Qualifiers Phase - II

Jayesh Choudhari

Advisor: Dr. Anirban Dasgupta

IIT Gandhinagar

May 11, 2016

Overview

1. Problem Statement
2. Models of influence
3. Submodularity
4. Greedy Approximation Algorithm
5. Submodularity of models
6. General Models
7. Non-Progressive Case
8. General Marketing Strategies
9. Experiments
10. Appendix

Problem Statement

Suppose we have a social network, and we want to spread some meme throughout the network (e.g. for marketing). We are able to possibly seed a small number of users with the meme and then onwards it propagates through the network using word of mouth.

How do we formalize this question, and come up with cost-effective ways of seeding to be able to reach out to the maximum number of users?

Introduction

- A social network – Graph of relationships and interactions within a group of individuals
- A meme – idea, video, innovation – an element of a culture or system of behaviour
    - cell phone among college students
    - adoption of new drug within medical profession
    - rise of political movement, etc.

Problem Setting

Given:

- Suppose we are given the estimates for the influence between the individuals and
- a budget for k nodes to be chosen

Aim:

- To trigger a cascade of influence such that maximum number of nodes are influenced

But:

how should we choose the few key (k) individuals to use for seeding this process?

Operational view of Social Network

- Social Network – Directed Graph
- Nodes adopting the idea are “Active” and nodes not adopting the idea are “Inactive”
- An “Inactive” node gets activated only because of its neighbors
- $\sigma(A) =$ *Expected number of active nodes at the end of process*, where A is the initial set of active nodes (targeted k nodes)

Models of Influence

Basic models of influence

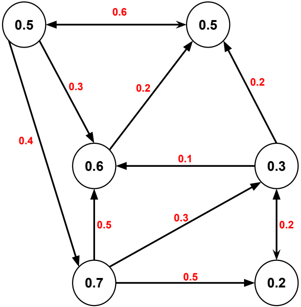
Linear Threshold Model (LTM)

- Node v is influenced by its neighbor according to the weight $b_{v,w} \in [0,1]$
- Each node v has a threshold $\theta_v \in U \sim [0,1]$
- Given the initial set of active nodes A_0 and the thresholds, the diffusion process unfolds deterministically in discrete steps:
  - in step t all nodes active in step $t-1$ remain active
  - each currently inactive node becomes active iff the total weight of the active neighbors is θ_v

$$\sum_{w \rightarrow v : w \;\; active} b_{v,w} \geq \theta_v$$

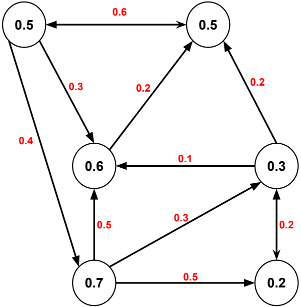
- The process runs until no more activations are possible
- $\sigma(A)$ is calculated as the expectation over all possible threshold values

