Maximizing the Spread of Influence through a Social Network

Kempe, Kleinberg, and Tardos (2003)

Paper Overview

- Introduction
- History
- Diffusion Models
- Paper Contributions
- Approach
  - Diffusion Models (+ Assumptions)
  - Example Network
  - Efficient Approximation
- Experimental Results
- Generalization
- Non progressive processes
- Marketing Strategies

Introduction

Problem: Need to try to convince a people to adopt a product/behavior

Constraints: Let's say you have a limited budget

Approach?

History

Traditional Methods:

- Mass Marketing
- Direct Marketing

Customers don’t necessarily make decisions in a vacuum

Diffusion of ideas

Social networks play a fundamental role as a medium for the spread of information

Dynamics of adoption is important

Prior research work in diffusion processes:

- “viral marketing” effects in the success of new products
- adoption of various strategies in game theoretic settings
- cascading failures in power systems.

Diffusion Models

Activation of Nodes (Users) in a directed Social Network Graph (G)

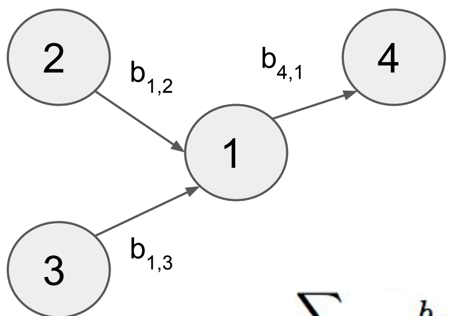
Models considered:

- Independent Cascades
- Linear Threshold

Initial Assumption: Progressive

Diffusion Models: Linear Threshold

(Granovetter and Schelling)



