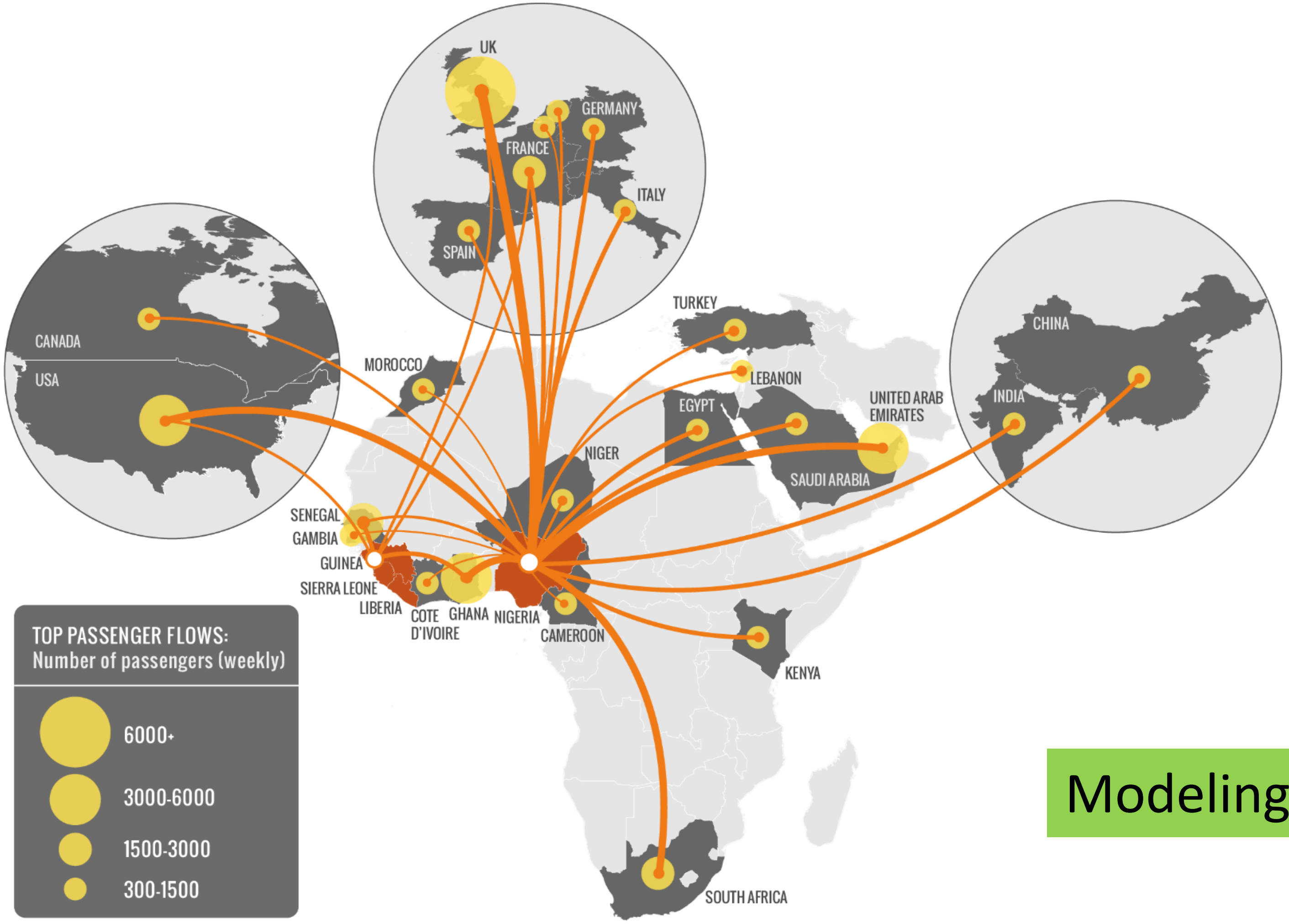
Diffusion in Networks

Luchon Summer School, 2015

Panayiotis Tsaparas
University of Ioannina, Greece

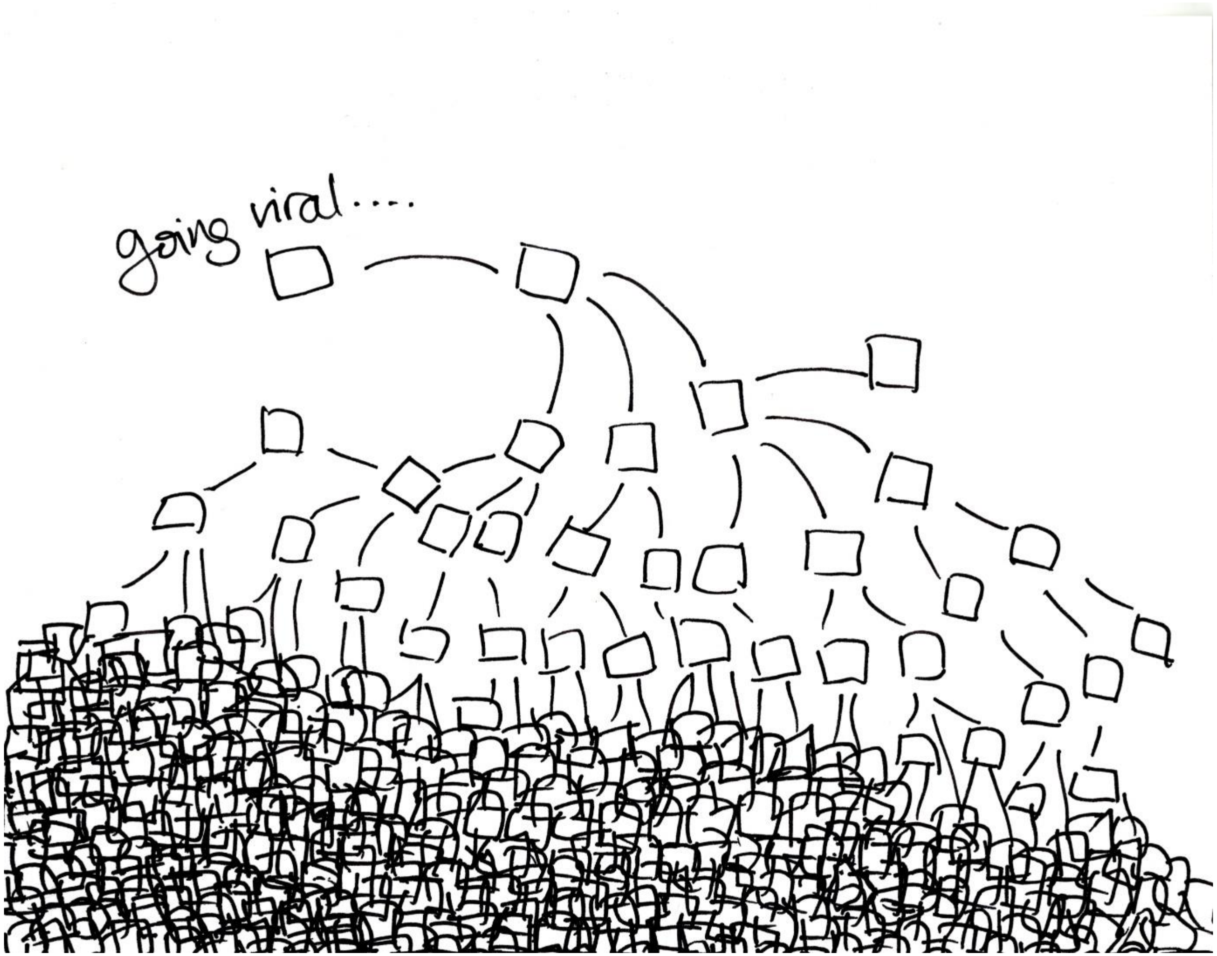
Diffusion: the process by which a piece of information spreads and reaches individuals through interactions in a netowork.

Why do we care?



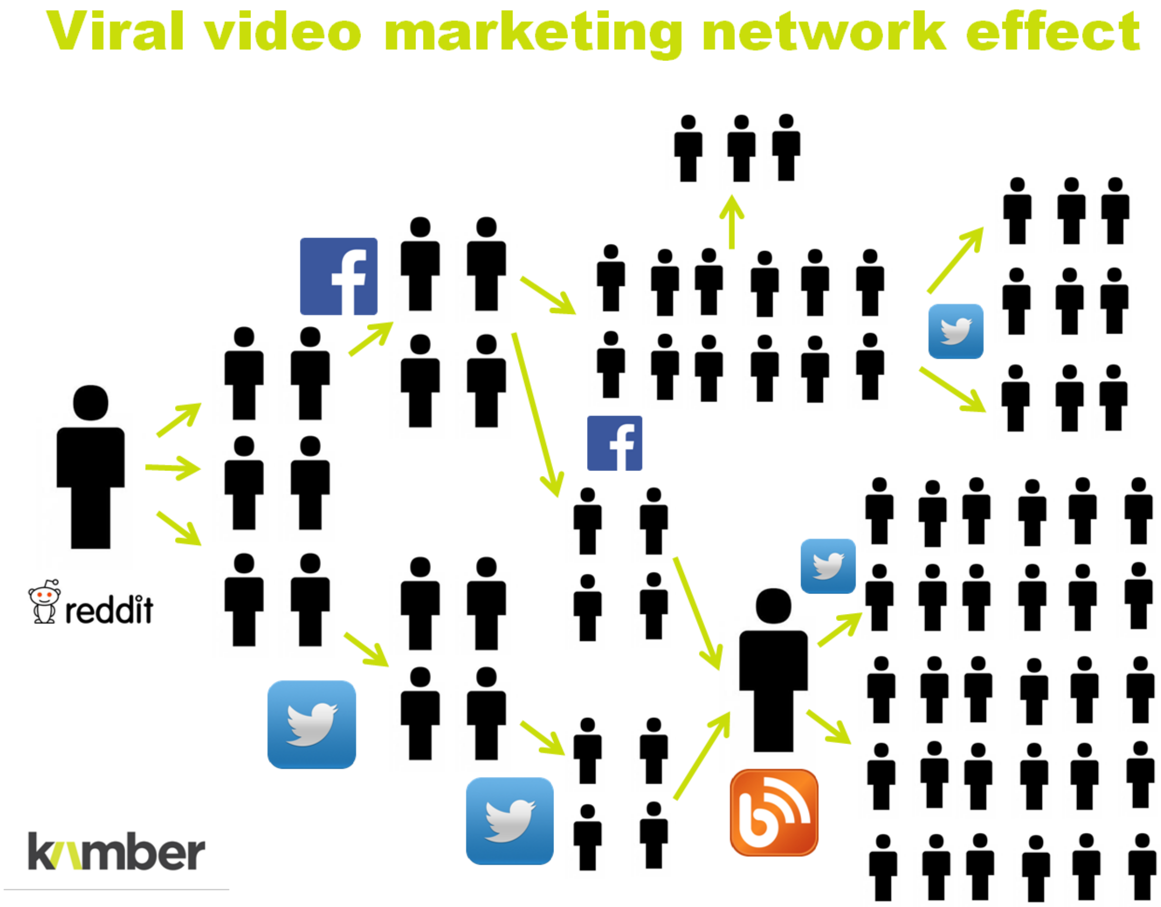
Modeling epidemics

Why do we care?



Viral marketing

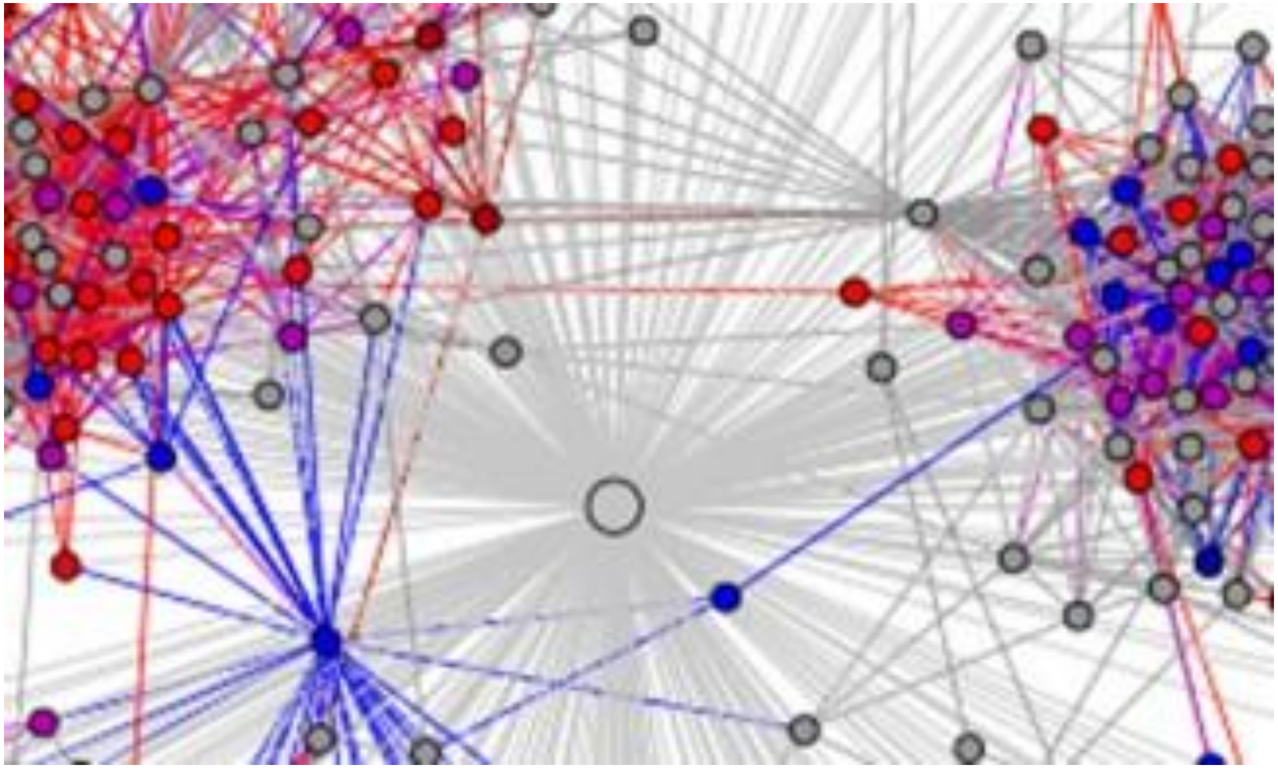
Why do we care?



Viral marketing

Why do we care?

Opinion Formation



Outline

- Epidemic models
- Influence maximization
- Opinion formation models

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

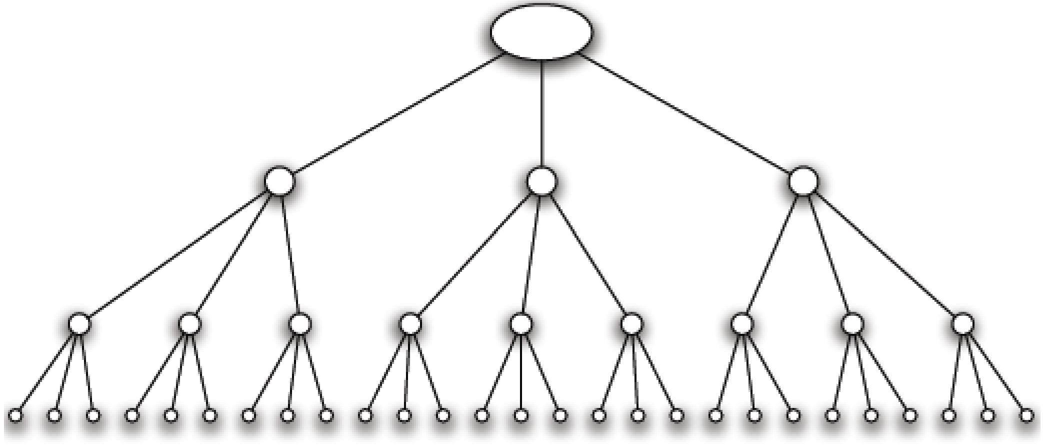
Epidemics

- 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

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.

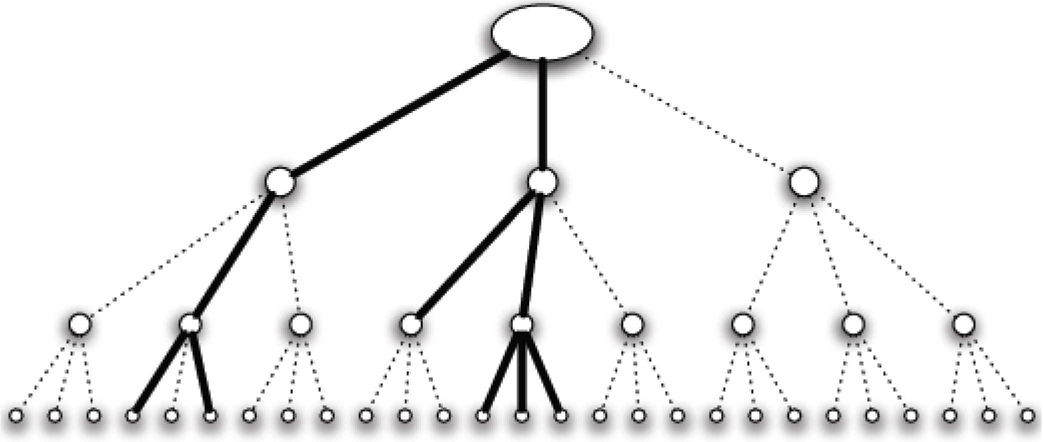


Contact network is a tree with branching factor k

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

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



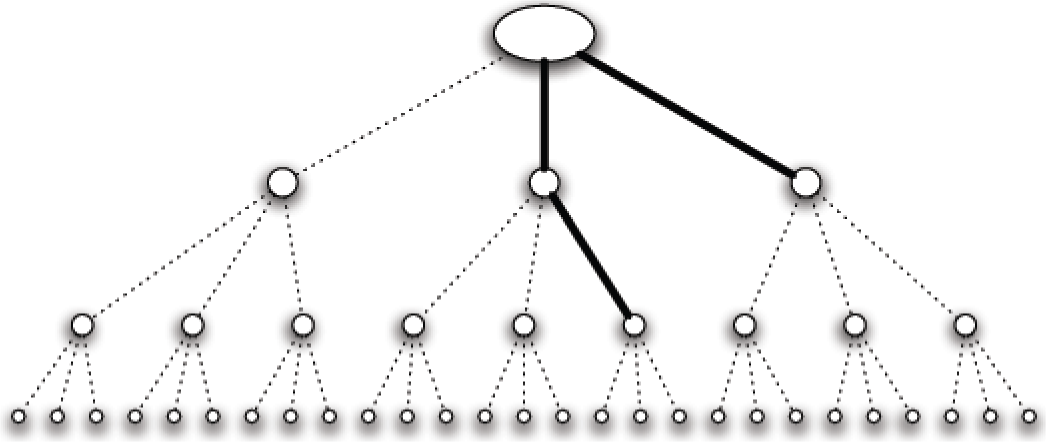
An aggressive epidemic with high infection probability

The epidemic survives after three steps

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

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 mild epidemic with low infection probability

The epidemic dies out after two steps

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Basic Reproductive Number

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

$$R_0 = kp$$

- Claim: (a) If $R_0 < 1$, then with probability 1, the disease dies out after a finite number of waves. (b) If $R_0 > 1$, then with probability greater than 0 the disease persists by infecting at least one person in each wave.