Example

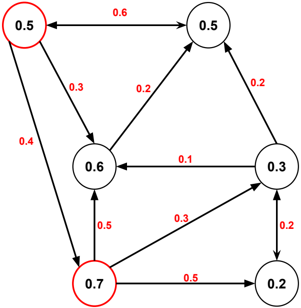


Step-0

Example

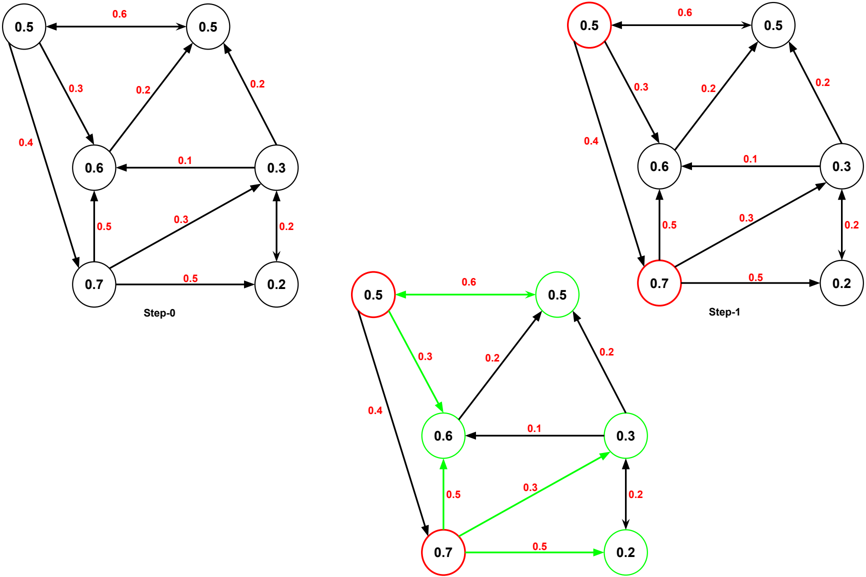


Step-0



Step-1

Example



Step-0

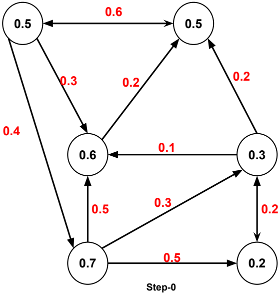
Step-1

Step-2

Independent Cascade Model (ICM)

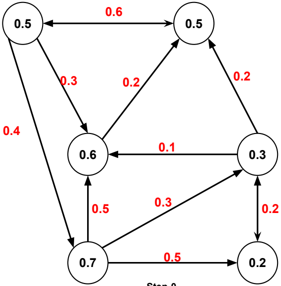
- Given the initial set of active nodes A_0 the diffusion process unfolds as follows:
    - when a node v first becomes active at time t it gets a single chance to activate its neighbor w with a probability $p_{v,w}$ – independently of the history
    - if v succeeds, w becomes active at step $t+1$; whether or not v succeeds, v doesn't gets any more chance to activate w
- The process runs until no more activations are possible
- $\sigma(A)$ is calculated as the expectation over all possible outcomes of success or failure