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

Activation condition

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

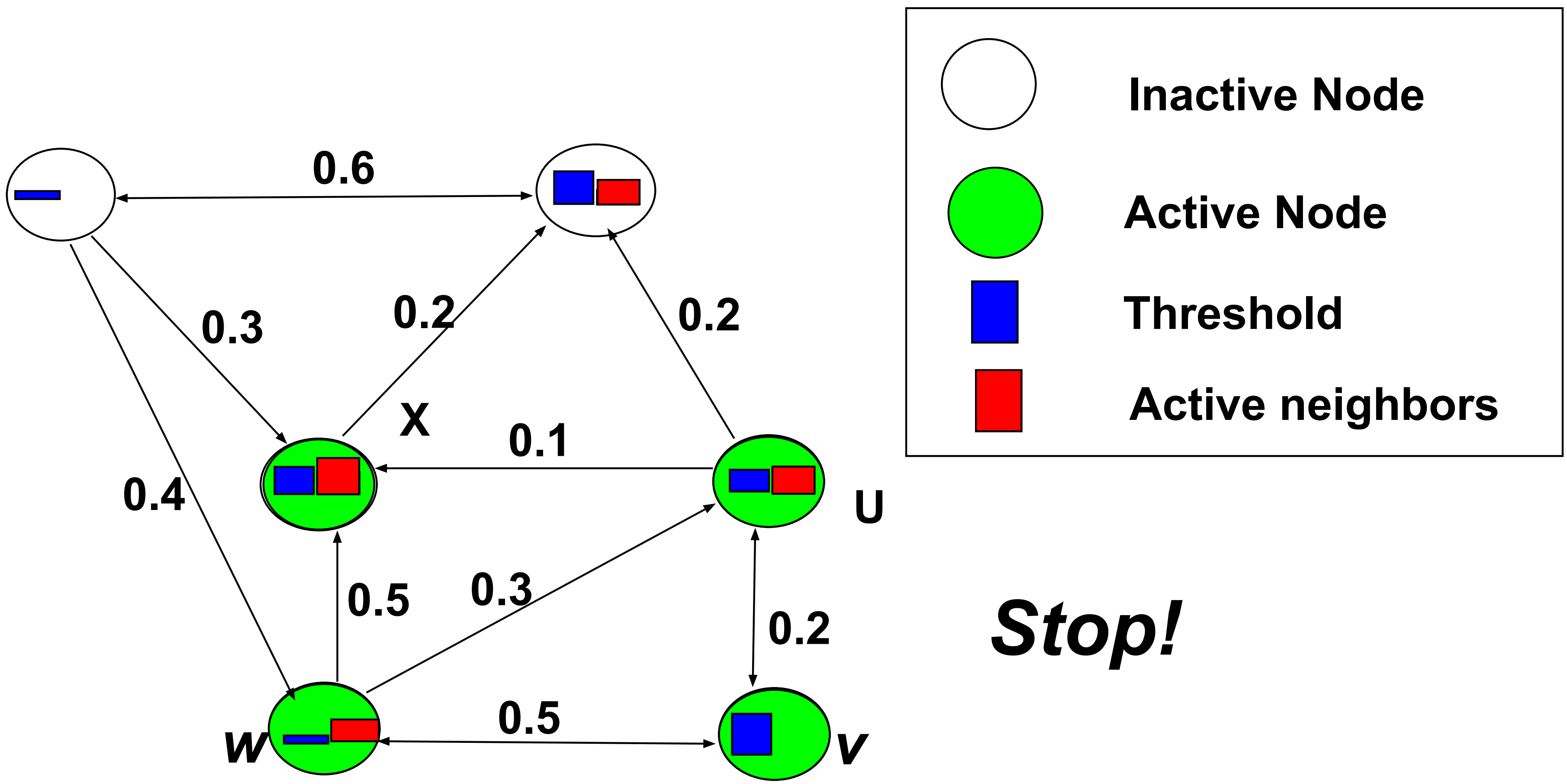
Model Parameters

- Edge weight $b_{v,w}$: Neighbor influence
- Threshold θ_v : tendency to adopt innovation

Diffusion proceeds in discrete time steps:

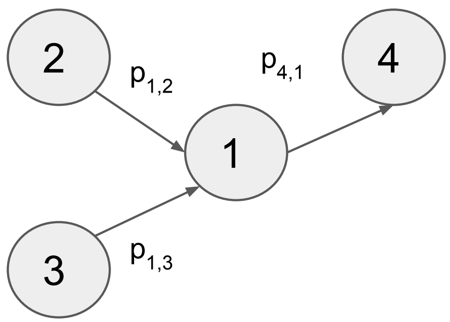
- Select initial set of active nodes A_0
- Nodes active at time step *t* are also active at time step *t+1*
- Check activation condition and update state A_t at every time step

Diffusion Models: Linear Threshold



Stop!

