Reminder

- TA's announcement on course project report 1
- PRA4 due 3/14
- HW4 due 3/21
- Course project progress report 2 due 3/26
- Come to OH for course project discussion!

Artificial Intelligence Methods for Social Good

Lecture 16:

Influence Maximization and Case Study on

HIV Prevention Among Homeless Youth

Instructor: Fei Fang

feifang@cmu.edu

Outline

- Influence Maximization Problem
- Discussion
- Monte Carlo Tree Search
- Case Study: HIV Prevention Among Homeless Youth

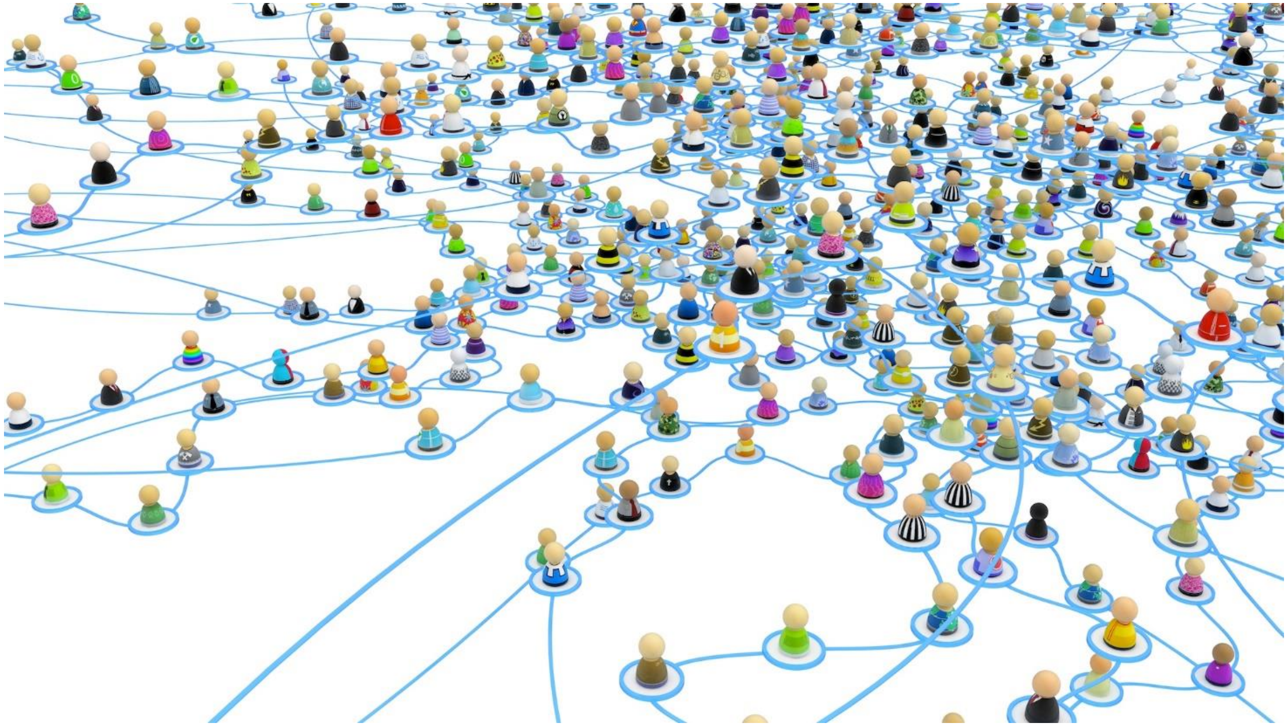
Learning Objectives

- Understand the concept of
  - Submodular function
- Describe
  - Independent Cascade Model
  - Linear Threshold Model
  - Influence Maximization Problem
  - Greedy Algorithm for Influence Maximization Problem
- For the case study, briefly describe
  - Significance/Motivation
  - Task being tackled, i.e., what is being predicted/estimated
  - Data usage, i.e., what data is used and how it is processed
  - Domain-specific considerations
  - AI method used
  - Evaluation process and criteria

Social Networks

https://kampp.org/2019/10/30/social-media-management/

Social Networks

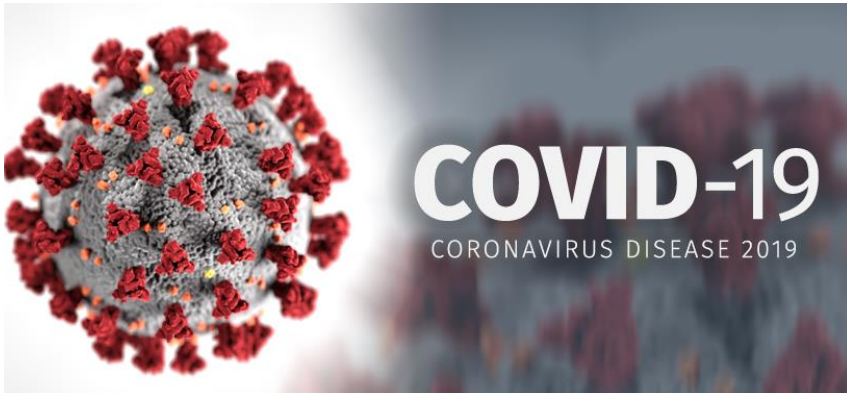


https://thenextweb.com/socialmedia/2013/11/24/facebook-grandparents-need-next-gen-social-network/

Propagation Process

- Viral propagation
  - Virus/Rumors
  - Get infected immediately and spread automatically
  - Individual agent does not make decisions

    Is it really the case?



- Decision-based models
  - Individual agent makes decisions
  - Influence and adoption

https://www.dshs.state.tx.us/coronavirus/
https://sloanreview.mit.edu/article/the-power-of-product-recommendation-networks/

Propagation Process

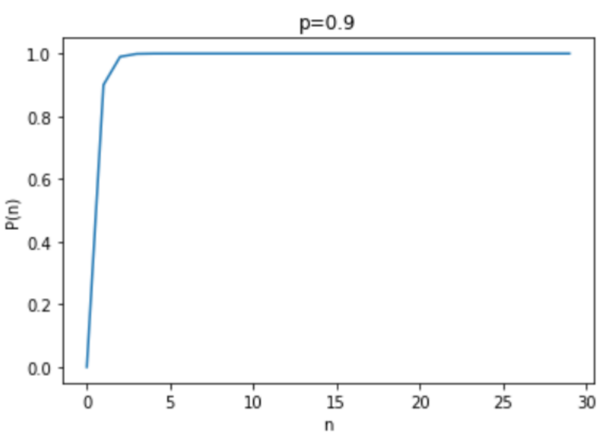
- General operational view:
  - A social network is represented as a (un)directed graph

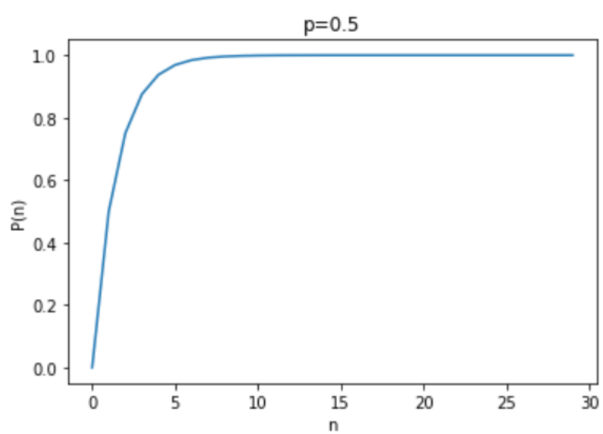
  - Nodes start either active or inactive
  - Active node may trigger activation of neighboring nodes
  - Monotonicity assumption: active nodes never deactivate

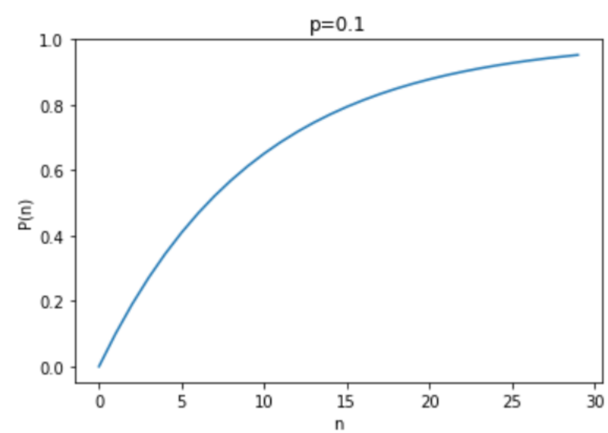
https://thenextweb.com/contributors/2017/08/21/blockchain-can-make-social-networks-private-profitable/

Influence Response Function

- Influence Response Function
  - Independent Draws
    - n friends recommend it to me
    - $P(n) = 1 - (1 - p)^n$
    - Diminishing return (concave function)

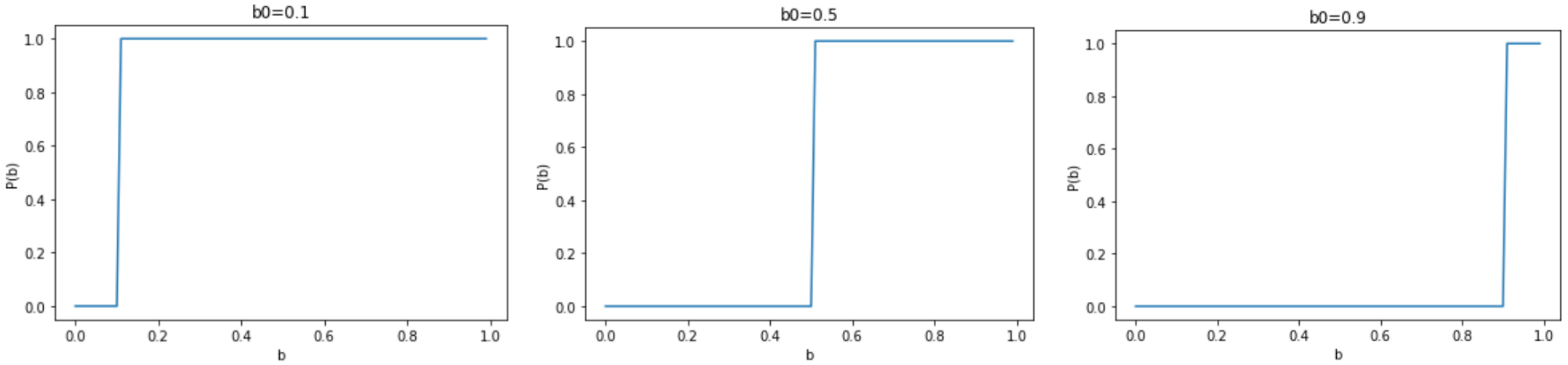






Influence Response Function

- Influence Response Function
  - Independent Draws
  - Linear Threshold
    - Many of my friends bought the item, reaching a critical mass
    - $P(b) = \delta(b > b_0)$



Influence Propagation Models

- Independent Cascade Model (Goldenberg, 2001)
  - Initial set of active nodes
  - Discrete time steps
  - When a node v just becomes active (activated in the last time step), it has a single chance of activating each currently inactive neighbor w (if failed, no second trial)
  - The activation attempt succeeds with probability $p_{v,w}$
  - Process runs until no more activations possible

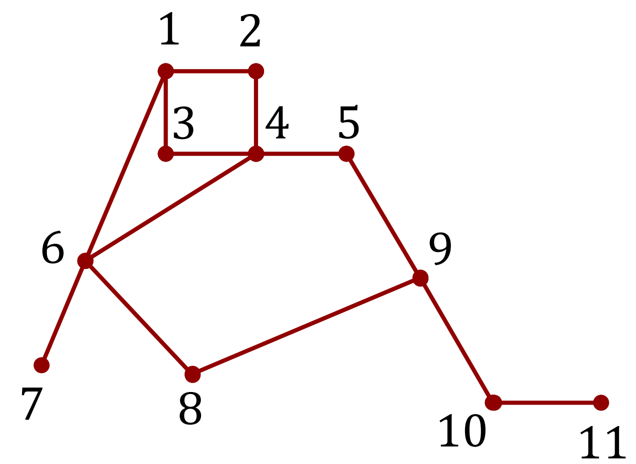
Independent Cascade Model

Independent Cascade Model (Goldenberg, 2001)

Initial set of active nodes A_0
For $t = 1 \ldots T$
    $A_t \leftarrow \emptyset$
    For $v \in A_{t-1}$
        For $w \in neighbor(v)$ and $w \notin \cup_{\tau=0}^{t} A_\tau$
            If $rand(\cdot) < p_{v,w}$, then
                $A_t \leftarrow A_t \cup \{w\}$