Example

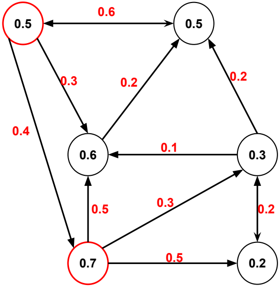


Step-0

Example

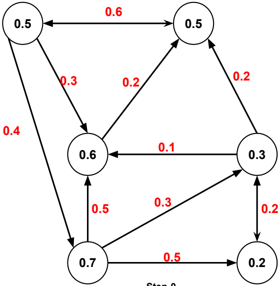


Step-0

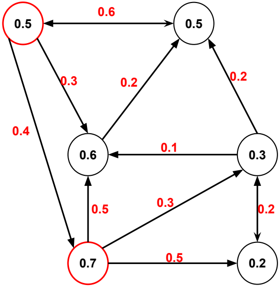


Step-1

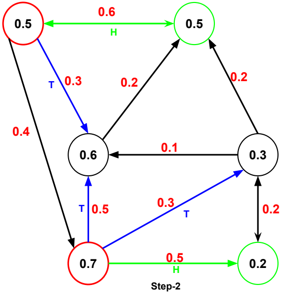
Example



Step-0

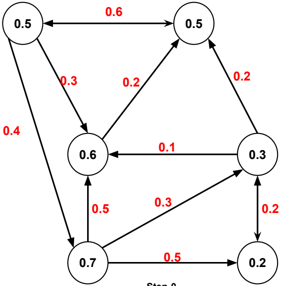


Step-1

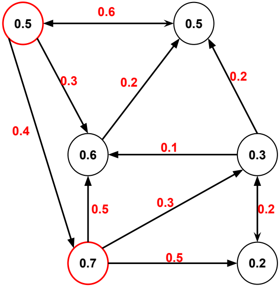


Step-2

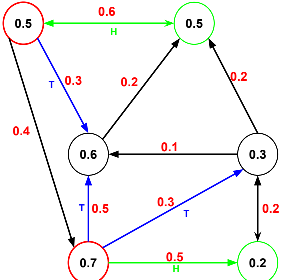
Example



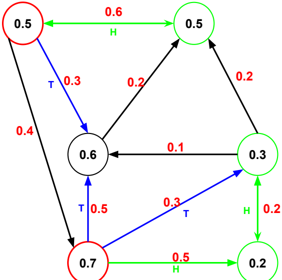
Step-0



Step-1

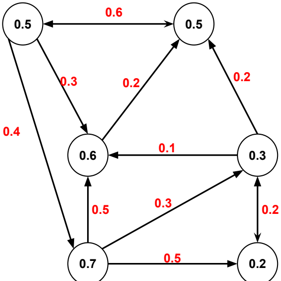


Step-2

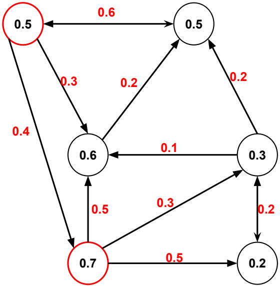


Step-3

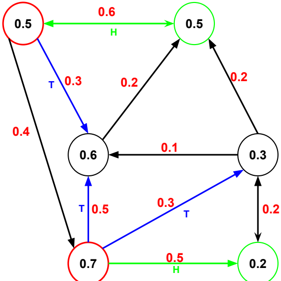
Example



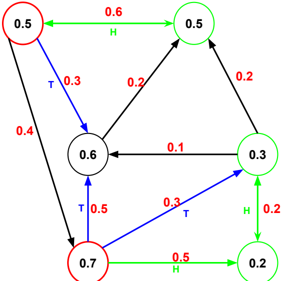
Step-0



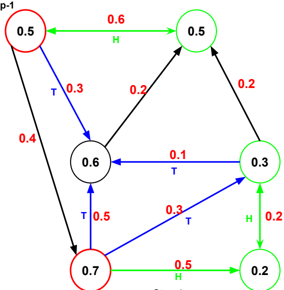
Step-1



Step-2



Step-3



Step-4

Influence Maximization is NP-Hard for LTM

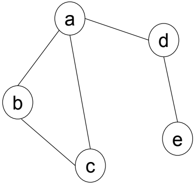
$VC \leq_p$ *Influence maximization (LTM)*

Influence Maximization is NP-Hard for LTM

$$VC \leq_p \textit{Influence maximization (LTM)}$$

Instance of Vertex Cover (VC) problem:

- Undirected graph $G = (V, E)$, $|V| = n$, integer k

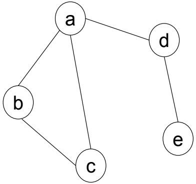


Influence Maximization is NP-Hard for LTM

$VC \leq_p$ *Influence maximization (LTM)*

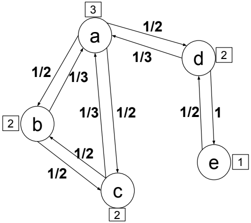
Instance of Vertex Cover (VC) problem:

- Undirected graph $G = (V, E)$, $|V| = n$, integer k



Instance of Influence Maximization Problem:

- Direct all the edges in both directions
- Assign each edge $e = (u, v)$ a weight of $1/deg(v)$

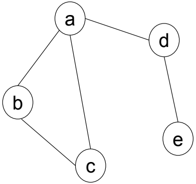


Influence Maximization is NP-Hard for LTM

$VC \leq_p$ *Influence maximization (LTM)*

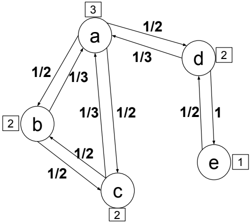
Instance of Vertex Cover (VC) problem:

- Undirected graph $G = (V, E)$, $|V| = n$, integer k



Instance of Influence Maximization Problem:

- Direct all the edges in both directions
- Assign each edge $e = (u, v)$ a weight of $1/deg(v)$



If there is a VC S in G of size k, then $\sigma(A) = n$, where $A_0 = S$

Influence Maximization is NP-Hard for ICM

$SC \leq_p$ *Influence maximization (ICM)*

Influence Maximization is NP-Hard for ICM

$SC \leq_p$ *Influence maximization (ICM)*

Instance of Set Cover (SC) problem:

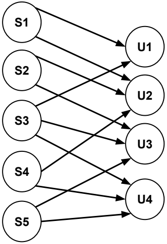
- Collection of subsets $S_1, \ldots, S_m$ of ground set $U = \{u_1, \ldots, u_n\}$
- Select k of the subsets such that $\cup_{i=1}^{k} S_i = U$

Influence Maximization is NP-Hard for ICM

$$SC \leq_p \textit{Influence maximization (ICM)}$$

Instance of Set Cover (SC) problem:

