



Opinion formation models literature

- Long list of models
  - Ising model
    - *Claudio Castellano, Santo Fortunato, and Vittorio Loreto. 2009. Statistical physics. of social dynamics. Rev. Mod. Phys. 81 (May 2009), 591–646.*
  - Voter model
    - *Holley and Liggett. 1975. Ergodic Theorems for Weakly Interacting Infinite Systems and the Voter Model. The Annals of Probability 3, 4 (1975), 643–663.*
  - DeGroot Model
    - *DeGroot. 1974. Reaching a consensus. JASA 69, 345 (1974), 118–121*
  - Friedkin-Johnson model
    - *Friedkin and Johnsen. 1990. Social influence and opinions. Journal of Mathematical Sociology 15, 3-4 (1990), 193–206.*
  - Bounded Confidence models
    - *Deffuant, Neau, Amblard, and Weisbuch. Mixing beliefs among interacting agents.Advances in Complex Systems. 2000.*
    - *Krause. A discrete nonlinear and non–autonomous model of consensus formation. Communications in difference equations. 2000.*
  - Axelrod cultural dynamics
    - *Axelrod. The dissemination of culture: A model with local convergence and global polarization. Journal of conflict resolution. 1997.*
  - … and multiple variants of those…

De Groot opinion formation model

- Every user i has an opinion $z_i \in [0,1]$
- The opinion of each user in the network is iteratively updated, each time taking the average of the opinions of its neighbors and herself

$$z_i^t = \frac{w_{ii} z_i^{t-1} + \sum_{j \in N(i)} w_{ij} z_j^{t-1}}{w_{ii} + \sum_{j \in N(i)} w_{ij}}$$

  – where $N(i)$ is the set of neighbors of user i.

DeGroot opinion formation model

- This iterative process converges

$$z_i = \frac{w_{ii} z_i + \sum_{j \in N(i)} w_{ij} z_j}{w_{ii} + \sum_{j \in N(i)} w_{ij}}$$

- We can show that the process will converge to consensus
- At convergence $z_i = z^*$ for all nodes i

What about personal biases?

- People tend to cling on to their personal opinions

The Friedkin and Johnsen opinion formation model

- Every user i has an intrinsic opinion $s_i \in [0,1]$ and an expressed opinion $z_i \in [0,1]$
- The public opinion z_i of each user in the network is iteratively updated, each time taking the average of the expressed opinions of its neighbors and the intrinsic opinion of herself

$$z_i^t = \frac{w_{ii} s_i + \sum_{j \in N(i)} w_{ij} z_j^{t-1}}{w_{ii} + \sum_{j \in N(i)} w_{ij}}$$

The Friedkin and Johnsen opinion formation model

- The FJ model also converges but not to a consensus
- At convergence:

$$z_i = \frac{w_{ii} s_i + \sum_{j \in N(i)} w_{ij} z_j}{w_{ii} + \sum_{j \in N(i)} w_{ij}}$$

Opinion formation as a game

- Assume that network users are rational (selfish) agents
- Each user has a personal cost for expressing an opinion

$$c(z_i) = w_{ii}(z_i - s_i)^2 + \sum_{j \in N(i)} w_{ij}\left(z_i - z_j\right)^2$$

Inconsistency cost: The cost for deviating from one's intrinsic opinion

Conflict cost: The cost for disagreeing with the opinions in one's social network

- Each user is selfishly trying to minimize her personal cost.

D. Bindel, J. Kleinberg, S. Oren. *How Bad is Forming Your Own Opinion?* Proc. 52nd IEEE Symposium on Foundations of Computer Science, 2011.

Opinion formation as a game

- The opinion z_i that minimizes the personal cost of user i

$$z_i = \frac{w_{ii}s_i + \sum_{j \in N(i)} w_{ij} z_j}{w_{ii} + \sum_{j \in N(i)} w_{ij}}$$

- In linear algebra terms (assume 0/1 weights):

$$(L + I)\boldsymbol{z} = \boldsymbol{s} \;\Rightarrow \boldsymbol{z} = (L + I)^{-1}\boldsymbol{s}$$

where L is the Laplacian of the graph.

Reminder: The Laplacian is the negated adjacency matrix with the degree on the diagonal $L = D - A$, where D is a diagonal matrix with the degrees

Understanding opinion formation

- To better study the opinion formation process we will show a connection between opinion formation and absorbing random walks.

Random Walks on Graphs

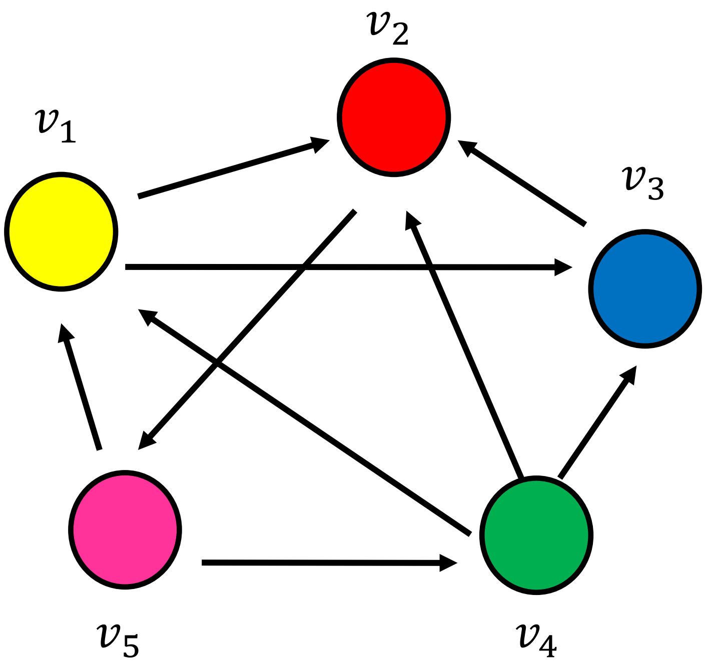
- A random walk is a stochastic process performed on a graph

- Random walk:
  - Start from a node chosen uniformly at random with probability $\frac{1}{n}$.
  - Pick one of the outgoing edges uniformly at random
  - Move to the destination of the edge
  - Repeat.

- Made very popular with Google's PageRank algorithm.

The Transition Probability matrix

$$\mathrm{A} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix}$$



$$\mathrm{P} = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \end{bmatrix}$$

$P[i,j] = 1/d_{out}(i)$: Probability of transitioning from node i to node j.

Node Probability vector

- The vector $p^t \;=\; (p_1^t, p_2^t, \dots\;, p_n^t)$ that stores the probability of being at node v_i at step t
  - p_i^0= the probability of starting from state i (usually) set to uniform
- We can compute the vector p^t at step t using a vector-matrix multiplication

$$p^t = p^{t-1}\,P = p^0 P^t$$

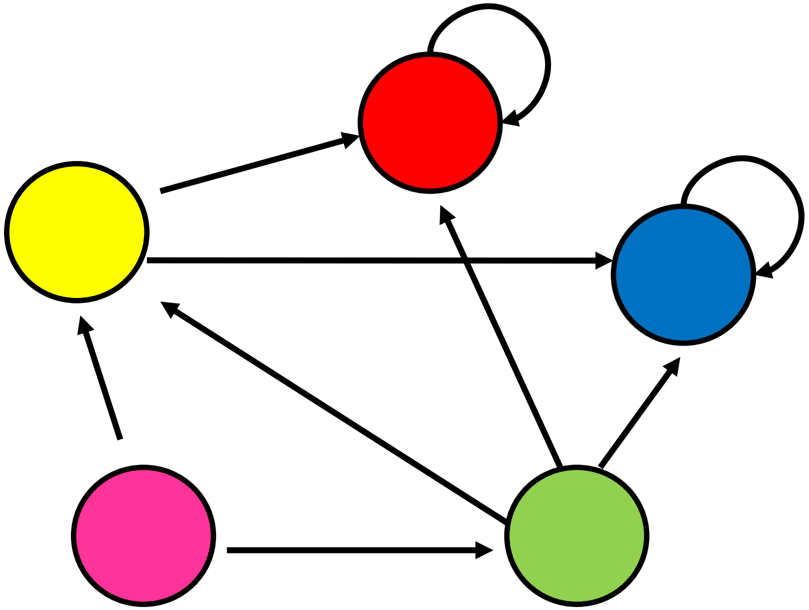
- After many steps $p^t \to \pi$ the probability converges to the stationary distribution π

Stationary distribution

- The stationary distribution of a random walk with transition matrix P, is a probability distribution π, such that $\pi = \pi P$
- The stationary distribution is independent of the initial vector if the graph is strongly connected, and not bipartite.
- All the rows of the matrix P^{∞} are equal to the stationary distribution π
- The stationary distribution is an eigenvector of matrix P
  - the principal left eigenvector of P – stochastic matrices have maximum eigenvalue 1
- The probability π_i is the fraction of times that we visited state i as $t \rightarrow \infty$

Random walk with absorbing nodes

- Absorbing nodes: nodes from which the random walk cannot escape.