Output the set of all nodes activated $A \leftarrow \cup_{t=0}^{T} A_t$

$p_{v,w} = 1, A_0 = \{4\}$



$A_1 =$ $A_2 =$

$A_3 =$ $A_4 =$

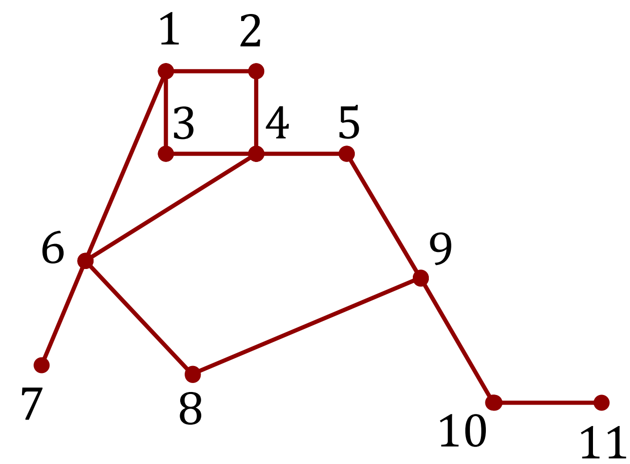
Independent Cascade Model

Independent Cascade Model (Goldenberg, 2001)

Initial set of active nodes A_0
For $t = 1 \dots T$
    $A_t \leftarrow \emptyset$
    For $v \in A_{t-1}$
        For $w \in neighbor(v)$ and $w \notin \cup_{\tau=0}^{t} A_\tau$
            If $rand(\cdot) < p_{v,w}$, then
                $A_t \leftarrow A_t \cup \{w\}$

Output the set of all nodes activated $A \leftarrow \cup_{t=0}^{T} A_t$

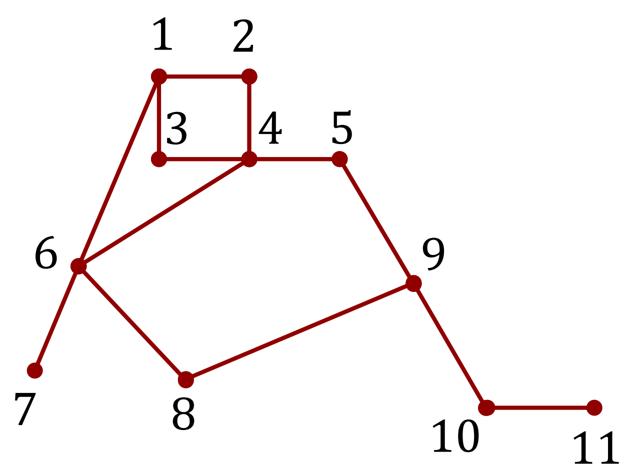
$p_{v,w} = 1, A_0 = \{4\}$



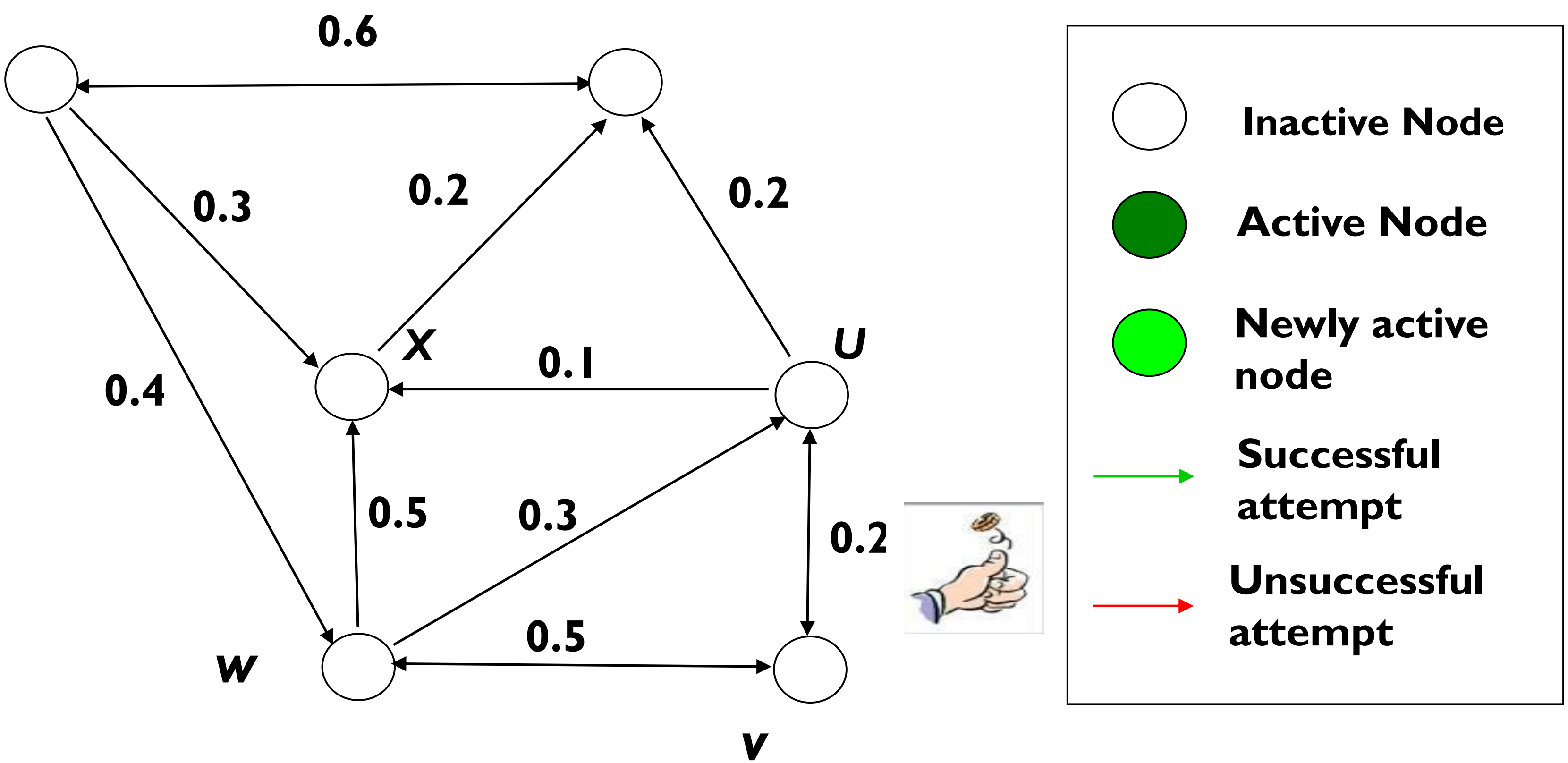
$A_1 = \{2,3,5,6\}$ $A_2 = \{1,7,8,9\}$
$A_3 = \{10\}$ $A_4 = \{11\}$

Poll 1

- How many time steps are needed to achieve global cascade in the following example with $p_{v,w} = 1$ and $A_0 = \{1\}$?
  - A: 2
  - B: 3
  - C: 4
  - D: 5
  - E: None of the above
  - F: I don't know



Independent Cascade Model Example with $p_{v,w} < 1$



Stop!

www.cs.cmu.edu/~xiaonanz/Maximizing-the-Spread-of-Influence.ppt

Influence Propagation Models

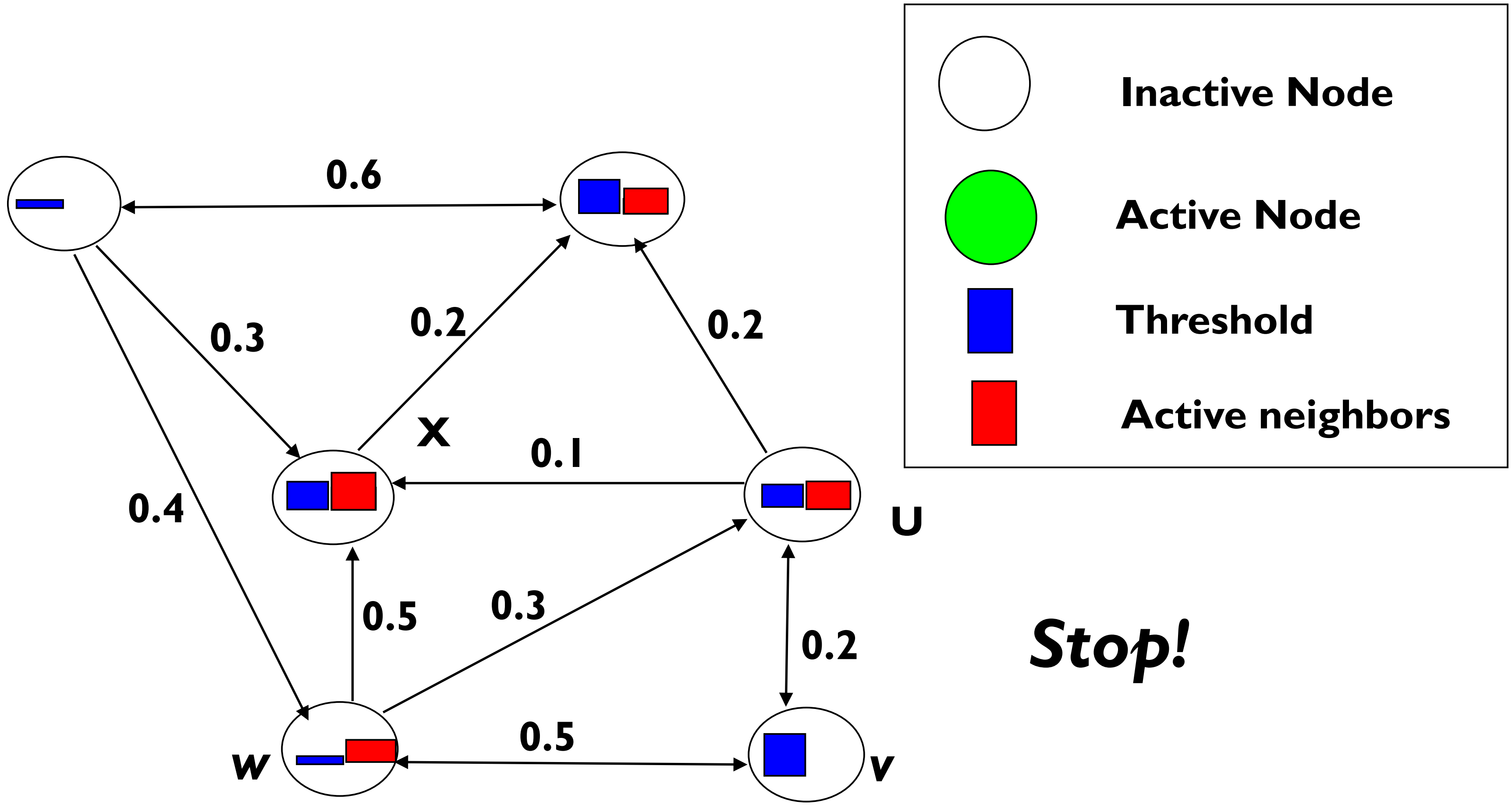
- Linear Threshold Model (M. Granovetter, 1978, T. Schelling, 1970, 1978)
  - Initial set of active nodes, Discrete time steps
  - Each node v has a threshold θ_v
  - Each edge has a weight b_{vw} indicating the influence of node v to node w

$$\sum_{w \in N(v)} b_{vw} \leq 1$$

  - A node v becomes active when total weight of active neighbors exceeds threshold θ_v

$$\sum_{w \in N(v) \text{ and } w \text{ is active}} b_{vw} \geq \theta_v$$

Linear Threshold Model Example



Stop!

www.cs.cmu.edu/~xiaonanz/Maximizing-the-Spread-of-Influence.ppt

Influence Maximization Problem

- How to select initial nodes A_0 to maximize influence $\sigma(A_0)$, under the constraint that A_0 has no more than K nodes

$$\max_{A_0} \sigma(A_0)$$
$$\text{s.t. } |A_0| \leq K$$

- The problem is NP-Hard (Kempe, Kleinberg & Tardos, 2003, 2005)

