Interest-based Information Diffusion in Dynamic Social Network

组员: 付茜, 顾容菲, 占钰霞

2020/5/20

Background

Influence Maximizing

Premise of viral marketing is that by initially targeting a few “influential” members of the network, we can trigger a cascade of influence. How should we choose the few key individuals to use for seeding this process so that the number of final influential members will be maximized ?

Government

- Dispel rumors
- Publish reports

Enterprise

- Viral marketing
- Product promotion
- “Word of mouth”

Static Social Network

- Social network: G (directed graph); User: node (active or inactive);
- Node's tendency to become active increases monotonically as more of its neighbors become active.

Liner Threshold Model

Node v is influenced by neighbor w with weight $b_{v,w}$, each node v has threshold θ_v .

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

Independent Cascade Model

When node v first becomes active in step t, it is given a single chance to activate each currently inactive neighbor w, and succeeds with probability $p_{v,w}$.

Kempe, David, Jon Kleinberg, and Eva Tardos. "Maximizing the spread of influence through a social network." *knowledge discovery and data mining* (2003): 137-146.

Static Social Network

Submodular function

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

for for all elements v and all pairs of sets $S \subseteq T$.

Greedy algorithm

For a non-negative, monotone submodular function f, 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

$$f(S) \geq (1 - 1/e) \cdot f(S^*)$$

In other words, S provides a $(1 - 1/e)$ approximation.

Dynamic Social Network

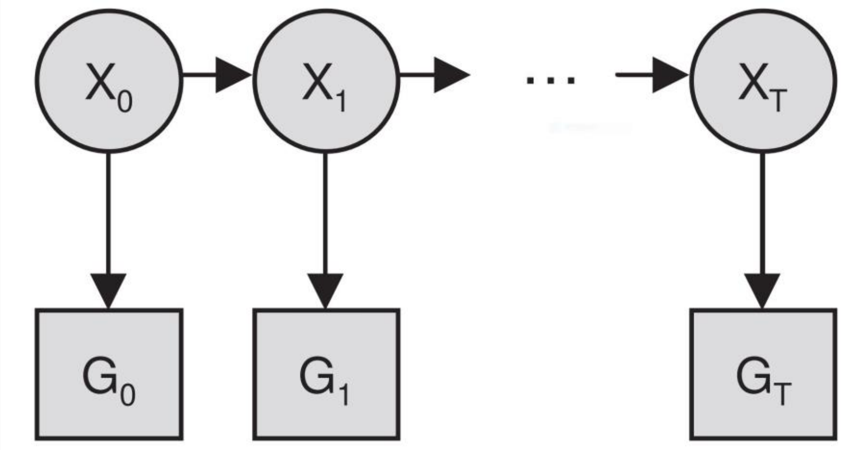
Disadvantage of static social network

- Diffusion model of static social network cannot simulate the reality well ;
- The real social network is complicated and time-varying.

What dynamic means

- Topology of network changes over time ;
- Some node may join in and some may vanish ;
- Edges between nodes may change due to their behavior.

Latent Space Model



Main Idea

- Embed an evolving social network graph $(G_0, G_1, \ldots, G_t)$ in p dimensional latent space $(X_0, X_1, \ldots, X_t)$.
- Entities can move in latent space between timesteps according to some Gaussian distribution.
- Change of latent space satisfy standard Markov assumption.

Tracking Problem: estimate X_t as a function of current observed social network graph G_t and previously estimated positions in latent space X_{t-1}.

Sarkar, Purnamrita, and Andrew W. Moore. "Dynamic social network analysis using latent space models." *Sigkdd Explorations* 7.2 (2005): 31-40.

