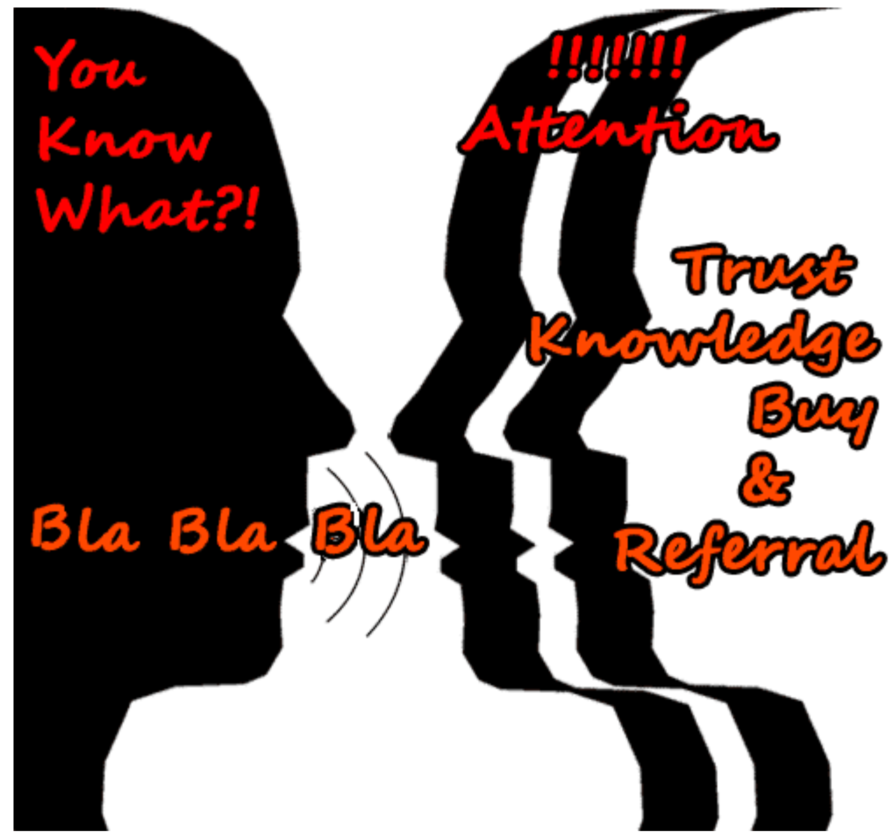
Spread of Influence through a Social Network

Adapted from :

http://www.cs.washington.edu/affiliates/meetings/talks04/kempe.pdf

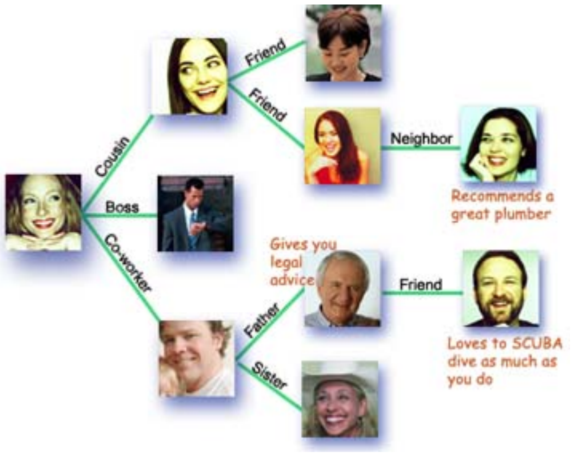
Influence Spread

- We live in communities and interact with our friends, family and even strangers.
- In the process, we influence each other.

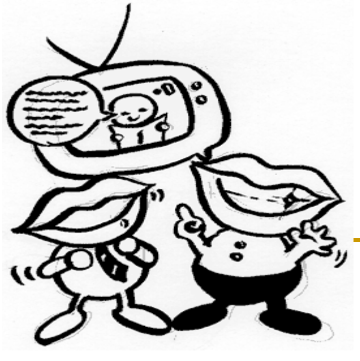


Social Network and Spread of Influence

- Social network plays a fundamental role as a medium for the spread of INFLUENCE among its members
  - Opinions, ideas, information, innovation…



- Direct Marketing takes the "word-of-mouth" effects to significantly increase profits (Gmail, Tupperware popularization, Microsoft Origami …)