Submodular Functions

- Submodular Functions
  - $f: 2^N \to \mathbb{R}$ is submodular if
    - For sets S, T where $S \subset T, \forall v \notin T$

$$f(S \cup \{v\}) - f(S) \geq f(T \cup \{v\}) - f(T)$$

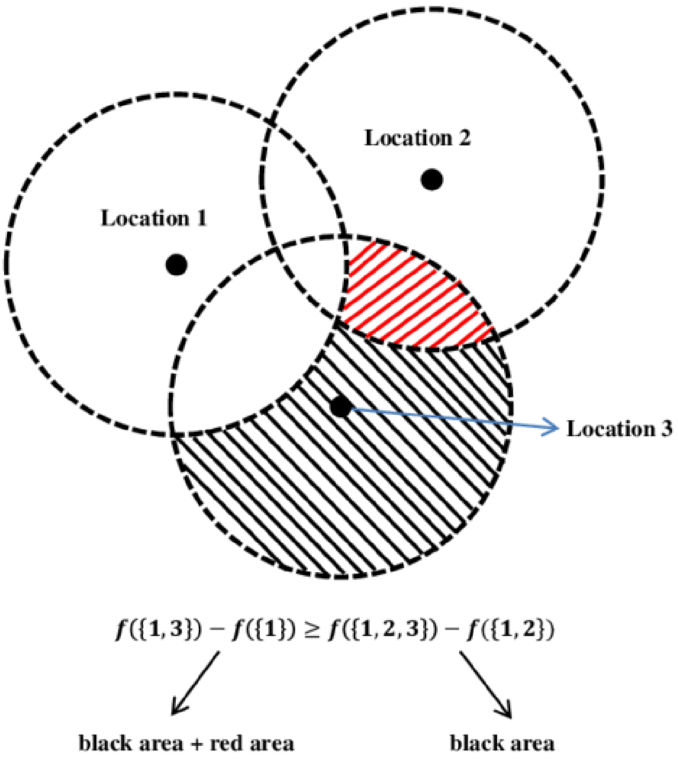
    - If $f(S \cup \{v\}) \geq f(S), \forall S, \forall v$, we say f is monotone
  - Diminishing return (similar to concave function): Marginal value is decreasing as the set gets larger
    - Define marginal value of v given S as

$$f_S(v) = f(S \cup \{v\}) - f(S)$$

    - f is submodular iff $f_S(v) \geq f_T(v)$ for sets S, T where $S \subset T$

Submodular Functions

- Example: Sensor Coverage Problem
  - Similar to maximum coverage problem

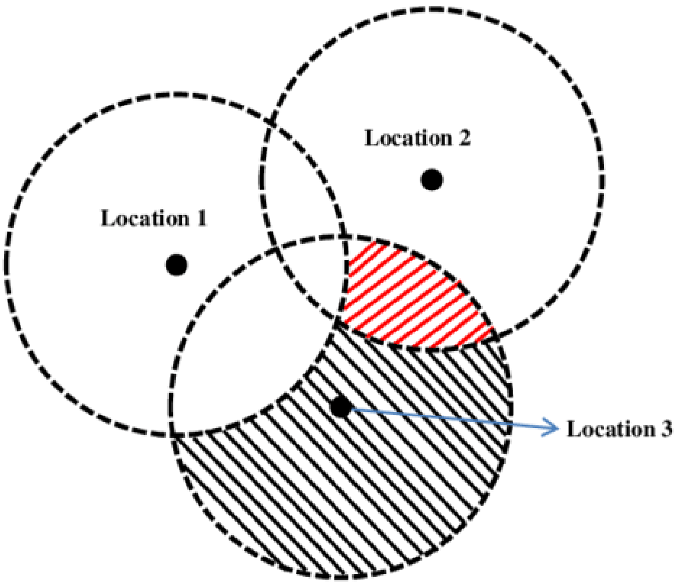


Greedy Algorithm for Problems with Submodularity

- Greedy algorithm leads to $1 - \frac{1}{e}$ approximation for submodular monotone function
  - For Maximum Coverage problem: Greedily pick the subset that covers most uncovered elements in each step
    - $U = \{1,2, \dots, 6\}$
    - $A = \{A_1, A_2, \dots, A_N\}$ is the set of subsets of U, i.e., $A_i \subset U$
      - $A_1 = \{1,3,5\}, A_2 = \{2,4,6\}, A_3 = \{1,6\}, A_4 = \{5,6\}$
    - $f: 2^N \to \mathbb{R}$
      - $f(S)$ where $S \subset A$ is the number of elements in U that is covered by any $A_i \in S$

Greedy Algorithm for Problems with Submodularity

- Example: Sensor Coverage Problem
  - Similar to maximum coverage problem



$$f(\{1,3\}) - f(\{1\}) \geq f(\{1,2,3\}) - f(\{1,2\})$$

black area + red area

black area

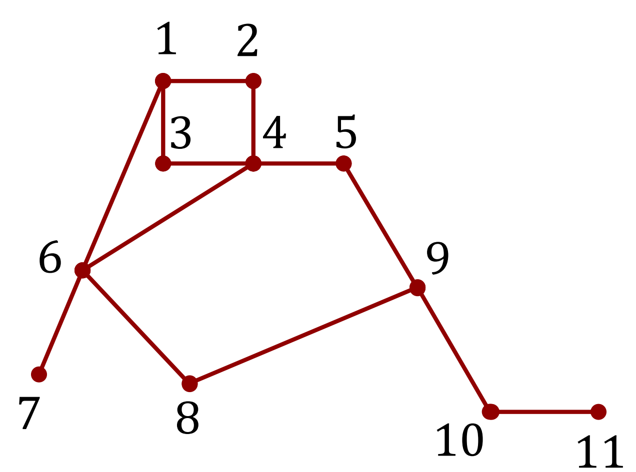
Greedy Algorithm for Influence Maximization

- Theorem: For both LTM and ICM, $\sigma(A_0)$ is a submodular function (Kempe, Kleinberg & Tardos, 2003)
- Also, it is easy to show that $\sigma(A_0)$ is monotone
- So greedy algorithm is a $1-\frac{1}{e}$ approximation for influence maximization problem

Greedy Algorithm for Influence Maximization
$A_0 \leftarrow \emptyset$ For $iter = 1..K$ Select $v = \underset{v' \in V \setminus A_0}{\text{argmax}}(\sigma(A_0 \cup \{v'\}) - \sigma(A_0))$ $A_0 \leftarrow A_0 \cup \{v\}$

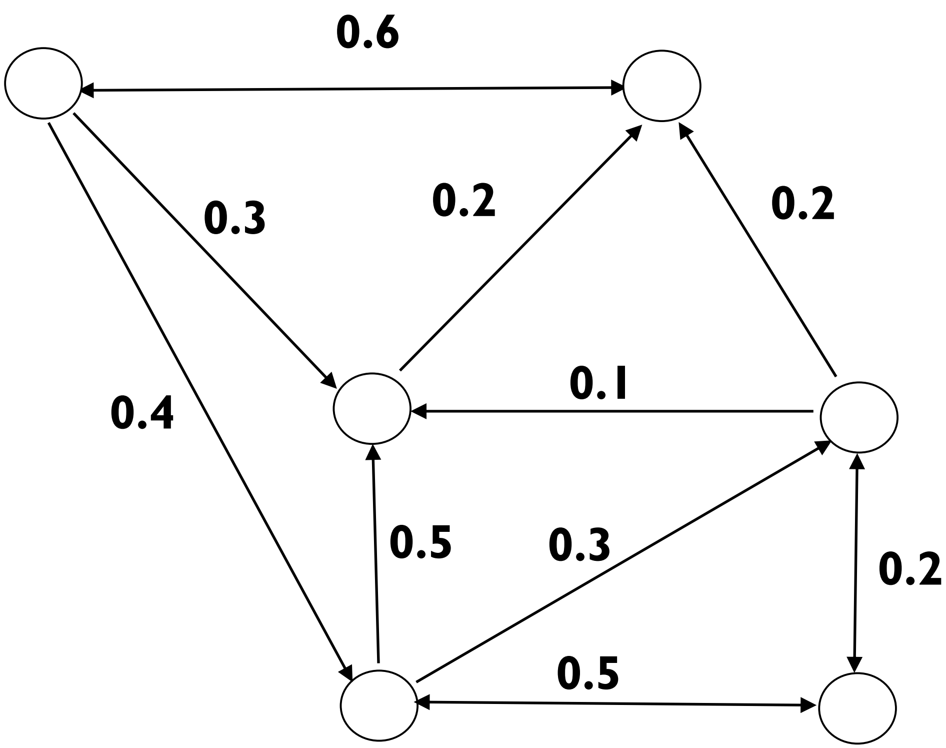
Greedy Algorithm for Influence Maximization

- Under ICM, if you can only activate one node to trigger the propagation process, which node should be selected to maximize influence with $p_{v,w} = 1$?
- If $p_{v,w} < 1$ and you can choose two nodes, which nodes will be chosen following the greedy algorithm?



Greedy Algorithm for Influence Maximization

- Under LTM, which 2 nodes will be chosen with the greedy algorithm?

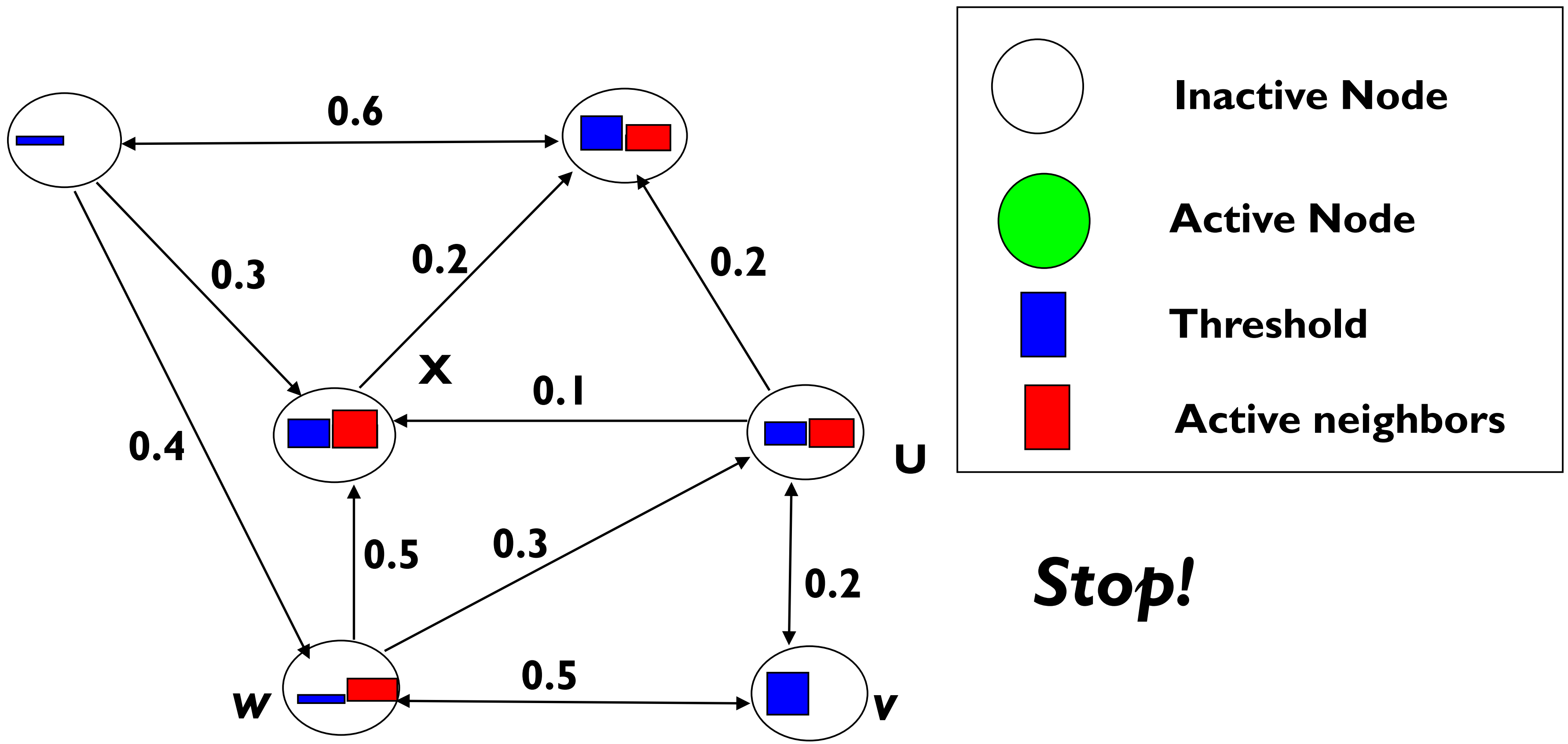


$\sigma(A_0)$ is the expected number of nodes being activated in the end

www.cs.cmu.edu/~xiaonanz/Maximizing-the-Spread-of-Influence.ppt

Greedy Algorithm for Influence Maximization

- Which 2 nodes will be chosen?



Stop!

www.cs.cmu.edu/~xiaonanz/Maximizing-the-Spread-of-Influence.ppt

Outline

- Influence Maximization Problem
- Discussion
- Monte Carlo Tree Search
- Case Study: HIV Prevention Among Homeless Youth

Discussion

- What are the possible applications of the models and algorithms introduced today?
- How to extend the current problem definition of influence maximization problem to reflect some characteristics of real-world problems?
- What are other significant problems that need to be solved based on the propagation model?