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

Proof

- X_n : number of infected nodes after n steps
- $q_n = \Pr[X_n \geq 1]$: probability that there exists at least 1 infected node after n steps
- $q^* = \lim q_n$: the probability of having infected nodes as $n \to \infty$
- We want to show that if $R_0 < 1, q^* = 0$ while if $R_0 > 1, q^* > 0$.

Proof

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

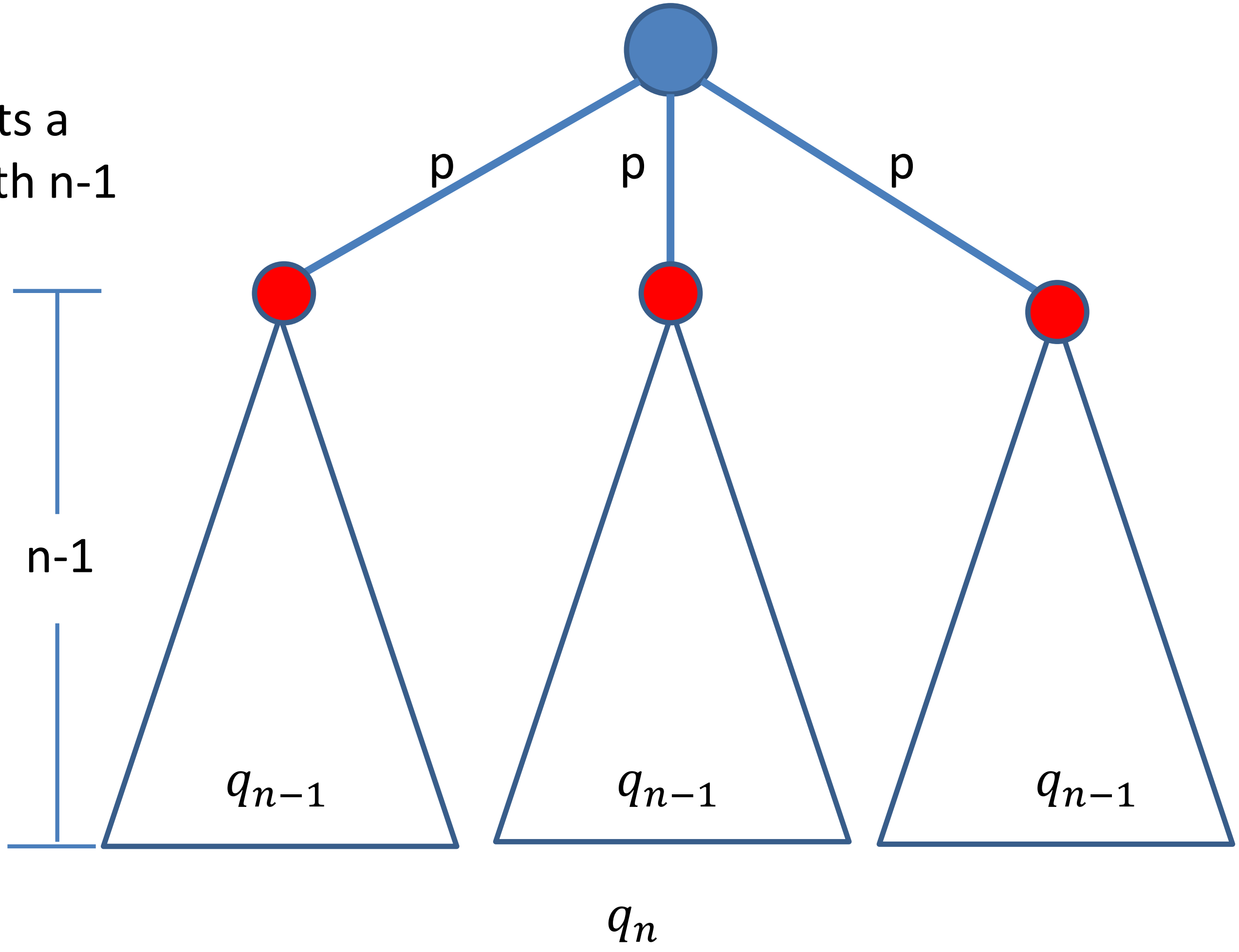
$$q_n = 1 - (1 - pq_{n-1})^k$$

if

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

then

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



We also have: $q_0 = 1$.

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

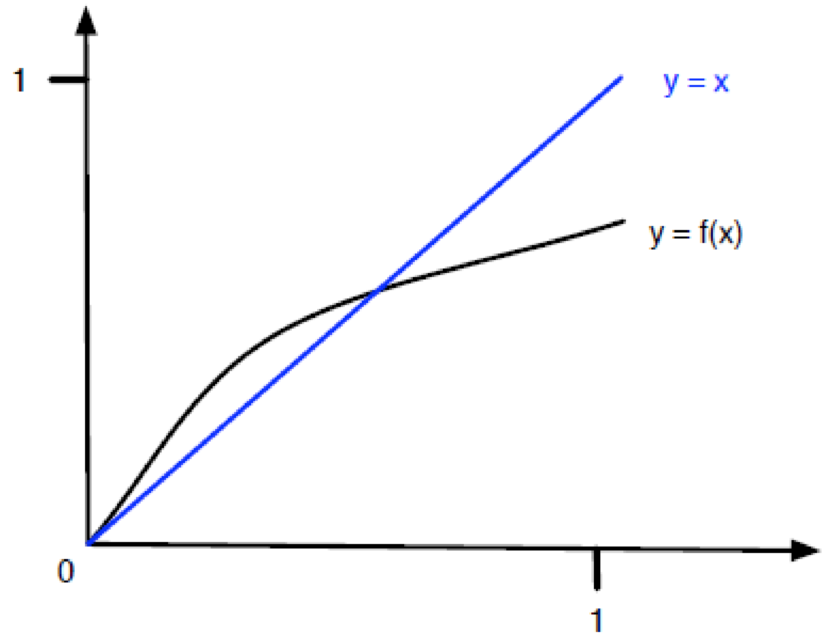
We want to find where this series converges

Proof

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

Proof

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

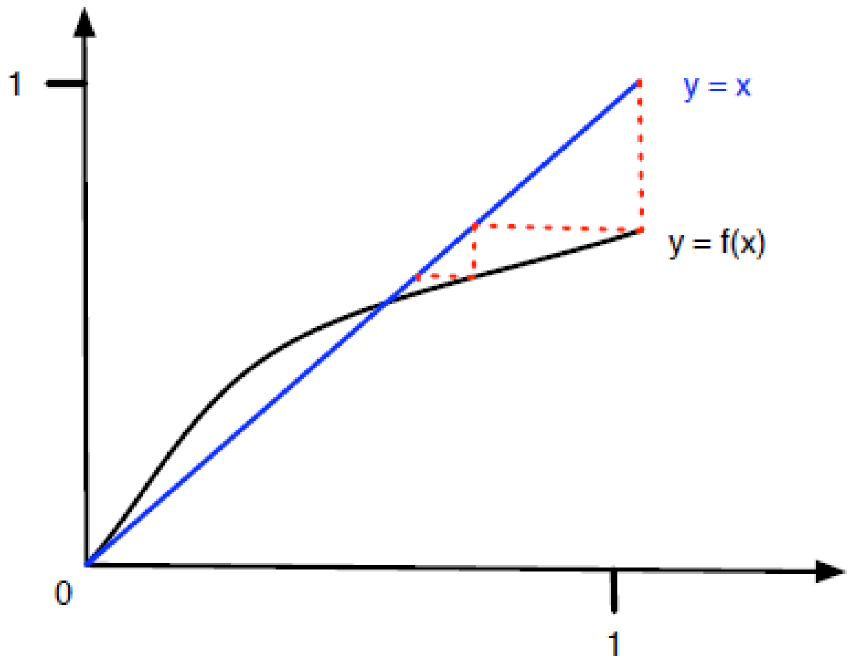


$f(x)$ crosses the line $y = x$ at some point

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Proof

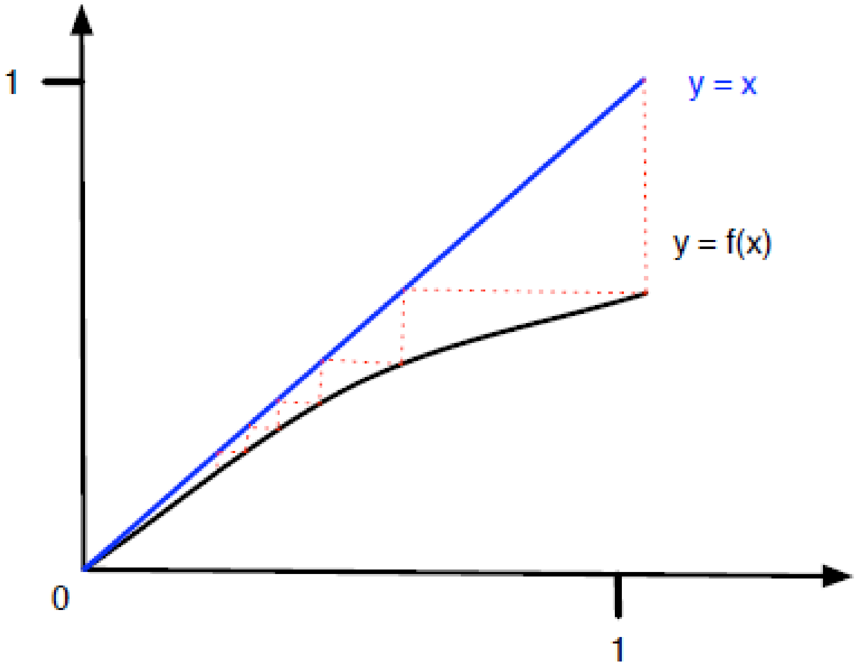
- Starting from the value 1, repeated applications of the function $f(x)$ will converge to the value $q^* = q_n = f(q_n)$



D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Proof

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



D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Branching process

- Assumes no network structure, no triangles or shared neihgbors

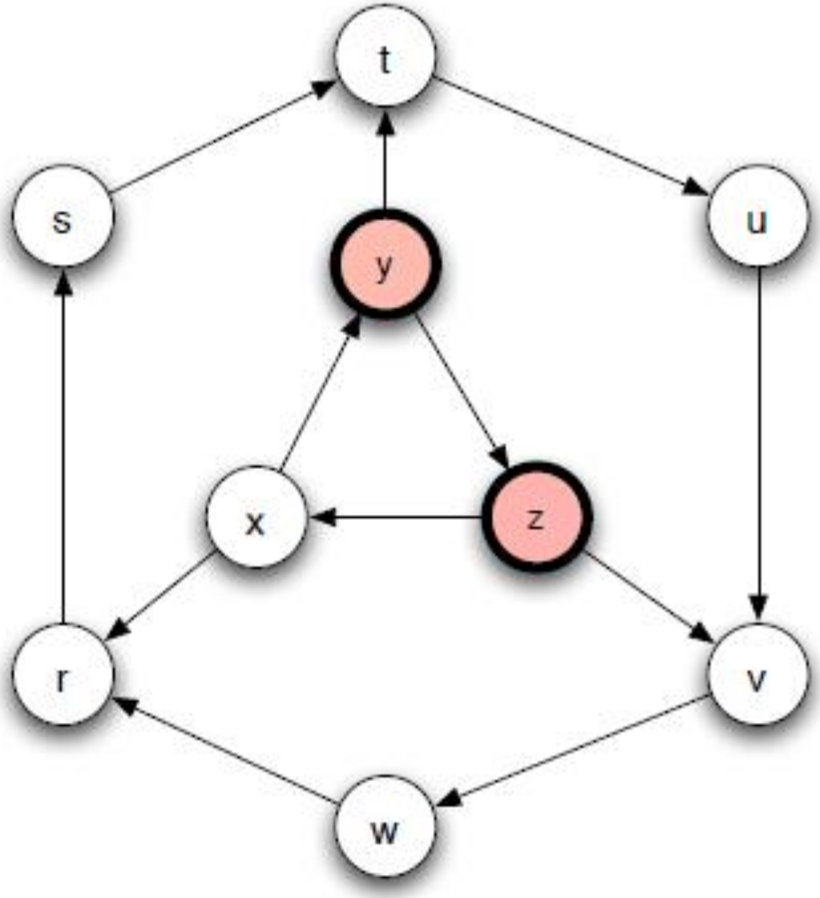
The SIR model

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

The SIR process

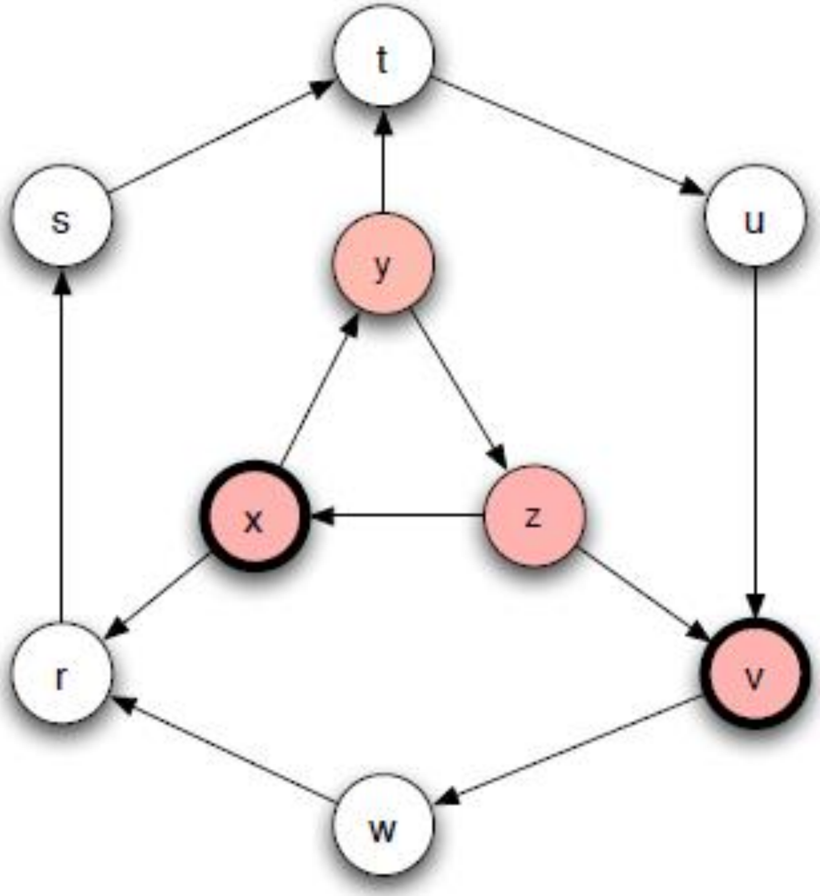
- Initially all nodes are in state S(usceptible), except for a few nodes in state I(nfected).
- An infected node stays infected for t_I steps.
  - Simplest case: $t_I = 1$
- At each of the t_I steps the infected node has probability p of infecting any of its susceptible neighbors
  - p: Infection probability
- After t_I steps the node is Removed

Example



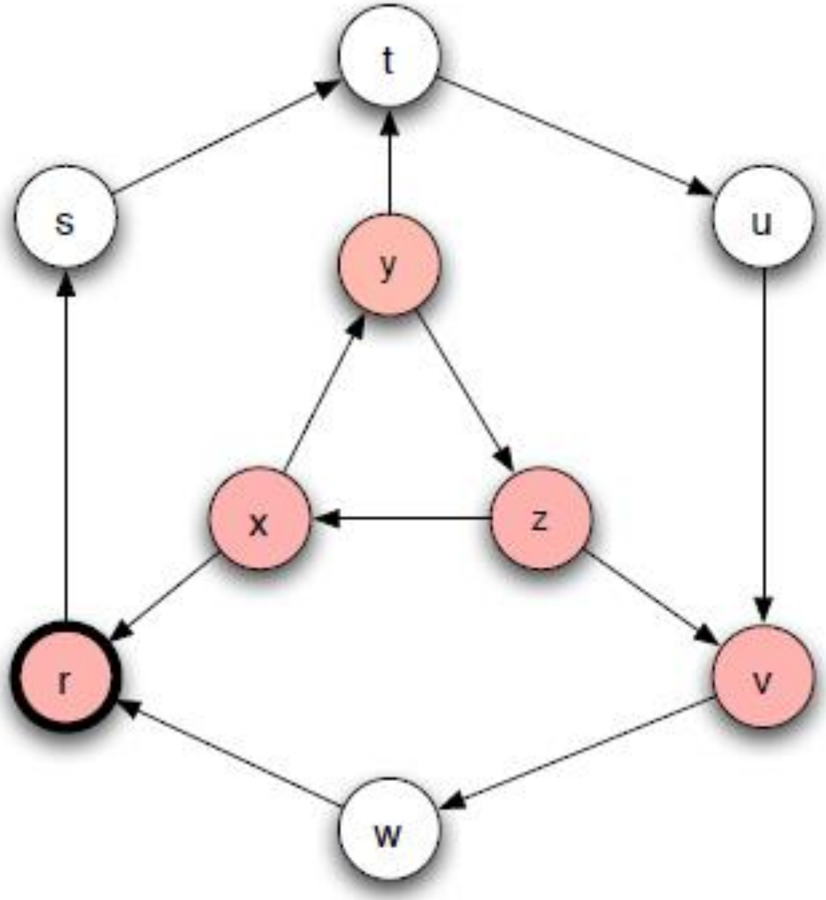
D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Example