Social Network and Spread of Influence

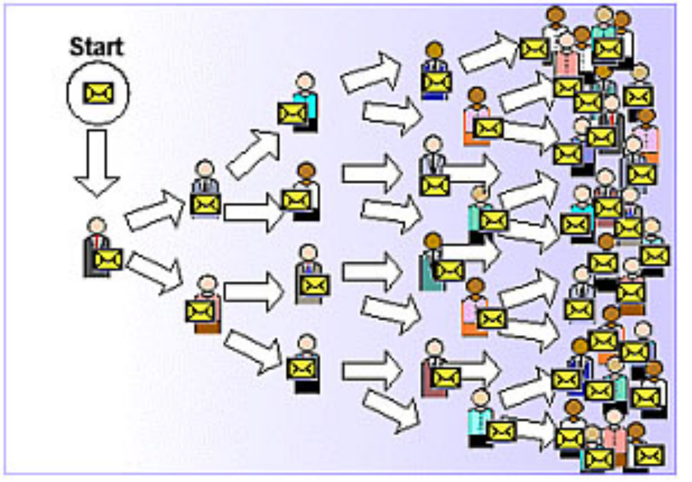
- Examples:
  - Hotmail grew from zero users to 12 million users in 18 months on a small advertising budget.
  - A company selects a small number of customers and ask them to try a new product. The company wants to choose a small group with largest influence.
  - Obesity grows as fat people stay with fat people (homofily relations)
  - Viral Marketing..

Viral Marketing

Identify influential customers

Convince them to adopt the product – Offer discount/free samples

These customers endorse the product among their friends



Problem Setting

- Given
  - a limited budget B for initial advertising (e.g. give away free samples of product)
  - estimates for influence between individuals
- Goal
  - trigger a large cascade of influence (e.g. further adoptions of a product)
- Question
  - Which set of individuals should B target at?
- Application besides product marketing
  - spread an innovation
  - detect stories in blogs (gossips)
  - Epidemiological analysis

What we need

- Models of influence in social networks.
- Obtain data about particular network (to estimate inter-personal influence).
- Devise algorithms to maximize spread of influence.

Outline

- Stochastic Models of influence diffusion
  - Linear Threshold
  - Independent Cascade
- Influence maximization problem
  - Algorithm

Outline

- Models of influence
  - Linear Threshold
  - Independent Cascade
- Influence maximization problem
  - Algorithm

Models of Influence

- Two basic classes of diffusion models: threshold and cascade
- General operational view:
  - A social network is represented as a directed graph, with each person (customer) as a node
  - Nodes start either **active** or **inactive**
  - An active node may trigger activation of neighboring nodes
  - **Monotonicity assumption**: active nodes never deactivate (not always true, e.g. epidemics (e.g. flu, covid), here more complex models are used)

Outline

- Models of influence
  - Linear Threshold
  - Independent Cascade
- Influence maximization problem
  - Algorithm

Linear Threshold Model

- A node v has (random) threshold $\theta_v \sim U[0,1]$
- A node v is influenced by each neighbor w according to a *weight* b_{vw} such that