Outline

- Influence Maximization Problem
- Discussion
- Monte Carlo Tree Search
- Case Study: HIV Prevention Among Homeless Youth

Monte Carlo Tree Search

- General framework to make online decision in sequential decision making problems
  - E.g., online planning in MDPs, to determine game plays in Go, chess, video games etc
- Not only applicable to MDPs, but also other domains that cannot be modeled as MDPs
  - The idea of Q value can still be used

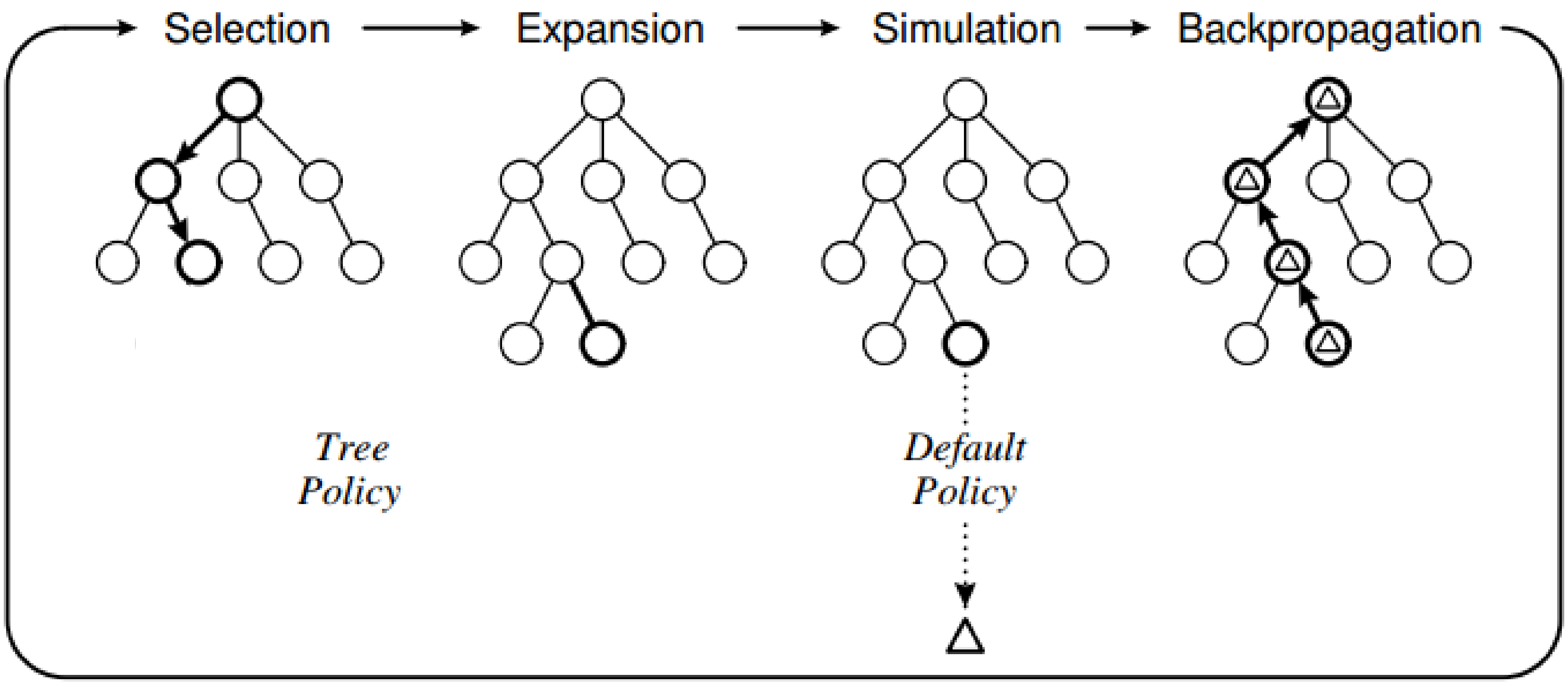
Monte Carlo Tree Search

- MCTS for single player setting: online planning in an unknown environment
- You are now in some state, need to choose an action, but you know nothing about the environment
- Helper: a simulator tells you your available actions, and reward after you take the action

Green player controlled by you
Actions={up, down, nothing}

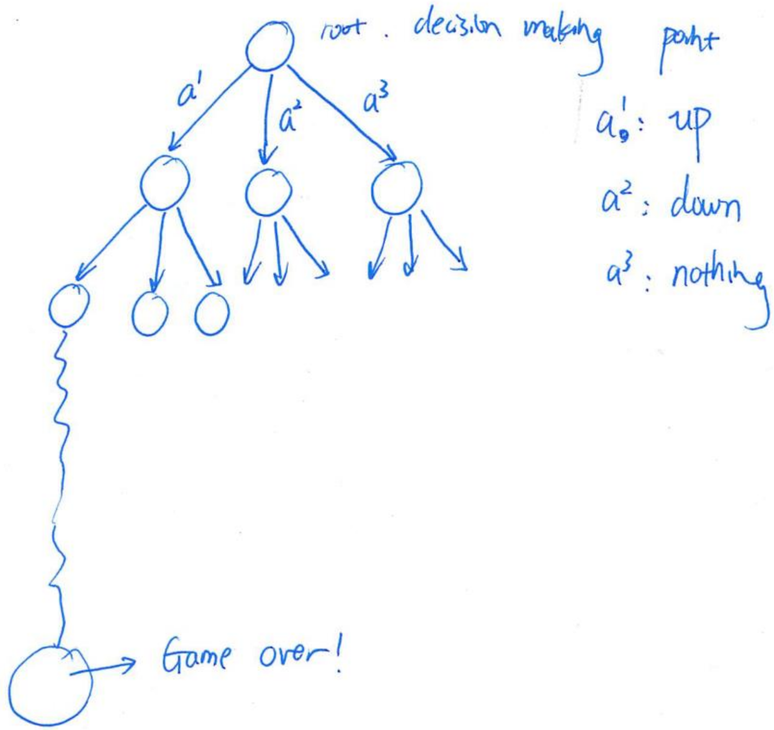
Monte Carlo Tree Search

- Build a search tree node by node
  - Node: state; Edge: available actions
- Repeat: Select→Expand→Simulate→Backpropagate



Monte Carlo Tree Search

- Build a search tree node by node
  - Node: state; Edge: available actions
- Repeat: Select→Expand→Simulate→Backpropagate

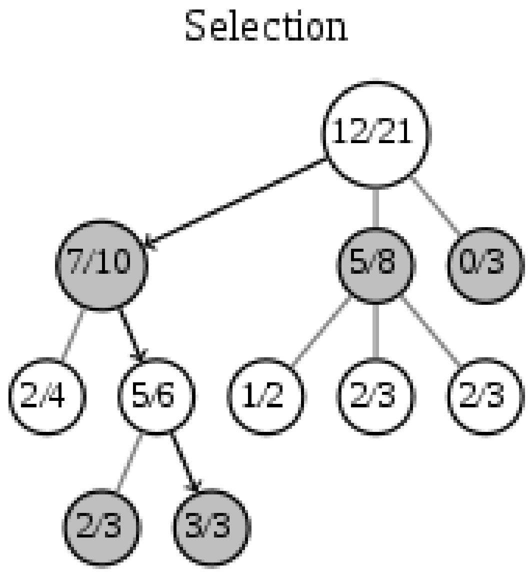


Monte Carlo Tree Search

- Simplest MCTS
  - In each iteration
    - Select: Choose the branch with the highest value
    - Expand: Add one node by randomly selecting an action
    - Simulate: Uniform random rollout
    - Backpropagate: update mean return (average accumulated reward) along the path
  - Output: action correspond to branch with highest value at the root node after K iterations

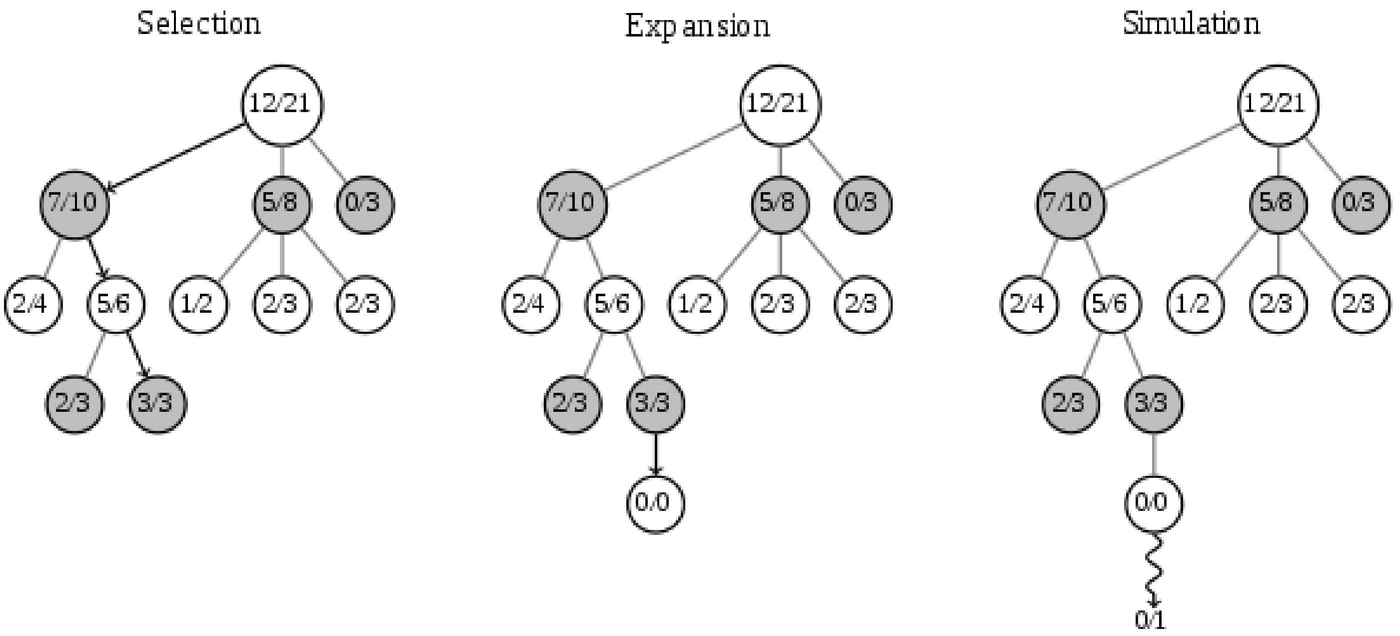
Monte Carlo Tree Search Example

Q: Assume the numbers in the nodes represent the mean return, which leaf node will be expanded when using the simplest MCTS?



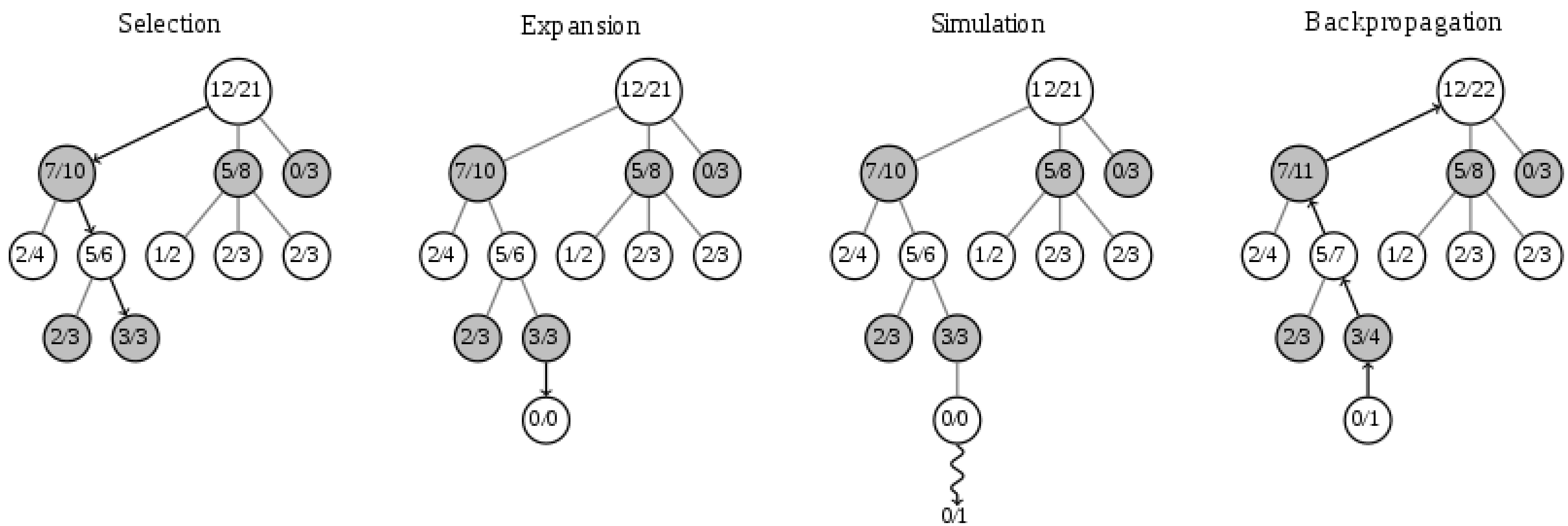
Monte Carlo Tree Search Example

Q: Assume the following tree is built for the Atari game, the numbers in the nodes represent the total number of times we win / the total number of times we visit the state, which nodes will be updated in the backpropagation step?



Monte Carlo Tree Search Example

Q: Assume the following tree is built for the Atari game, the numbers in the nodes represent the total number of times we win / the total number of times we visit the state, which nodes will be updated in the backpropagation step?