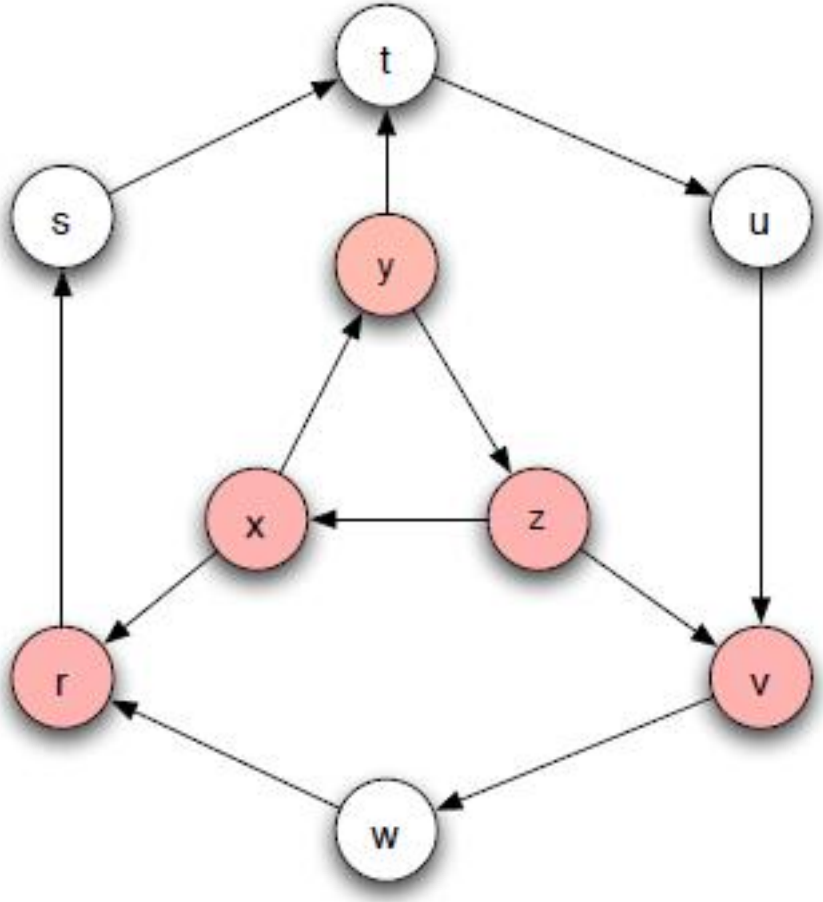
D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Example



D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Example



D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

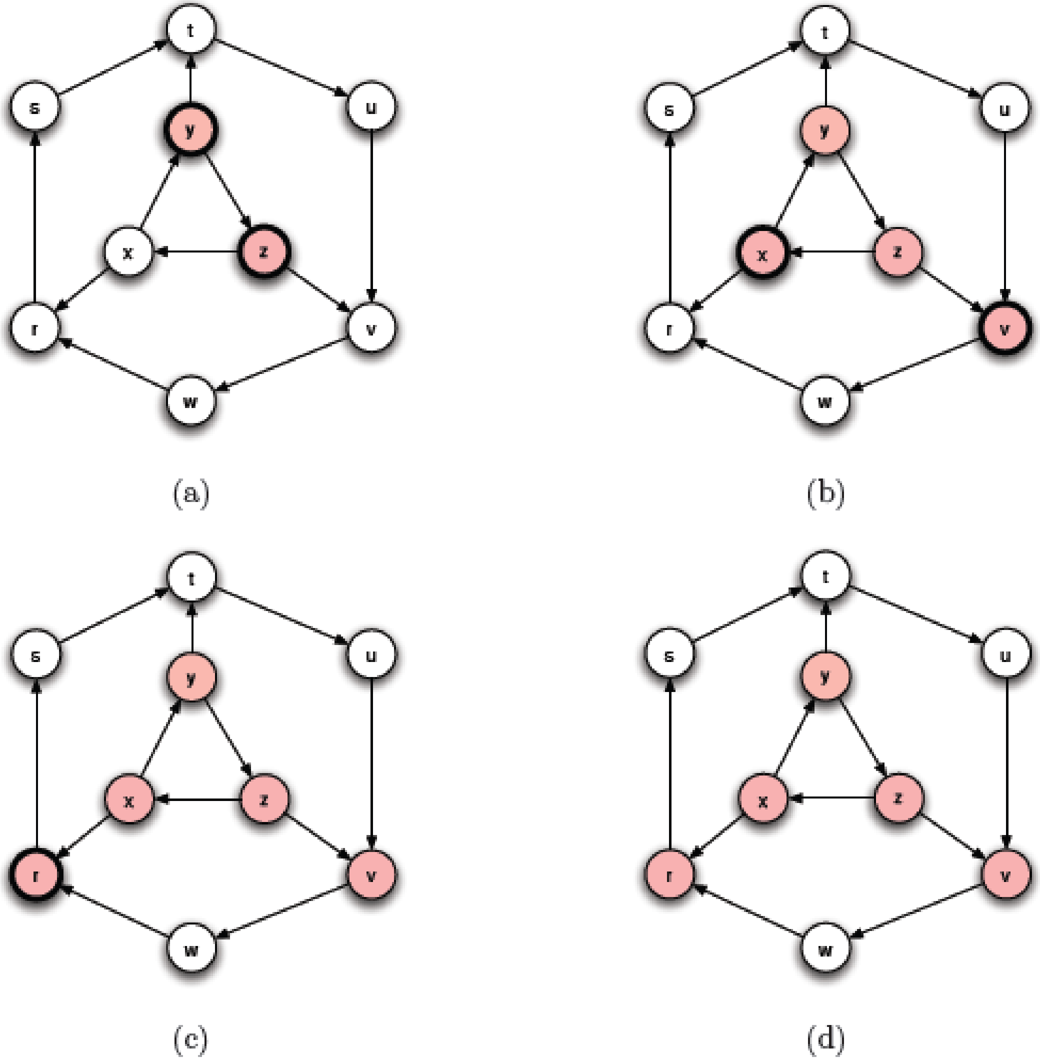


Figure 21.2: The course of an SIR epidemic in which each node remains infectious for a number of steps equal to $t_I = 1$. Starting with nodes y and z initially infected, the epidemic spreads to some but not all of the remaining nodes. In each step, shaded nodes with dark borders are in the Infectious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

SIR and the Branching process

- The branching process is a special case where the graph is a tree (and the infected node is the root)
  - The existence of triangles shared neighbors makes a big difference
- The basic reproductive number is not necessarily informative in the general case

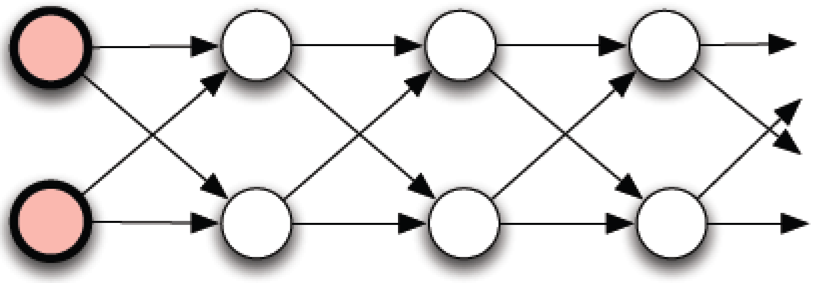


Figure 21.3: In this network, the epidemic is forced to pass through a narrow "channel" of nodes. In such a structure, even a highly contagious disease will tend to die out relatively quickly.

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

Percolation

- Percolation: we have a network of “pipes” which can curry liquids, and they can be either open, or closed
  - The pipes can be pathways within a material
- If liquid enters the network from some nodes, does it reach most of the network?
  - The network percolates

SIR and Percolation

- There is a connection between SIR model and percolation
- When a virus is transmitted from u to v, the edge (u,v) is activated with probability p
- We can assume that all edge activations have happened in advance, and the input graph has only the active edges.
- Which nodes will be infected?
  - The nodes reachable from the initial infected nodes
- In this way we transformed the dynamic SIR process into a static one.
  - This is essentially percolation in the graph.

Example

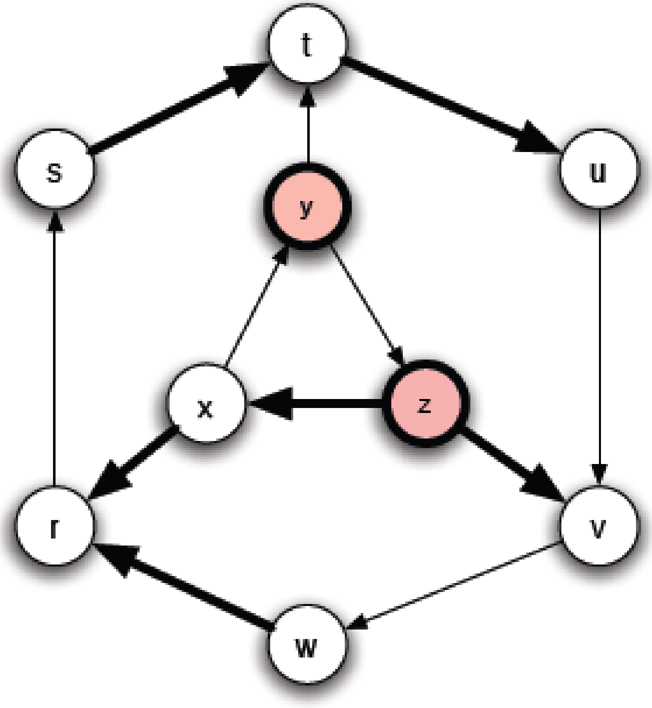


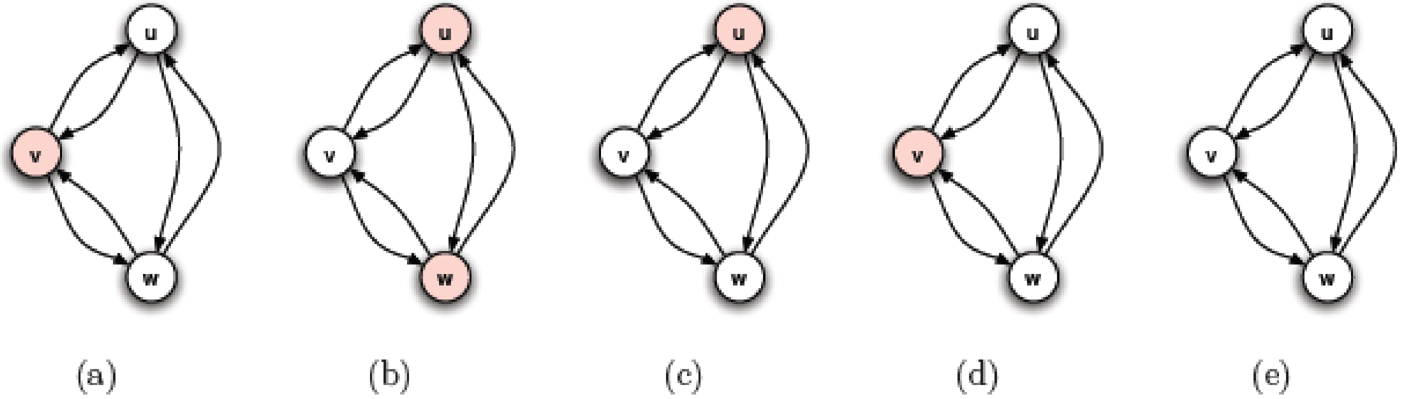
Figure 21.4: An equivalent way to view an SIR epidemic is in terms of *percolation*, where we decide in advance which edges will transmit infection (should the opportunity arise) and which will not.

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

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



(a) (b) (c) (d) (e)

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?

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

An eigenvalue point of view

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

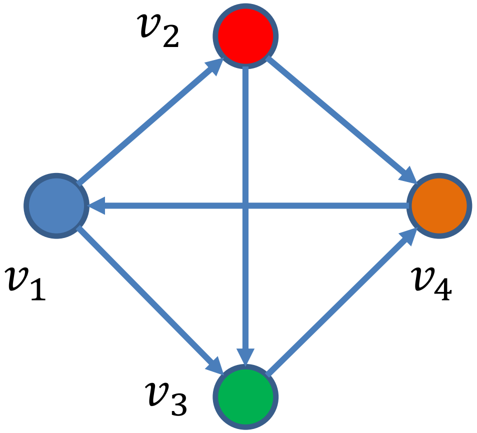
$$\lambda_1(A) \leq \frac{q}{p}$$

- 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

Reminder

- Adjacency matrix of a graph



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

- Eigenvalue of matrix A is a value λ such that $Ax = \lambda x$

Multiple copies model

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

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

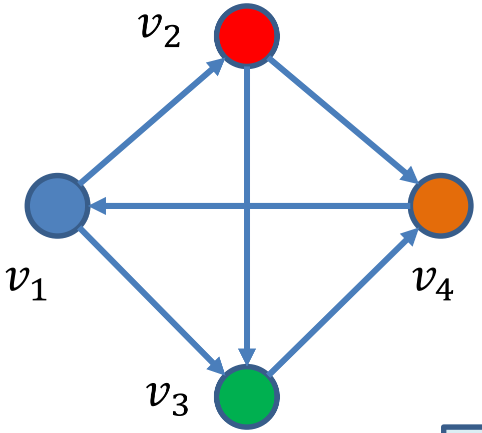
```
For each node i
    For each of the $v_i^t$ virus copies at node $i$
        the copy is copied to a neighbor $j$ with prob $p$
        the copy dies with probability $q$
```

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

Analysis

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

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



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

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

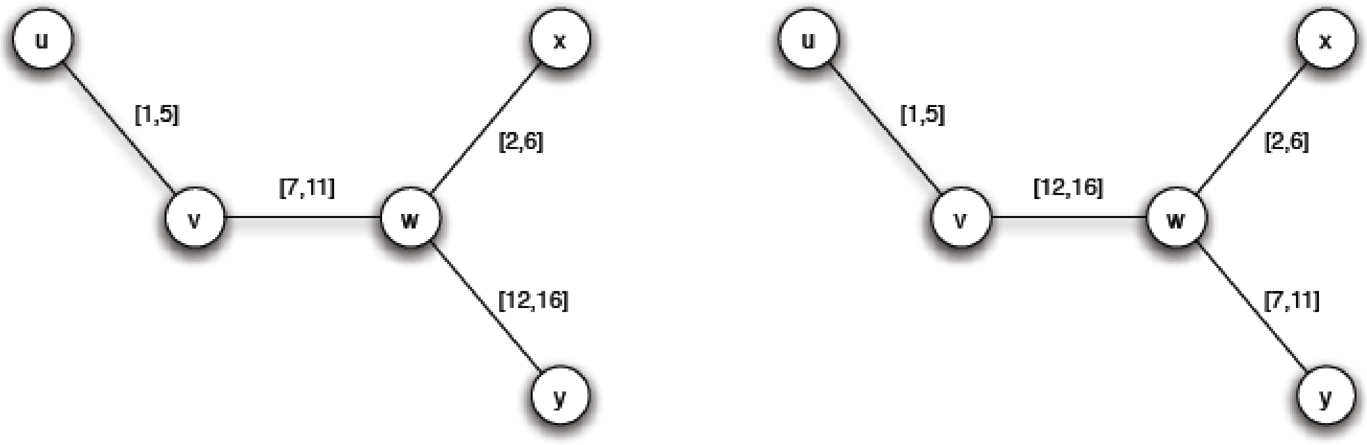
Probability that the copy from node v_4 survives at v_4

Analysis

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

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.*

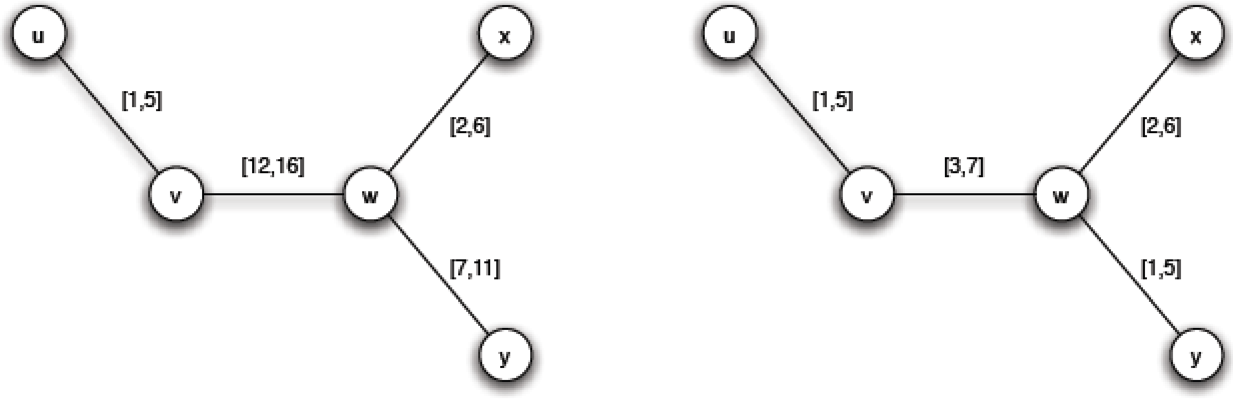
(b) *The same network as in (a), except that the timing of the w-v and w-y partnerships have been reversed.*

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.

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

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.

D. Easley, J. Kleinberg. Networks, Crowds and Markets: Reasoning about a highly connected world.

References

- D. Easley, J. Kleinberg. *Networks, Crowds and Markets: Reasoning about a highly connected world*. Cambridge University Press, 2010 – Chapter 21
- Y. Wang, D. Chakrabarti, C. Wang, C. Faloutsos. *Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint*. SRDS 2003
- G. Giakkoupis, A. Gionis, E. Terzi, P. Tsaparas. *Models and algorithms for network immunization*. Technical Report C-2005-75, Department of Computer Science, University of Helsinki, 2005.

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?

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.