$$\sum_{w \text{ neighbor of } v} b_{v,w} \leq 1$$

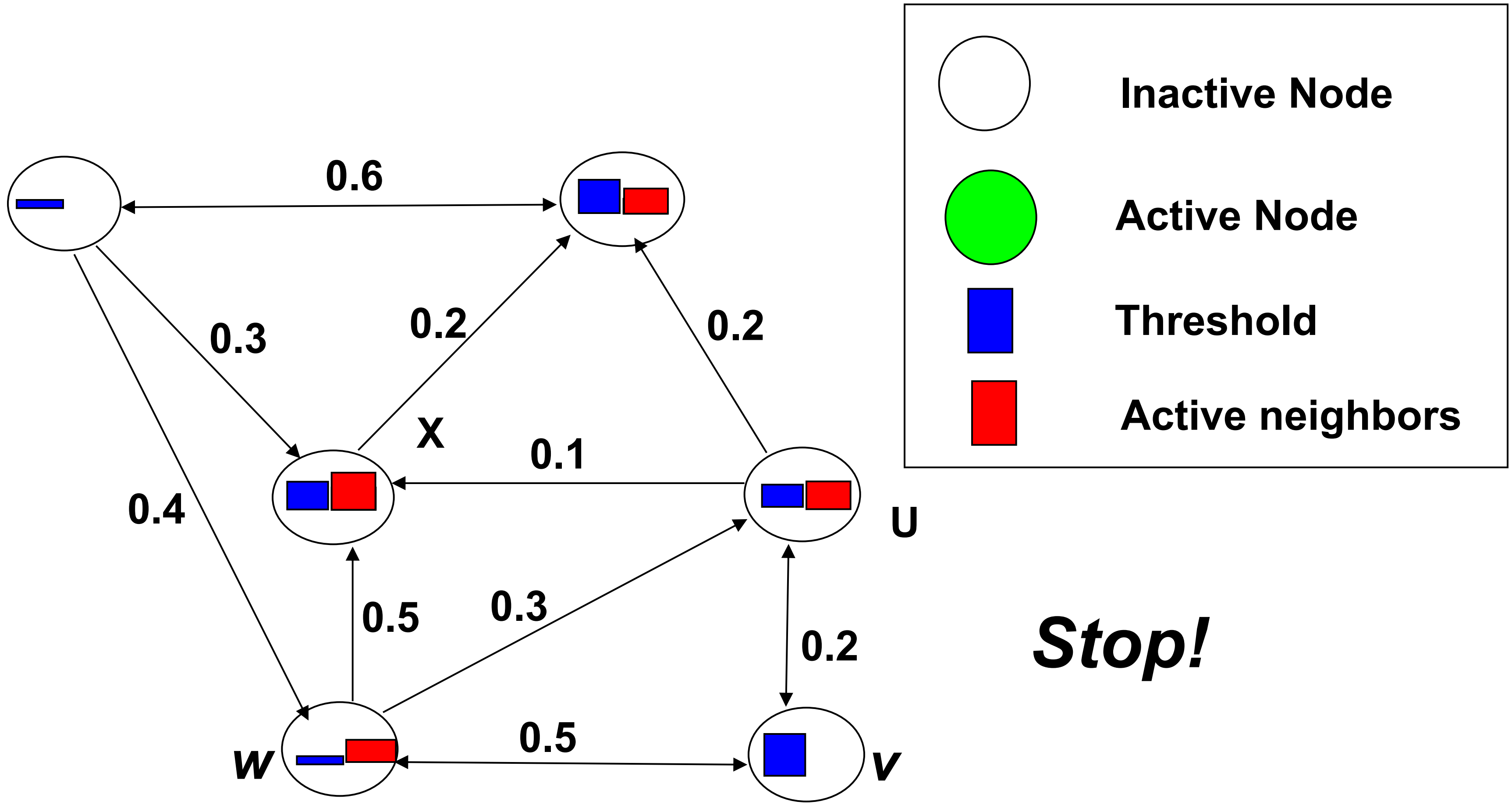
- A node v becomes **active** when at least

(weighted) θ_v fraction of its neighbors are active

$$\sum_{w \text{ active neighbor of } v} b_{v,w} \geq \theta_v$$

Similar to perceptron model..

Example (weights on edges are the $b_{u,v}$)



Stop!