Recap: Upper Confidence Bound in MAB

- UCB1 Algorithm:
  - Always choose the arm with the highest upper confidence bound defined as $\mu_{UB}^k = \widehat{\mu_k} + \sqrt{\frac{2 \ln t}{N(k)}}$
  - Intuition: If μ_{UB}^k is large, either arm k is a good arm or $N(k)$ is small (not enough data is gathered)
  - General principle: optimism in the face of uncertainty

Monte Carlo Tree Search

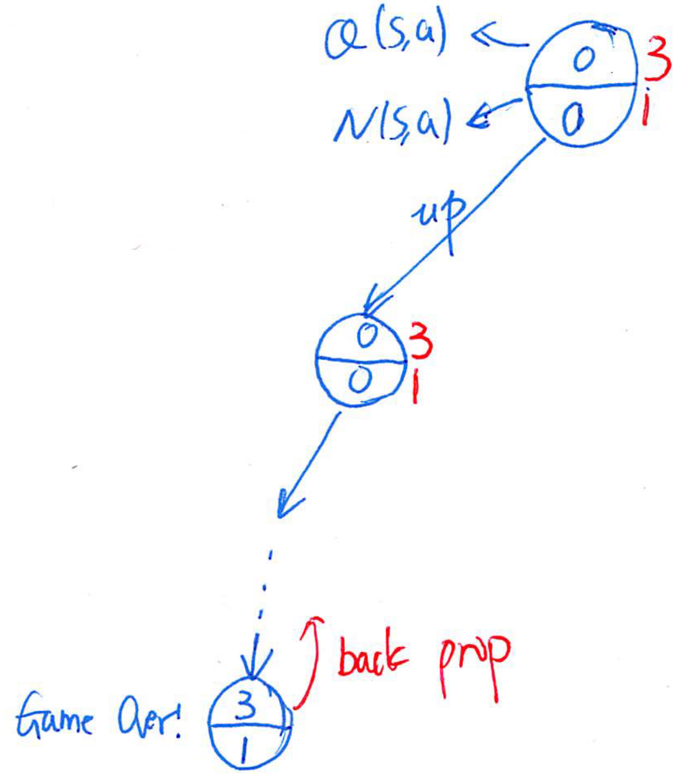
- Upper Confidence Bounds for Trees (UCT)
  - For each node, keep track of estimated action value and visit count: $Q(s,a)$ and $N(s,a)$
  - Select: Balance exploration vs exploitation:
    - If some actions never been chosen, randomly choose among them
    - Choose branch with highest Upper Confidence Bounds (UCB):

$$Q^{\oplus}(s,a) = Q(s,a) + c\sqrt{\frac{\ln N(s)}{N(s,a)}}$$

Monte Carlo Tree Search

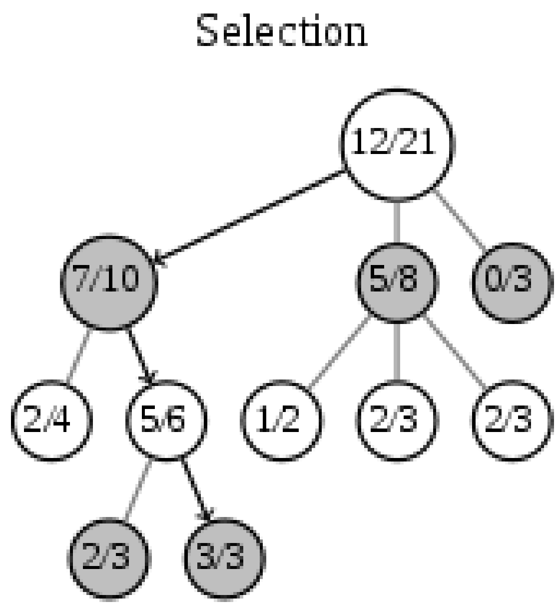
- Upper Confidence Bounds for Trees (UCT)
  - For each node, keep track of estimated action value and visit count: $Q(s,a)$ and $N(s,a)$
  - Select: Balance exploration vs exploitation:
    - If some actions never been chosen, randomly choose among them
    - Choose branch with highest Upper Confidence Bounds (UCB):

$$Q^{\oplus}(s,a) = Q(s,a) + c\sqrt{\frac{\ln N(s)}{N(s,a)}}$$



Monte Carlo Tree Search Example

Q: Assume the numbers in the nodes represent the Q(s,a) / N(s,a), which leaf node will be expanded when using UCT with $c = 10000$?



$$Q^{\oplus}(s,a) = Q(s,a) + c\sqrt{\frac{\ln N(s)}{N(s,a)}}$$

Monte Carlo Tree Search

- More advanced MCTS
  - Other advanced options:
    - Simulate: Terminate after T_0 steps and estimate the reward
    - Expand: Add more nodes to the tree
    - Output: Optimal action at root node, as well as Q and N in the subtree corresponds to the optimal action
    - Initialize search tree with domain knowledge

Outline

- Influence Maximization Problem
- Discussion
- Monte Carlo Tree Search
- Case Study: HIV Prevention Among Homeless Youth

Overview

HIV Prevention Among Homeless Youth

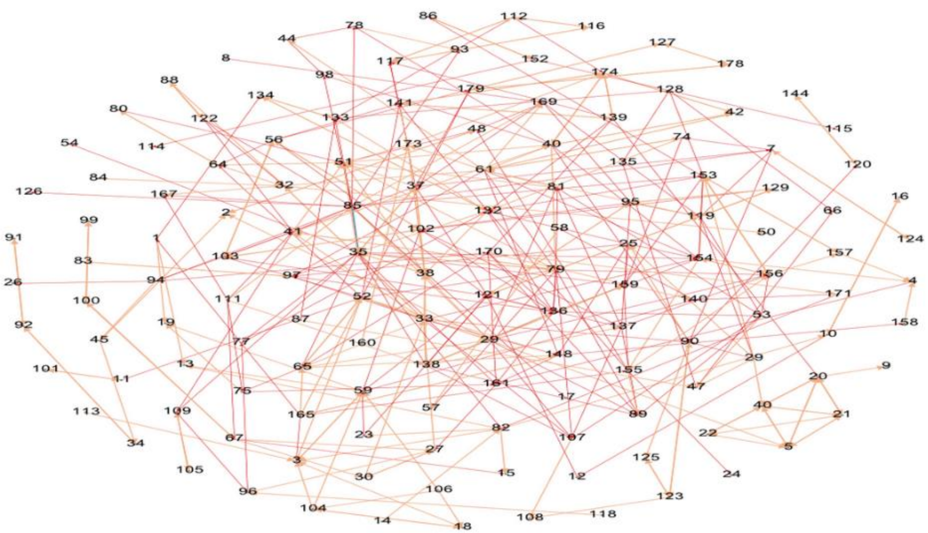
HIV Prevention Among Homeless Youth

- Organize interventions among homeless youth.
  - Try to raise awareness about HIV prevention practices.
  - Urge them to adopt safer behaviors.
  - Encourage them to spread the message among their social circles.

S.P.Y
safe place for youth

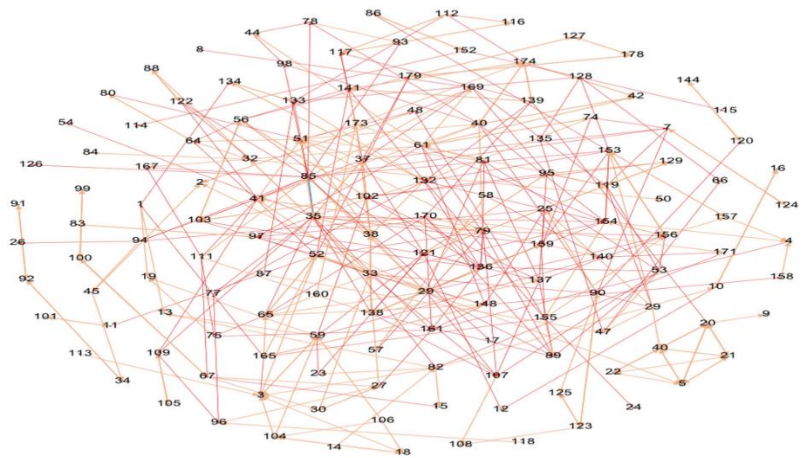
Resource Constraints with Homeless Shelters

- Homeless shelters operate under resource constraints
  - Cannot intervene on every homeless youth themselves.
  - Rely on word-of-mouth effects among homeless youth.
  - Maximize number of youth who get informed about HIV.



Influence Maximization Problem

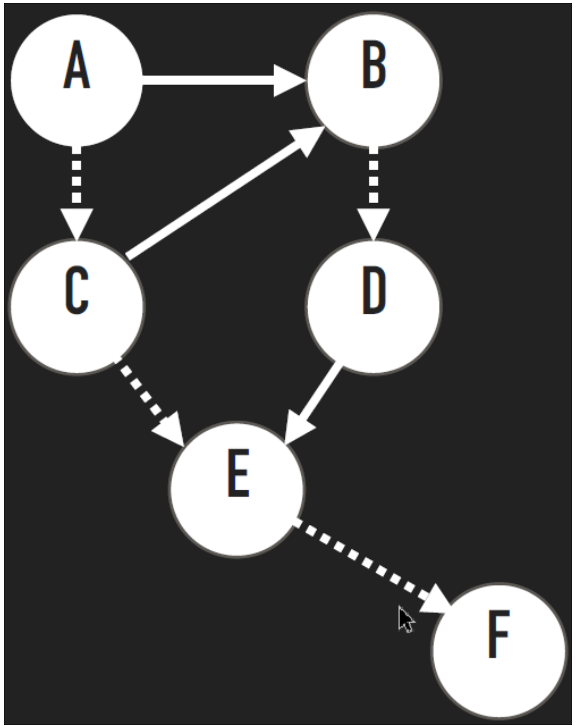
- Given social network G and influence model I, choose K nodes to maximize expected influence spread



- Existing algorithms
  - With theoretical guarantees
    - Greedy [Kempe et al. 2003]
    - CELF [Leskovec et al. 2007]
    - TIM [Borgs et al. 2012, Tang et al. 2014]
  - Efficient Heuristics
    - IRIE [Jung et al. 2011]
    - Sketch based heuristic [Cohen et al. 2015]

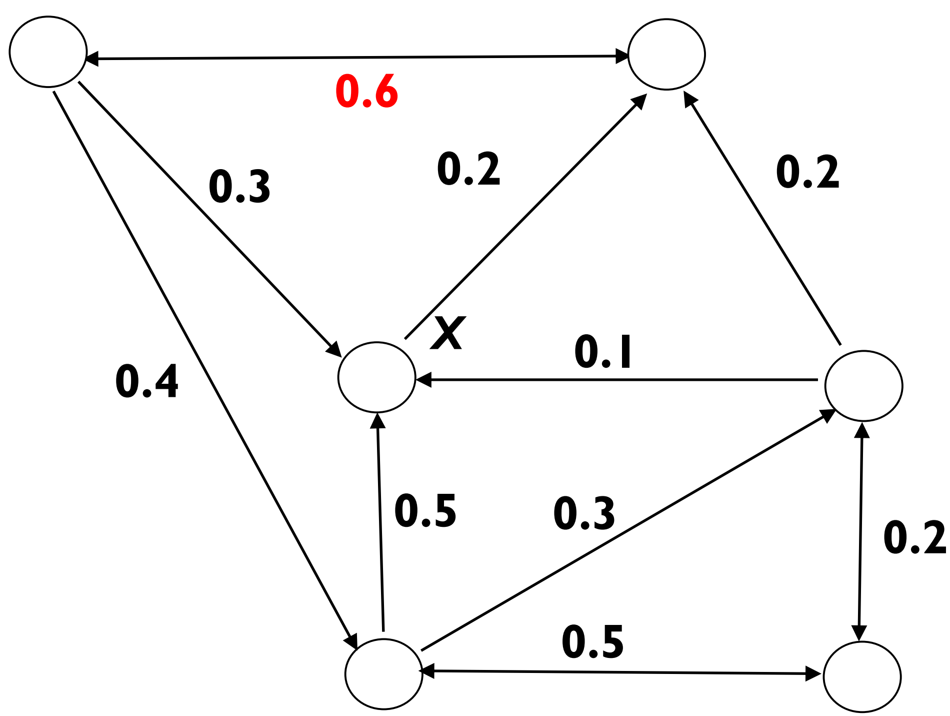
Crucial Practical Challenge: Uncertainty

- Uncertainty in problem input
  - Uncertainty in social network structure



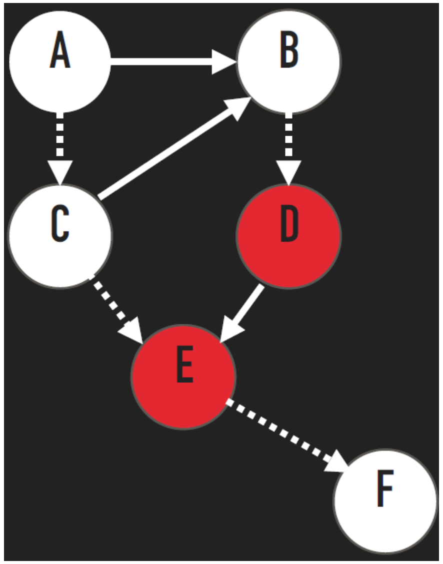
Crucial Practical Challenge: Uncertainty