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

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: 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

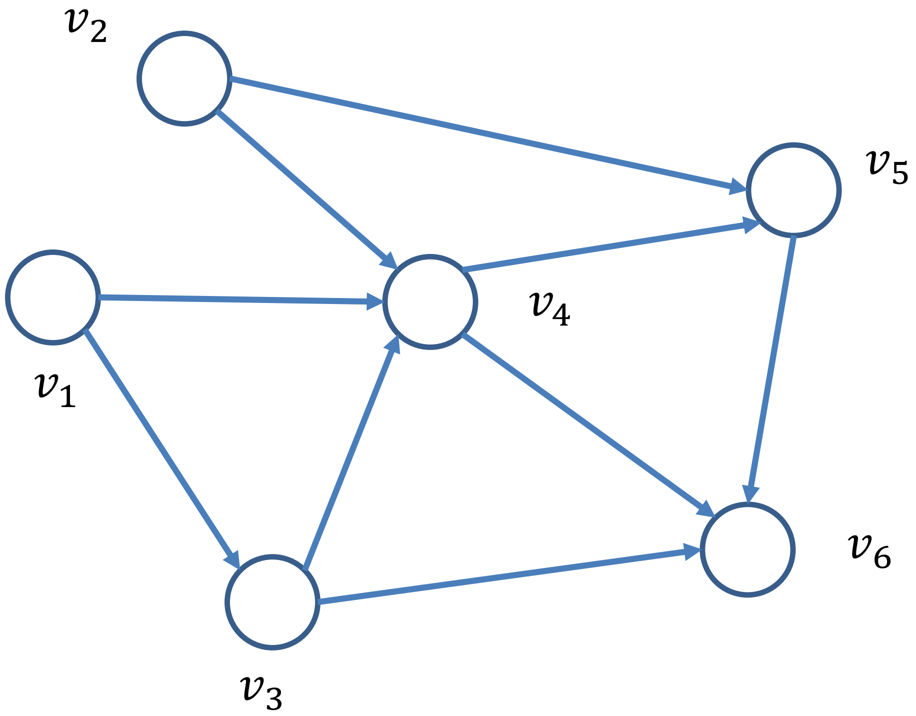
Proof idea

- 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

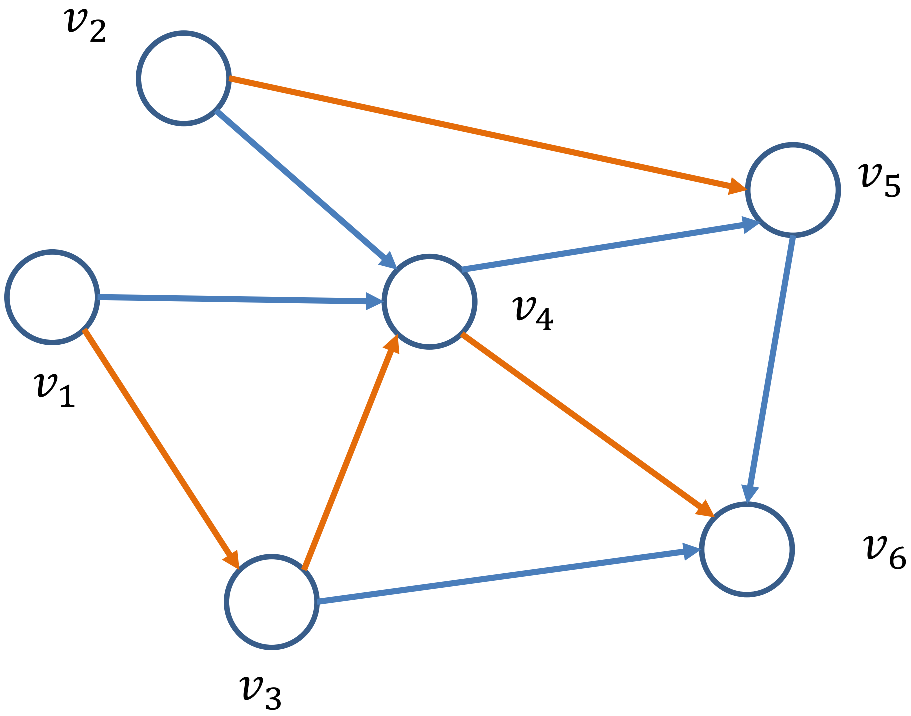
- Consider the special case of a DAG (Directed Acyclic Graph)
  - There is a topological ordering of the nodes $v_0, v_1, \dots, v_n$ such that edges go from left to right
- Consider node v_i in this ordering and assume that S_i is the set of neighbors of v_i that are active.
- What is the probability 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 graph.

Example



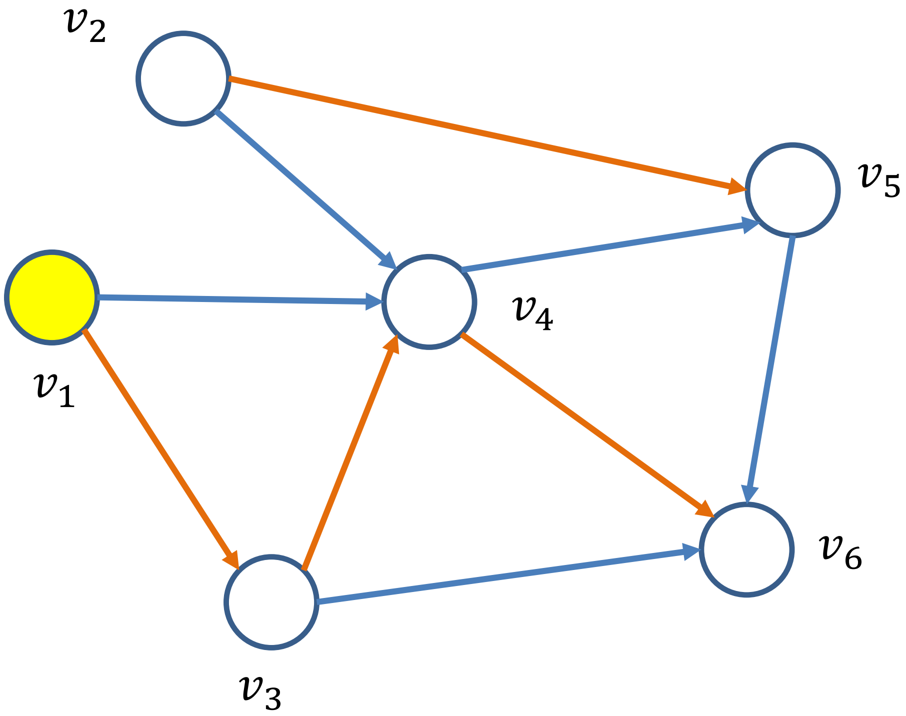
Assume that all edge weights incoming to any node sum to 1

Example



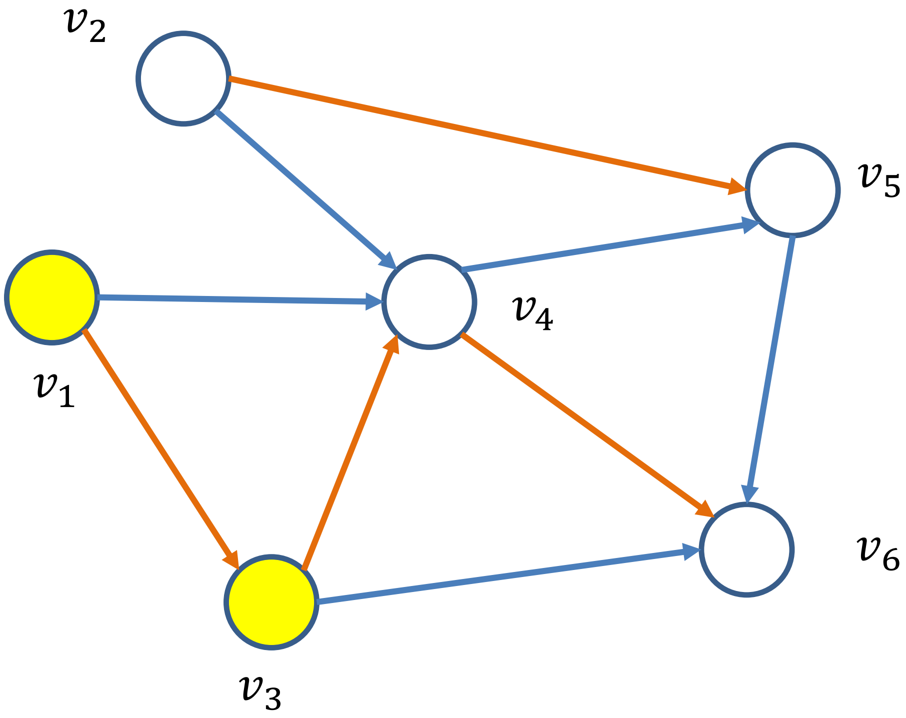
The nodes select a single incoming edge with probability equal to the weight (uniformly at random in this case

Example



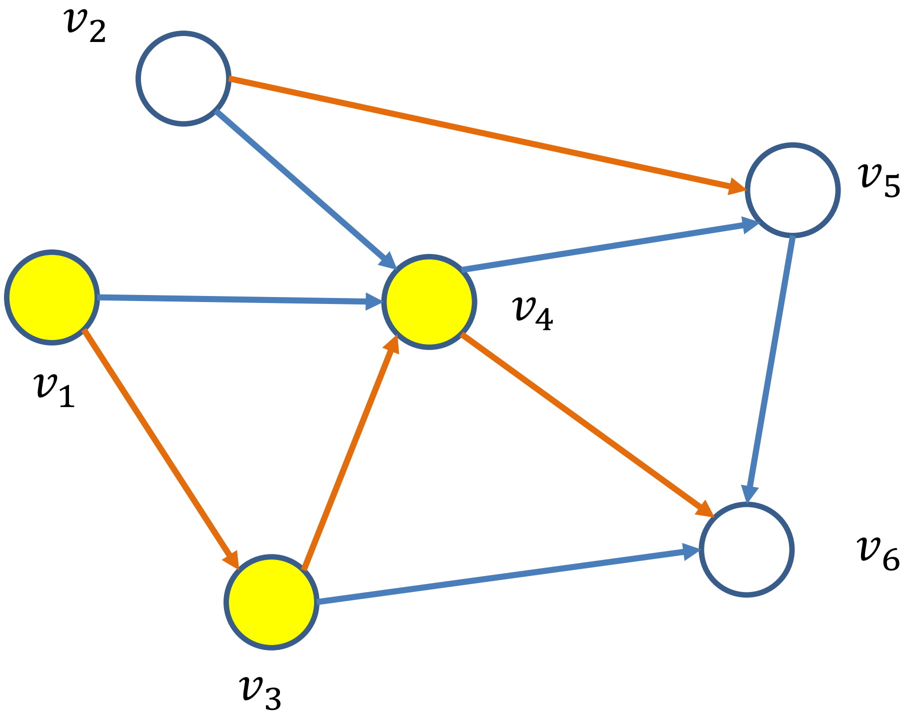
Node v_1 is the seed

Example



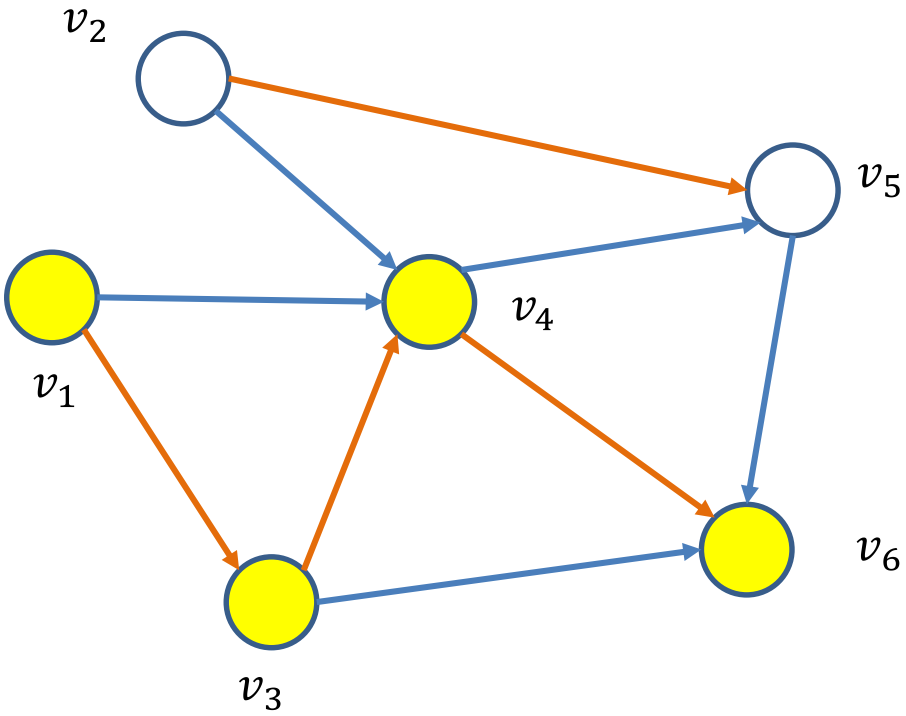
Node v_3 has a single incoming neighbor, therefore for any threshold it will be activated

Example



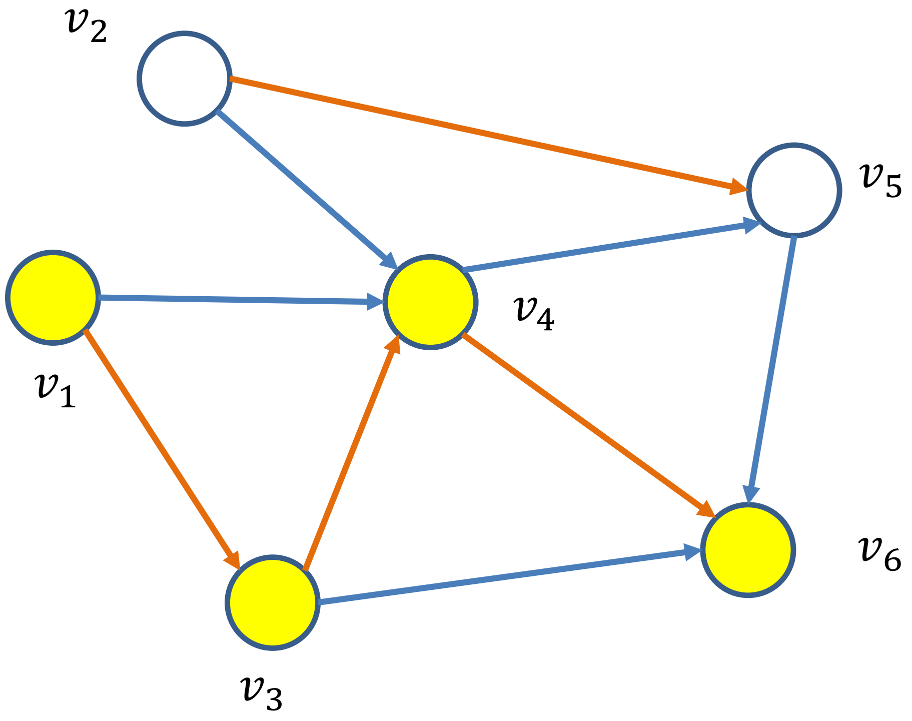
The probability that node v_4 gets activated is 2/3 since it has incoming edges from two active nodes.
The probability that node v_4 picks one of the two edges to these nodes is also 2/3

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).

Experiments

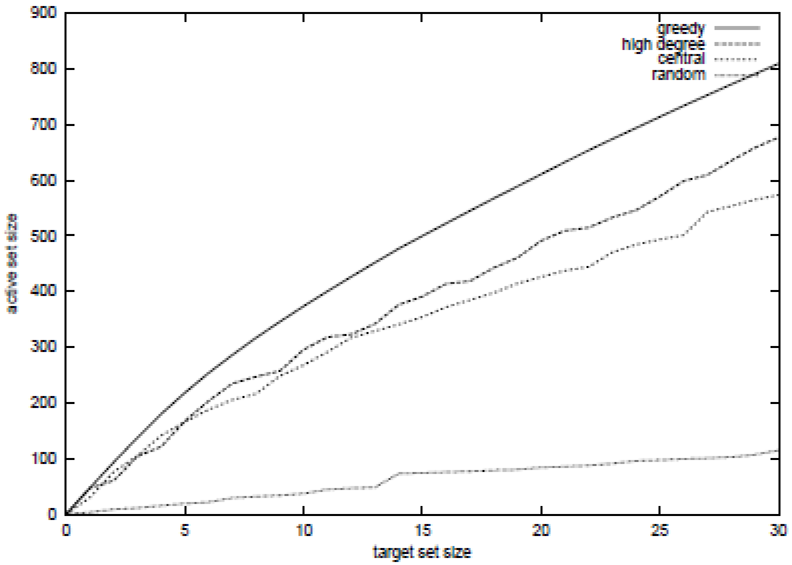


Figure 2: Results for the weighted cascade model

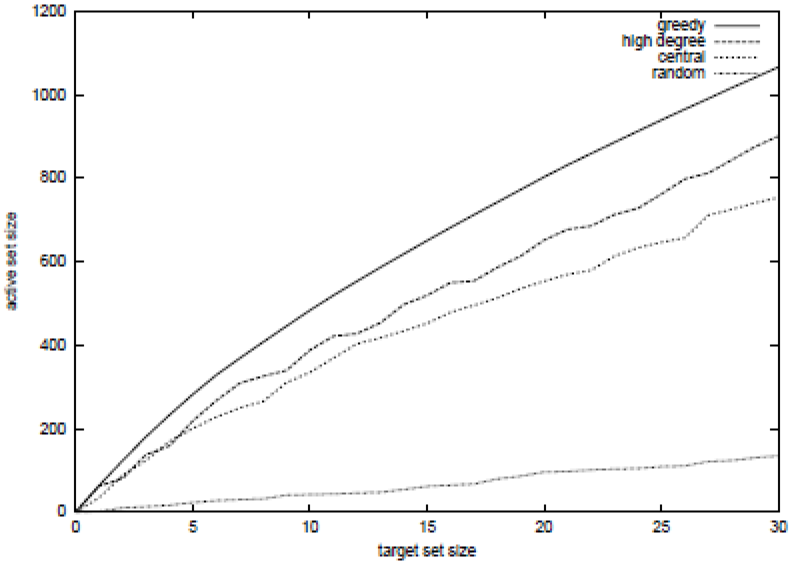
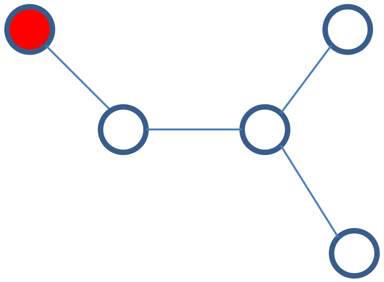
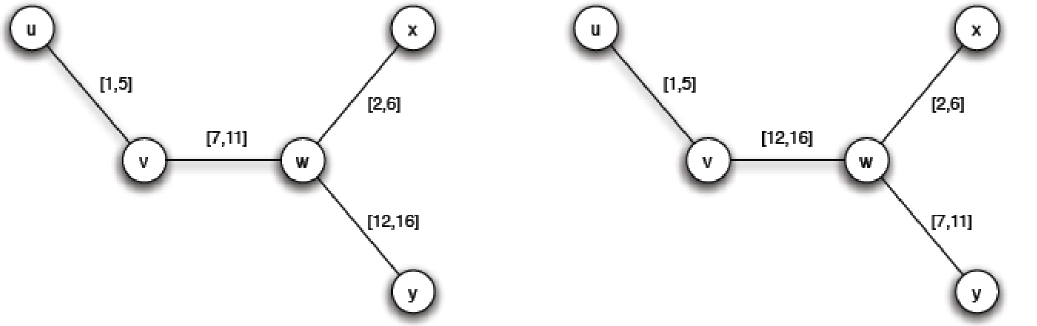


Figure 1: Results for the linear threshold model

Another example

- What is the spread from the red node?