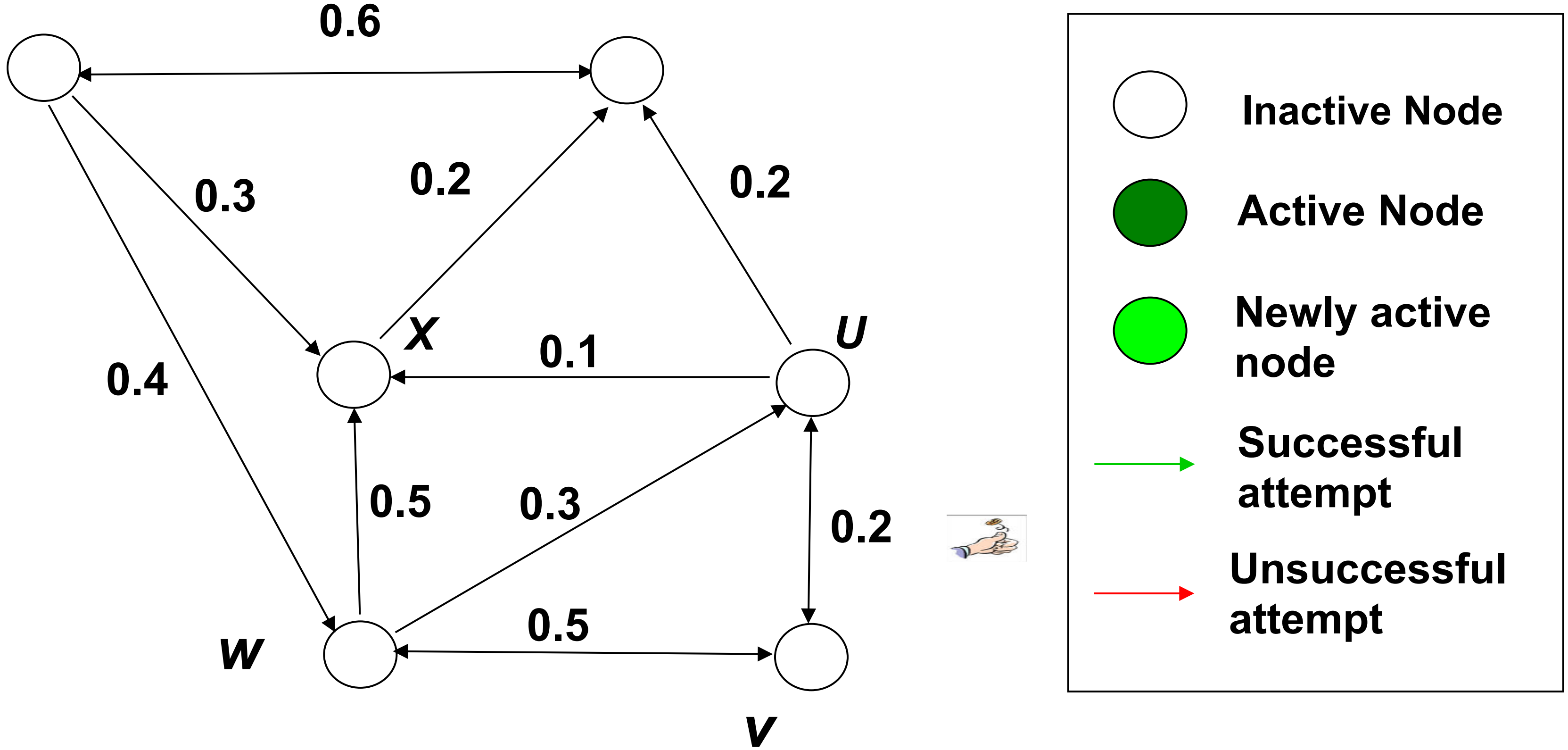
Outline

- Models of influence
  - Linear Threshold
  - Independent Cascade
- Influence maximization problem
  - Algorithm