Latent Space Model

$$X_t = \arg\max P(X \,|G_t, X_{t-1}) = \arg\max \underline{P(G_t|X)}\ \underline{P(X|X_{t-1})}$$

$P(G_t|X_t)$: Observation Model

$$\log P(G_t|X_t) = \sum_{i\sim j} \log p(i \sim j) + \sum_{i\nsim j} \log p(i \nsim j)$$

$$= \sum_{i\sim j, d_{i,j}\le r_{i,j}} \log p(i \sim j) \ + \#\big(i \sim j, d_{i,j} > \ r_{i,j}\big) \log \rho$$

$$= \sum_{i\nsim j, d_{i,j}\le r_{i,j}} \log p(i \nsim j) + \#\big(i \nsim j, d_{i,j} > \ r_{i,j}\big) \log(1-\rho)$$

$P(X_t|X_{t-1})$: Transition Model

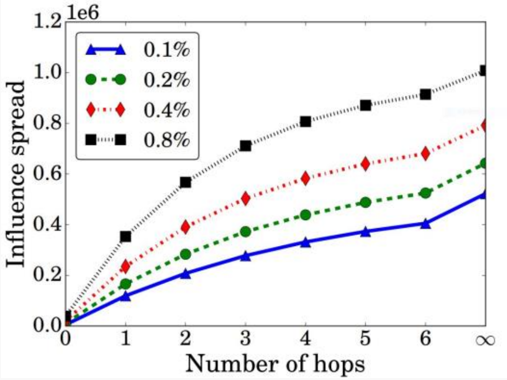
$$\log P(X_t|X_{t-1}) = -\sum_{i=1}^{n} \frac{\left| X_{i,t} \ - \ X_{i,t-1}\right|^2}{2\sigma^2}$$

$$+ \ some\ const$$

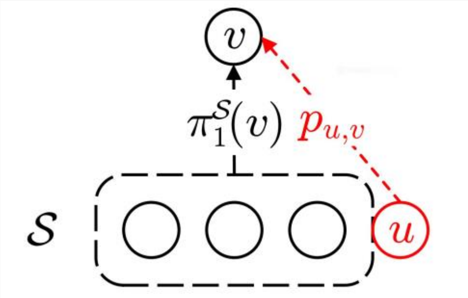
Hop-Based Approach

Motivation

- Traditional Monte-Carlo method: slow and not scalable ;
- Newly heuristic algorithms: poor solution and not theoretical ;
- Hop-based approach: efficient & effective ! Majority of influence spread is produced within the first few hops of propagation.



One-hop of propagation

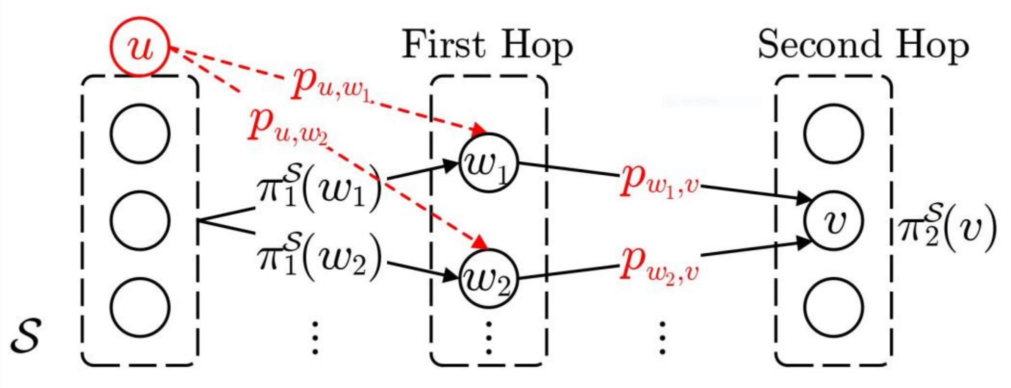


S : seed set; $\mathcal{I}_v$: inverse neighbor of node v ;
$\pi_h^S(v)$: probability for node v to be activated within h hops.

$$\pi_1^S(v) = \begin{cases} 1 & if\ v \in S\,; \\ 1 - \prod\limits_{\omega \in \mathcal{I}_v \cap S} (1 - p_{\omega,v}) & otherwise. \end{cases}$$

Tang, Jing, Xueyan Tang, and Junsong Yuan. "Influence Maximization Meets Efficiency and Effectiveness: A Hop-Based Approach." *advances in social networks analysis and mining* (2017): 64-71.