- Uncertainty in problem input
  - Uncertainty in social network structure
  - Uncertainty in specification of influence model



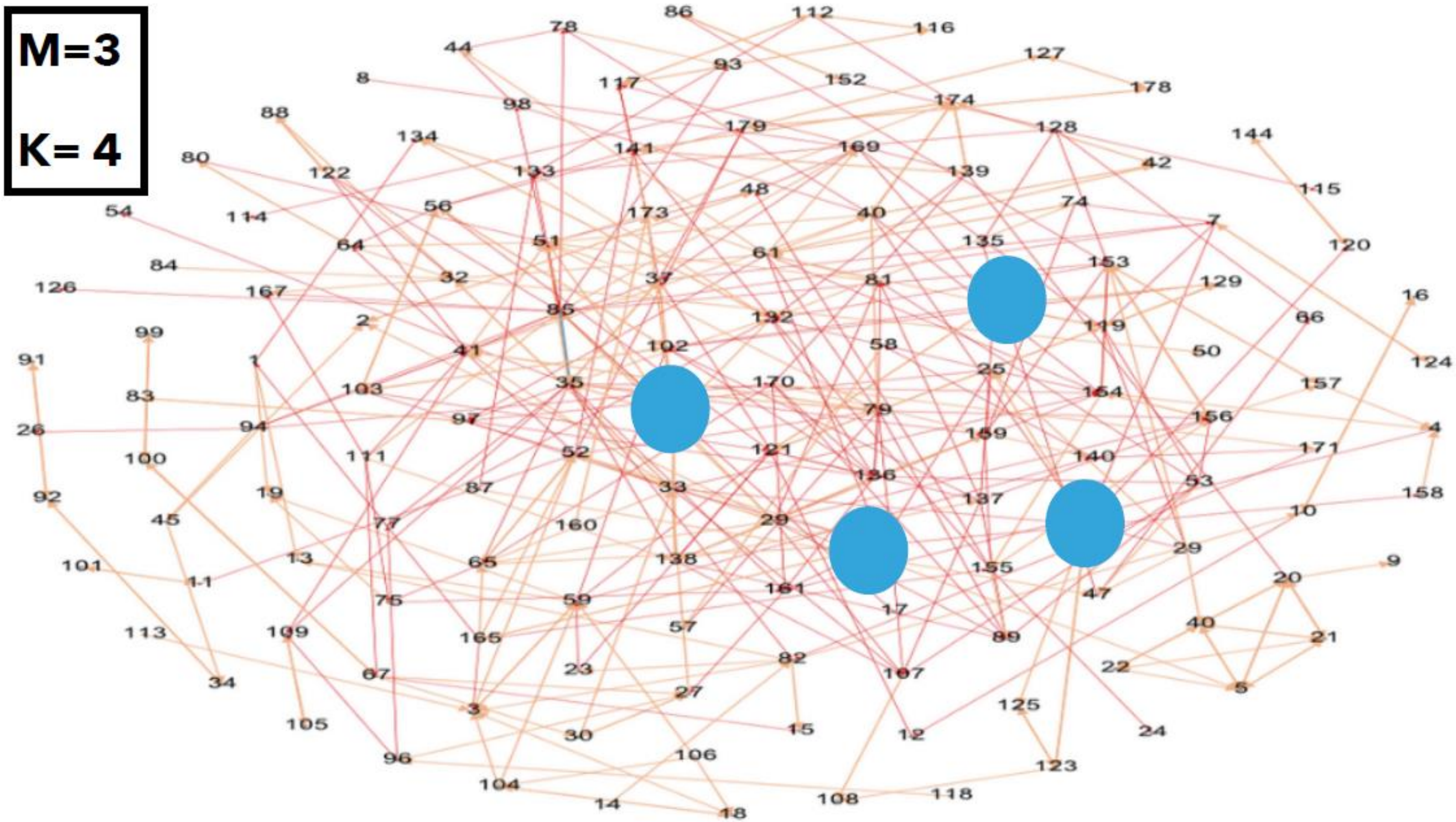
Crucial Practical Challenge: Uncertainty

- Uncertainty in problem input
  - Uncertainty in social network structure
  - Uncertainty in specification of influence model
- Uncertainty during problem execution
  - Uncertainty about state of influence of nodes



Especially important if need to choose seed nodes sequentially

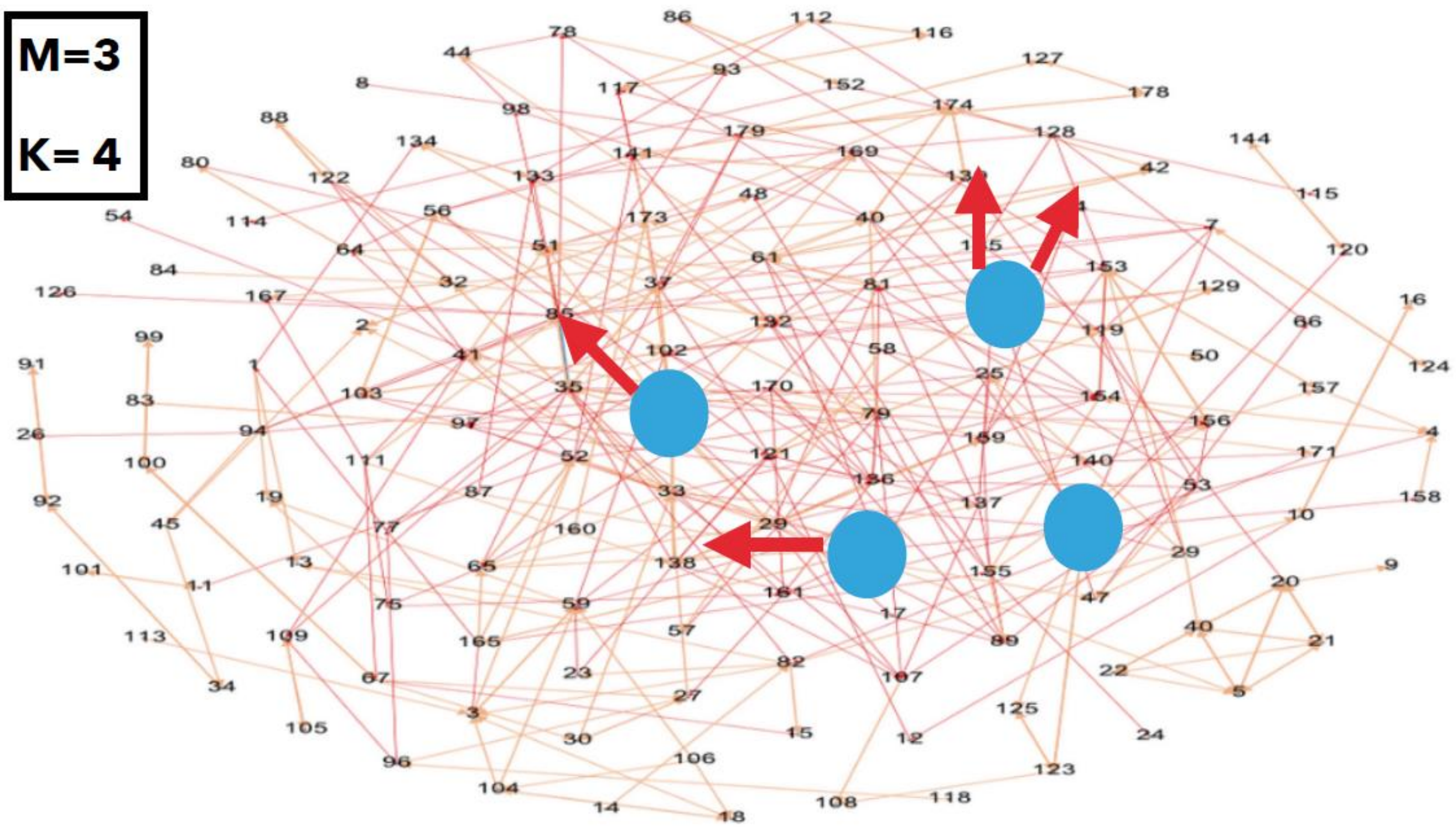
Crucial Practical Challenge: Multi-Stage Selection