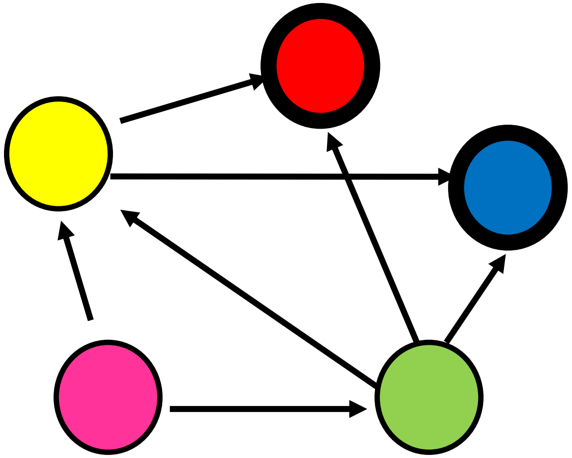


- Two absorbing nodes: the red and the blue.

P. G. Doyle, J. L. Snell. *Random Walks and Electrical Networks*. 1984

Absorption probability

- In a graph with more than one absorbing nodes a random walk that starts from a non-absorbing (transient) node t will be absorbed in one of them with some probability
  - For a transient node t we can compute the probability of absorption at an absorbing node s



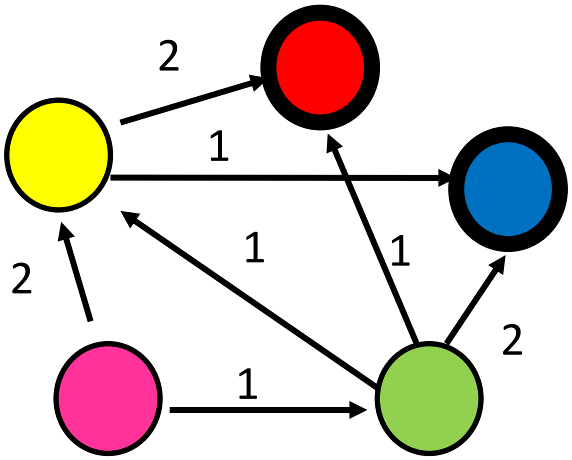
Absorption probabilities

- Computing the probability of being absorbed:
  - The absorbing nodes have probability 1 of being absorbed in themselves and zero of being absorbed in another node.
  - For the non-absorbing nodes, take the (weighted) average of the absorption probabilities of your neighbors
    - if one of the neighbors is the absorbing node, it has probability 1
  - Repeat until convergence (= very small change in probs)

$$P(Red|Pink) = \frac{2}{3}P(Red|Yellow) + \frac{1}{3}P(Red|Green)$$

$$P(Red|Green) = \frac{1}{4}P(Red|Yellow) + \frac{1}{4}$$

$$P(Red|Yellow) = \frac{2}{3}$$



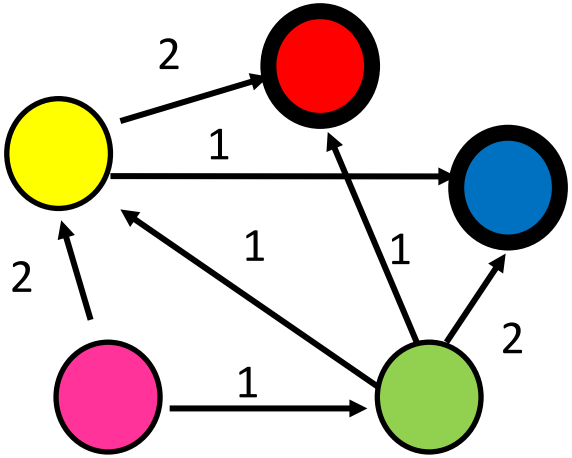
Absorption probabilities

- Computing the probability of being absorbed:
  - The absorbing nodes have probability 1 of being absorbed in themselves and zero of being absorbed in another node.
  - For the non-absorbing nodes, take the (weighted) average of the absorption probabilities of your neighbors
    - if one of the neighbors is the absorbing node, it has probability 1
  - Repeat until convergence (= very small change in probs)

$$P(Blue|Pink) = \frac{2}{3}P(Blue|Yellow) + \frac{1}{3}P(Blue|Green)$$

$$P(Blue|Green) = \frac{1}{4}P(Blue|Yellow) + \frac{1}{2}$$

$$P(Blue|Yellow) = \frac{1}{3}$$



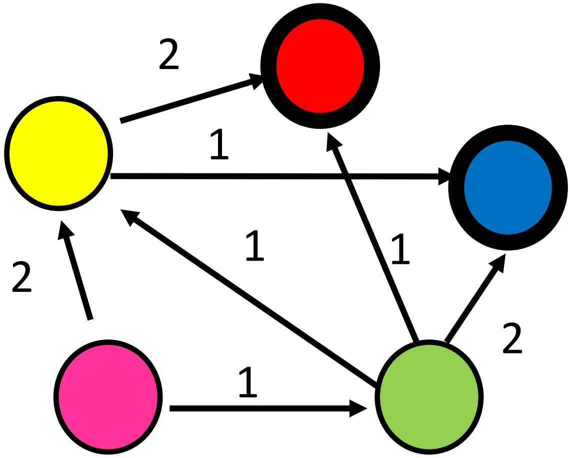
Absorption probabilities

- Computing the probability of being absorbed:
  - The absorbing nodes have probability 1 of being absorbed in themselves and zero of being absorbed in another node.
  - For the non-absorbing nodes, take the (weighted) average of the absorption probabilities of your neighbors
    - if one of the neighbors is the absorbing node, it has probability 1
  - Repeat until convergence (= very small change in probs)

General equation for the probability of transient node t being absorbed at absorbing node a

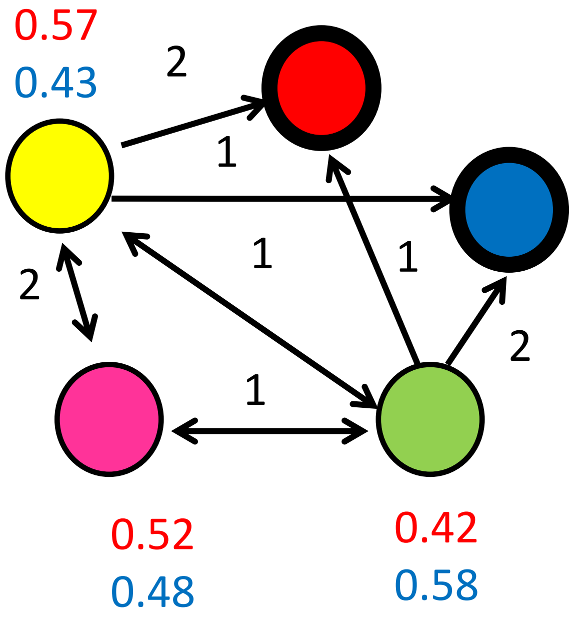
$$P(a|t) = \sum_{(t,x)\in E} P[t,x]P(a|x)$$