Diffusion Models: Independent Cascades



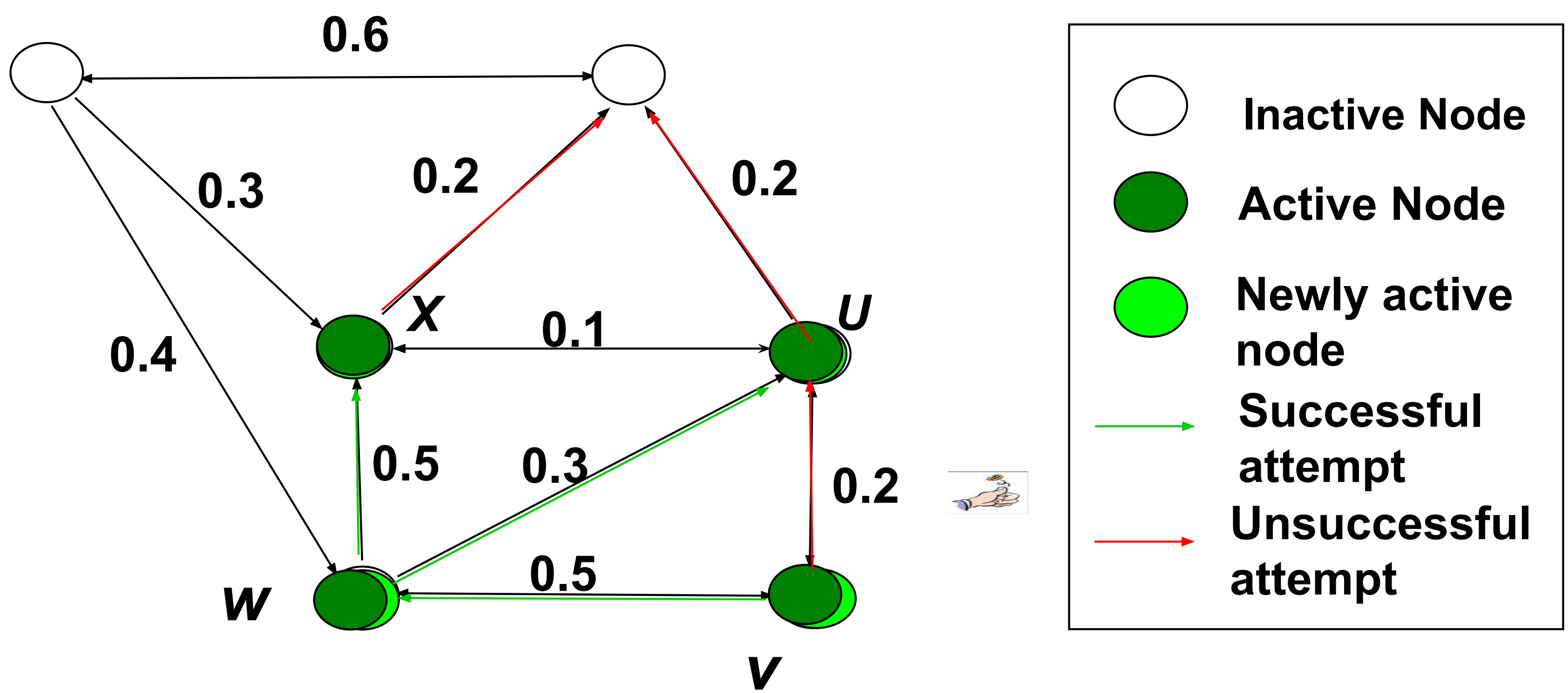
Model Parameters

- Edge weight $p_{v,w}$: Activation success probability

Diffusion proceeds in discrete time steps:

- Select initial set of active nodes A_0
- Nodes active at time step t are also active at time step $t+1$
- Active node v given a single chance to activate neighbor w
- Multiple neighbors' attempted sequenced in random order

Diffusion Models: Independent Cascades



Stop!

Influence Maximization Problem

Domingos-Richardson framework: Find a k-node subset A_0 of maximum influence $\sigma(A_0)$

Questions:

- How is influence defined?
- What is a submodular function?
- Why is this important?

THIS IS NP-HARD!

Paper Contributions

- First provable approximation guarantee to within a factor of $(1 - 1/e - \varepsilon)$
- Experimental validation: Comparison with popular heuristics
- Extensions
  - General Framework
  - Complex Marketing Actions
  - Non-Progressive

A general approach for reasoning about the performance guarantees of algorithms for these types of influence problems in social network

Approximation Guarantees

We want to find k-element set A_0 for which $\sigma(A_0)$ is maximized

THEOREM: For a non-negative, monotone submodular function σ, let S be a set of size k obtained by selecting elements one at a time, each time choosing an element that provides the largest marginal increase in the function value. Let S^* be a set that maximizes the value of f over all k-element sets.

Then $\sigma(S) \geq (1-1/e) \cdot \sigma(S^*)$; in other words, S provides a (1−1/e) approximation. (Nemhauser, Wolsey, Fisher, 78)

Influence function σ submodularity proof?

Approximation Guarantees

Assumptions:

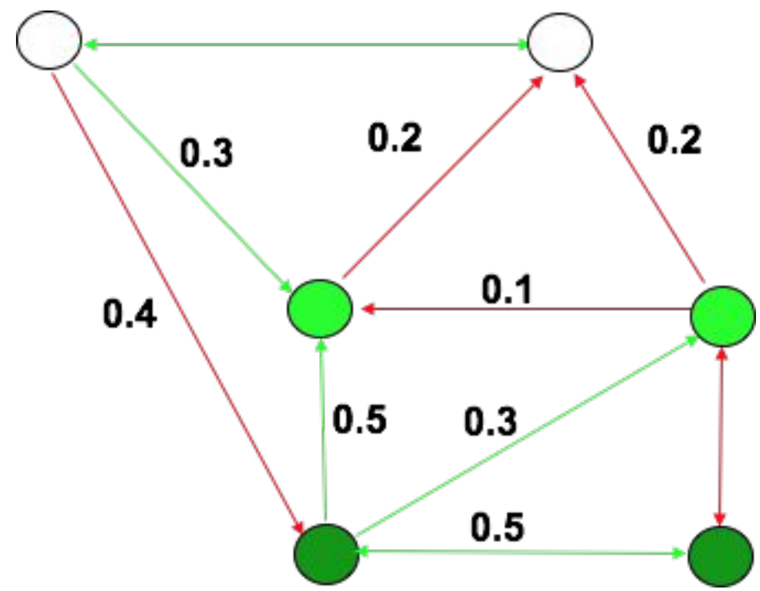
- σ is non-negative
- Submodular function: $\sigma(S \cup \{v\}) - \sigma(S) \geq \sigma(T \cup \{v\}) - \sigma(T)$
- Monotone: $\sigma(S \cup \{v\}) \geq \sigma(S)$

Submodularity: Independent Cascades

Pre-flip coins for each pair of nodes (v,w) using $p_{v,w}$

- X: State of edges (live/blocked)
- $R(v,X)$: Set of nodes with live-edge paths from x
- $\sigma_x(S \cup \{v\}) - \sigma_X(S) = R(v,X \cup_{u \in S} R(u, X)$
- $\sigma_X(S \cup \{v\}) - \sigma_X(S) \geq \sigma_X(T \cup \{v\}) - \sigma_X(T)$

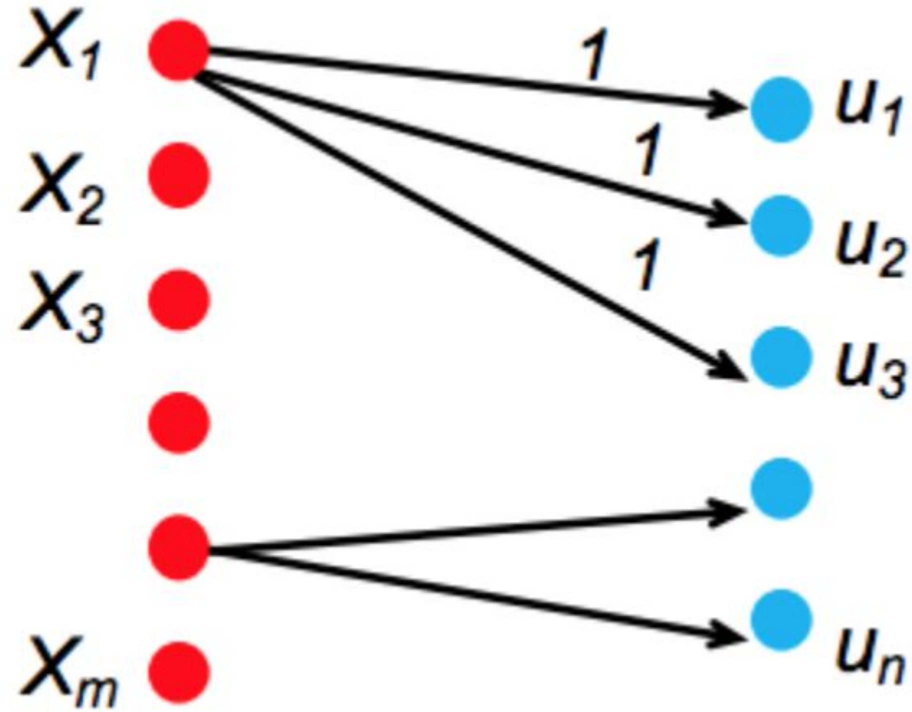
$$\sigma(A) = \sum_{\text{outcomes } X} \text{Prob}[X] \cdot \sigma_X(A)$$



NP-Hardness: Independent Cascades

Reduce to set-cover problem

- Sets $X_1,...,X_m$
- Nodes $u_1,...,u_n$
- Edge between X_i and u_j if u_j in X_i
- Set of k nodes A $\sigma(A) >= n+k$



Optimization is NP-Hard

Submodularity Proof: Linear Threshold

Slightly more complicated: Activation dependent on aggregate

Trigger Graph Model: Pick a live incoming edge for each node v using $b_{v,w}$.