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

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

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

Evolving network

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

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

Time

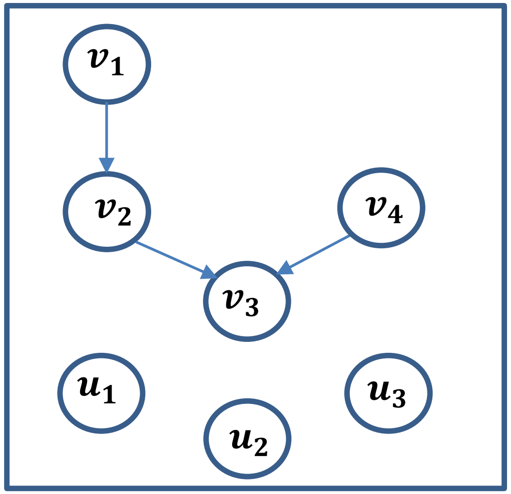
- How does the evolution of the network relates to the evolution of the diffusion?
  - How much physical time does a diffusion step last?
- Assumption: The two processes are in sync. One diffusion step happens in on one graph snapshot
- Evolving IC model: at time-step t, the infectious nodes try to infect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

Submodularity

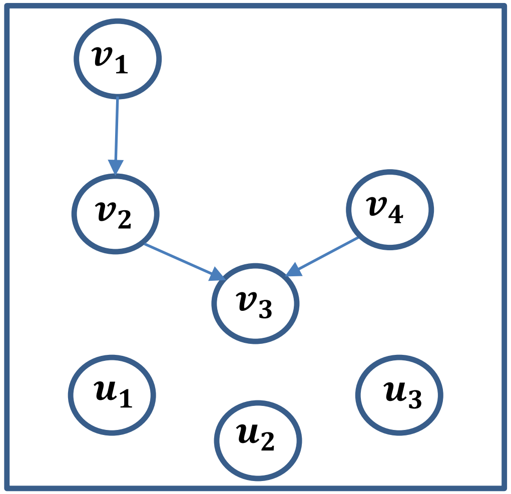
- Will the spread function remain monotone and submodular?

- No!

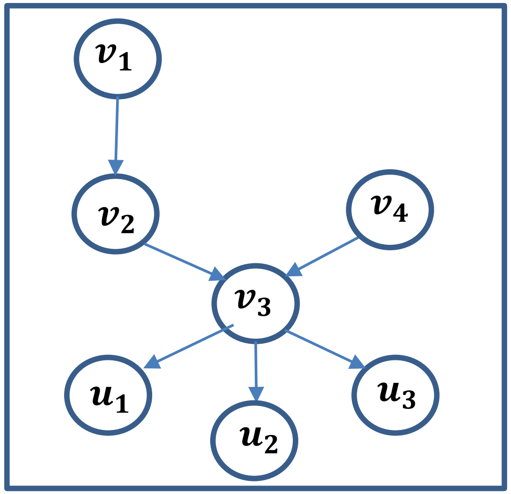
Monotonicity for the EIC model



G^1

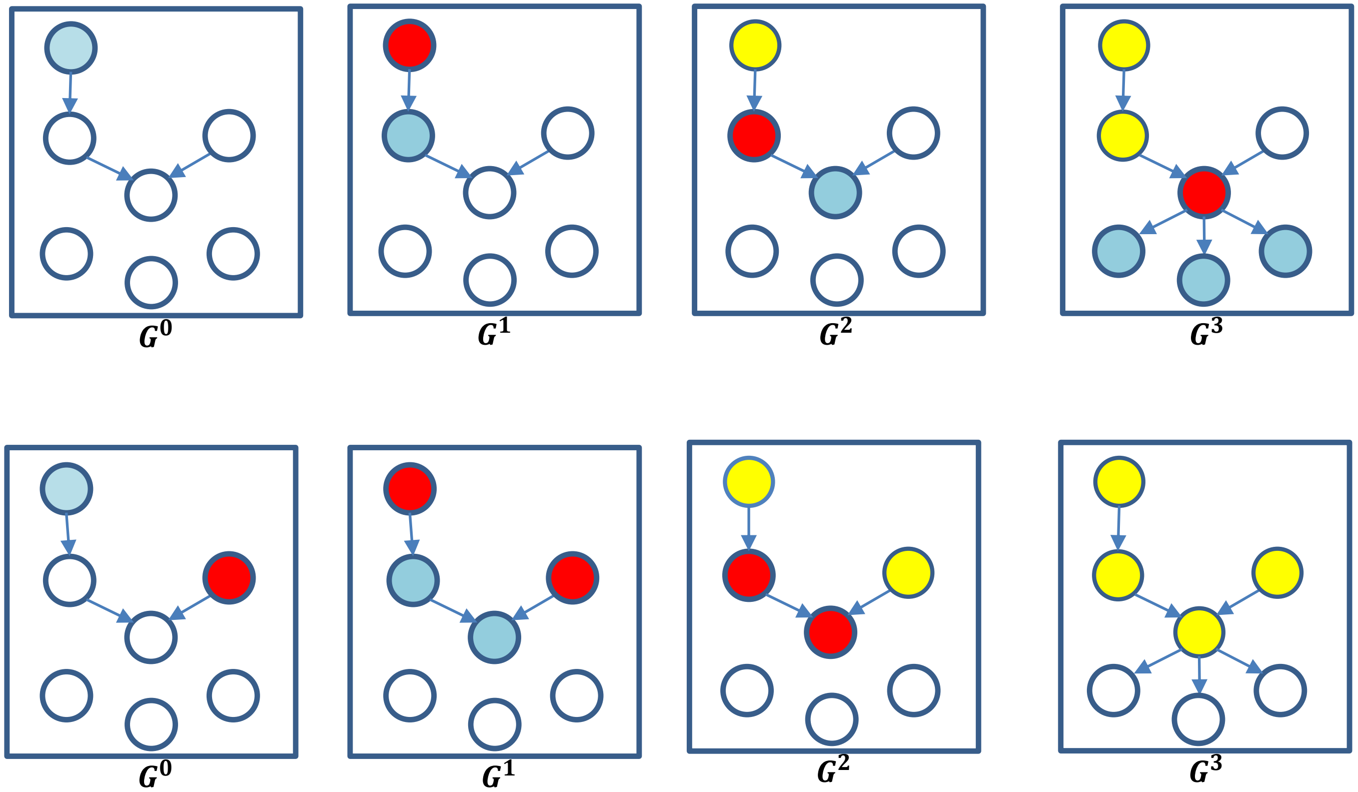


G^2



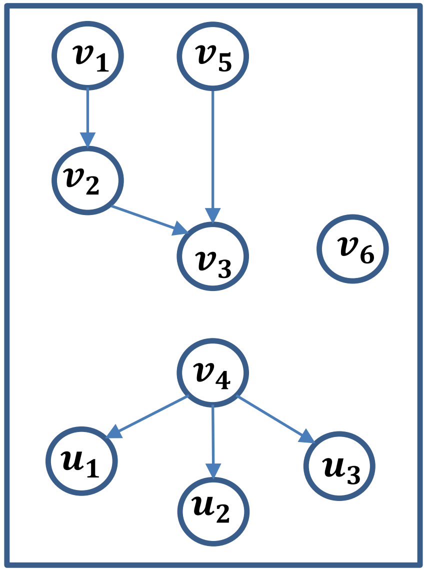
G^3

Monotonicity for the EIC model

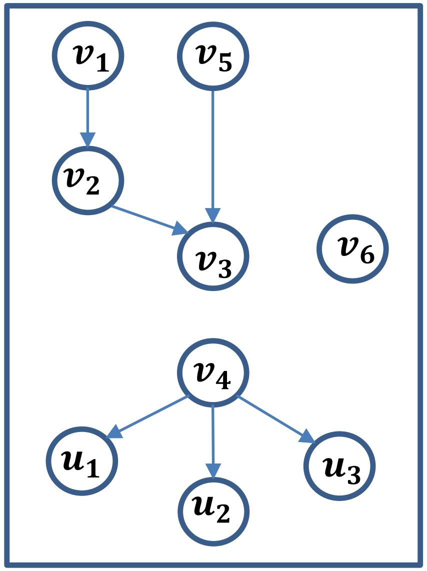


The spread is not monotone in the case of the Evolving IC model

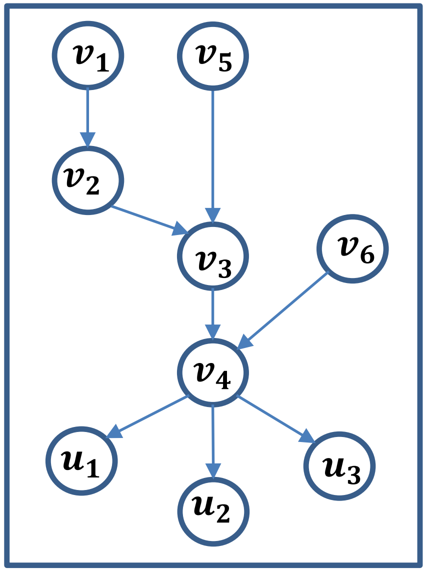
Submodularity for the EIC model



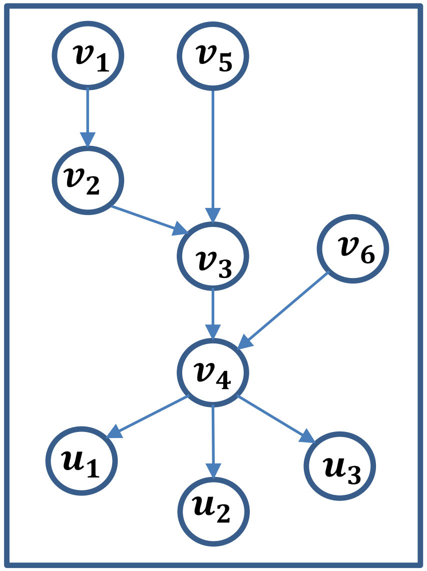
G^1



G^2

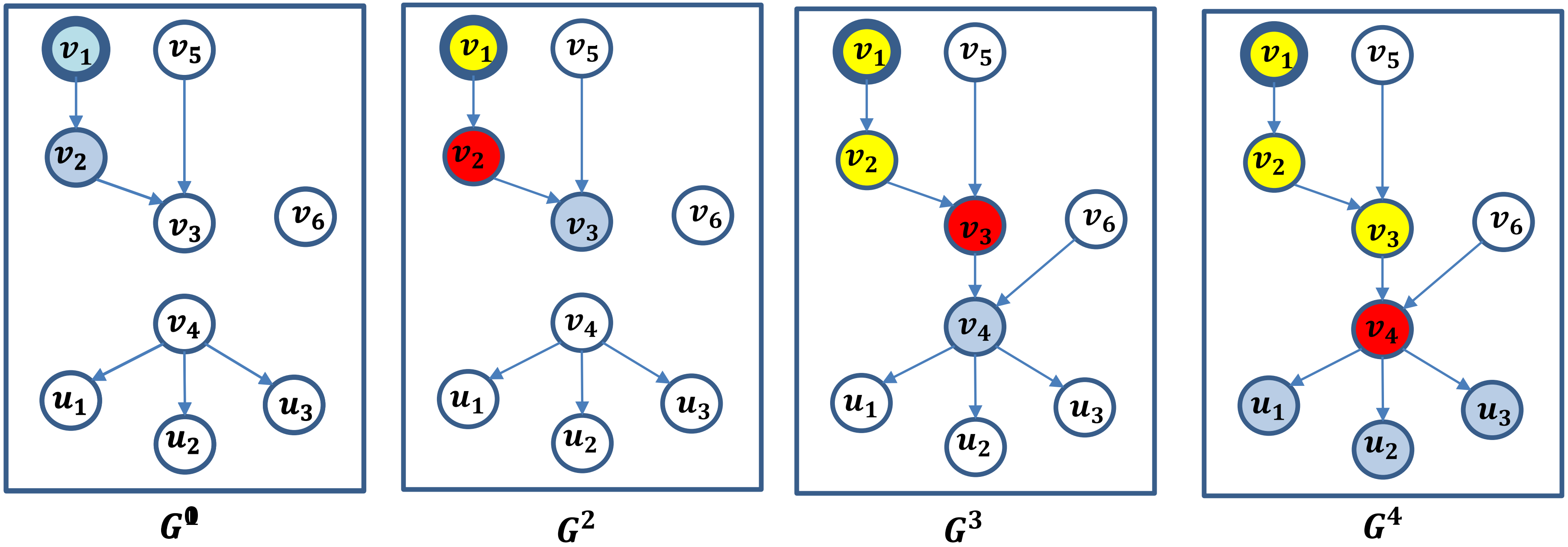


G^3



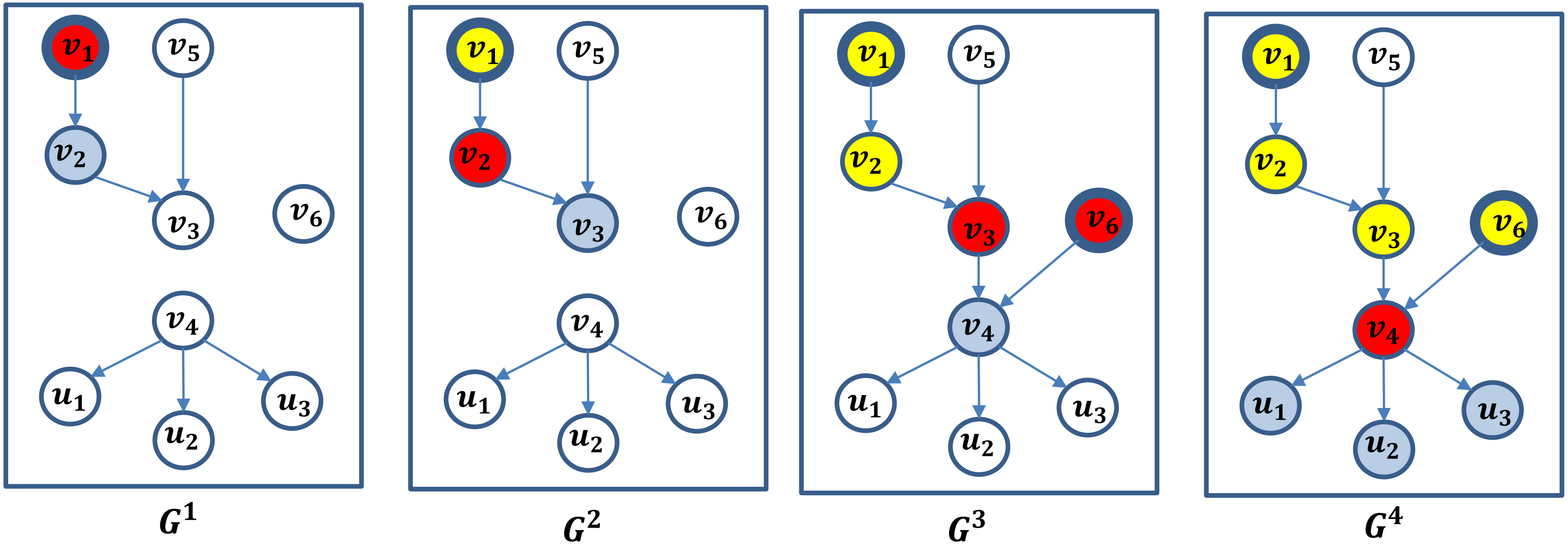
G^4

Submodularity for the EIC model



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

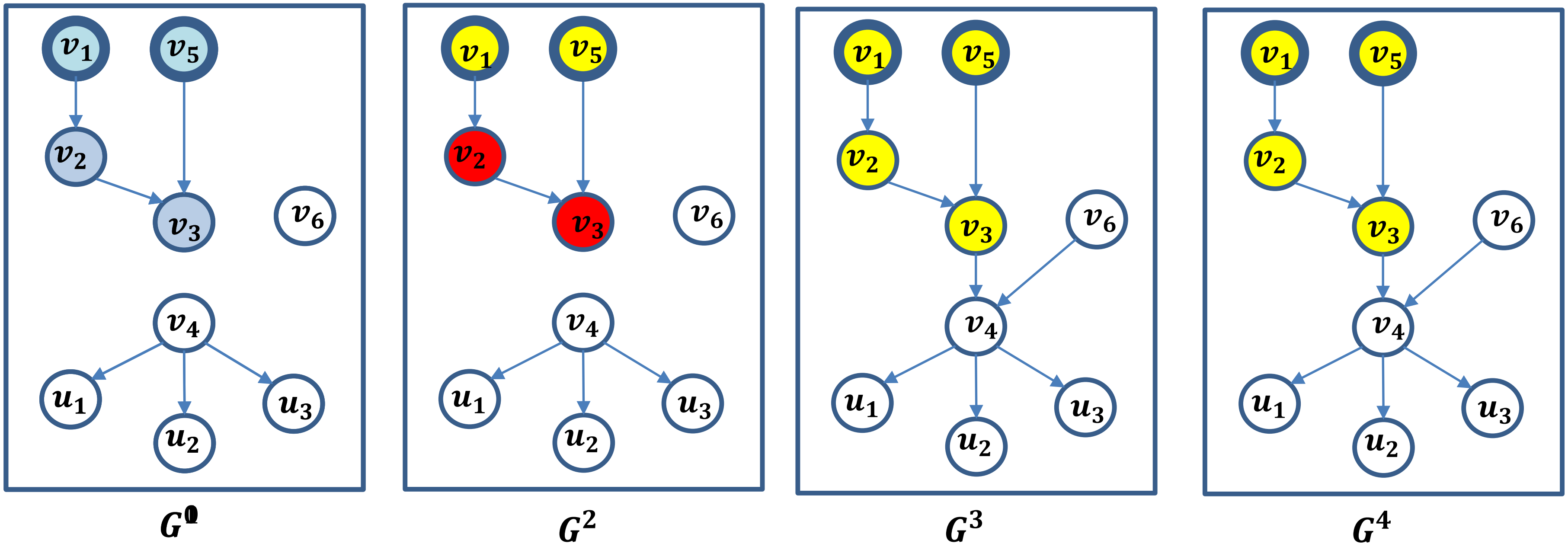
Submodularity for the EIC model



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

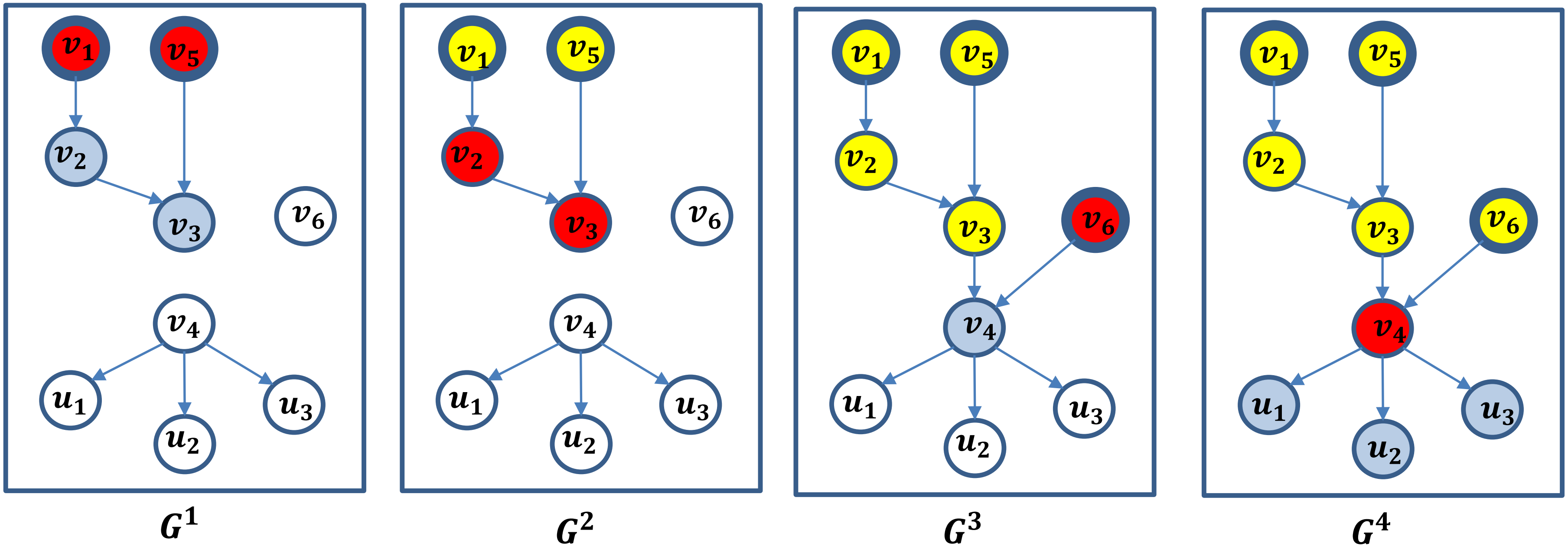
Adding node v_6 at time $t = 3$ does not increase the spread