The weighted average of the neighbors



Absorption probabilities

- Compute the absorption probabilities for red and blue



Absorption probabilities

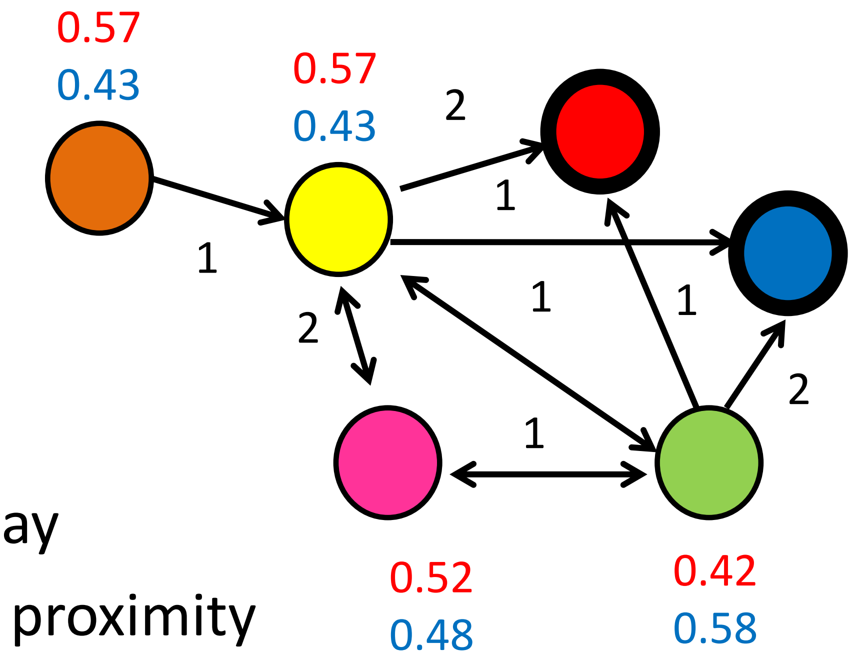
- The absorption probability has several practical uses.
- Given a graph (directed or undirected) we can choose to make some nodes absorbing.
  - Simply direct all edges incident on the chosen nodes towards them and create a self-loop.
- The absorbing random walk provides a measure of proximity of transient nodes to the chosen nodes.
  - Useful for understanding proximity in graphs
  - Useful for propagation in the graph
    - E.g, on a social network some nodes are malicious, while some are certified, to which class is a transient node closer?

Penalizing long paths

- The orange node has the same probability of reaching red and blue as the yellow one

$$P(Red|Orange) = P(Red|Yellow)$$

$$P(Blue|Orange) = P(Blue|Yellow)$$



- Intuitively though it is further away
- The probability does not capture proximity

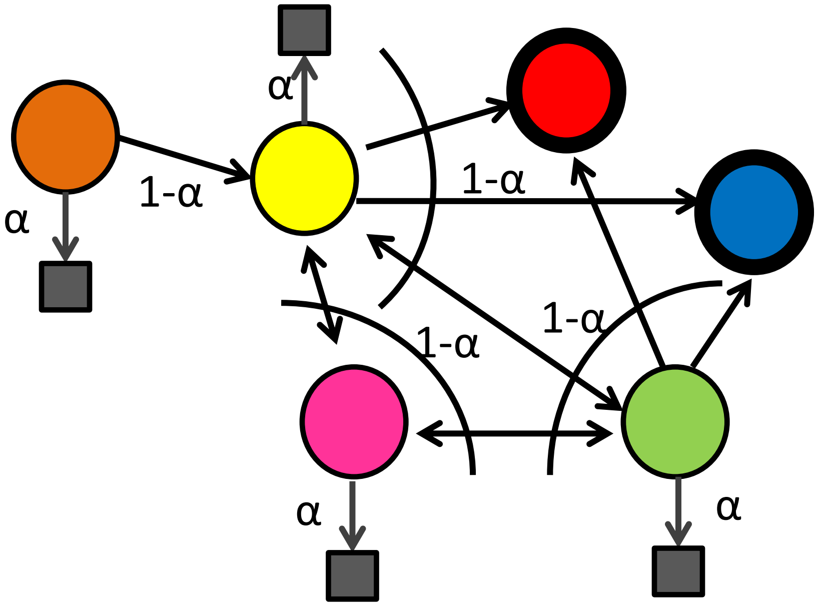
Penalizing long paths

- Add a universal absorbing node to which each node gets absorbed with probability α.

With probability α the random walk dies

With probability (1-α) the random walk continues as before

The longer the path from a node to an absorbing node the more likely the random walk dies along the way, the lower the absorbtion probability



$$P(Red|Green) = (1-\alpha)\left(\frac{1}{4}P(Red|Yellow) + \frac{1}{4}P(Red|Pink) + \frac{1}{4}\right)$$

Linear Algebra

- The transition matrix of the absorbing random walk looks like this

$$P = \begin{array}{c} \begin{array}{cc} \text{T} & \text{A} \end{array} \\ \left[\begin{array}{c|c} P_{TT} & P_{TA} \\ \hline 0 & I \end{array}\right] \end{array}$$

T: transient

A: absorbing

- P_{TT}: transition probabilities between transient nodes
- P_{TA}: transition probabilities from transient to absorbing nodes
- Computing the absorption probabilities corresponds to iteratively multiplying matrix P with itself

Linear Algebra

- After many iterations:

$$P^{\infty} = \begin{bmatrix} 0 & Q \\ 0 & I \end{bmatrix}$$

- The matrix Q holds the absorption probabilities
    - $Q[i,k]$ = The probability of being absorbed in absorbing state a_k when starting from transient state t_i

$$Q = P_{TA} + P_{TT}P_{TA} + P_{TT}^2 P_{TA} + \cdots$$

Linear algebra

- The fundamental matrix

$$F = P_{TT} + P_{TT}^2 + \cdots = \sum_{i=1}^{\infty} P_{TT}^i = (1 - P_{TT})^{-1}$$

- $F[i,j]$ =The sum of probabilities of visiting transient state t_j when starting from state t_i after any number of steps
- Also: The expected number of visits to transient state t_j when starting from state t_i after any number of steps
- The transient-to-absorbing matrix Q

$$Q = FP_{T\mathrm{A}}$$

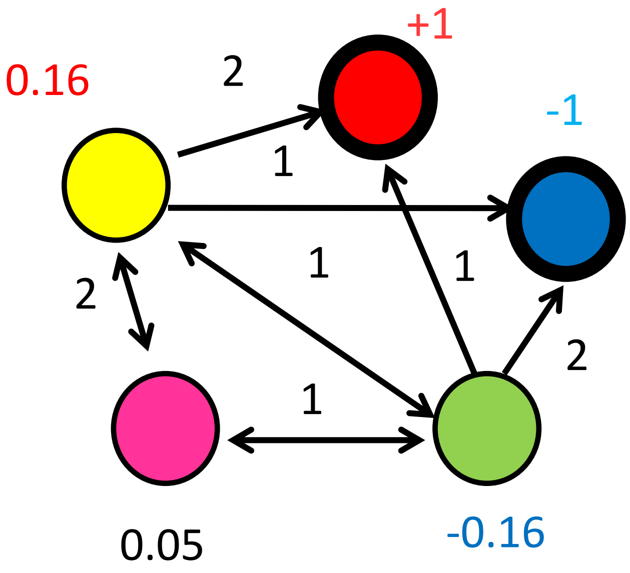
Propagating values

- Assume that Red has a positive value and Blue a negative value
- We can compute a value for all transient nodes in the same way we compute probabilities
  - This is the expected value at the absorbing node for the random walk that starts from a transient node

$$V(Pink) = \frac{2}{3}V(Yellow) + \frac{1}{3}V(Green)$$

$$V(Green) = \frac{1}{5}V(Yellow) + \frac{1}{5}V(Pink) + \frac{1}{5} - \frac{2}{5}$$

$$V(Yellow) = \frac{1}{6}V(Green) + \frac{1}{3}V(Pink) + \frac{1}{3} - \frac{1}{6}$$



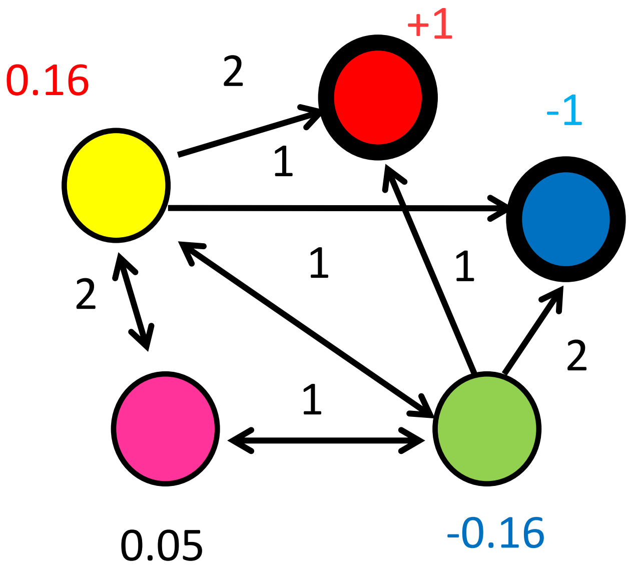
Propagating values

- Assume that Red has a positive value and Blue a negative value
- We can compute a value for all transient nodes in the same way we compute probabilities
  - This is the expected value at the absorbing node for the non-absorbing node