- X: State of edges (live/blocked)
- $R(v,X)$: Set of nodes with live-edge paths from x

Claim: Trigger graph model = Linear threshold model

Submodularity Proof: Linear Threshold

- State of model is the pair (A_{t-1}, A_t)
- Show that transition probabilities between states are same in both models
- Both models start in same state
- Distribution over states is always identical
- Pr(v becomes active at time = t+1)
    - In model 1: Chance that influence weights in $A_t \setminus A_{t-1}$ push it over threshold given not already exceeded.
    - In model 2: Chance that its live edge comes from $A_t \setminus A_{t-1}$ and not $A_{t-1}, A_{t-2}, \ldots, A_0$
    - In both cases it is the same: $$\frac{\sum_{u \in A_t \setminus A_{t-1}} b_{v,u}}{1 - \sum_{u \in A_{t-1}} b_{v,u}}$$

NP-Hardness: Linear Threshold

Show to be equivalent to vertex-cover problem

Optimization is NP-Hard

Experimental Results

Run on co-authorship network from the complete list of papers in the high energy physics theory section of the e-print arXiv (2003)

10748 nodes (researchers), and 53000 edges (co-authorship)

Greedy Algorithm VS

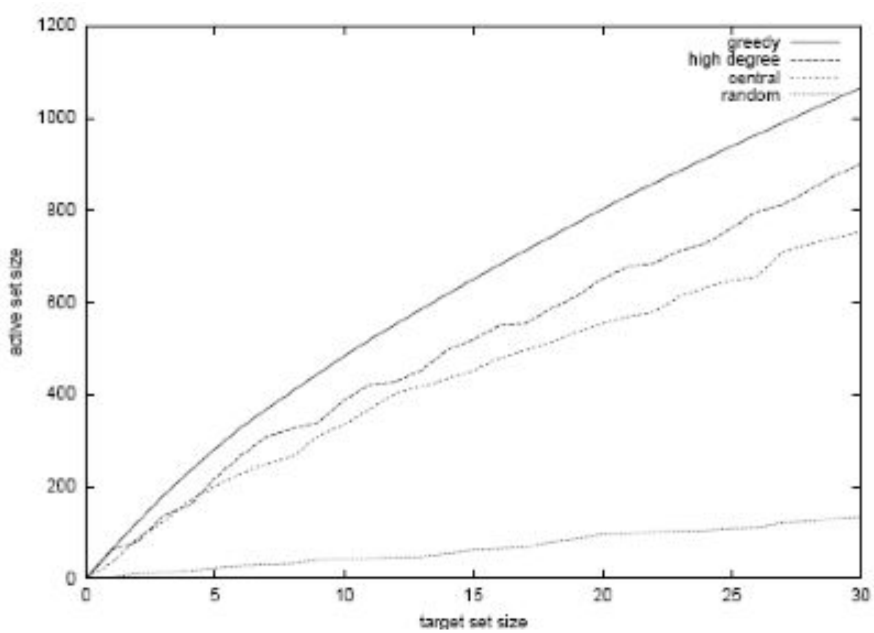
- Structural measures
  - Degree Centrality
  - Distance Centrality
- Random selection

Experimental Results

- Linear Threshold Model: multiplicity of edges as weights. Weight of edge $(u \rightarrow v) = C_{uv} / d_u$
- Independent Cascade Model:
  - Case 1: uniform probabilities p on each edge (parallel edges?)
  - Case 2: edge from u to v has probability $1/ d_v$ of activating v.

