An Optimal Allocation Approach to Influence Maximization Problem on Modular Social Network

Tianyu Cao, Xindong Wu, Song Wang, Xiaohua Hu

ACM SAC 2010
outline

- Social network
  - Definition and properties
  - Social contagion
- Models of Influence
  - Linear threshold model
  - Independent cascade model
- Influence maximization problem
  - Greedy algorithm and the CELF optimized version
  - Degree discount algorithm
  - OASNET algorithm
- Experiments
Social Network Definitions

- A **social network** is a social structure made of individuals (or organizations) called "nodes," which are tied (connected) by one or more specific types of interdependency, such as friendship, kinship, financial exchange, dislike, sexual relationships, or relationships of beliefs, knowledge or prestige. (Wiki)

- Alternatively: people + interactions.
Social Network Examples

The University Karate Club Network[8]
Social Network Example

Modularity and structure

Clauset et al., 2006 [7]: NCAA football
Social Network Example

The Structure of Romantic and Sexual Relations at "Jefferson High School"

Each circle represents a student and lines connecting students represent romantic relations occurring within the 6 months preceding the interview. Numbers under the figure count the number of times that pattern was observed (i.e. we found 63 pairs unconnected to anyone else).
Social Network Example

Malaysian Blogosphere Division[9]
Social Network Properties

- Power law degree distribution. \( P(k) = c k^{-r} \)
- Positive assortativity. Similar degree nodes connecting to each other.
- High clustering coefficient. Your friends tend to know each other.
- Short diameter. (six degree of separation)
- Community structure.
Social Contagion

- Yawning is contagious.
Social contagion

- “I’ll have what she is having.”
Models of contagion/diffusion

- Initially some nodes are active:
- Active nodes spread their influence on the other nodes, and so on ...

Leskovec & Faloutsos, WWW 2008
Models of Influence

- Two basic classes of diffusion models: threshold[5] and cascade [1]

- 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
Linear Threshold Model

- A node $v$ has random threshold $\vartheta_v \sim U[0,1]$
- 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) $\vartheta_v$ fraction of its neighbors are active

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

Stop!
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

Stop!
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?
Influence Maximization Problem

- Influence of node set $S$: $f(S)$
  - expected number of active nodes at the end, if set $S$ is the initial active set

- Problem:
  - Given a parameter $k$ (budget), find a $k$-node set $S$ to maximize $f(S)$
  - Constrained optimization problem with $f(S)$ as the objective function
First Solution

• Greedy Algorithm[1]
  – 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?[1]
  – 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.
Problems with Greedy Algorithm

- The greedy algorithm in [1] is slow. Its time complexity is $O(\text{knms})$, where $n$ is the number of nodes, $m$ is the number of edges and $s$ is the times of simulation.
- It does not assume any topological structure of the social network. There might be some space for optimization given a certain topological property.
CELF optimized Greedy Algorithm

- The function $f(S)$ is sub-modular for both the independent cascade model and the linear threshold model.\[1\]
  
  $f(TU\{v\}) - f(T) \leq f(SU\{v\}) - f(S)$ if $S$ is a subset of $T$. 
Recall the greedy algorithm.

For \( k \) iterations:

Add node \( v \) to \( S \) that maximizes \( f(S + v) - f(S) \).

CELF optimized greedy algorithm [2]

Keep an ordered list of marginal benefits \( b_i \) from previous iteration

Re-evaluate \( b_i \) only if \( b_i \) in previous iteration is larger than the maximum marginal benefits for current iteration
In the second iteration, the nodes b, e, c do not need to be evaluated because their gain in first iteration is smaller than the gain of d in the current iteration.[10]
CELF optimized Greedy Algorithm

- **Pros**
  - It produces the same result as the greedy algorithm with much less time. (Experiments show that it is 700 times faster than the original greedy algorithm)

- **Cons**
  - It does not improve the running time complexity. For a large network it is still slow.
Degree Discount Heuristics

- The degree heuristics is a good baseline compared to other heuristics (random, distance centrality).
- If a node $u$ has been selected as a seed, when considering selecting a new seed $v$ based on its degree, we should not count the edge $\{v, u\}$ towards its degree.[2]
Node 6 is selected in the second iteration because it has a higher degree than node 5, whose degree is discounted by 1.
Degree Discount heuristics