General equation for value propagation:

$$v(i) = \sum_{(i,j)\in E} P[i,j]v(j)$$

The value of i is the weighted average of the values of its neighbors



Linear algebra

- Computation of values is essentially multiplication of the matrix Q with the vector of values of the absorbing nodes

$$\boldsymbol{v} = Q\boldsymbol{s}$$

  - $\boldsymbol{s}$: vector of values of the absorbing nodes
  - $\boldsymbol{v}$: vector of values of the transient nodes

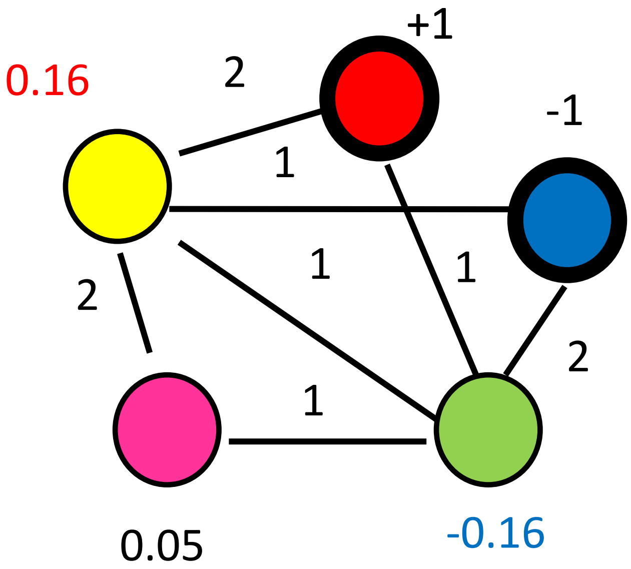
Electrical networks and random walks

- Our graph corresponds to an electrical network
- There is a positive voltage of +1 at the Red node, and a negative voltage -1 at the Blue node
- There are resistances on the edges inversely proportional to the weights (or conductance proportional to the weights)
- The computed values are the voltages at the nodes

$$V(Pink) = \frac{2}{3}V(Yellow) + \frac{1}{3}V(Green)$$

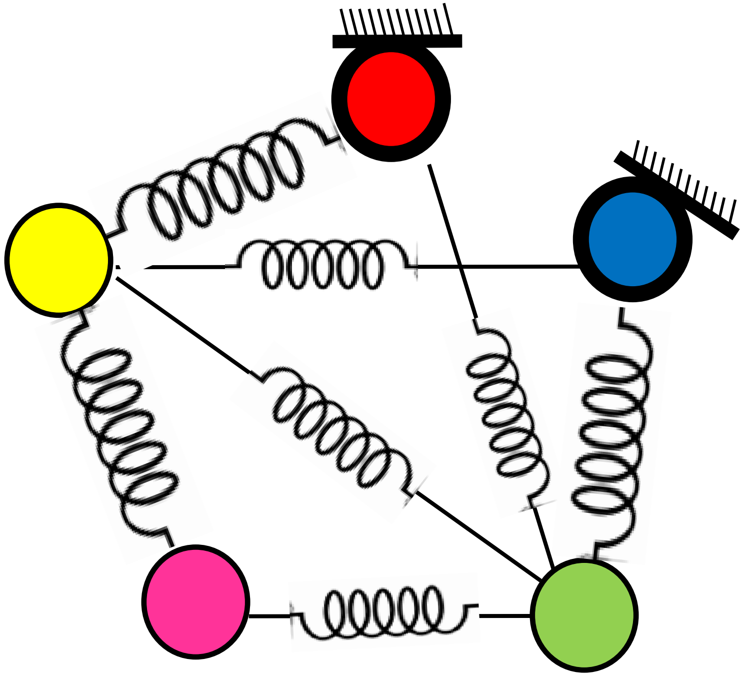
$$V(Green) = \frac{1}{5}V(Yellow) + \frac{1}{5}V(Pink) + \frac{1}{5} - \frac{2}{5}$$

$$V(Yellow) = \frac{1}{6}V(Green) + \frac{1}{3}V(Pink) + \frac{1}{3} - \frac{1}{6}$$



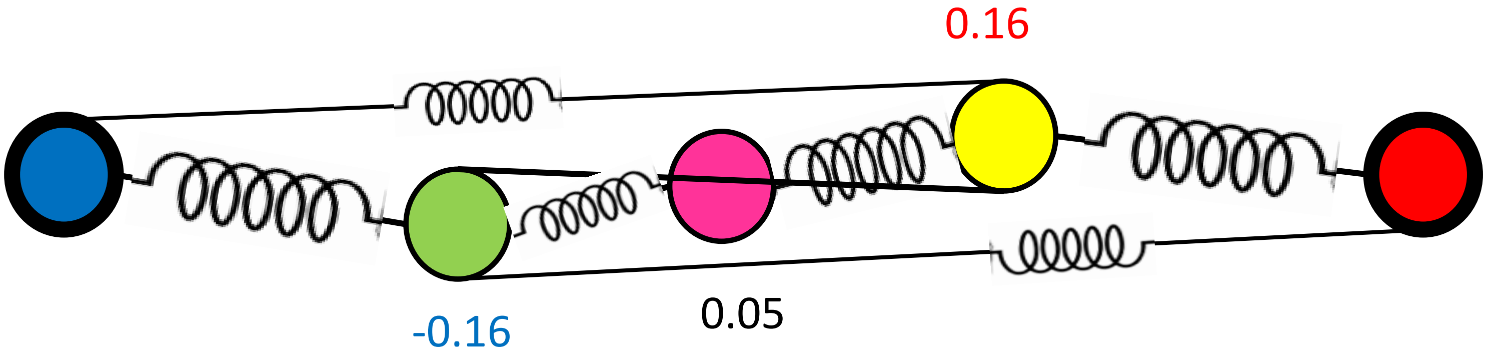
Springs and random walks

- Our graph corresponds to a spring system
- The Red node is pinned at position +1, while the Blue node is pinned at position -1 on a line.
- There are springs on the edges with hardness proportional to the weights
- The computed values are the positions of the nodes on the line



Springs and random walks

- Our graph corresponds to a spring system
- The Red node is pinned at position +1, while the Blue node is pinned at position -1 on a line.
- There are springs on the edges with hardness proportional to the weights
- The computed values are the positions of the nodes on the line

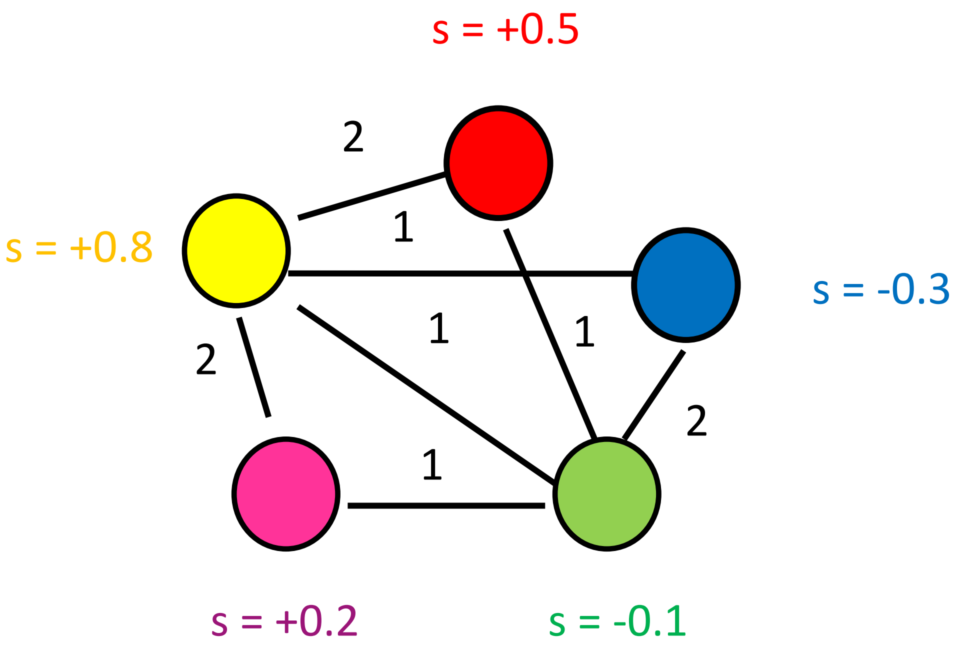


Label Propagation and Transductive Learning

- If we have a graph of relationships and some labels on some nodes we can propagate them to the remaining nodes
  - Make the labeled nodes to be absorbing and compute the absorption probabilities for the rest of the graph
  - E.g., a social network where some people are tagged as spammers
  - E.g., the movie-actor graph where some movies are tagged as action or comedy.
- This is a form of semi-supervised learning/classification
  - We make use of the unlabeled data, and the relationships
- It is also called transductive learning because it does not produce a model, but just labels the unlabeled data that is at hand.
  - Contrast to inductive learning that learns a model and can label any new example

Back to opinion formation

- The value propagation we described is closely related to the opinion formation process/game we defined.
  - Can you see how we can use absorbing random walks to model the opinion formation for the network below?