For $\sigma(A)$: Simulate the process 10000 times for each targeted set, re-choosing thresholds or edge outcomes pseudo-randomly from [0, 1] every time

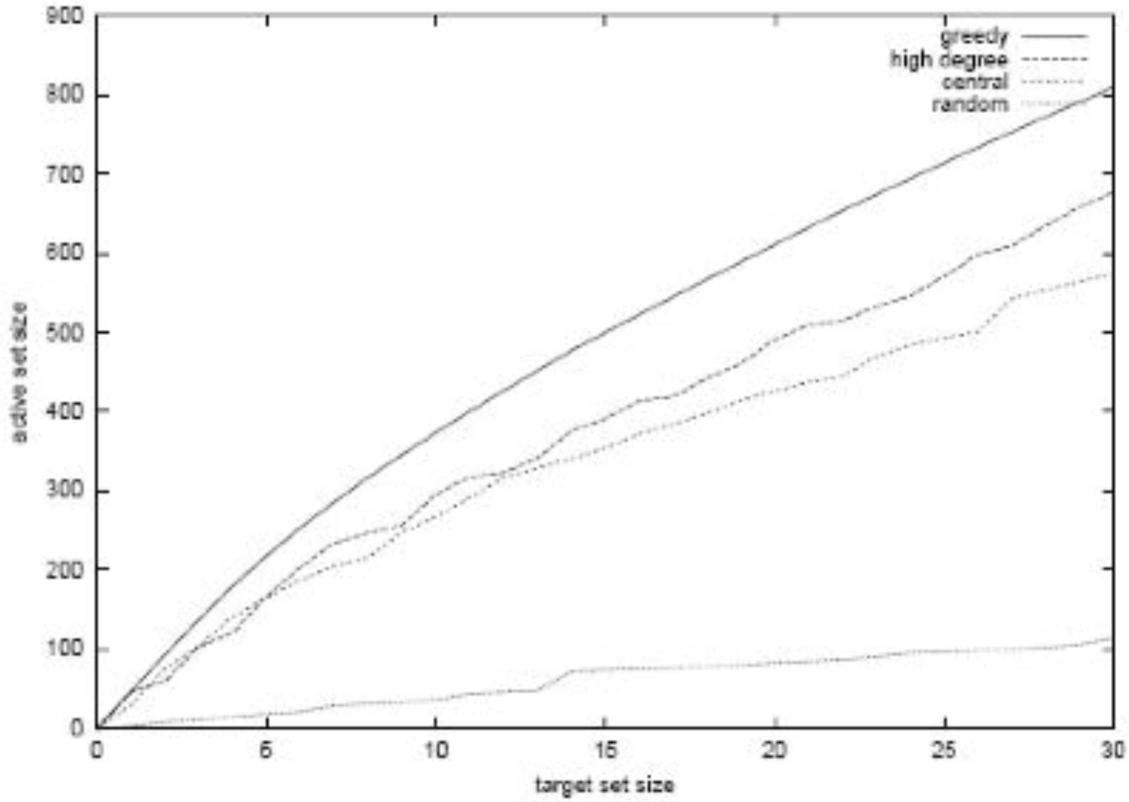
Linear Threshold Model



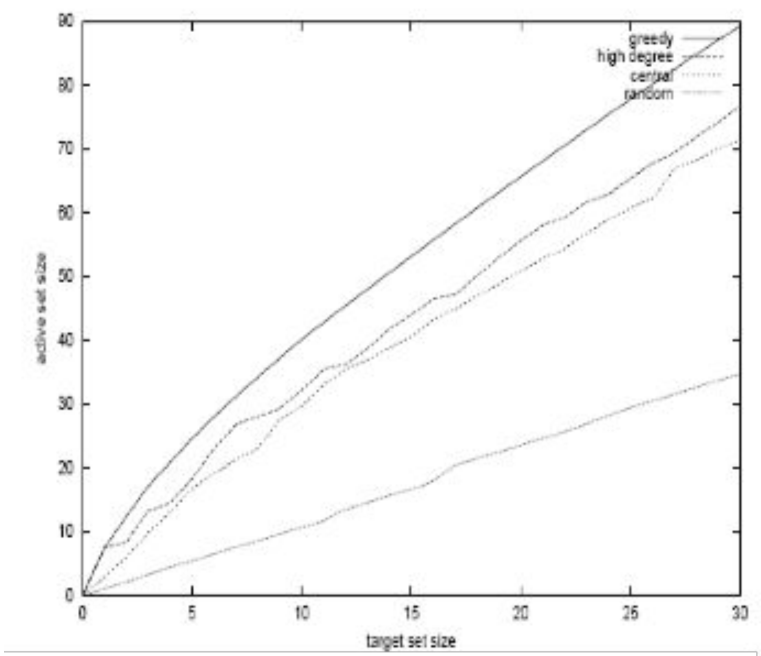
Greedy algorithm outperforms

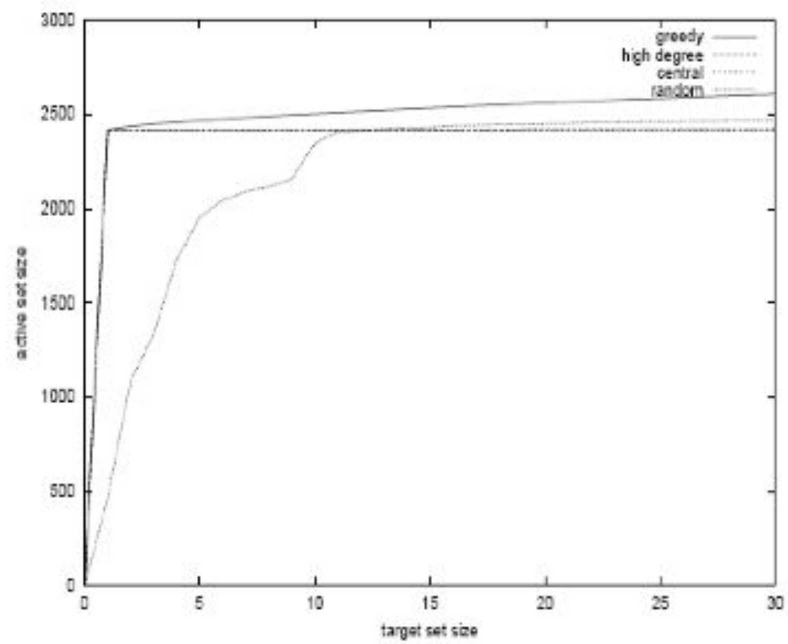
- Degree centrality node heuristic by about 18%
- Distance centrality heuristic by over 40%

Independent Cascade Model: Case 2



Independent Cascade Model : Case 1





Generalizations of Diffusion Models

- Generalized threshold
  - Define monotone threshold function $f_v(S)$ such that node v is activated for $f_v(S) \geq \theta_v$
  - Linear Threshold: $\sum_{w \text{ active neighbor of } v} b_{v,w} \geq \theta_v.$
- General cascade
  - Define success probability probability $p_v(u, S)$
  - Independent cascade $p_{v,u}$ independent of S

These can be shown to be equivalent

NP-hard to approximate in general

Non-progressive processes

- Nodes can switch back to inactivity
- Can be reduced to progressive case
- k interventions rather than k nodes