Crucial Practical Challenge: Multi-Stage Selection

M=3

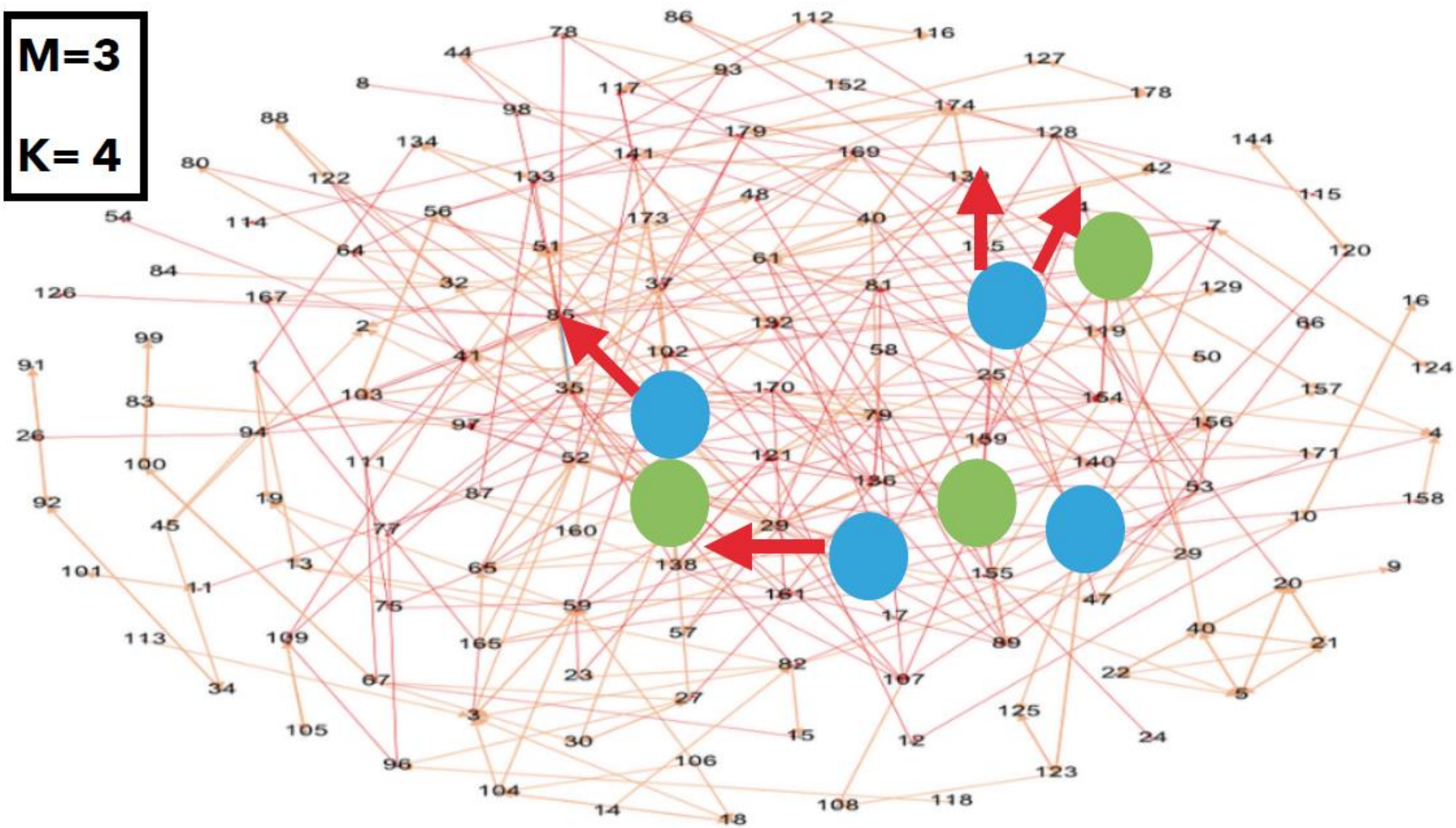
K= 4



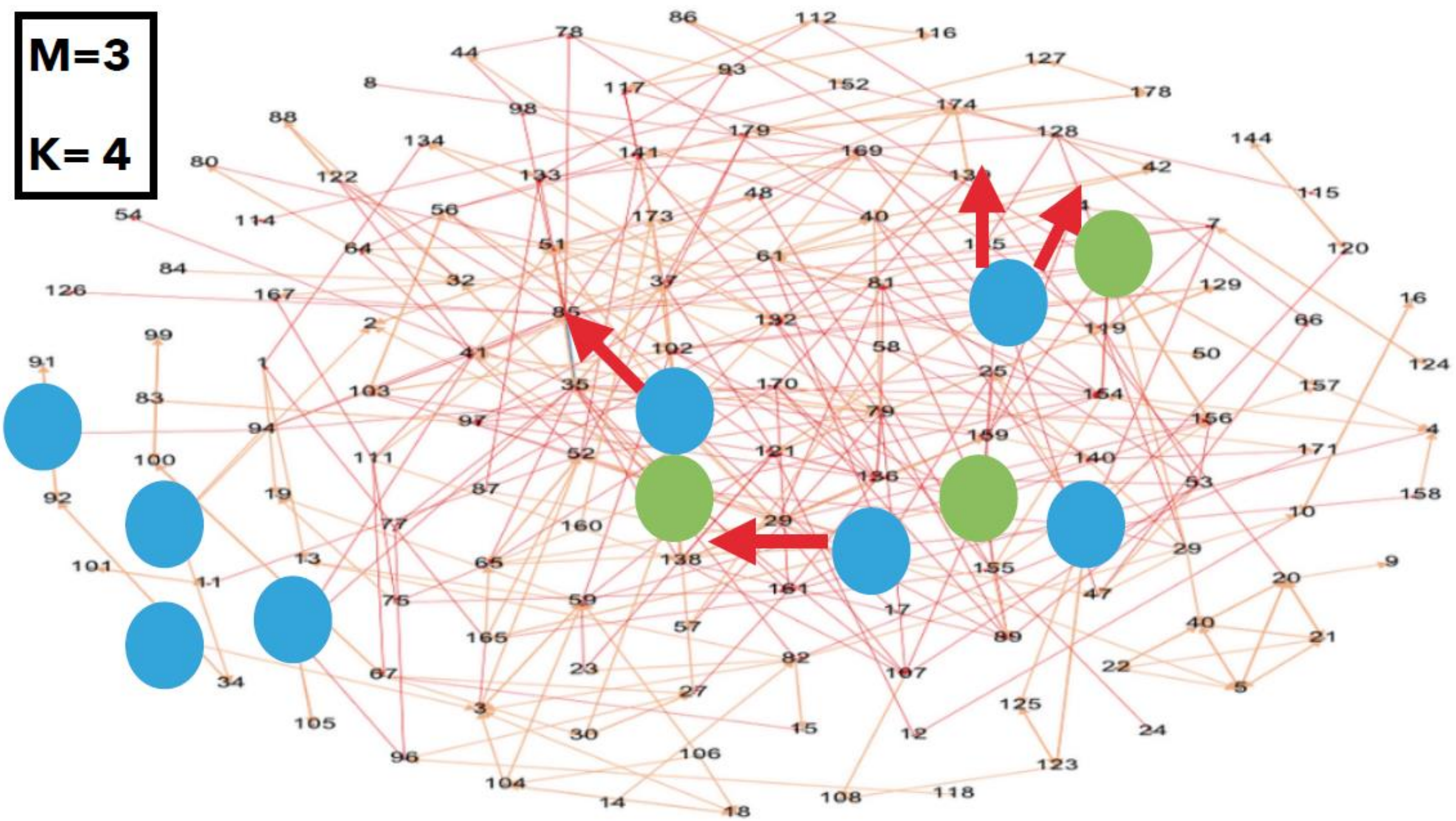
Crucial Practical Challenge: Multi-Stage Selection

M=3

K= 4

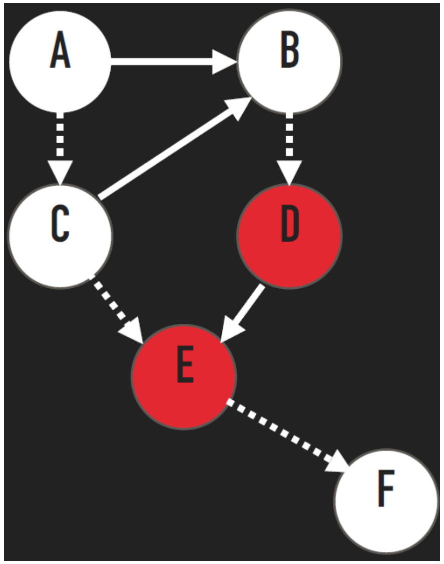


Crucial Practical Challenge: Multi-Stage Selection



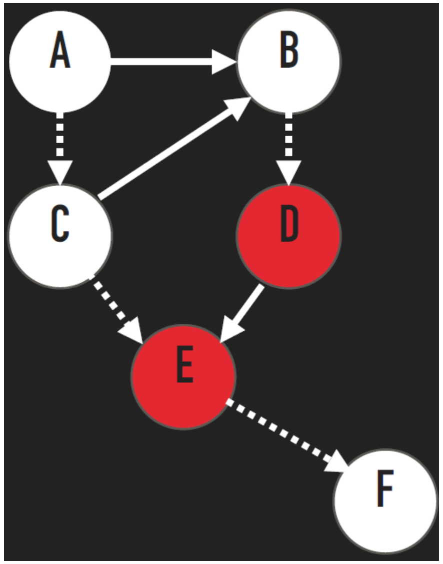
Problem Definition

- Dynamic Influence Maximization Under Uncertainty (DIME)
  - Input: A social network $G = (V, E)$
    - Consider uncertainty in social network structure
    - Each edge is associated with an existence probability $u(e)$, or even a distribution
  - Influence propagation model
    - Similar to ICM but with nodes get multiple chances to influence their un-influenced neighbors
    - Each edge is associated with propagation probability $p(e)$



Problem Definition

- Task
  - Choose a sequential adaptive plan (policy)
    - Picking a subset of nodes for M rounds
      - M interventions organized by shelter
    - Size of each subset is K
      - Maximum capacity of shelter
    - Maximize expected influence spread
      - Assuming the influence spreads for L time steps in each round
- Uncertainty about edge and state
  - Does not observe exact influence state
  - Can observe existence of edges adjacent to seed nodes



Theoretical Results

- THEOREM 1 (AAMAS'16A)
  - DIME Problem is NP-Hard
  - i.e., very hard to solve exactly
- THEOREM 2 (AAMAS'16A)
  - For any $\epsilon > 0$, it is impossible for an algorithm to guarantee a $n^{-1+\epsilon}$ approximation to the full information optimal solution.
  - i.e., very hard to approximate in polynomial time
- THEOREM 3 (AAMAS'16A)
  - The influence function of DIME is not adaptive sub-modular.
  - Thus, greedy alg. may not perform well empirically

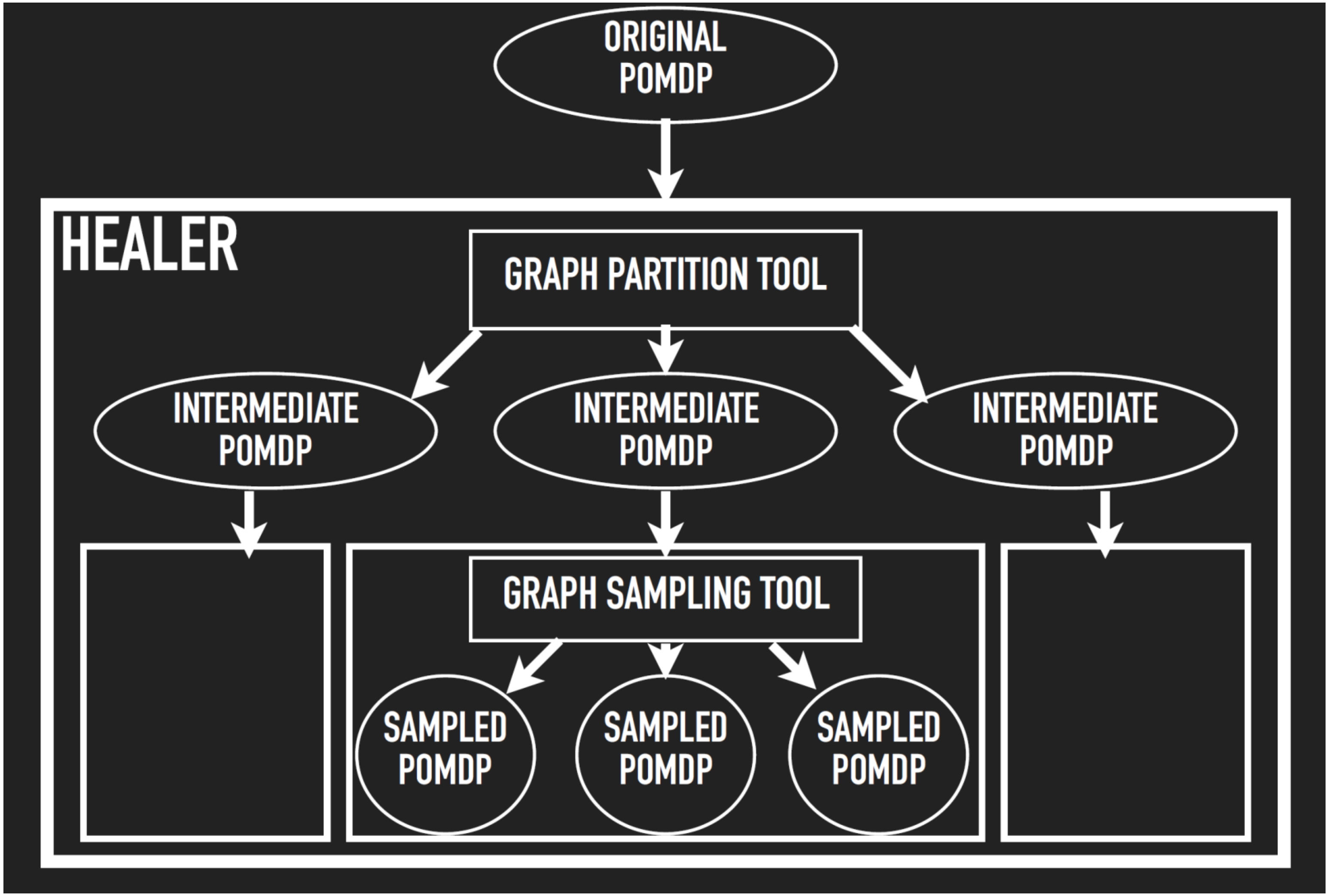
Discussion

- Can we use greedy algorithm for influence maximization to solve DIME? What are the limitations of doing so?

- Formulate DIME as a POMDP. What are the states, actions, observations?

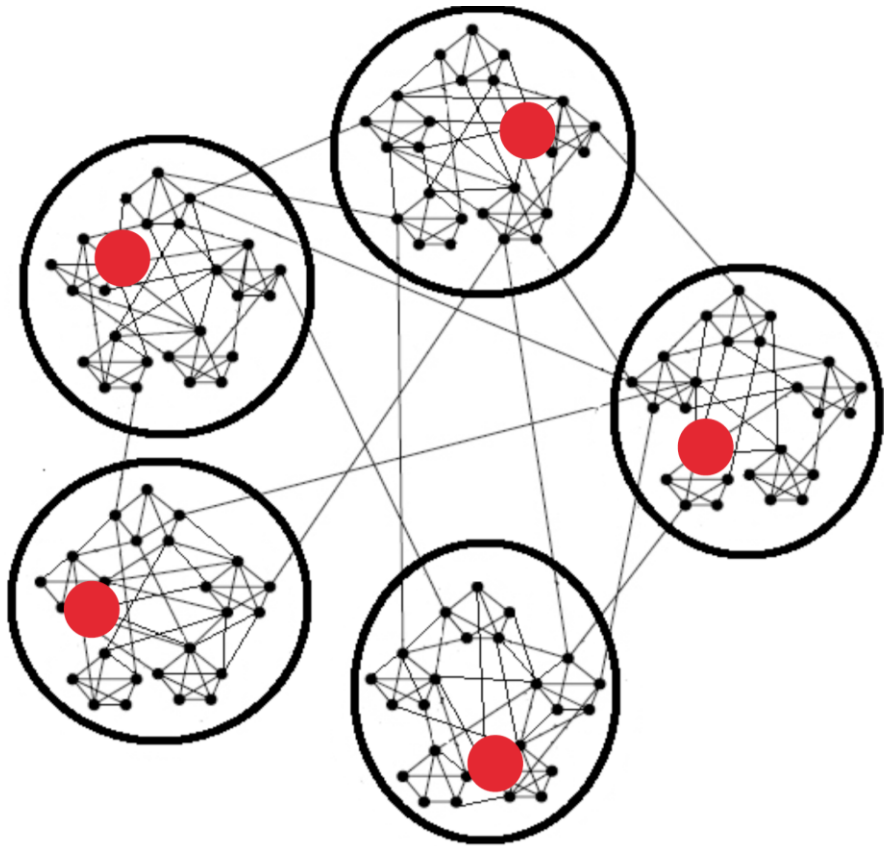
- Can we use Monte Carlo Tree Search to solve DIME? How? What are the potential issues?