- Collection of subsets $S_1, \ldots, S_m$ of ground set $U = \{u_1, \ldots, u_n\}$
- Select k of the subsets such that $\cup_{i=1}^{k} S_i = U$



Instance of Influence maximization problem:

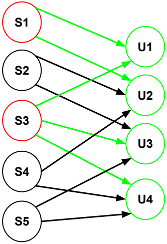
- Directed bipartite graph of $n + m$ nodes
- $p_{i,j} = 1$ if $u_j \in S_i$

Influence Maximization is NP-Hard for ICM

$SC \leq_p$ *Influence maximization (ICM)*

Instance of Set Cover (SC) problem:

- Collection of subsets $S_1, \ldots, S_m$ of ground set $U = \{u_1, \ldots, u_n\}$
- Select k of the subsets such that $\cup_{i=1}^{k} S_i = U$



Instance of Influence maximization problem:

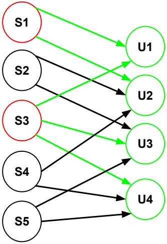
- Directed bipartite graph of $n + m$ nodes
- $p_{i,j} = 1$ if $u_j \in S_i$

Influence Maximization is NP-Hard for ICM

$SC \leq_p$ *Influence maximization (ICM)*

Instance of Set Cover (SC) problem:

- Collection of subsets $S_1, \ldots, S_m$ of ground set $U = \{u_1, \ldots, u_n\}$
- Select k of the subsets such that $\cup_{i=1}^{k} S_i = U$



Instance of Influence maximization problem:

- Directed bipartite graph of $n + m$ nodes
- $p_{i,j} = 1$ if $u_j \in S_i$

- Activating k nodes corresponding to sets in Set cover activates n nodes of U
- If $\sigma(A) \geq n + k$, then set cover problem is solvable

Submodular Functions

Submodularity

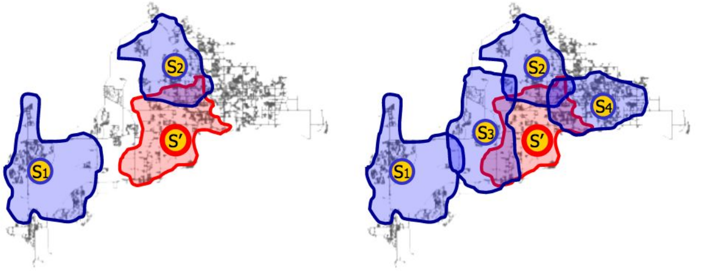
If Ω is a finite set, a submodular function is a set function $f : 2^{\Omega} \to \mathbb{R}$, where 2^{Ω} denotes the power set of Ω, which satisfies one of the following equivalent definitions.

1. For every $X, Y \subseteq \Omega$ with $X \subseteq Y$ and every $v \in \Omega \backslash Y$ we have that $f(X \cup \{v\}) - f(X) \geq f(Y \cup \{v\}) - f(Y)$.
2. For every $S, T \subseteq \Omega$ we have that $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$.

Submodularity

If Ω is a finite set, a submodular function is a set function $f : 2^{\Omega} \rightarrow \mathbb{R}$, where 2^{Ω} denotes the power set of Ω, which satisfies one of the following equivalent definitions.

1. For every $X, Y \subseteq \Omega$ with $X \subseteq Y$ and every $v \in \Omega \backslash Y$ we have that $f(X \cup \{v\}) - f(X) \geq f(Y \cup \{v\}) - f(Y)$.
2. For every $S, T \subseteq \Omega$ we have that $f(S) + f(T) \geq f(S \cup T) + f(S \cap T)$.