Theorem: The non-progressive influence maximization problem on G over a time horizon τ is equivalent to the progressive influence maximization problem on the layered graph G^{τ} . Node v is active at time t in the non-progressive process if and only if v_t is activated in the progressive process.

General Marketing Strategies

- m: #marketing actions M_i
- Different nodes may respond to marketing actions in different ways
- Marketing strategy vector **x**
- $h_v(\boldsymbol{x})$: probability that node v is activated by strategy **x**
  - Non-decreasing
  - $h_v(\mathbf{x}+\mathbf{a}) - h_v(\mathbf{x}) \leq h_v(\mathbf{y}+\mathbf{a}) - h_v(\mathbf{y})$

General Marketing Strategies

Expected revenue from final activated set *σ(A)*

$$g(\mathbf{x}) \;=\; \sum_{A \subseteq V} \sigma(A) \cdot \prod_{u \in A} h_u(\mathbf{x}) \cdot \prod_{v \notin A} (1 - h_v(\mathbf{x})).$$

Maximize this using a hill-climbing algorithm

THEOREM 6.1. When the hill-climbing algorithm finishes with strategy x, it guarantees that g(x) ≥ (1 − e − k·γ k+δ·n) · g(xˆ), where xˆ denotes the optimal solution subject to P i xˆi ≤ k

Questions?