HEALER Algorithm Overview



Key Idea 1: Graph Partitioning

- Choose a small subset of nodes in each cluster of nodes

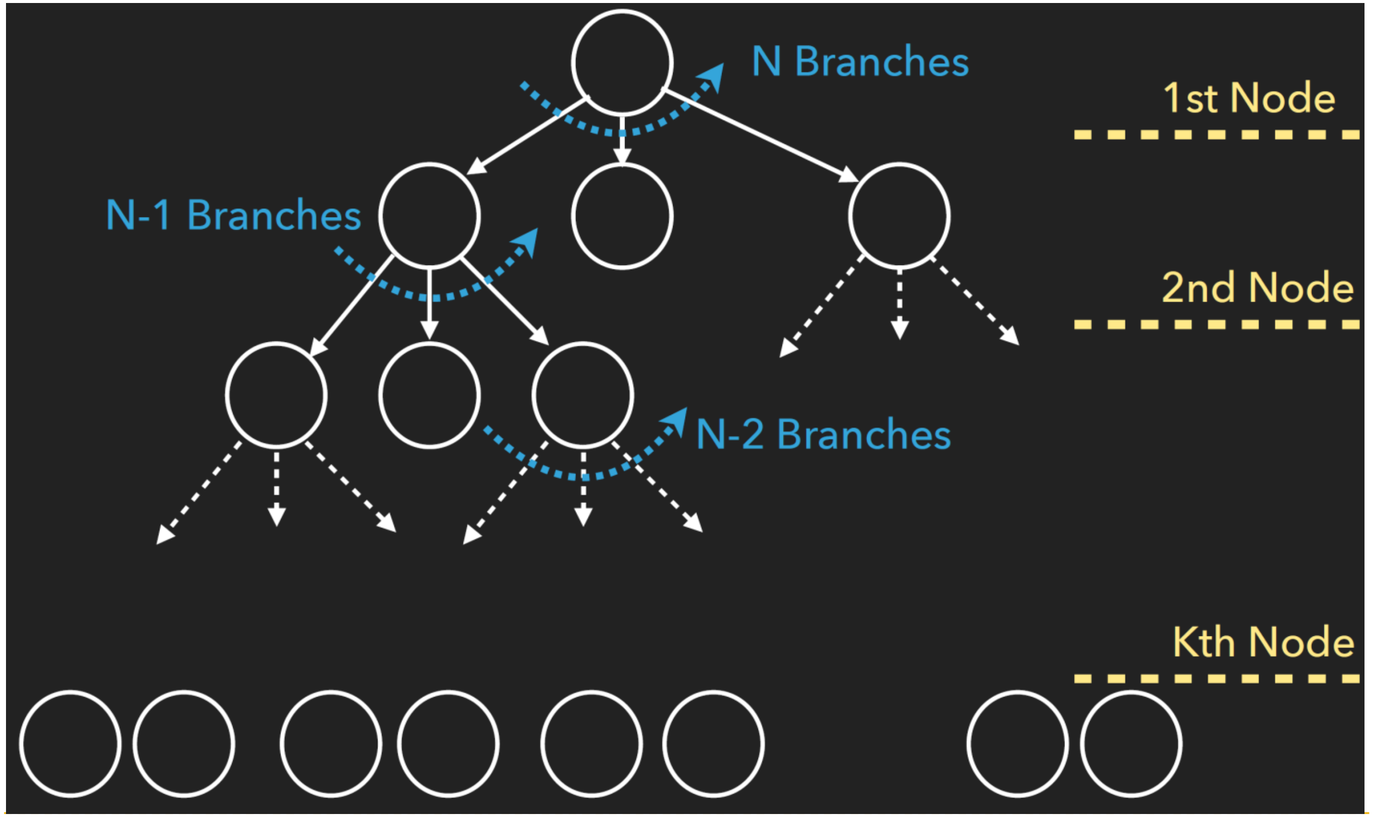


Key Idea 2: Graph Sampling

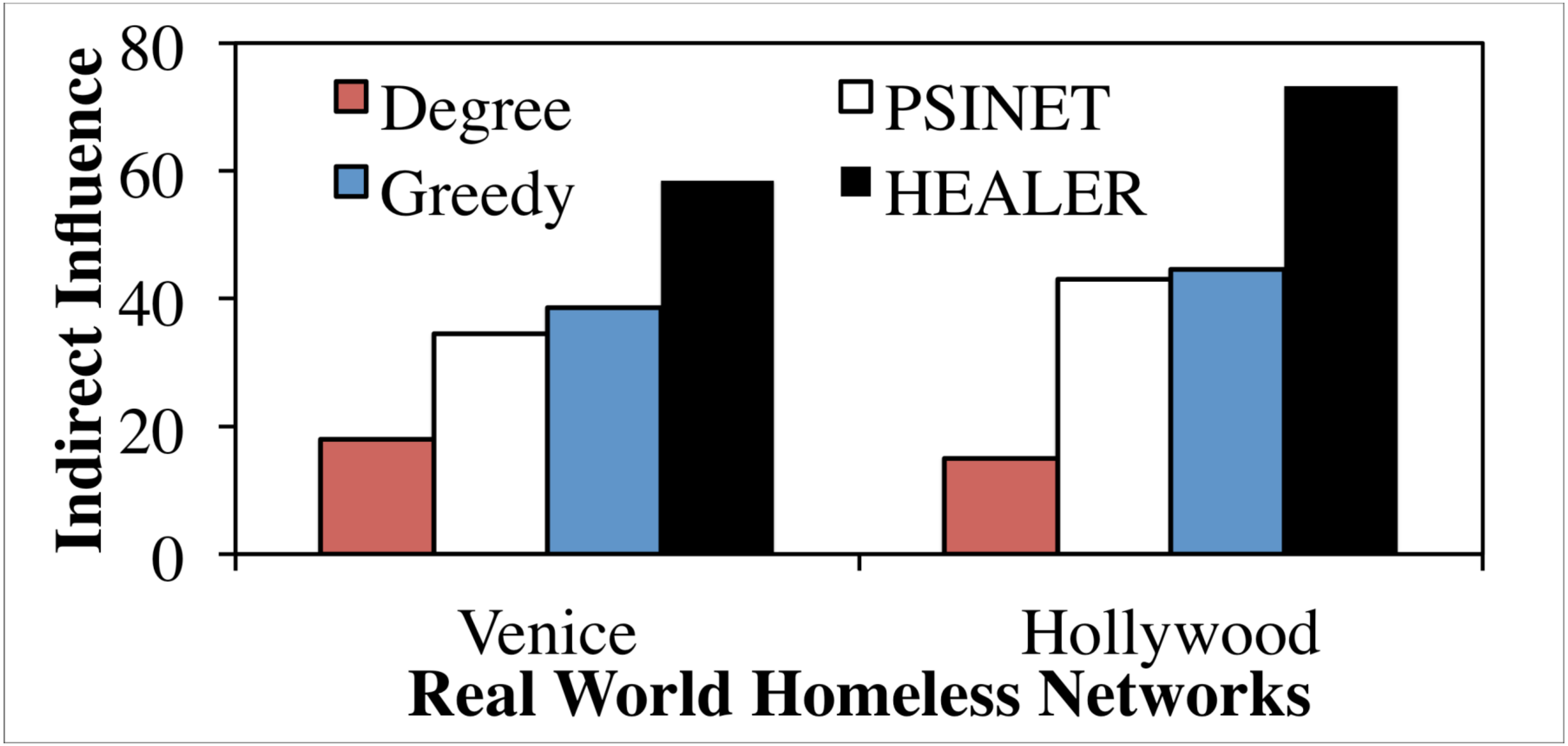
- We don’t know the actual graph structure because of the uncertainty of edge existence
- Then draw a few samples!
- Choose subset of nodes that can be good in expectation for the sampled graphs

Key Idea 3: Monte-Carlo Tree Search + Multi-Armed bandit

- In each step, choose one node
- Maintain an MAB at each search tree node



Evaluation


Indirect Influence
80
60
40
20
0
Degree
Greedy
PSINET
HEALER
Venice
Hollywood
Real World Homeless Networks

Evaluation

Poll 2

- Why greedy algorithm for influence maximization (referred to as the IM problem) cannot be directly applied to the problem of spreading HIV-related information among homeless youth (referred to as IS problem)?
  - A: Greedy algorithm has a good approximation bound but does not perform well empirically
  - B: In this IS problem, we need to select seed nodes in several batches sequentially, which is different from the IM problem
  - C: There are uncertainties in the IS problem
  - D: The greedy algorithm cannot scale
  - E: None of the above
  - F: I don't know

Acknowledgment

- The slides are prepared based on lecture slides of Leonid Zhukov and guest lecture of Amulya Yadav

References

- *Using Social Networks to Aid Homeless Shelters: Dynamic Influence Maximization under Uncertainty (Links to an external site.)*
- *Influence Maximization in the Field: The Arduous Journey From Emerging to Deployed Application (Links to an external site.)*
- *Uncharted but not Uninfluenced: Influence Maximization with an Uncertain Network*

References

- Maximizing the spread of influence through a social network (Links to an external site.)
- *Submodular Functions: Extensions, Distributions, and Algorithms. A Survey (Links to an external site.)*
- *Information and Influence Propagation in Social Networks*

Backup Slides

Propagation Process

- Discuss: when would you adopt a recommendation from your friends?

- How to model a social network and influence in the network?

Submodular Functions

- Submodular Functions
  - Alternative definition: f is submodular iff $\forall S, T$

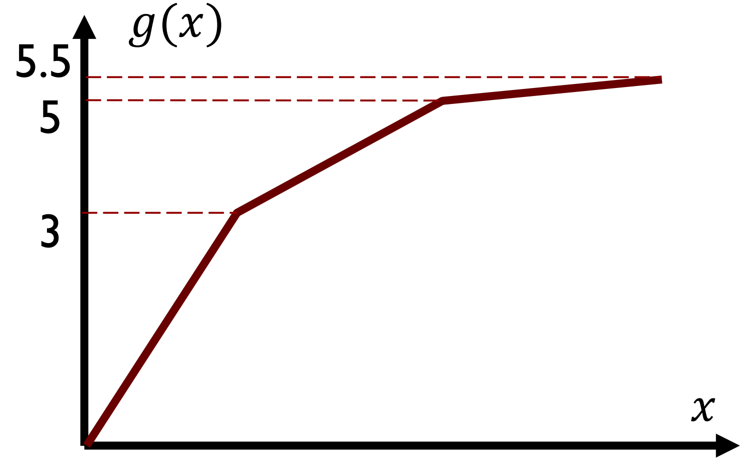
$$f(S \cup T) + f(S \cap T) \leq f(S) + f(T)$$

  - After-class exercise: Prove that these two definitions are equivalent

Submodular Functions

- Example: Team of defensive resources (ground vehicle and unman aerial vehicle)
  - $A = \{GV1, GV2, UAV1, UAV2, UAV3\}$
  - $f(S)$ where $S \subset A$ is described by

$$f(S) = g(\#GV) + g(\#UAV)$$



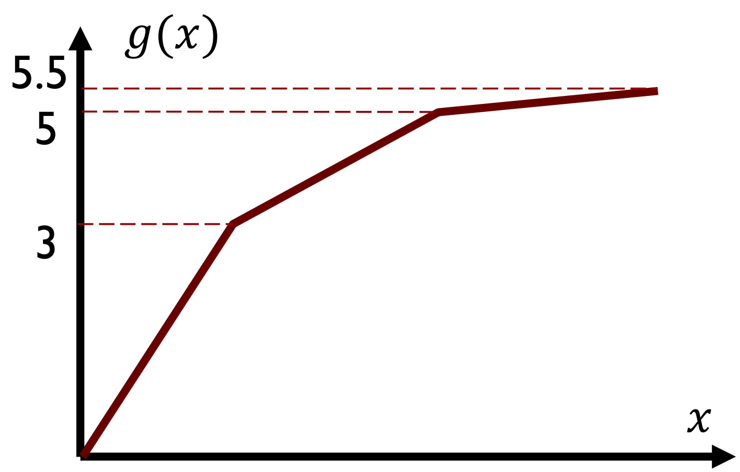
Submodular Functions

- Example: Maximum Coverage problem
  - $U = \{1,2,\dots,m\}$ is the set of elements
    - E.g., $m = 6$
  - $A = \{A_1, A_2, \dots, A_N\}$ is the set of subsets of U, i.e., $A_i \subset U$
    - For example, $A_1 = \{1,3,5\}, A_2 = \{2,4,6\}, A_3 = \{1,6\}, A_4 = \{5,6\}$
  - $f: 2^N \rightarrow \mathbb{R}$
    - $f(S)$ where $S \subset A$ is the number of elements in U that is covered by any $A_i \in S$

Greedy Algorithm for Problems with Submodularity

- Example: Team of defensive resources (ground vehicle and unman aerial vehicle)
  - $A = \{GV1, GV2, UAV1, UAV2, UAV3\}$
  - $f(S)$ where $S \subset A$ is described by

$$f(S) = g(\#GV) + g(\#UAV)$$



Poll

- Let θ_0 = common threshold, N_0 = common number of neighbors. $b_{vw} = \frac{1}{N_0}$. Consider the following three scenarios
  - S1: $\theta_0 = a, N_0 = c$
  - S2: $\theta_0 = a + 0.1, N_0 = c$
  - S3: $\theta_0 = a, N_0 = c + 1$
- When $c > 1$, what is ordering of the probability of getting global cascade following the LTM model under these three scenarios?
  - A: S1>=S2>=S3
  - B: S3>=S2>=S1
  - C: S2>=S1, S3>=S1, relationship between S2, S3 is unknown
  - D: None of the above
  - E: I don't know