(a) *Adding* s' *to set* $\{s_1, s_2\}$ (b) *Adding* s' *to superset* $\{s_1, \ldots, s_4\}$

Ref:http://www.cs.cmu.edu/~dgolovin/papers/submodular_survey12.pdf

Submodularity Approximation

Theorem ([Nemhauser et. al, 1978])

Let $\sigma(\cdot)$ be a non-negative monotone submodular function. Then the greedy algorithm that (for k iterations) adds an element with the largest marginal increase in $\sigma(\cdot)$ produces a $k-element$ set A such that

$$\sigma(A) \geq (1 - 1/e) \cdot max_{|A^*| \leq k} \sigma(A^*)$$

Submodularity Approximation

Theorem ([Nemhauser et. al, 1978])

Let $\sigma(\cdot)$ be a non-negative monotone submodular function. Then the greedy algorithm that (for k iterations) adds an element with the largest marginal increase in $\sigma(\cdot)$ produces a $k - element$ set A such that

$$\sigma(A) \geq (1 - 1/e) \cdot max_{|A^*| \leq k} \sigma(A^*)$$

Also, using $(1 + \delta)$-approximate values for the function to be optimized, gives $(1 - 1/e - \epsilon)$-approximation, where ϵ depends on δ and goes to 0 as $\delta \rightarrow 0$.

Greedy Approximation Algo

Algorithm 1 Greedy Approximation Algorithm

```
Start with $A = \emptyset$
while $|A| \leq k$ do
    For each node $x$ use repeated sampling to approximate $\sigma(A \cup x)$ to within $(1 \pm \epsilon)$ approximation
    Add the node with the largest estimate for $\sigma(A \cup x)$ to $A$
end while
Output set of nodes $A$
```

Claim

If the diffusion process starting with A is simulated independently at least

$$\Omega\left(\frac{n^2}{\epsilon^2} ln(1/\delta)\right)$$

times, then the $\sigma_{sim}(A)$ over these simulations is a $(1+\epsilon)$-approximation to $\sigma(A)$, with probability at least $(1-\delta)$

Submodularity for Models

Submodularity for ICM

- An active node v flips a bias coin with $Pr(head) = p_{v,w}$ (It doesn't matter whether the coin was flipped at the time when v got active or the start of the whole process)
- (Continuing with the same reasoning) we can assume that all the coins corresponding to the edges are flipped at the start of the process (each independently), and the results are checked later in the event when v is active and w is still inactive
- Live Edges – for which the coin flip indicated an activation will be successful and the remaining edges are termed as Blocked Edges

It is easy to see that

A node v is active if and only if there is a path from some node in A to v only through "live" edges

Submodularity for ICM

Consider a probability space in which each sample point specifies one possible set of outcomes for all the coin flips on the edges.

- Let X: sample point in the space
- $\sigma_X(A)$: Number of activated nodes when A is the initial activation set and X is one of the set of outcomes
- $R(v, X)$: Set of all nodes that can be reached from v on a path consisting entirely of "live" edges

$$\therefore \sigma_X(A) = \cup_{v \in A} R(v, X)$$

Submodularity for ICM

Claim

For a fixed outcome X, the function $\sigma_X(\cdot)$ is submodular

Proof

Let S and T be the two sets of nodes such that $S \subseteq T$

$$\sigma_X(S \cup v) - \sigma_X(S) = \#nodes(R((S \ \cup \ v), X)) - \cup_{u \in S} R(u, X)$$

$$\#nodes(R((S \cup v), X) - \cup_{u \in S} R(u, X)) \geq \#nodes(R((T \cup v), X) - \cup_{u \in T} R(u, X))$$

Therefore,

$$\sigma_X(S \cup v) - \sigma_X(S) \geq \sigma_X(T \cup v) - \sigma_X(T)$$

Finally,

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

General Models

General Threshold Model

- Decision of node v to become active can be based on *arbitrary monotone function* of the set of active neighbors
- Each node v has associated with a *monotone threshold function* f_v mapping subsets of $v's$ neighbors to a real numbers in [0,1], and $f(\emptyset) = 0$
- Each node v has a threshold $\theta_v \in U \sim [0, 1]$

Node v becomes active at time t iff $f_v(S) \geq \theta_v$, where S is the set of neighbors of v that are active in time $t - 1$

General Cacade Model

- Define $p_v(u, S) \in [0, 1]$, as an increamental function, where S is set of neighbors of v that have already tried and failed to activate v and $u \notin S$
- We are only interested in the cascade models defined by an increamental functions that are *order independent* in the following sense:
  let $S = \{u_1, u_2, \ldots, u_{|S|}\}$, and π, ψ are two arbitrary permutations of $1, 2, \ldots, |S|$, then

$$\prod_{i=1}^{|S|}(1 - p_v(u_{\pi(i)}, \{u_{\pi(1)}, \ldots u_{\pi(i-1)}\})) = \prod_{i=1}^{|S|}(1 - p_v(u_{\psi(i)}, \{u_{\psi(1)}, \ldots u_{\psi(i-1)}\}))$$