Submodularity for the EIC model



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

Submodularity for the EIC model

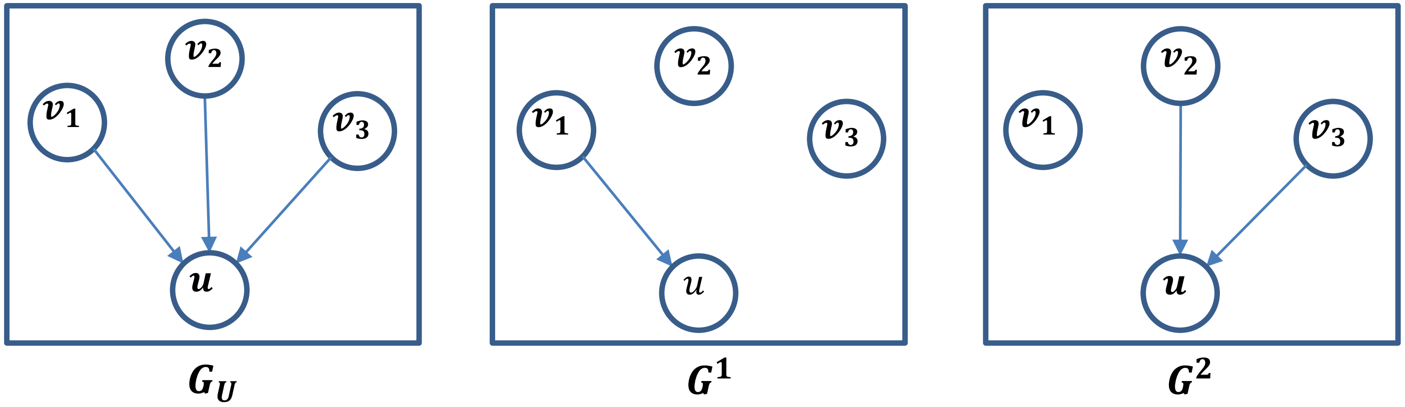


Activating nodes v_1 and v_5 at time $t = 0$ has spread 4

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

Evolving LT model

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



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

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

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

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

Improvements

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

      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

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})

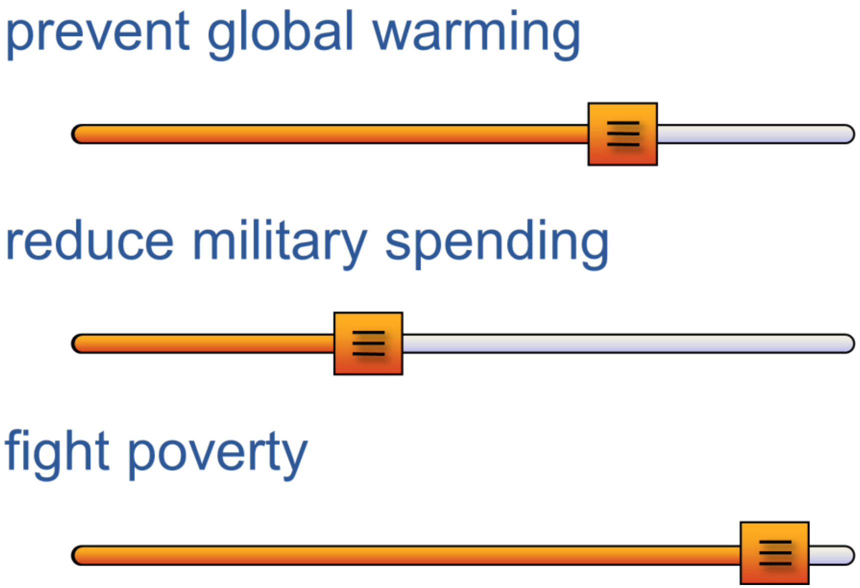
Diffusion of opinions

- The network can also diffuse opinions.
  - What people believe about an issue, a person, an item, is shaped by their social network
- Opinions 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.

What is an opinion?

- An opinion is a real value
  - In our models a value in the interval [0,1]

  (0: negative, 1: positive)



How are opinions formed?

- Opinions change over time

How are opinions formed?

- And they are influenced by our social network

An opinion formation model (De Groot)

- 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

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

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

- This iterative process converges to a consensus

What about personal biases?

- People tend to cling on to their personal opinions

Another opinion formation model (Friedkin and Johnsen)

- 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

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

Opinion formation as a game

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

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

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

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

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

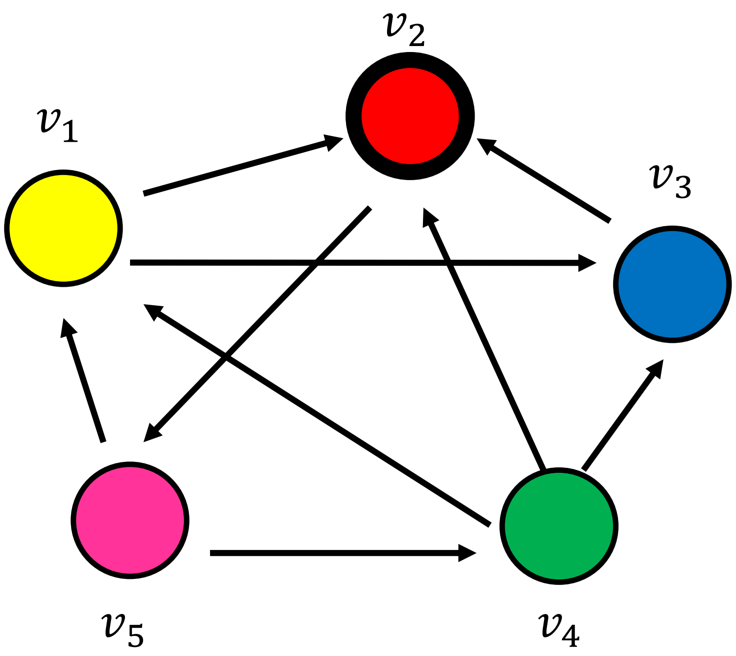
- A random walk is a stochastic process performed on a graph

- 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.

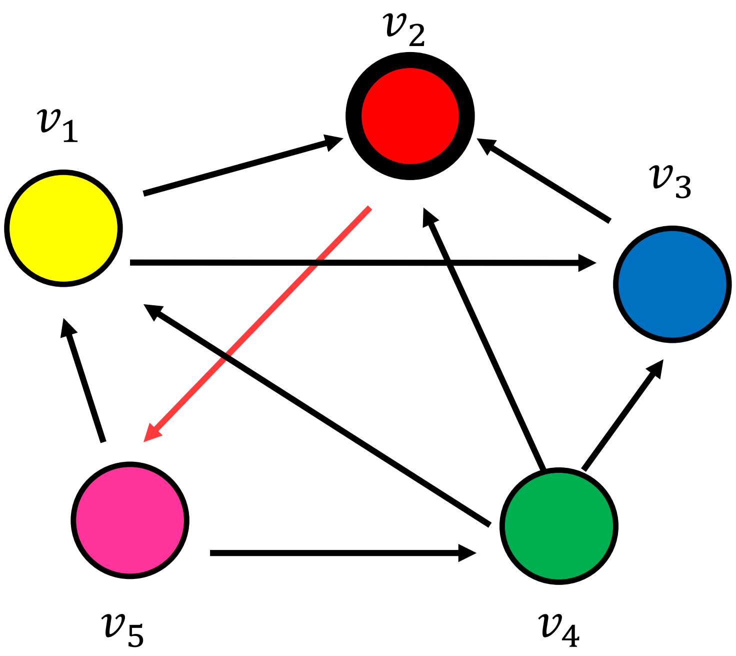
Example

- Step 0



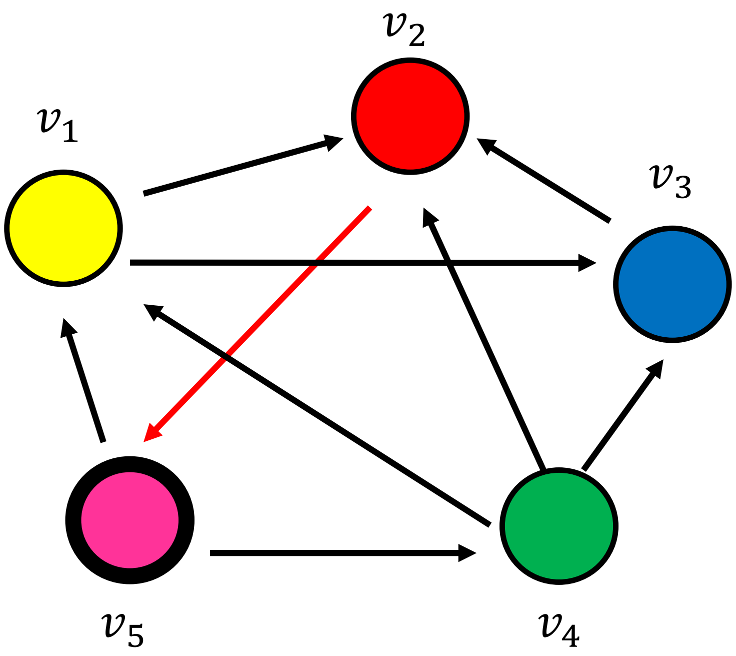
Example

- Step 0



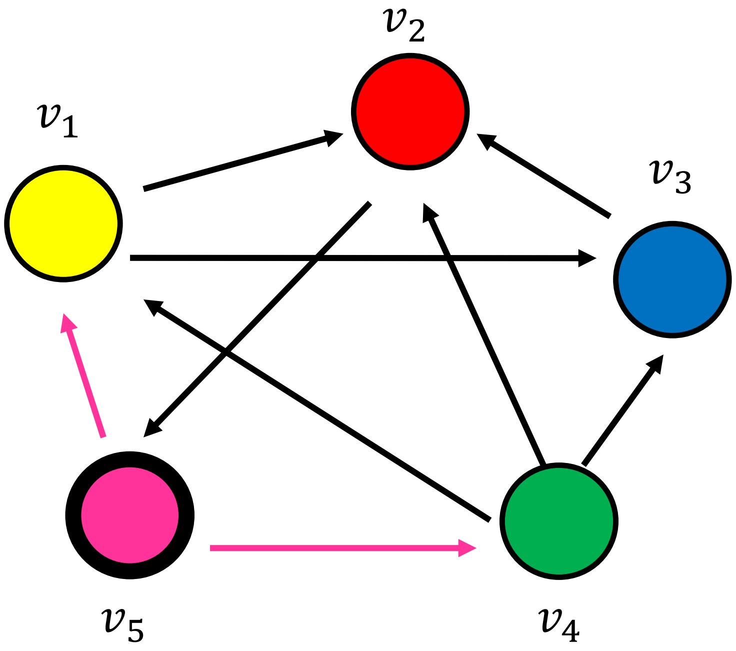
Example

- Step 1



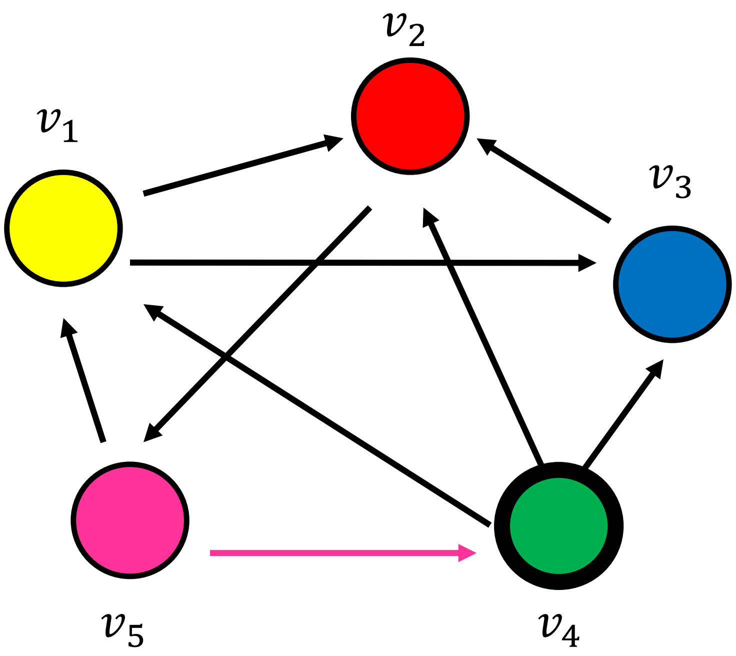
Example

- Step 1



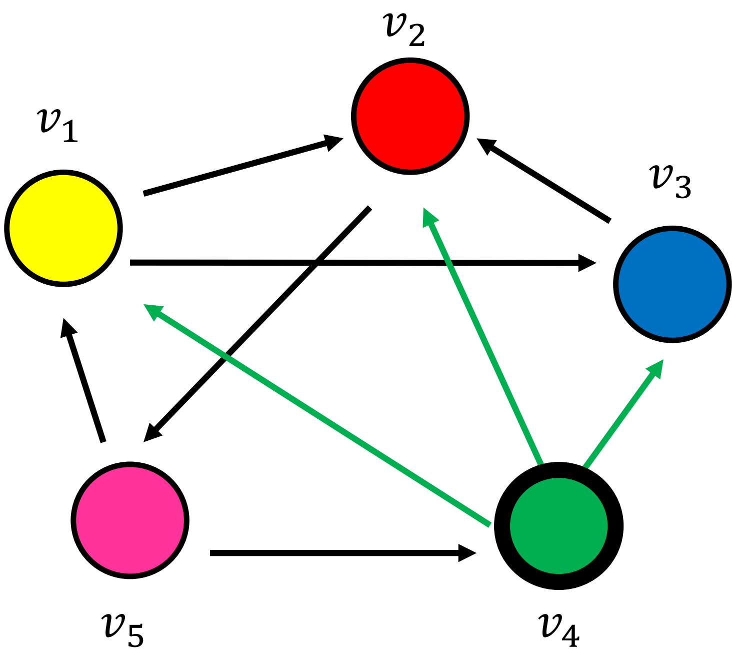
Example

- Step 2



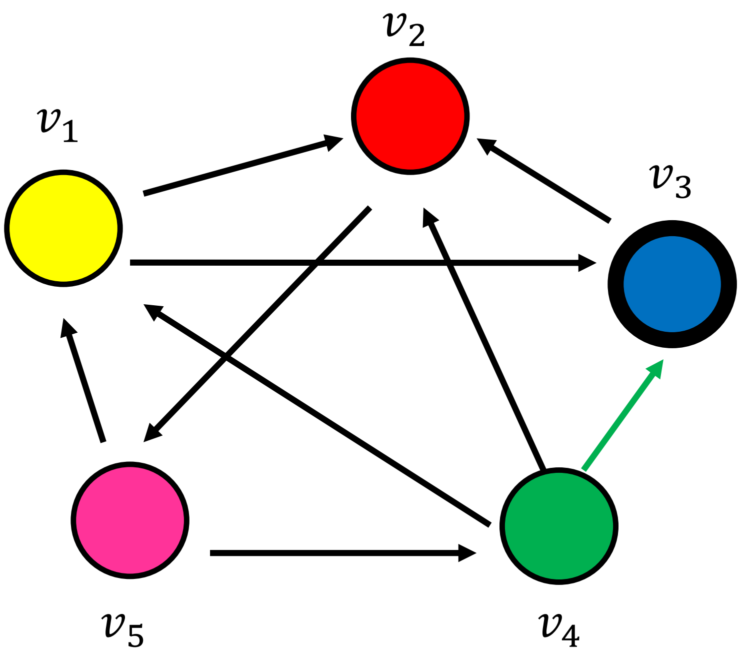
Example

- Step 2



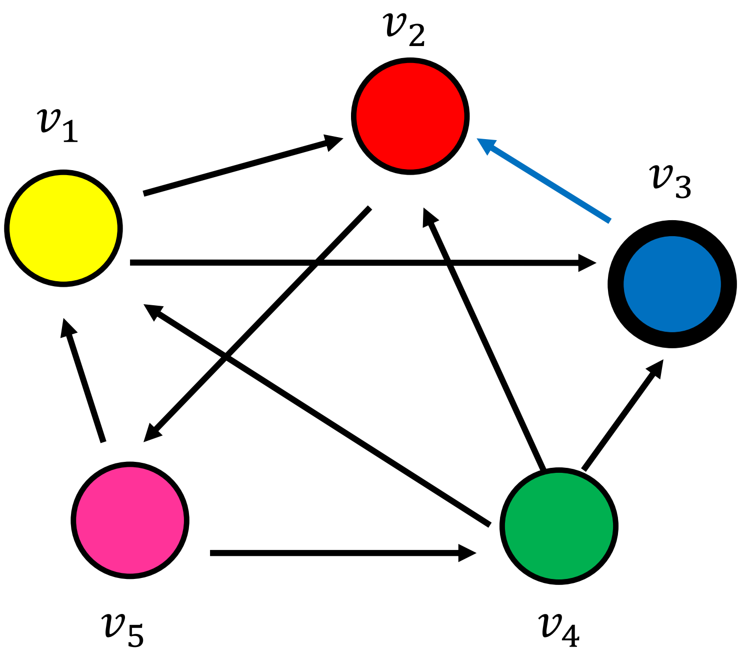
Example

- Step 3



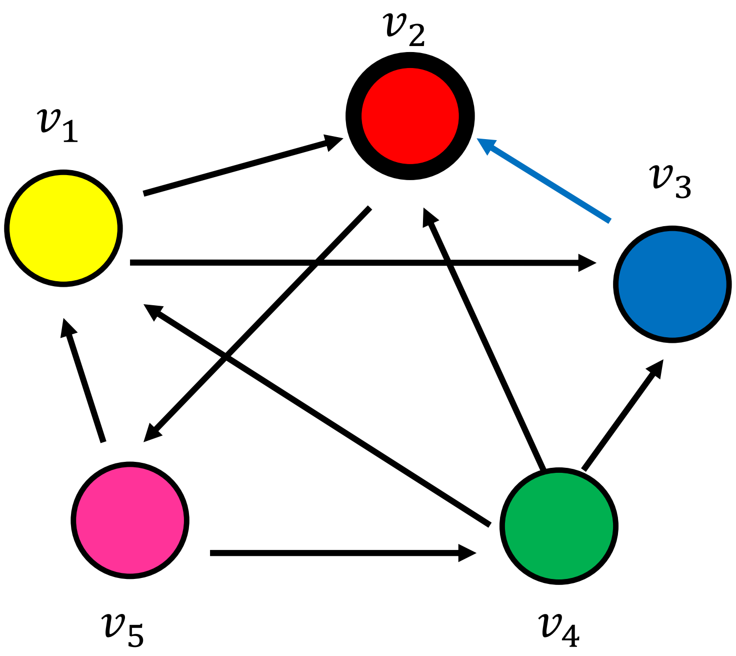
Example

- Step 3



Example

- Step 4…



Random walk

- Question: what is the probability p_i^t of being at node i after t steps?