Reminder:

$$z_i = \frac{s_i + \sum_{j \in N(i)} w_{ij} z_j}{1 + \sum_{j \in N(i)} w_{ij}}$$

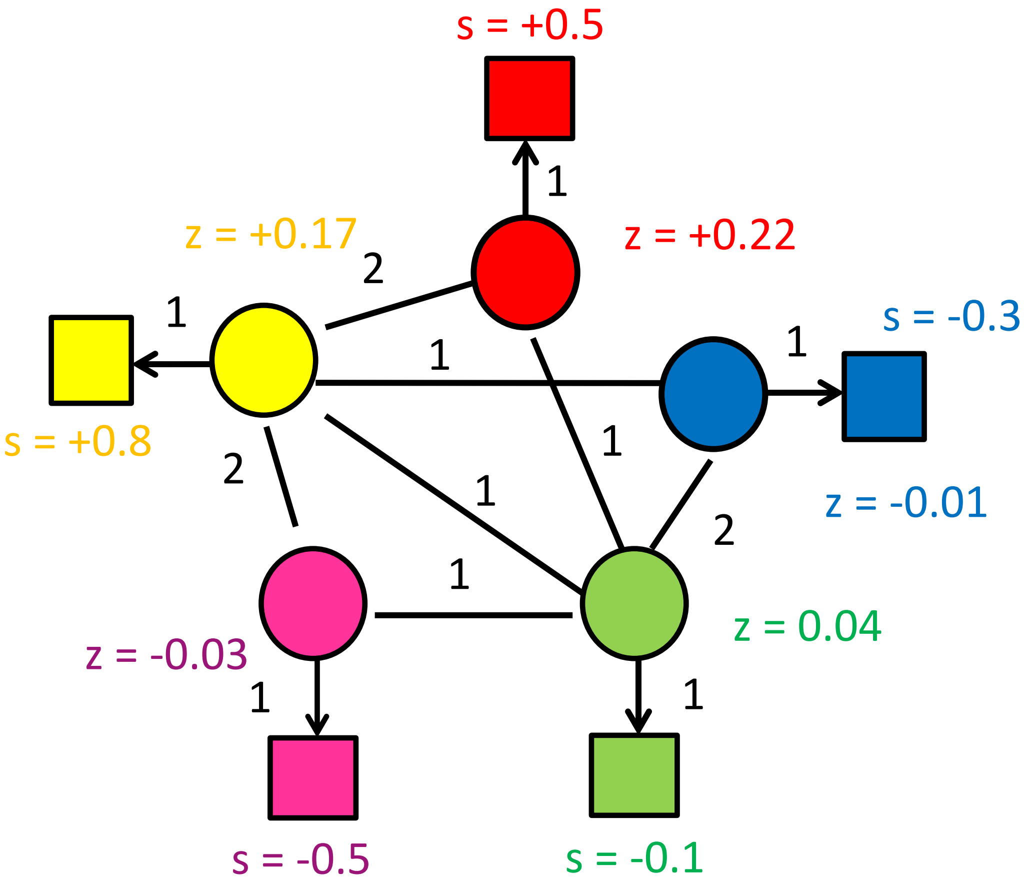
Opinion formation and absorbing random walks

Add to the network one absorbing node per user with value the intrinsic opinion of the user

Connect each transient node to her absorbing node with weight w_{ii}

The expressed opinion for each node is computed using the value propagation we described

- Repeated averaging



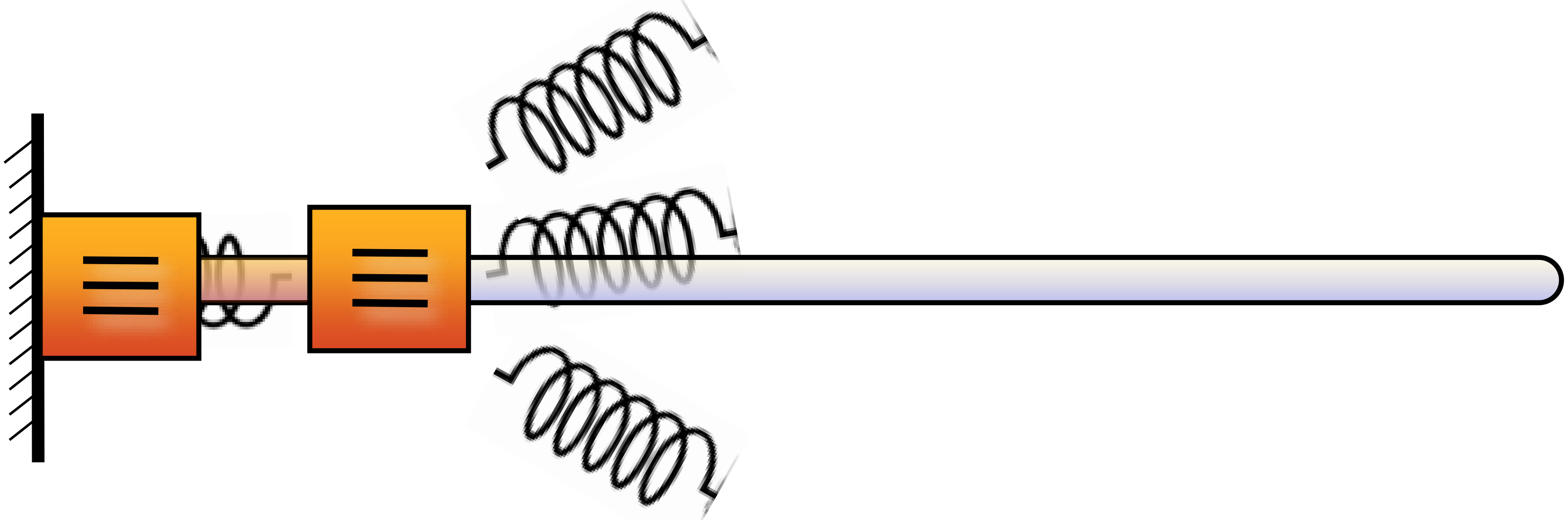
$$v(red) = \frac{0.5 + 2 \cdot v(yellow) + v(green)}{4}$$

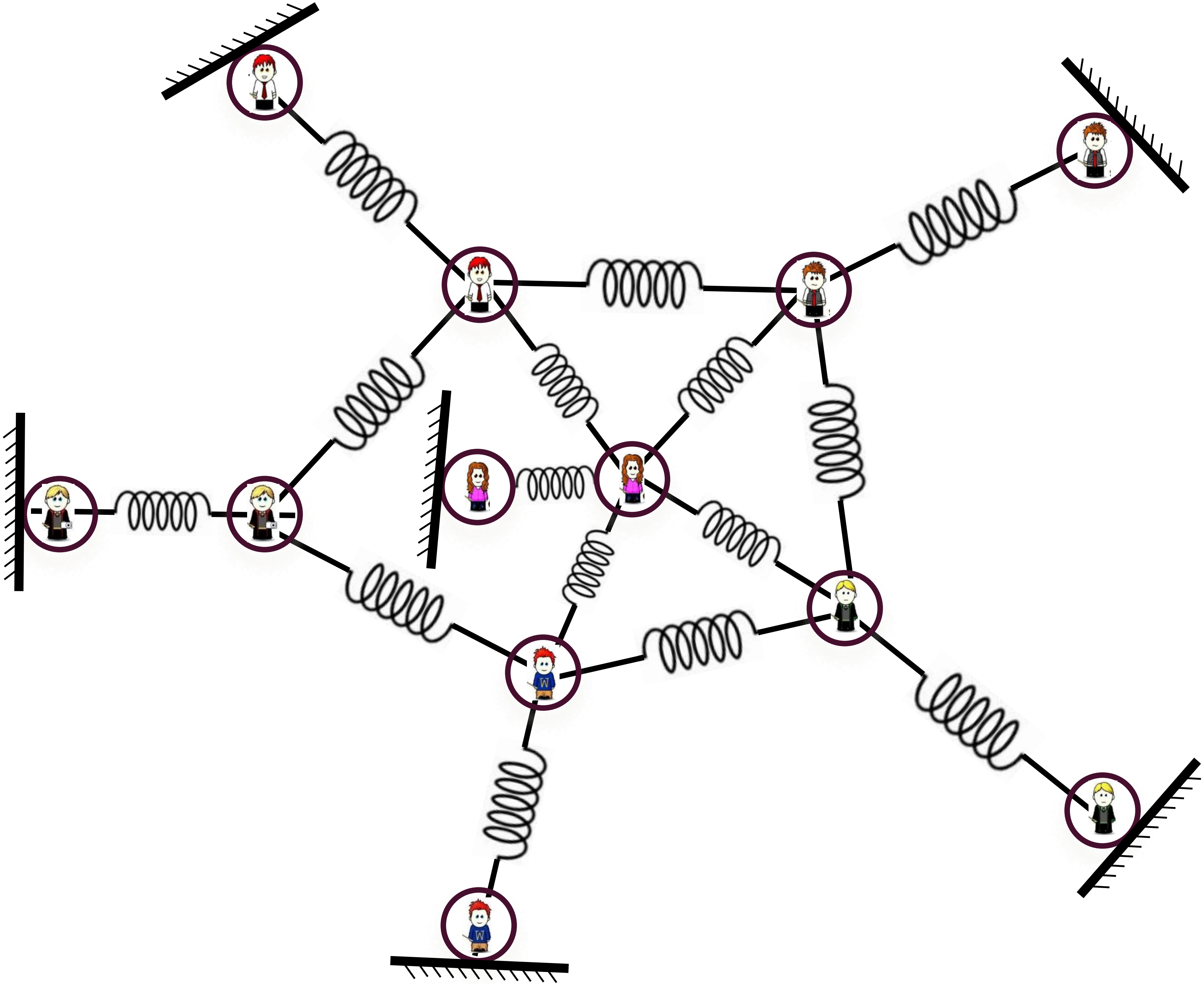
$$z_{red} = \frac{0.5 + 2 \cdot z_{yellow} + z_{green}}{4}$$