Independent Cascade Model

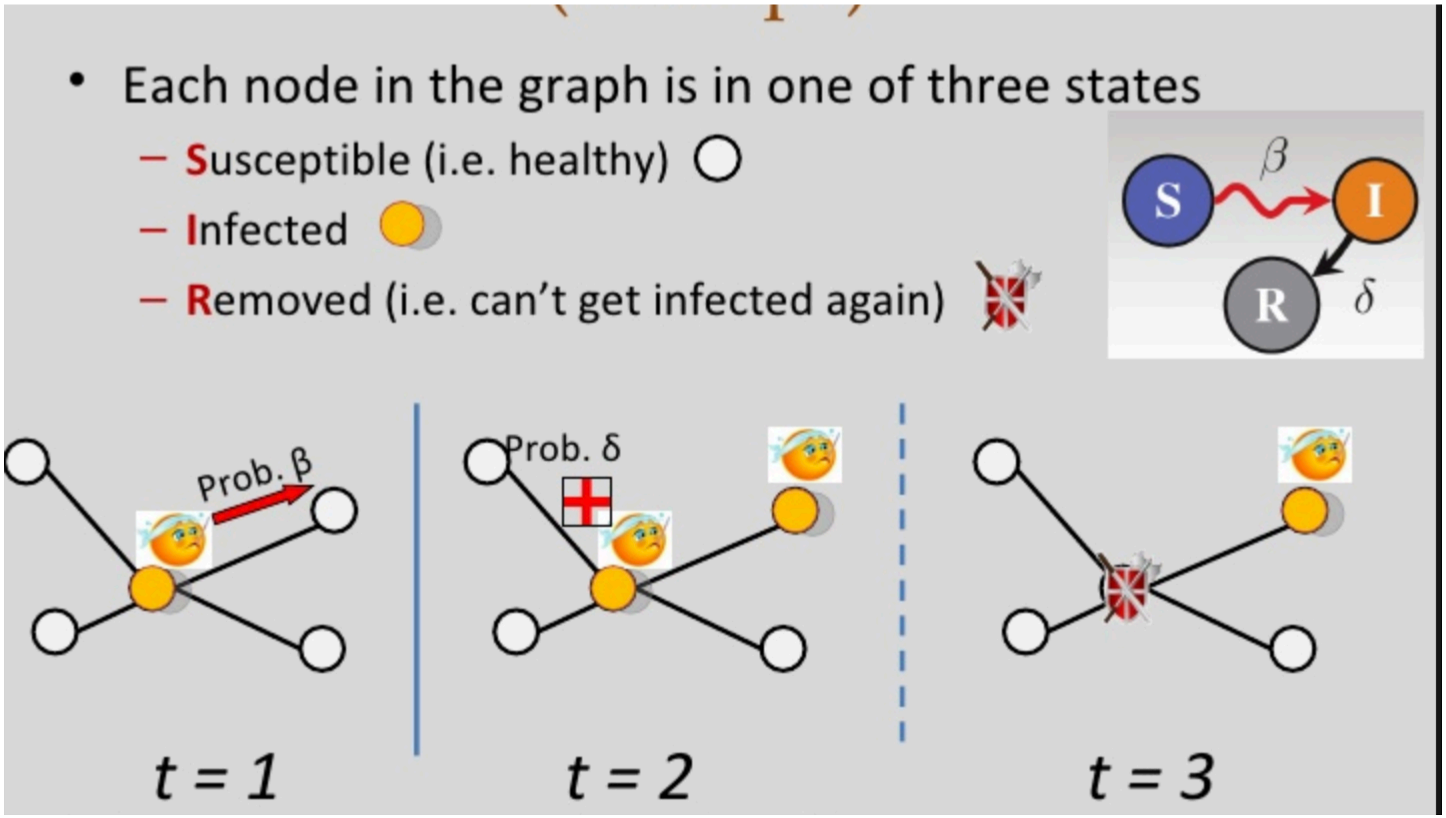
- When node *v* becomes active, it has a **single** chance of activating **each** currently inactive neighbor *w*.
- The activation attempt succeeds with probability p_{vw} .

Example (edges are now weighted with probabilities)



Stop!

Virus Diffusion model (simple SIR model)



Outline

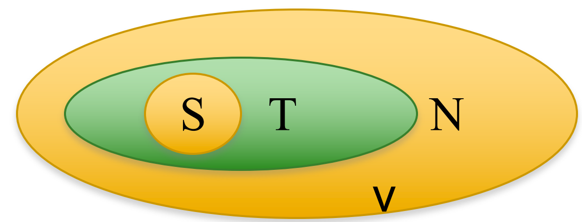
- Models of influence
  - Linear Threshold
  - Independent Cascade
  - SIR
- Influence maximization problem
  - Algorithm

Influence Maximization Problem

- Influence of node set S: *f(S)*
  - expected number of active nodes at **steady state**, if set S is the initial **active set** in t0
- Problem:
  - Given a parameter *k* (**budget**), find a ***k*-node set *S*** to maximize *f(S)*
  - This can be casted as a constrained optimization problem with *f(S)* as the objective function
  - Of course for disease epidemics the problem is to **minimize**

f(S): properties (to be demonstrated)

- Non-negative (obviously)
- Monotone: $f(S+v) \geq f(S)$
- Submodular:



  - Let N be a finite set
  - A set function $f: 2^N \mapsto \Re$ is submodular *iff*

$$\forall S \subset T \subset N, \forall v \in N \setminus T,$$

$$f(S+v) - f(S) \geq f(T+v) - f(T)$$

N\T =N-T

Intuitive explanation: The **difference** in the value of the function that a **single element** (v) makes when added to an input set decreases as the size of the input set increases. (Also called *diminishing returns*)

Bad News

- For a submodular function *f*, if *f* only takes non-negative value, and is monotone, finding a *k*-element set *S* for which *f(S)* is maximized is an **NP-hard** optimization problem.
- **It is NP-hard to determine** the optimum for influence maximization for both independent cascade model and linear threshold model.