Equivalence

Consider an instance of general threshold model with f_v as threshold functions

- Given that the nodes in S have tried and failed, then node $v's$ threshold $\theta_v \in (f_v(S), 1]$
- Also, $\theta_v \in U \sim [0, 1]$

Therefore,

$$p_v(u, S) = \frac{f_v(S \cup u) - f_v(S)}{1 - f_v(S)}$$

where, u is neighbor of v and is not in S yet

Equivalence

- Consider node v in cascade model and let $S = \{u_1, u_2, \dots, u_k\}$ be its neighbors
- Assume that the nodes in S try to activate v in order $u_1, u_2, \dots, u_k$ and let $S_i = \{u_1, u_2, \dots, u_i\}$. Then,

$$Pr[v\ not\ activated] = \prod_{i=1}^{k}(1 - p_v(u_i, S_{i-1}))$$

Then, **Threshold Model's** activation function $f_v(S)$ can be given as in terms of probabilities from **Cascade Model**

$$f_v(S) = 1 - \prod_{i=1}^{k}(1 - p_v(u_i, S_{i-1}))$$

Equivalence

Thus, each individual node becomes active with the same probability under both the processes, i.e.

$$\frac{f_v(S \cup u) - f_v(S)}{1 - f_v(S)} = 1 - \prod_{i=1}^{k}(1 - p_v(u_i, S_{i-1}))$$

Non-Progressive Process

Non-Progressive Case

- Nodes can become active from inactive as well as inactive from active
- **Model Formulation** **(Simplest way):**
  - Each node v chooses a *new threshold* at each time step t as $\theta_v^{(t)}$
  - Node v will be active in step t iff $f_v(S_{t-1}) \geq \theta_v^{(t)}$, where S_{t-1} is the set of neighbors active at time step $(t-1)$
  - **Objective Function:** $\sigma(A) = \sum_t (\#nodes\ active\ in\ step\ t)$

Non-Progressive Case

- τ – Time horizon for which the process will run
- *Intervention:* Activation of a particular node v at time $t \leq \tau$
- The question is which $k - interventions$ should be made to maximize the influence if the process it to run for τ time steps?

Layered graph

- Given a graph $G = (V, E)$ and the time limit τ, $G^\tau = (\tau \cdot |V|, \tau \cdot |E|)$
- Let the t^{th} layer of G^τ be

$$L_t = \{v_t | v \in V\}$$

- For each node v_t the influence function is defined as

$$f'_{v_t}(S) = f_v(\{u | u_{t-1} \in S\})$$

Non-Progressive Case

Theorem (Non-progressive Influence maximization)

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 iff v_t is activated in the progressive process.

General Marketing Strategies

General Marketing Strategies

- In general there might be m different number of marketing actions M_i available, each of which may affect some subset of nodes by increasing their probabilities of activation
- The more we spend on any action the stronger its effect will be; different nodes may respond to marketing actions in different ways
- Let $x = [x_1, x_2, \ldots, x_m]$ be the investment vector corresponding to marketing action $M = [M_1, M_2, \ldots, M_m]$, such that the total investments does not exceed some budget
- $Pr[node\ v\ will\ become\ active] = h_v(x)$
- Assume that function $h_v(x)$ is non-decreasing and satisfies "diminishing returns" property, i.e. $\forall x \geq y$ and $a \geq 0$,

$$h_v(x+a) - h_v(x) \leq h_v(y+a) - h_v(y)$$

General Marketing Strategies

- As a function of marketing strategy x, each node v becomes active with a probability $h_v(x)$, resulting in a (random) set of initial active nodes A
- Given the initial active set A, try to maximize the expected size of the final active set $\sigma(A)$
- The expected revenue of the marketing strategy x is then,

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

Hill-Climbing Algorithm

- To maximize g, we assume that we can evaluate the function at any point x approximately, and find the direction i with approximately maximal gradient
- Let e_i denote the unit vector along the i^{th} direction and divide each unit of total budget k into equal parts of size δ