It is equal to the expected intrinsic opinion at the place of absorption

Opinion of a user

- For an individual user u
  - u's absorbing node is a stationary point
  - u's transient node is connected to the absorbing node with a spring.
  - The neighbors of u pull with their own springs.




W
W

Opinion maximization problem

- Public opinion:

$$g(z) = \sum_{i \in V} z_i$$

- Problem: Given a graph G, the given opinion formation model, the intrinsic opinions of the users, and a budget k, perform k interventions such that the public opinion is maximized.

- Useful for image control campaign.

- What kind of interventions should we do?

Possible interventions

1. Fix the expressed opinion of k nodes to the maximum value 1.
   - Essentially, make these nodes absorbing, and give them value 1.

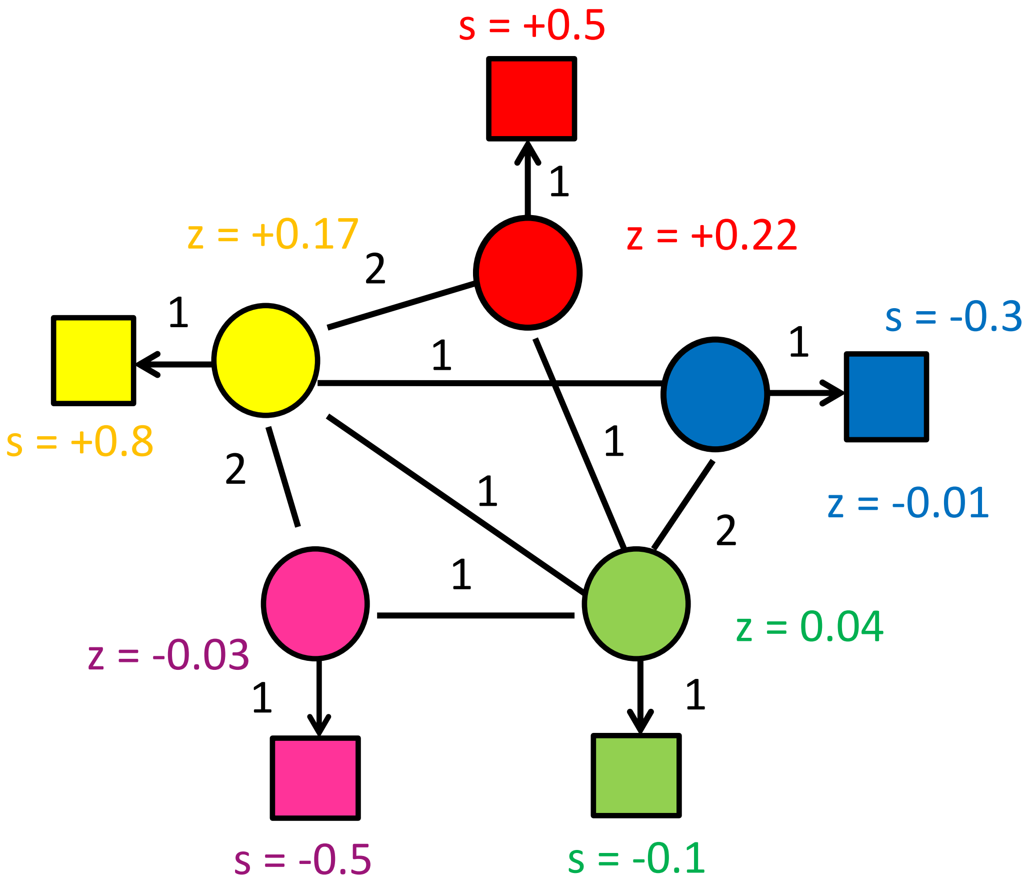
2. Fix the intrinsic opinion of k nodes to the maximum value 1.
   - Easy to solve, we know exactly the contribution of each node to the overall public opinion.

3. Change the underlying network to facilitate the propagation of positive opinions.
   - For undirected graphs this is not possible

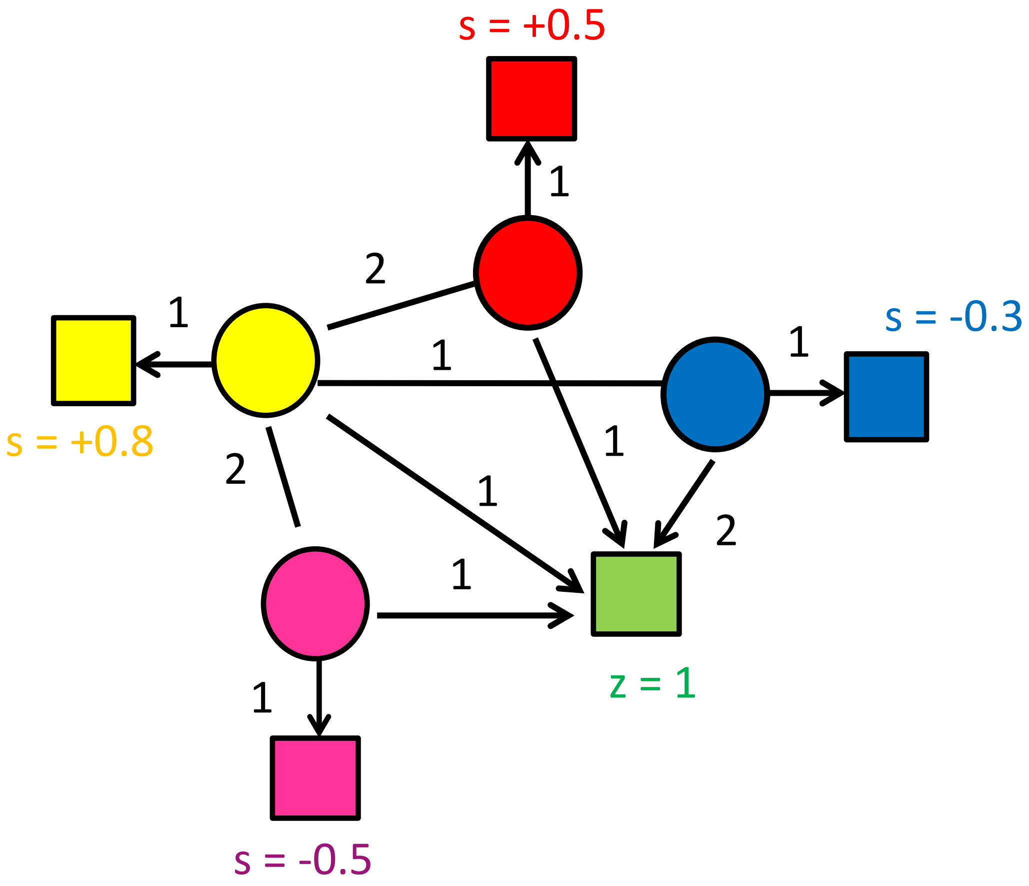
$$g(z) = \sum_i z_i = \sum_i s_i$$

   - The overall public opinion does not depend on the graph structure!
   - What does this mean for the wisdom of crowds?

