CS 59000-ENS: Influence Maximization

Assefaw Gebremedhin
Purdue University
agebreme@purdue.edu

http://www.cs.purdue.edu/homes/agebreme/Networks

Spring 2013
Marketing

- Mass marketing
- Directed marketing
- Viral marketing
Viral marketing

- Identify influential customers
- Provide them with incentives to adopt
- Hope their influence will cause cascade effect of adoption
Maximizing the spread of influence

Choose the most influential $k$ people such that a cascade starting from those nodes reaches the largest population

Domingos and Richardson, KDD 2001 and 2002
How to find $k$ such people?

We look at the work of Kempe, Kleinberg and Tardos, KDD 2003
Diffusion models

- Basic elements of an operational model
  - Directed graph
  - Active and inactive nodes
  - Progressive (and non-progressive) cases

- A number of diffusion models
  - Independent Cascade Model
    - Based on interacting particle systems
  - Linear Threshold Model
    - A node’s tendency to adopt increases monotonically as more of its neighbors become adopters
  - And their extensions
Independent Cascade Model

- We have a directed graph $G=(V,E)$
- Start with an initial set of active nodes $S$
- Process unfolds in discrete time steps
- When node $v$ first becomes active in step $t$, it is given a single chance to activate each currently inactive neighbor $w$; it succeeds with probability $p_{vw}$.
  - If $w$ has multiple newly activated neighbors, their attempts are sequenced randomly.
- If $v$ succeeds, $w$ becomes active in step $t+1$
- Process runs until no more activations are possible
Independent Cascade Model
Linear Threshold Model

- Directed graph \( G=(V,E) \)
- Start with an initial set of active nodes \( S \)
- A node \( v \) is influenced by each neighbor \( w \) according to a weight \( b_{vw} \) such that the sum of the weights over all neighbors is \( \leq 1 \).

Dynamics of the process:

- Each node chooses a threshold \( t_v \) uniformly from \([0,1]\)
  - This represents the weighted fraction of \( v \)'s neighbors that must become active in order for \( v \) to become active
- Diffusion unfolds deterministically in discrete steps
  - In step \( t \), all nodes that were active in step \( t-1 \) remain active
  - And any node \( v \) for which the total weight of its active neighbors is at least \( t_v \) becomes active
Maximizing the spread of influence under ICM and LTM

Given budget $k$, select a set $S$ of $k$ nodes so as to maximize $f(S)$ where

$f(S)$: expected number of nodes active at the end, if set $S$ is targeted for initial activation

The problem is NP-hard (Reduction to Set Cover)
Influence Maximization is NP-hard

- **Set Cover:**
  - Given a universe of elements $U = \{u_1, \ldots, u_n\}$ and sets $X_1, \ldots, X_m$ contained in $U$
  - Are there $k$ sets among $X_1, \ldots, X_m$ s.t their union is $U$?

- **Goal:** encode SC as an instance of IM

- **Build a bipartite $X$-to-$U$ graph**
  - For each $X_i$, create a directed edge $(X_i, u)$ for each $u$ contained in $X_i$.
  - Put weight 1 on the edge

- There exists a set $S$ of size $k$ with $f(S) = k + n$ iff there exists a size $k$ set cover

$X_1 = \{u_1, u_2, u_3\}$
Approximation Method

- **Greedy Method**

  For \( k \) iterations:
  
  Add a node \( u \) to set \( S \) that maximizes \( f(S+u) - f(S) \)

- **Theorem**

  The greedy method is a \( (1-1/e) \) approximation for both ICM and LTM
Proof of approximation bound

1) Prove that $f(S)$ is
   a) **Monotone**
      If $S$ is contained in $T$ then $f(S) \leq f(T)$
   b) **Submodular** (diminishing returns)
      \[ f(S + v) - f(S) \geq f(T + v) - f(T) \]
      whenever $S$ contained in $T$

2) Invoke Nemhauser-Wolsey-Fisher Theorem (1978)
Diminishing returns

\[ \forall S \subseteq T \]

Adding \( u \) to \( T \) helps less than adding it to \( S \)!
Proof of approximation bound

1) Prove that $f(S)$ is
   a) **Monotone**
      
      If $S$ is contained in $T$ then $f(S) \leq f(T)$
      
      Easy to see (activating more nodes never hurts)
   b) **Submodular** (diminishing returns)
      
      $f(S+v) - f(S) \geq f(T+v) - f(T)$ whenever $S$ contained in $T$
      
      Will show this next…

2) Invoke Nemhauser-Wolsey-Fisher Theorem (1978)
Submodularity of diffusion process

- Flip all the coins at the beginning and record results
  (recall percolation from last lecture)
- Active nodes in the end are reachable via live edges
- Study reachability in the these graphs
- This problem is submodular
Submodularity of Reachability

\[ g(T + v) - g(T) \subseteq g(S + v) - g(S) \text{ when } S \subseteq T. \]
Evaluating $f(S)$

- Not only one, but many scenario
- Is it still submodular?
  - Yes, since linear (nonnegative) combination of submodular functions is submodular

- How does one evaluate $f(S)$?
  - Efficient computation still an open question.

- Good estimates can be obtained by simulation (MC)
  - Repeating the diffusion process often enough
Experiments

- Collaboration network (co-authorships in papers in arXiV high-energy physics theory):
  - 10,748 nodes
  - 53,000 edges

- Independent Cascade Model
  - Case 1: inform probabilities on each edge
  - Case 2: edge from v to w has probability $1/\text{deg}(w)$ of activating w

- Simulated the process 10,000 times for each targeted set
  - Each time re-choosing edge outcomes randomly

- Compared with three other heuristics
  - Degree centrality
  - Distance centrality
  - Random
Results

$p_{uv} = 1\%$

$p_{uv} = 10\%$

Degree and distance centrality: captures only structure
Greedy: captures dynamics
Results

\[ p_{uv} = \frac{1}{\text{deg}(v)} \]