Hill Climbing Algorithm

Algorithm 2 Hill Climbing Algorithm

```
Start with $x^{(0)} = \mathbf{0}$
for all rounds $t = 0, \ldots, k \cdot \delta^{-1}$ do
    Let $i_t$ be the direction maximising $g(x^{(t)} + \delta \cdot e_i) - g(x^{(t)})$
    Set $x^{(t+1)} = x^{(t)} + \delta \cdot e_i$
end for
```

Hill-Climbing Algorithm

Theorem (Hill Climbing Algo Approximation)

When the hill-climbing algorithm finishes with strategy x, with γ-approximate gradient values, it guarantees that

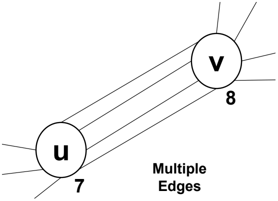
$$g(x) \geq \left(1 - e^{\frac{k\cdot\gamma}{k+\delta\cdot m}}\right) \cdot g(\hat{x})$$

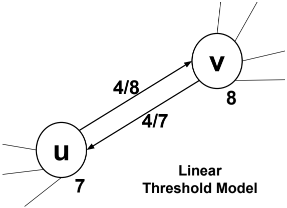
where, $g(x)$ is non-negative, monotone, and satisfies "diminishing returns", and $\hat{x}$ denotes the optimal solution subject to $\sum_i \hat{x}_i \leq k$.

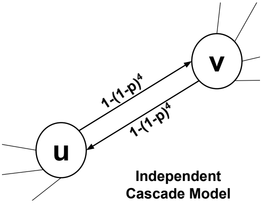
Experiments and Results

Experiments and Results - [Kempe et. al, 2003]

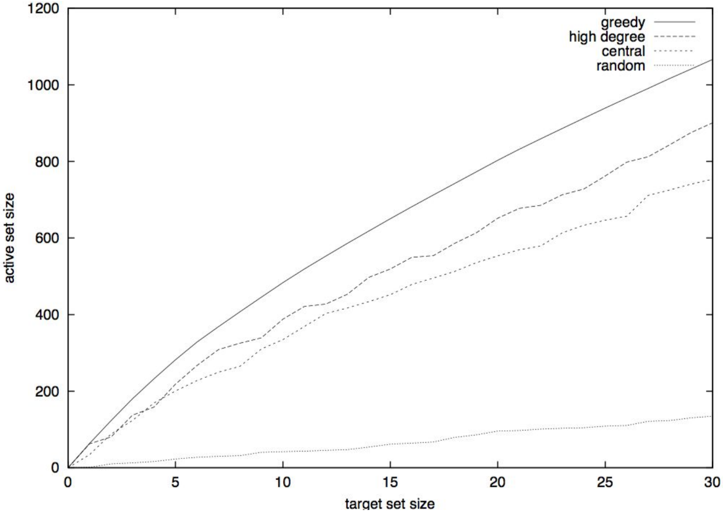
- Co-authorship dataset – High-Energy Physics Theory (2002)
- Edge $e = (u, v)$ from author u to author v if they have co-authored a paper
- $|V| = 10748$, Edges between 53000 pair of nodes
- Mulitple/parallel edges between 2 authors indicates strength of relationship