Good News

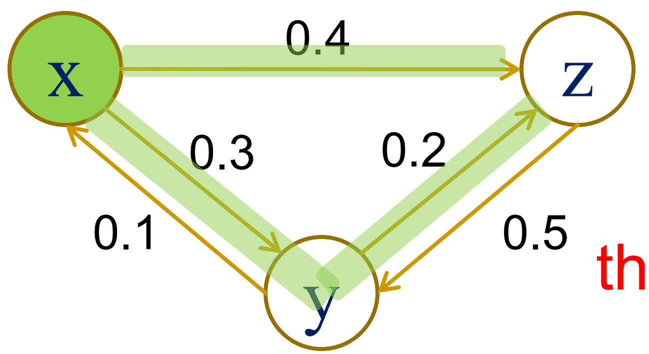
- We can use Greedy Algorithm!
  - Start with an empty set S
  - For k iterations:

    Add node *v* to S that maximizes $f(S + v) - f(S)$.
- How good (bad) it is?
  - Theorem: The greedy algorithm is a $(1 - 1/e)$ approximation.
  - The resulting set S activates at least $(1- 1/e) > 63\%$ of the number of nodes that any size-k set S could activate (so at **worst 63% of the optimum**).

Estimating Spread S(v) (Linear Threshold Model)

- We observe that the influence of a node x on node z can be computed by enumerating all simple paths starting from x and ending in z.

A simple path is a path that doesn't contain any cycle



Total influence of x on z is 0.46

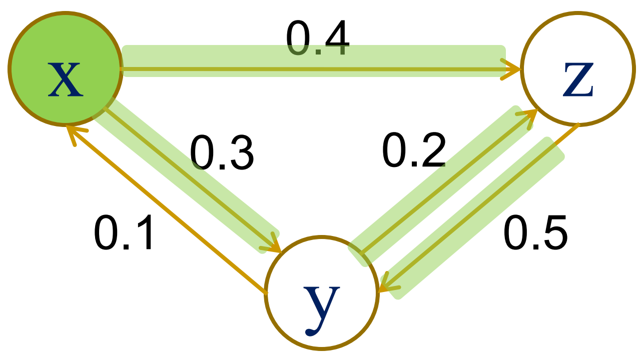
Influence of x on z through path x→y→z is 0.3 * 0.2 = 0.06

Influence of x on z through path x → z is 0.4

$$\Upsilon_{x,z} = 0.3 \cdot 0.2 + 0.4 = 0.46$$

Estimating Spread (Linear Threshold Model)

- Thus, the spread of a node can be computed by enumerating simple paths starting from the node.



Influence of x on x itself

Influence of x on y

Influence of x on z

Total influence of node x is 1.96

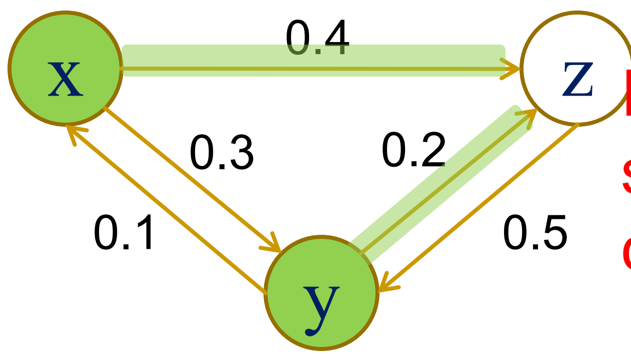
Influence Spread of node x is $\Upsilon_{x,x} + \Upsilon_{x,y} + \Upsilon_{x,z}$

$= 1 + (0.3 + 0.4 * 0.5) + (0.4 + 0.3 * 0.2) = 1.96$

Estimating Spread (Linear Threshold Model)

Theorem 1. *In the LT model, the spread of a set S is the sum of the spread of each node $u \in S$ on subgraphs induced by $V - S + u$. That is,*

$$\sigma(S) = \sum_{u \in S} \sigma^{V-S+u}(u)$$



Influence of node x in a subgraph that does not contain y

Let the seed set S = {x,y}, then influence spread of S is

$$\sigma(S) = \sigma^{V-y}(x) + \sigma^{V-x}(y) = 1 + 0.4 + 1 + 0.2 = 2.6$$

Performance

- Spread of influence algorithms have been demonstrated to obtain much better performance wrt simpler methods such as:
  - Page Rank – Top-k nodes with highest page rank.
  - High Degree – Top-k nodes with highest degree.
- Temporal complexity is an issue (several algorithms recently improved over "base" algorithm described here)