Fixing the expressed opinion



Fixing the expressed opinion



Opinion maximization problem

- The opinion maximization problem is NP-hard.
- The public opinion function is monotone and submodular
  - The Greedy algorithm gives a $\left(1-\frac{1}{e}\right)$-approximate solution
- In practice Greedy is slow. Heuristics that use random walks perform well.

A. Gionis, E. Terzi, P. Tsaparas. *Opinion Maximization in Social Networks*. SDM 2013

Additional models

- Ising model
- Voter model
- Bounded confidence models
- Axelrod cultural dynamics model

A Physics-based model

- The Ising ferromagnet model:
  - A user i is a "spin" s_i that can assume two values: ±1
  - The total energy of the system is

$$H = -\frac{1}{2}\sum_{\langle i,j\rangle} s_i s_j$$

Defined over the neighboring pairs

  - A spin is flipped with probability $\exp(-\frac{\Delta\mathrm{E}}{\mathrm{T}})$ where ΔE is the change in energy, and T is the "temperature" of the system.
- The model assumes no topology
  - Complete graph (all-with-all), or regular lattice.
- For low temperatures, the system converges to a single opinion

The Voter model

- Each user has an opinion that is an integer value
  - Usually opinions are in {0,1} but multiple opinion values are also possible.
- Opinion formation process:
  - At each step we select a user at random
  - The user selects one of its neighbors at random (including herself) and adopts their opinion
- The model can be proven to converge for certain topologies.

Bounded confidence model

- Confirmation bias: People tend to accept opinions that agree with them
  - "Why facts don't change our minds" (New Yorker)

- Bounded Confidence model: A user i is influenced by a neighbor j only if

$$\left|z_i - z_j\right| \leq \epsilon$$

for some parameter ϵ

Bounded Confidence models

- Defuant model: Given a parameter μ at time t, a randomly selected user i selects a neighbor j at random, and if $\left|z_i^t - z_j^t\right| \leq \epsilon$ their opinions are updated as:

$$z_i^{t+1} = z_i^t + \mu(z_j^t - z_i^t)$$
$$z_j^{t+1} = z_j^t + \mu(z_i^t - z_j^t)$$

Similar to Voter model

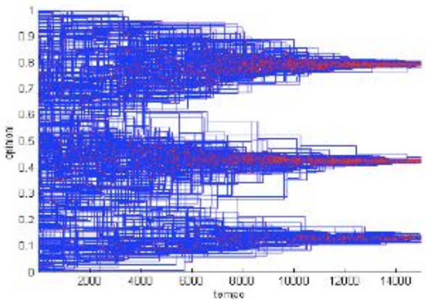
- Hegselmann-Krause (HK) model: Each node i updates their opinions as the average of the opinions of the neighbors that agree with them

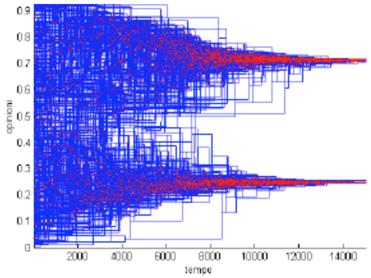
$$z_i^t = \frac{w_{ii} z_i^{t-1} + \sum_{j \in N(i): \left|z_i^t - z_j^t\right| \leq \epsilon} w_{ij} z_j^{t-1}}{w_{ii} + \sum_{j \in N(i): \left|z_i^t - z_j^t\right| \leq \epsilon} w_{ij}}$$

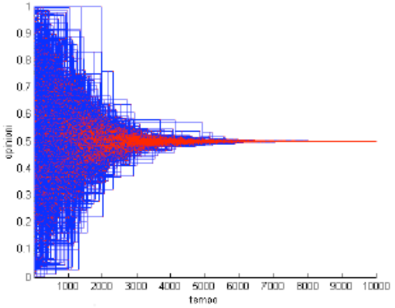
Similar to DeGroot model

Bounded Confidence models

- Depending on the parameter ϵ and the initial opinions, bounded confidence models can lead to plurality (multiple opinions), polarization (two competing opinions), or consensus (single opinion)







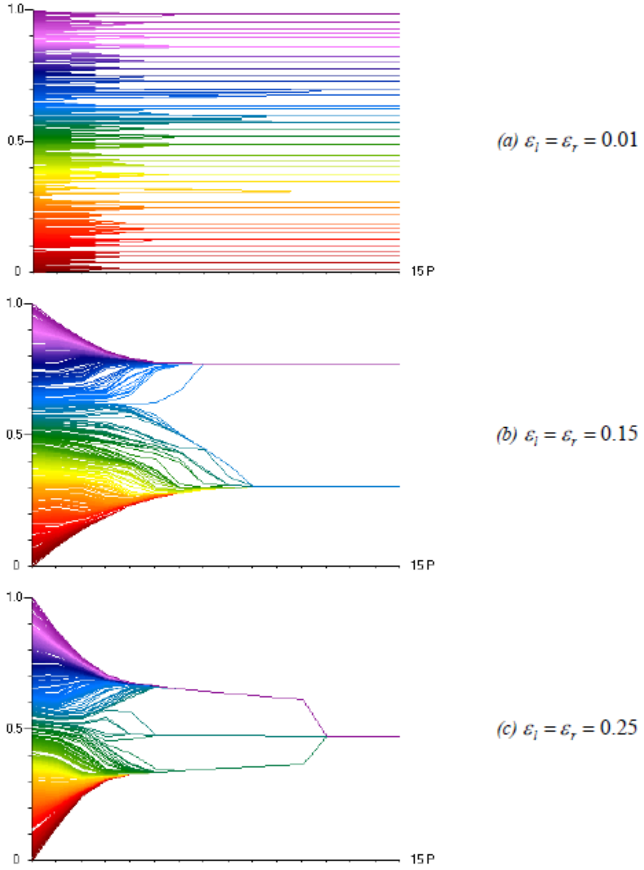


Figure 2: Stops while walking along the diagonal

Axelrod model

- Cultural dynamics: Goes beyond single opinions, and looks at different features/habits/traits
  - Tries to model the effects of social influence and homophily.
- Model:
  - Each user i has a vector σ_i of F features
  - A user i decides to interact with user j with probability

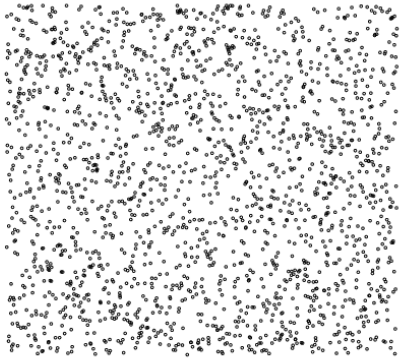
$$\omega_{ij} = \frac{1}{\mathrm{F}} \sum_{f=1}^{F} \delta(\sigma_i(f), \sigma_j(f))$$

Fraction of common features

  - If there is interaction, the user changes one of the disagreeing features to the value of the neighbor
- The state where all users have the same features is an equilibrium, but it is not always reached (cultural pockets)

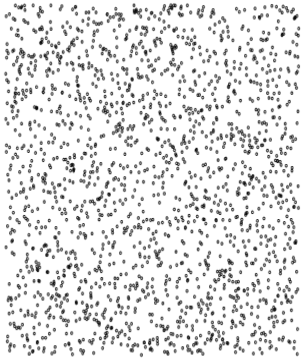
Empirical measurements

- There have been various experiments for validating the different models in practice

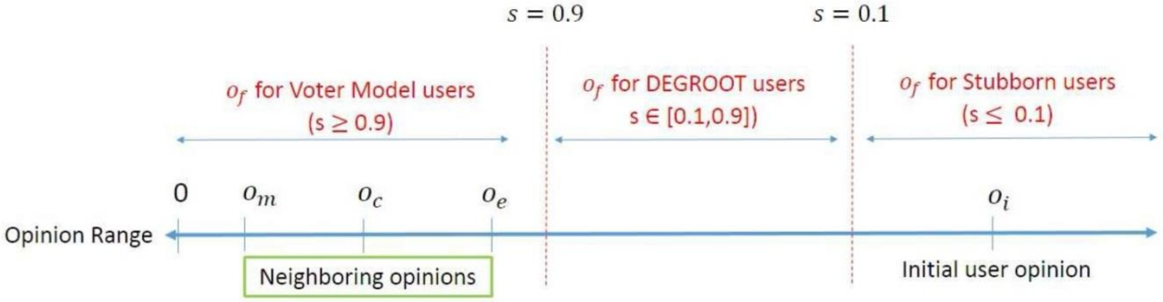


- Das, Gollapudi, Munagala (WSDM 2014)
  - User surveys:
    - estimate number of dots in images
    - Estimate annual sales of various brands.
  - For each survey:
    - Users asked to provide initial answers on all questions in the survey
    - Then, each user shown varying number of (synthetic) neighboring answers.
    - Users given opportunity to update their answers