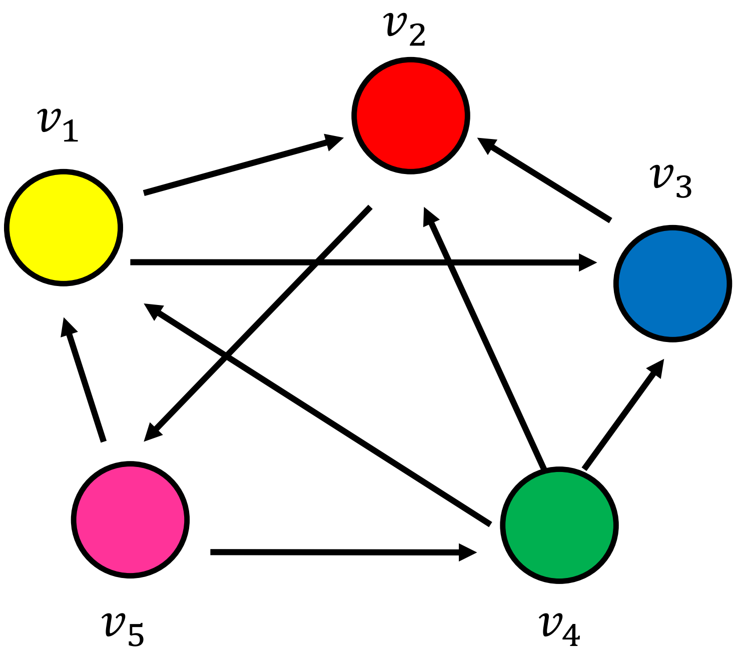
$$p_1^0 = \frac{1}{5} \qquad p_1^t = \frac{1}{3}p_4^{t-1} + \frac{1}{2}p_5^{t-1}$$

$$p_2^0 = \frac{1}{5} \qquad p_2^t = \frac{1}{2}p_1^{t-1} + p_3^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_3^0 = \frac{1}{5} \qquad p_3^t = \frac{1}{2}p_1^{t-1} + \frac{1}{3}p_4^{t-1}$$

$$p_4^0 = \frac{1}{5} \qquad p_4^t = \frac{1}{2}p_5^{t-1}$$

$$p_5^0 = \frac{1}{5} \qquad p_5^t = p_2^{t-1}$$



Markov chains

- A Markov chain describes a discrete time stochastic process over a set of states

$$S = \{s_1, s_2, \dots, s_n\}$$

according to a transition probability matrix $P = \{P_{ij}\}$

  - P_{ij} = probability of moving to state j when at state i

- Matrix P has the property that the entries of all rows sum to 1

$$\sum_j P[i,j] = 1$$

A matrix with this property is called stochastic

- State probability distribution: The vector $p^t = (p_1^t, p_2^t, \dots, p_n^t)$ that stores the probability of being at state s_i after t steps
- Memorylessness property: The next state of the chain depends only at the current state and not on the past of the process (first order MC)
  - Higher order MCs are also possible
- Markov Chain Theory: After infinite steps the state probability vector converges to a unique distribution if the chain is irreducible (possible to get from any state to any other state) and aperiodic

Random walks

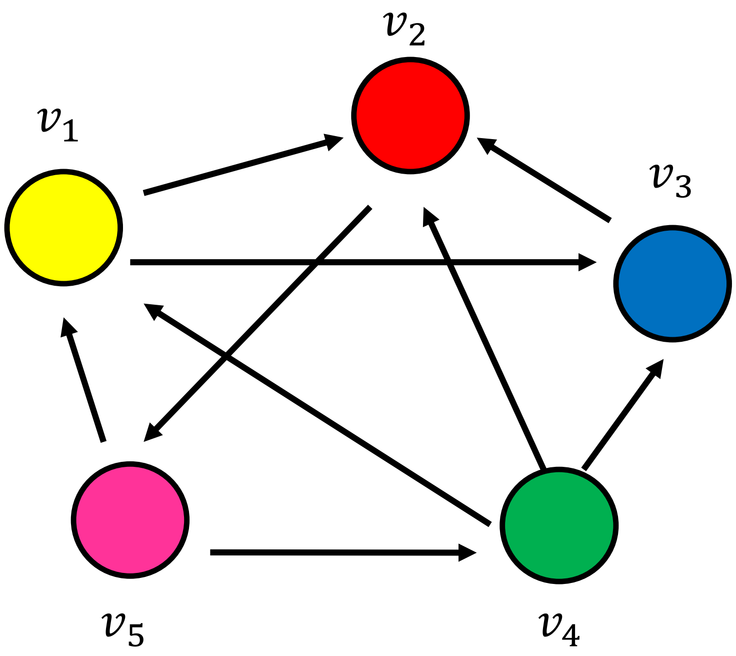
- Random walks on graphs correspond to Markov Chains
  - The set of states S is the set of nodes of the graph G
  - The transition probability matrix is the probability that we follow an edge from one node to another

$$P[i,j] = 1/\deg_{out}(i)$$

An example

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

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



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

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

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 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 $t \rightarrow \infty$

- Markov Chain Theory: The random walk converges to a unique stationary distribution independent of the initial vector if the graph is strongly connected, and not bipartite.

Computing the stationary distribution

- The Power Method

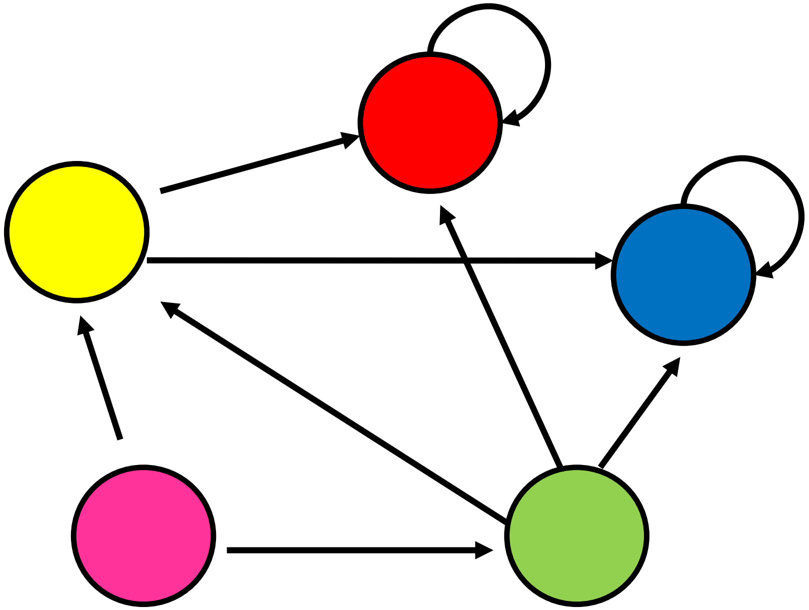
```
Initialize $q^0$ to some distribution
Repeat
    $q^t = q^{t-1}P$
Until convergence
```

- After many iterations $q^t \to \pi$ regardless of the initial vector q^0
- Power method because it computes $q^t = q^0P^t$

- Rate of convergence
  - determined by the second eigenvalue λ_2

Random walk with absorbing nodes

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

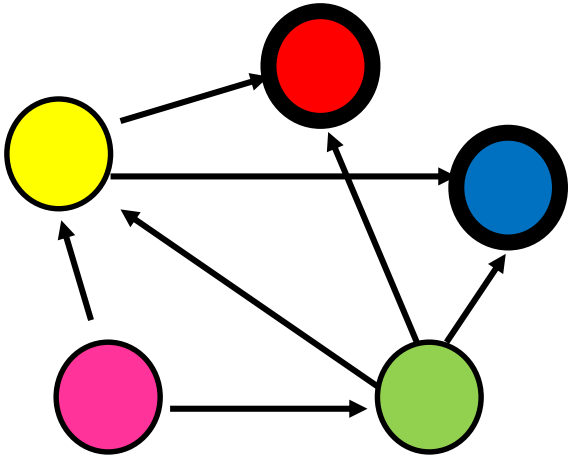


- Two absorbing nodes: the red and the blue.

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 node t we can compute the probabilities of absorption



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?

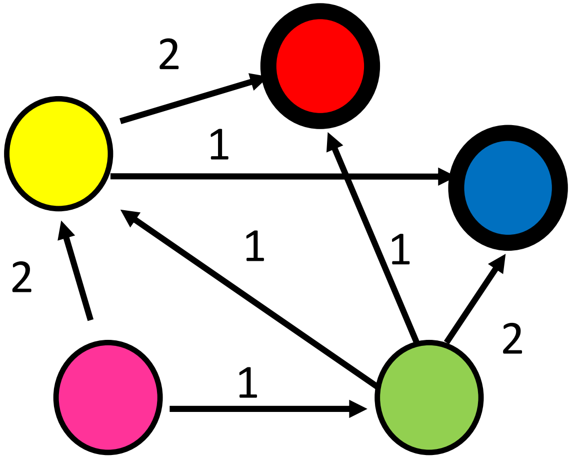
Absorption probabilities

- The absorption probability can be computed iteratively:
  - 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)

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

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

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



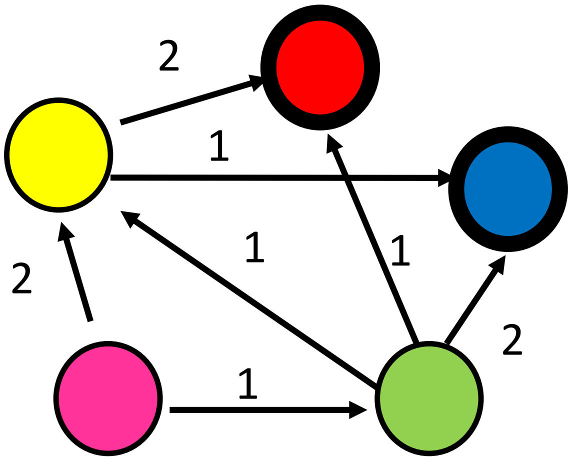
Absorption probabilities

- The absorption probability can be computed iteratively:
  - 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)

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

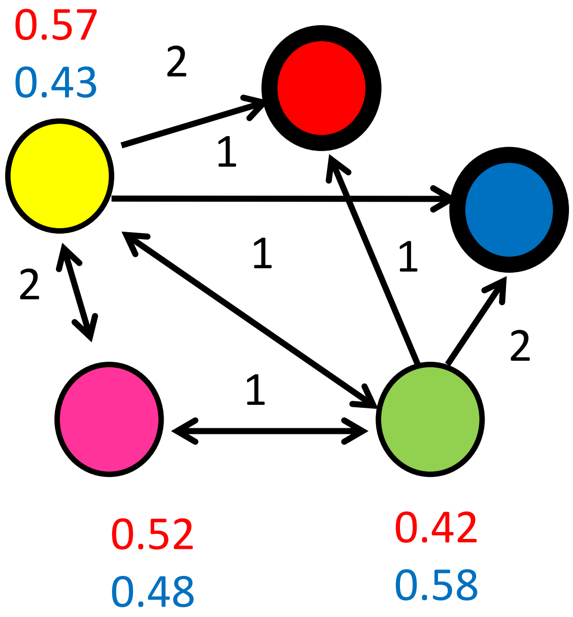
$$P(Blue|Green) = \frac{1}{4}P(Blue|Yellow) + \frac{1}{2}$$

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



Absorption probabilities

- Compute the absorption probabilities for red and blue



Linear Algebra

- Our matrix looks like this

$$P = \begin{bmatrix} P_{TT} & P_{TA} \\ 0 & I \end{bmatrix}$$

- P_{TT}: transition probabilities between transient nodes
- P_{TA}: transition probabilities from transient to absorbing nodes
- When computing the absorption probability to node i we essentially iteratively apply matrix P on the vector $(0, \dots, 1, \dots, 0)$

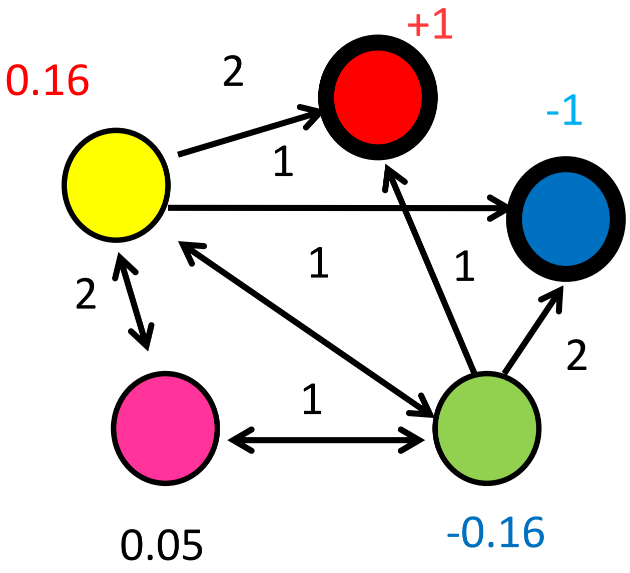
Propagating values

- 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

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

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

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



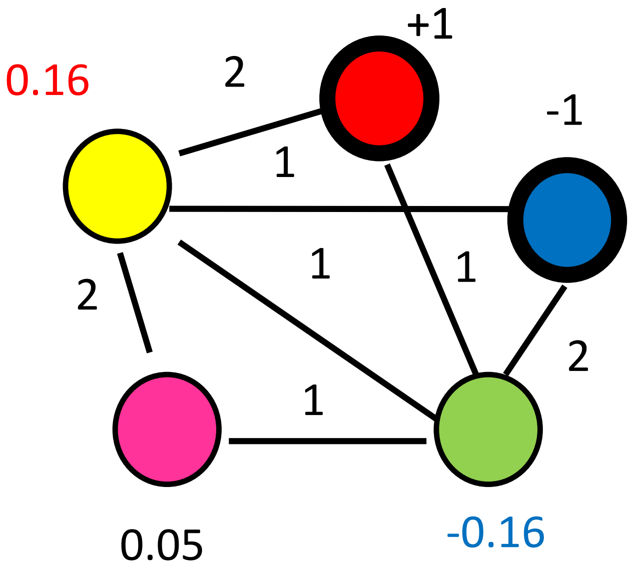
Electrical networks and random walks

- 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

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

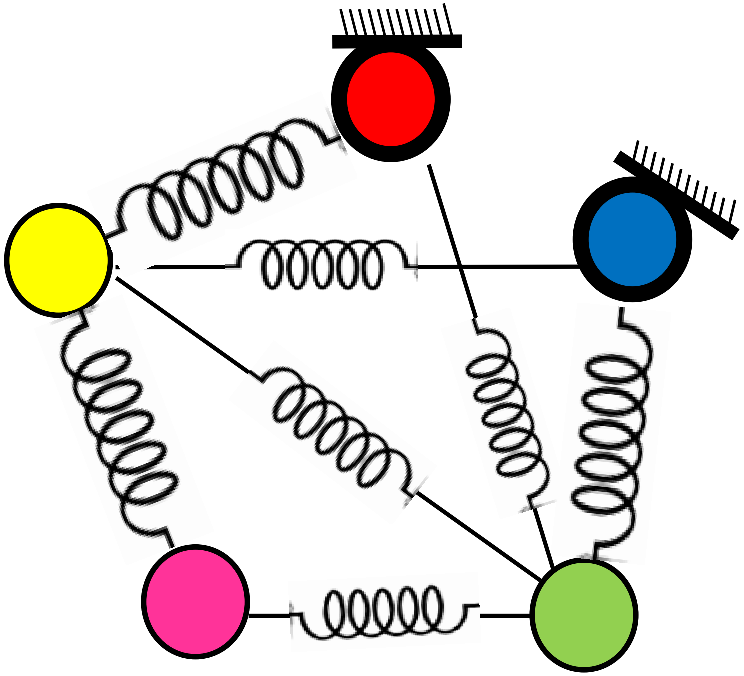
$$V(Green) = \frac{1}{5} V(Yellow) + \frac{1}{5} V(Pink) + \frac{1}{5} - \frac{2}{5}$$

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



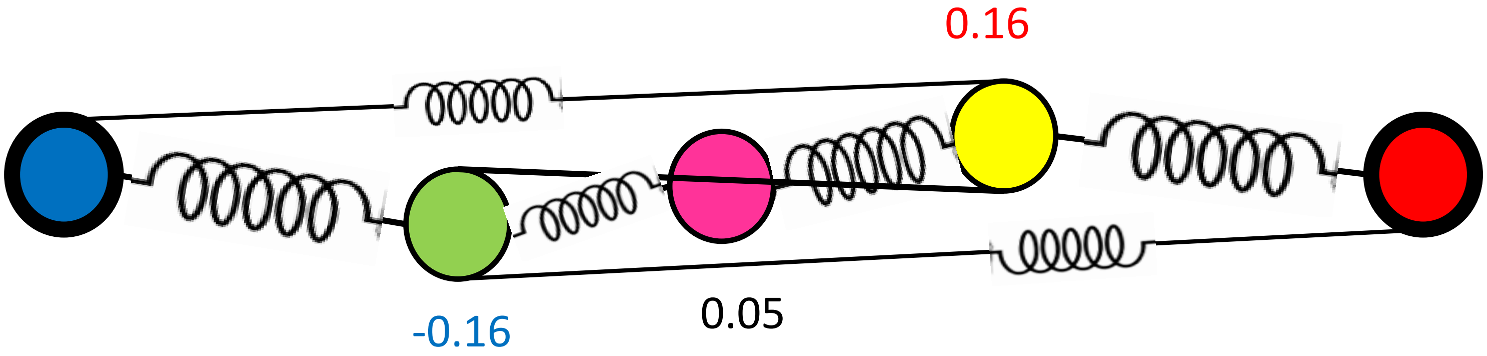
Springs and random walks

- Our graph corresponds to an 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