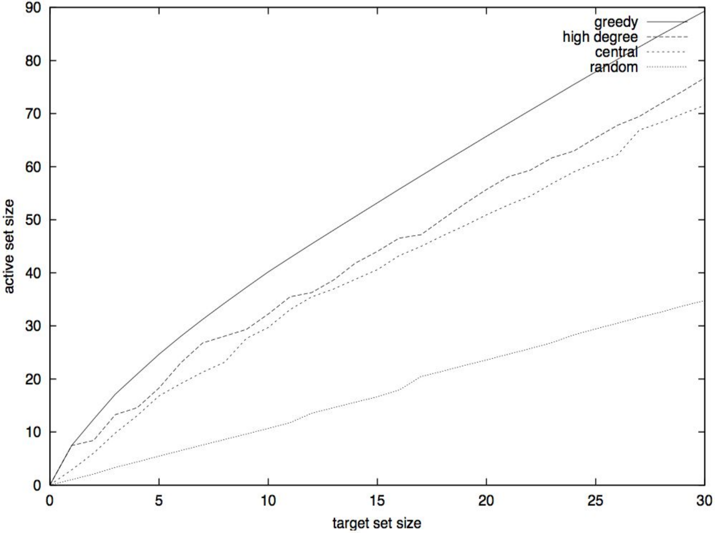


Results - LTM

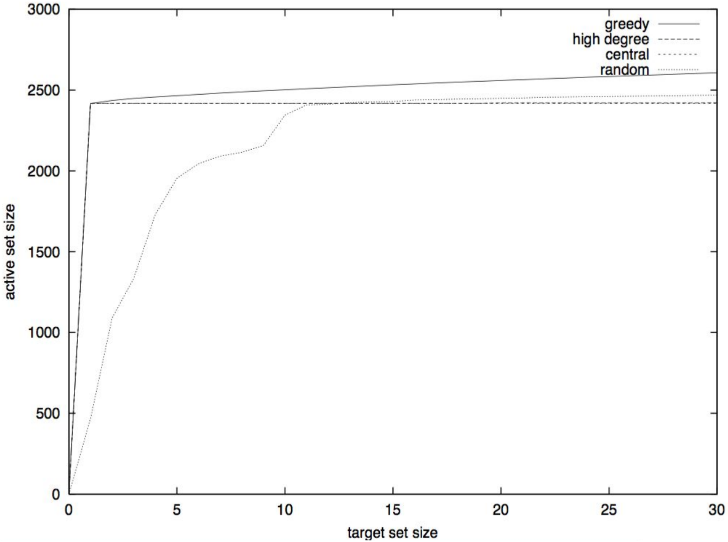


greedy
high degree
central
random
active set size
target set size
0
200
400
600
800
1000
1200
0
5
10
15
20
25
30

Results - ICM (p = 1%)



Results - ICM (p = 10%)



Questions?

References

David Kempe, Jon Kleinberg, Eva Tardos (2003)
Maximizing the Spread of Influence through Social Network
SIGKDD (2003).

G. L. Nemhauser, L. A. Wolsey, M. L. Fisher (1978)
An Analysis of Approximations for Maximizing Submodular Set Functions-I
Mathematical Programming 14, (1978), 265-294.

David Kempe, Jon Kleinberg, Eva Tardos (2015)
Maximizing the Spread of Influence through Social Network
Theory Of Computing 11(4), (2015), 105-147.

Thank You!

Submodularity for LTM

Triggering Set Technique:

- Let $b_{v,w} = 0$ when w is not a neighbor of v
- Suppose v picks at most one of its incoming edges at random with probability $b_{v,w}$ selects no edge with probability $1 - \sum_w b_{v,w}$
- Selected edges are “Live Edges”, and other edges are “Blocked Edges”

Submodularity for LTM

Claim

For a given targeted set A, the following two distributions over sets of nodes are the same

1. The distribution over active sets obtained by running the Linear Threshold process to completion starting from A; and
2. The distribution over sets reachable from A via live-edge paths, under the random selection of live edges defined above (Triggering set technique)

Submodularity for LTM

Proof: Part-I

- Let A_t set of active nodes at the end of iteration t, for $t = 0, 1, \dots,$ (A_0 is the set initially targeted)
- Probability that a node v becomes active at time $t+1$, given that it has not become active till time t is:

$$\frac{\sum_{u \in A_t \setminus A_{t-1}} b_{v,u}}{1 - \sum_{u \in A_{t-1}} b_{v,u}}$$

Submodularity for LTM

Proof: Part-II (live-edge paths)

Run the live-edge process as follows:

- Start with targeted set A
- For each node v, if $v's$ live edge is from A, v is reachable, else, keep the source of $v's$ live edge unknown if not from A
- Now, A'_1 is the new set of reachable nodes. Continue in the similar way to get the node sets $A'_2, A'_3, \ldots$
- If the node is not been reached by the end of stage t, then the probability that it is reachable in stage $t+1$ is

$$\frac{\sum_{u \in A_t \setminus A_{t-1}} b_{v,u}}{1 - \sum_{u \in A_{t-1}} b_{v,u}}$$

Thus, the probability of a node being active is equal in both the cases

Submodularity for LTM

Proving Submodularity:
Follows the similar proof as in case of Independent Cascade Model