- There is a more sophisticated discount mechanisms when the diffusion probability is small[2].
- In the IC model with propagation probability $p$, suppose that $d_v=O(1/p)$ and $t_v=o(1/p)$ for a vertex $v$. The expected number of additional vertices in $\text{Star}(v)$ influenced by selecting $v$ into the seed set is:

$$1+(dv-2tv-(dv-tv)tvp+o(tv))\cdot p$$

$$\begin{align*}
(1 - p)^{tv} \cdot (1 + (dv - t_v) \cdot p) \\
= (1 - tvp + o(tv)p) \cdot (1 + (dv - t_v) \cdot p) \\
= 1 + (dv - 2tv)p - (dv - t_v)tvp^2 + o(tv)p \\
\{\text{since } tvp = o(1)\} \\
= 1 + (dv - 2tv - (dv - t_v)tvp + o(t_v)p). \\
\end{align*}$$
Result on the above heuristics[2]

Figure 1: Influence spreads of different algorithms on the collaboration graph NetHEPT under the independent cascade model ($n = 15,233$, $m = 58,891$, and $p = 0.01$).
Result on the above heuristics[2]
Result on the above heuristics[2]

Figure 9: Influence spreads of different algorithms on the collaboration graph NetHEPT under the linear threshold model ($n = 15,233$ and $m = 58,891$).
Our motivation

- Make use of the modularity of social networks. Generally a social network exhibits some degree of modularity.
- Try to make an algorithm that scales better than greedy algorithm in terms of the time complexity.

Idea:
- Treat the seeds as a resource. Try to allocate the resources to the communities of a network in an intelligent way.
Growth function

- The growth function $F(k, G, M, \Gamma)$ maps the number of initial seeds $k$, the network $G$, the diffusion model $M$ and a base strategy of selecting initial active nodes (like high degree) to the expected number of active nodes when the diffusion process terminates.
- Relation with $f(S)$ in [1]. $\Gamma(k, G) = S$. 
Growth Function

- Condensed Matter physics network with diffusion probability 0.05
- Condensed Matter physics network with linear threshold model
Problem Reformulation

- We made a simplified assumptions. Communities are disconnected.
- Maximize $D=\sum F_i(k_i, G_i, M_i, \Gamma_i)$ s.t. $\sum k_i=k$.
- Solutions:
  - $OPT(k,n) = \text{MAX}\{F_n(i, G_n, M_n, \Gamma_n) + OPT(k-1,n-1)\}$ for $0 \leq i \leq k$. 
Adaptation to a real world network

- The assumption that all the communities are disconnected is too restrictive.
- Communities are usually weakly connected.
- It violates the requirement of the reformulated problem.
- Second assumption: The number of cross community edges is relatively small and they are far from the center of the community. It won’t have a large effect.
OASNET algorithm

- The steps of the algorithm.
- 1. Use an external community finding algorithm to find the communities of the network. Extract these communities from the network.
- 2. Simulate the diffusion process on each community to find its corresponding growth function.
- 3. Apply the optimization equation on the n growth functions to find an optimal allocation of n initial target nodes into n communities.
- 4. Return the seed as the union of $\Gamma(k_i, G_i)$.

Time complexity is $O(kms)$. 
Experiment Settings

- The greedy algorithm does not scale well to the cases of our experiments (too slow for networks over 30,000 nodes). We use four heuristics to compare.
- Degree heuristics (M5), equal allocation (M2), random allocation (M4), proportional allocation (M1).
Experiment Settings

- Datasets.
  1. Condense matter physics collaboration network[3];
  2. PGP giant component network[4];
  3. Lederberg citation network;
  4. Zewail Citation Network.
Condensed Matter physics network with diffusion probability 0.05
Experiment Results

PGP Network with diffusion probability 0.1
Experiment Results

PGP Network with diffusion probability 0.2
Experiment Results

Lederberg Citation Network with diffusion probability 0.1
Lederberg Citation Network with diffusion probability 0.2
Experiment Results

Zewail Citation Network with diffusion probability 0.1
Zewail Citation Network with diffusion probability 0.2
Experiment Results

Condensed matter physics Network with threshold model
Experiment Results

PGP Network with threshold model
Lederberg Citation Network with threshold model
Experiment Results

Zewail Citation Network with threshold model
Conclusions

- The proposed algorithm outperforms comparing heuristics in most cases.
- The difference between the proposed method and comparison heuristics are larger when the diffusion probability is higher for the independent cascade model.
References


References


References


Thanks