Online User Studies

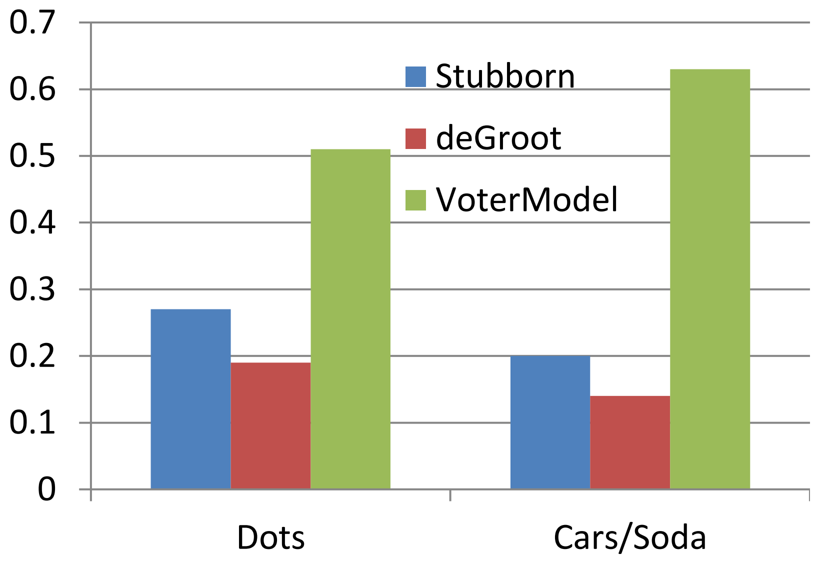


?

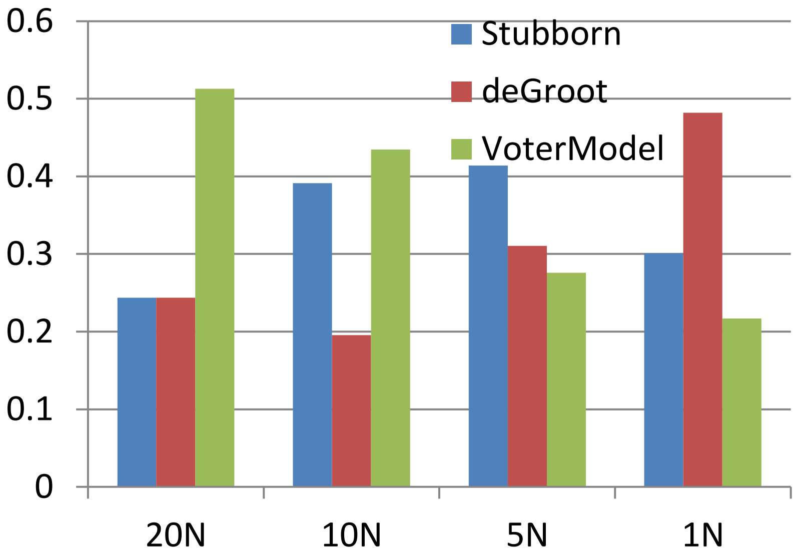


- Define $s = \frac{|o_i - o_f|}{|o_i - o_e|}$
  - (o_i: original opinion, o_f: final opinion, o_e: closest neighboring opinion)
- User behavior categorized as:
  - Stubborn ($s < 0.1$)
  - DeGroot ($0.1 < s < 0.9$)
  - Voter ($s > 0.9$)

Voter vs DeGroot



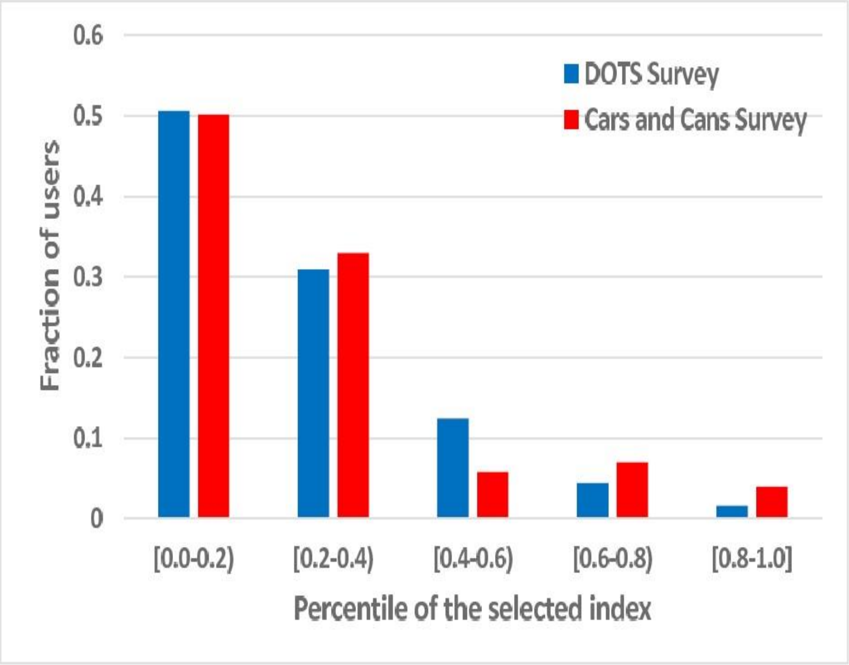
Distribution over stubborn, deGroot and voter



Effect of number of neighboring opinions

- Voter model is prevalent for large number of neighbors,
- DeGroot becomes more prevalent for smaller number of neighbors

Biased Conforming Behavior



- Adoption of neighboring opinions not uniform random (unlike Voter Model)
- Users give higher weights to “close by” opinions

Biased Voter Model

- Users update their opinions iteratively
- At each iteration, a user sorts the opinions of her neighbors in increasing order of distance from her own opinion
- With probability p she adopts the 1st closest, else with probability p she adopts the 2nd closest, else …
- If no opinion has been adopted, with probability α she keeps her own opinion
- With probability $1 - \alpha$ she adopts an opinion chosen uniformly at random between her own opinion and the closest neighboring opinion

Other problems related to opinion formation

- Modeling polarization
  - Understand why extreme opinions are formed and people cluster around them
- Modeling herding/flocking
  - Understand under what conditions people tend to follow the crowd
- Modeling the backfire effect
  - Understand when opposite opinions lead to strengthening your opinion, rather than moderating it
- Computational Sociology
  - Use big data for studying and modeling human social behavior.

R. Hegselmann, U. Krause. *Opinion Dynamics and Bounded Confidence. Models, Analysis, and Simulation*. Journal of Artificial Societies and Social Simulation (JASSS) vol.5, no. 3, 2002

Acknowledgements

- Many thanks to Evimaria Terzi, Aris Gionis and Sreenivas Gollapudi for their generous slide contributions.

References

- M. H. DeGroot. *Reaching a consensus*. J. American Statistical Association, 69:118–121, 1974.
- N. E. Friedkin and E. C. Johnsen. *Social influence and opinions*. J. Mathematical Sociology, 15(3-4):193–205, 1990.
- D. Bindel, J. Kleinberg, S. Oren. *How Bad is Forming Your Own Opinion?* Proc. 52nd IEEE Symposium on Foundations of Computer Science, 2011.
- P. G. Doyle, J. L. Snell. *Random Walks and Electrical Networks*. 1984
- Grinstead and Snell's Introduction to Probability
- A. Gionis, E. Terzi, P. Tsaparas. *Opinion Maximization in Social Networks*. SDM 2013
- R. Hegselmann, U. Krause. *Opinion Dynamics and Bounded Confidence. Models, Analysis, and Simulation*. Journal of Artificial Societies and Social Simulation (JASSS) vol.5, no. 3, 2002
- C. Castellano, S. Fortunato, V. Loreto. *Statistical Physics of Social Dynamics*, Reviews of Modern Physics 81, 591-646 (2009)
- A. Das, S. Gollapudi, K. Munagala, *Modeling opinion dynamics in social networks*. WSDM 2014