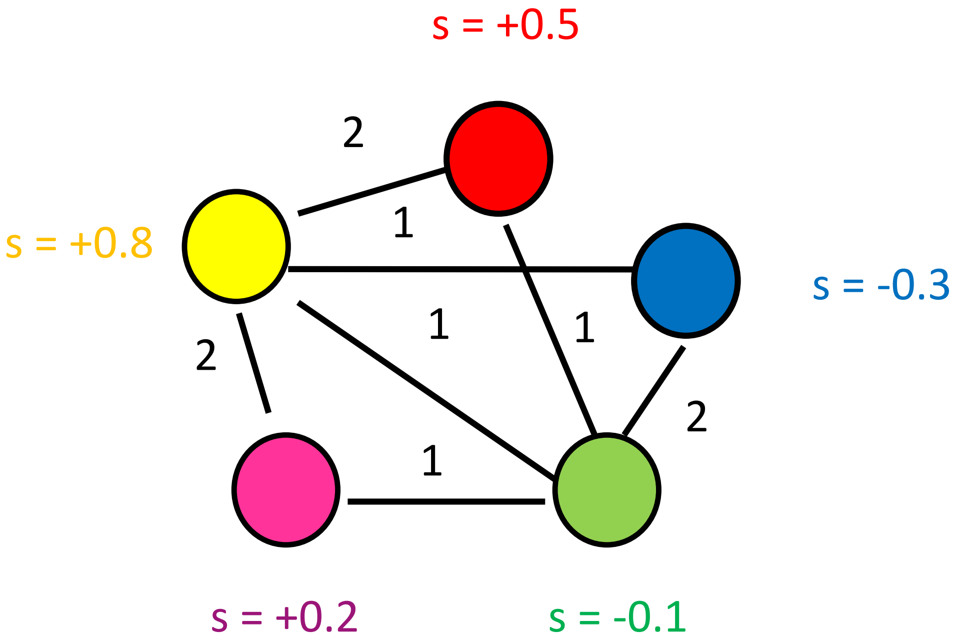
Springs and random walks

- Our graph corresponds to an 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



Back to opinion formation

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



Reminder:

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

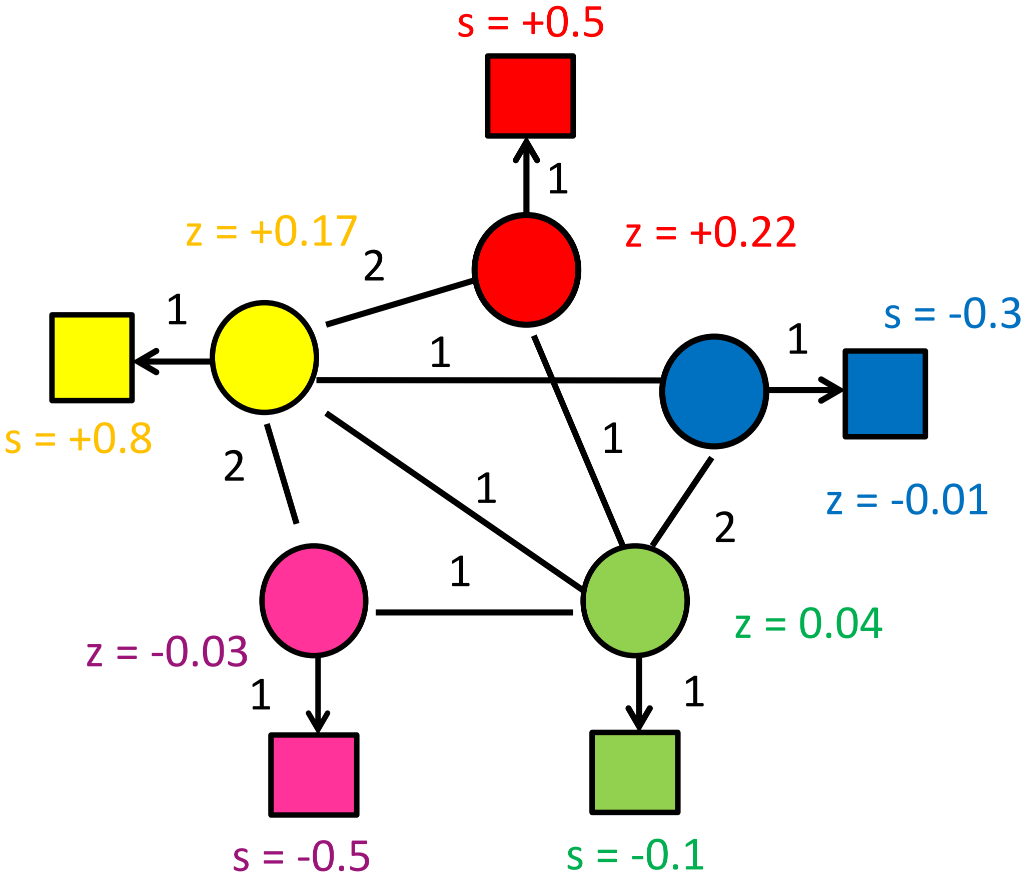
Opinion formation and absorbing random walks

One absorbing node per user with value the intrinsic opinion of the user

One transient node per user that links to her absorbing node and the transient nodes of her neighbors

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

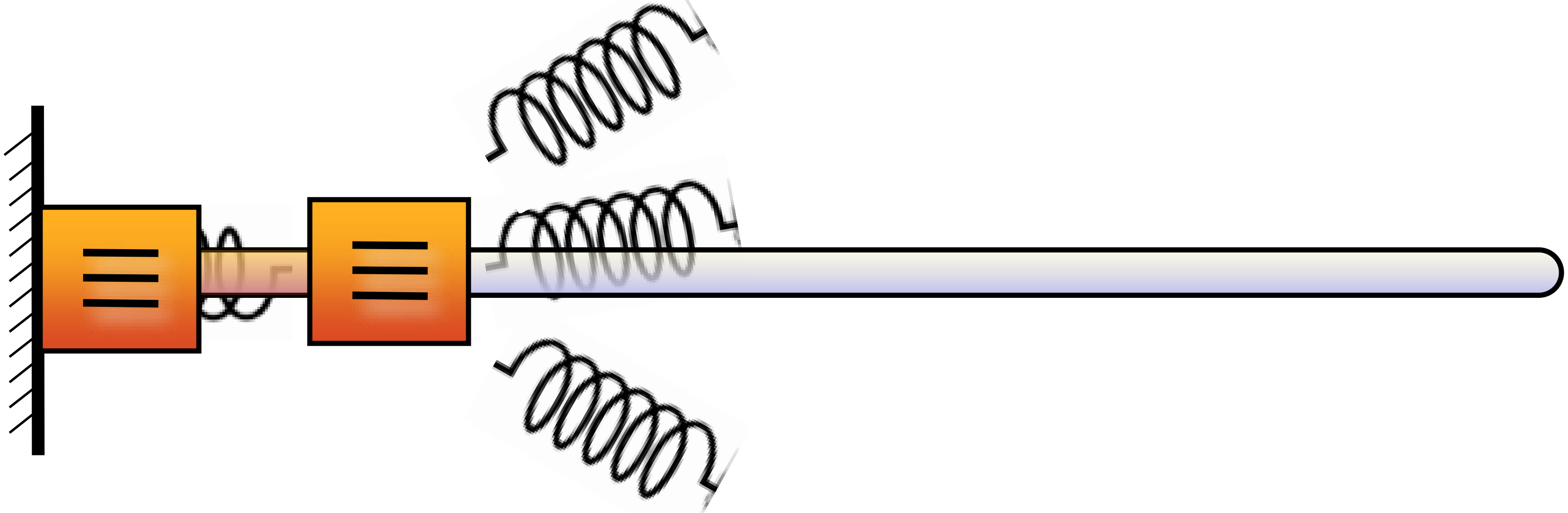
- Repeated averaging

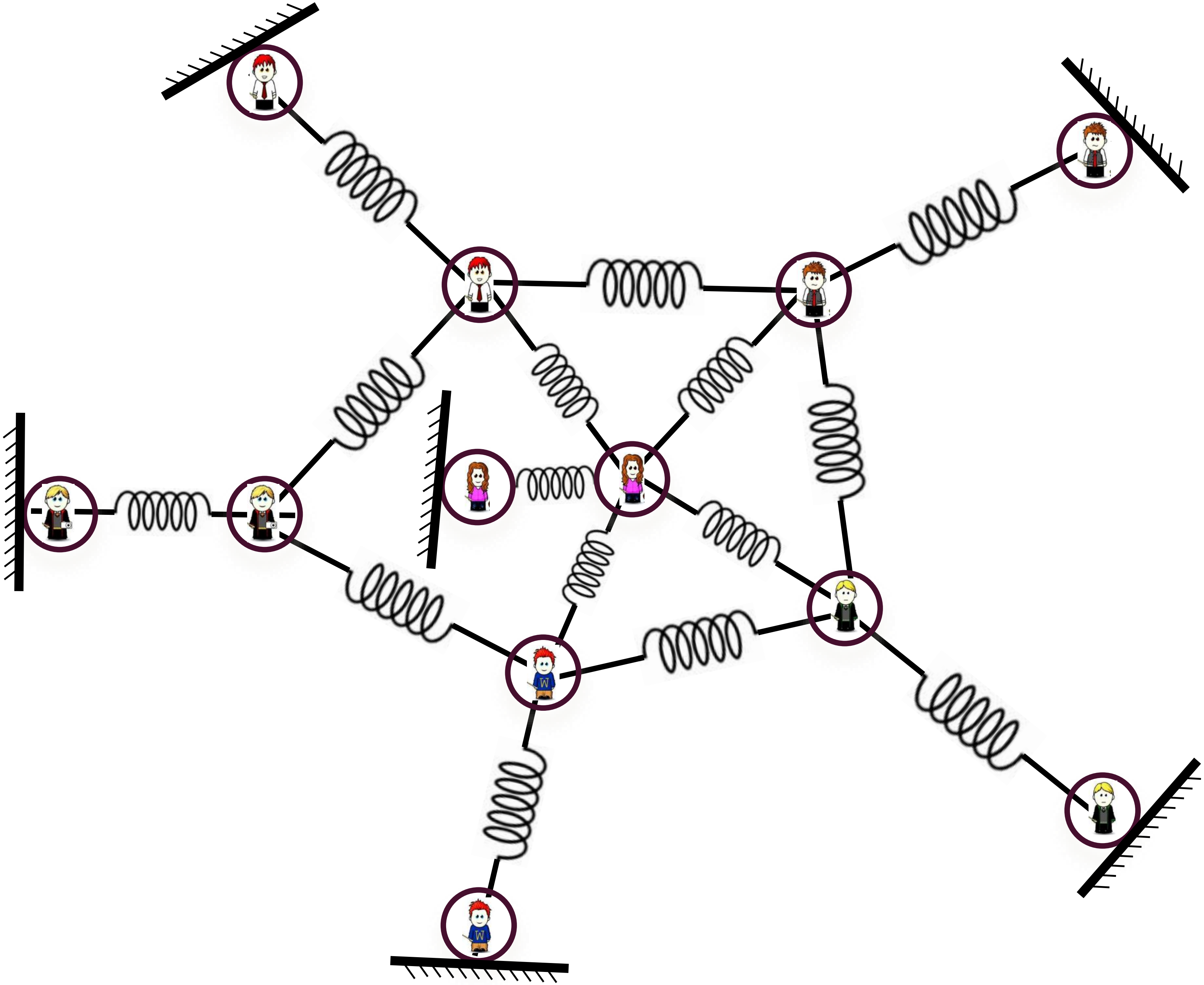


It is equal to the expected intrinsic opinion at the place of absorption

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.

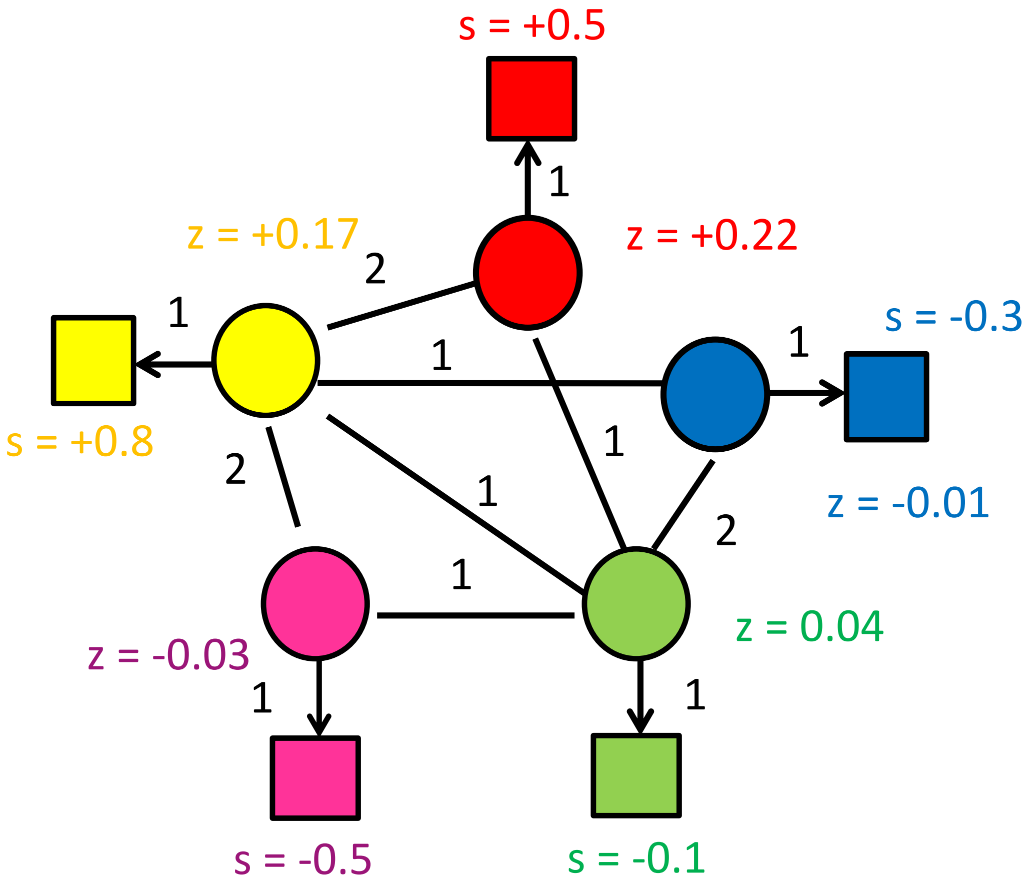
2. Fix the intrinsic opinion of k nodes to the maximum value 1.
   - Easy to solve, we know exactly the contribution of each node to the overall public opinion.

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

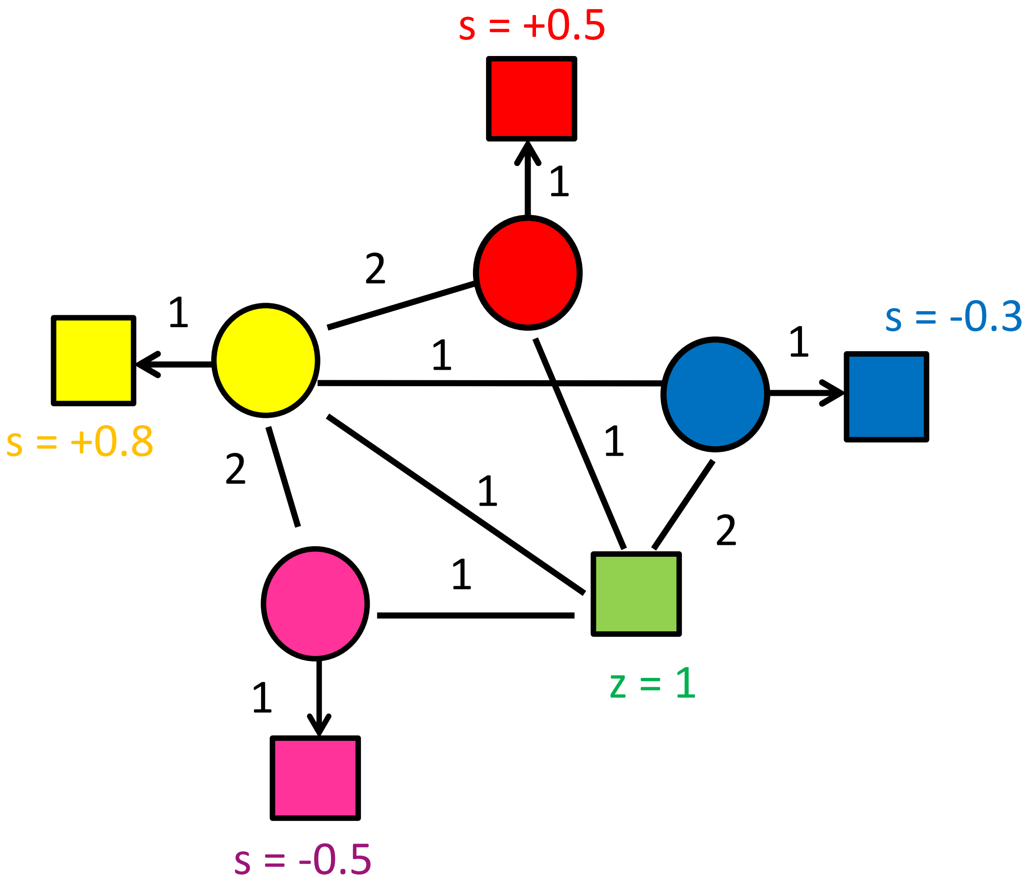
$$g(z) = \sum_i z_i = \sum_i s_i$$

   - 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 an $\left(1-\frac{1}{e}\right)$-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

Other problems related to opinion formation

- Modeling polarity
  - Understand why extreme opinions are formed and people cluster around them
- Modeling herding/flocking
  - Understand under what conditions people tend to follow the crowd
- Computational Sociology
  - Use big data for 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

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
- 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

Thank you!

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