Hop-Based Approach



h-hops of propagation

THEOREM : For any $h \geq 1$, the influence spread produced within h hops of propagation is submodular and monotone under the IC model.

Algorithm Analysis

Under IC model, the solution S_h returned by the hop-based greedy algorithm satisfy:

$$\sigma(S_h) \geq \left(\left(1 - \frac{1}{e}\right)\alpha\right) \cdot \sigma(S^*)$$

Where the ratio $\alpha \leq \frac{\sigma_h(S)}{\sigma(S)} \leq \frac{\sigma(S_h)}{\sigma(S)}$.

Experiments

Dataset: Stanford SNAP

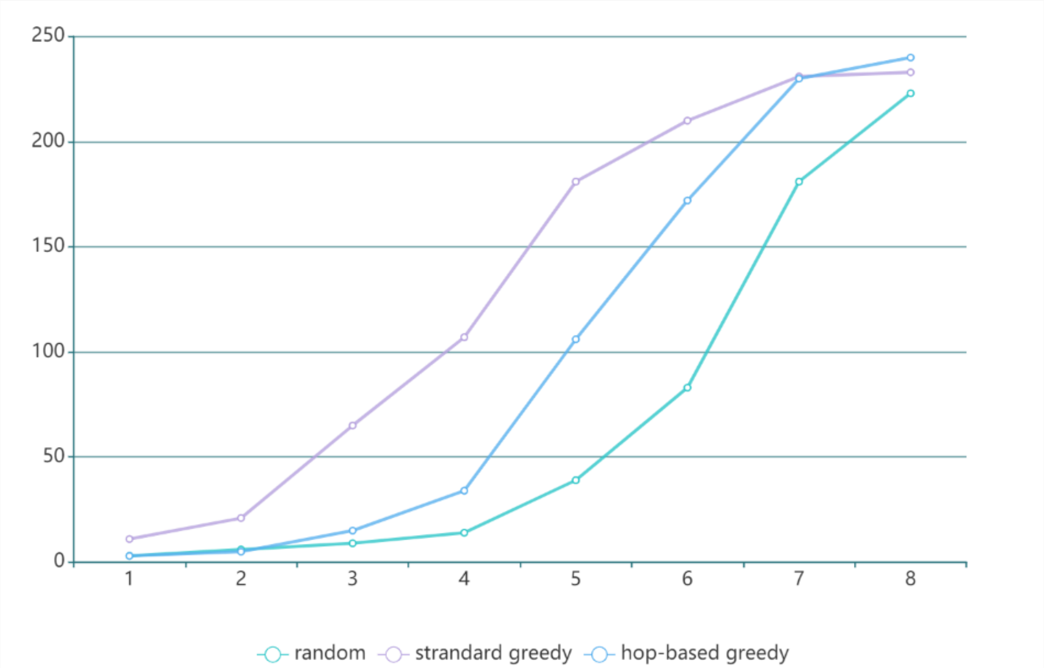
12831.circles	2012/11/8 5:44	CIRCLES 文件	1 KB
12831.edges	2012/11/8 5:44	EDGES 文件	40 KB
12831.egofeat	2012/11/8 5:44	EGOFEAT 文件	3 KB
12831.feat	2012/11/8 5:44	FEAT 文件	653 KB
12831.featnames	2012/11/8 5:44	FEATNAMES 文件	20 KB

- Every user (or say node) has 5 files in the dataset folder, including its circles, edges to other nodes and features.
- Considering the privacy of users, these file have been "anonymized".

Dataset statistics	
Nodes	81306
Edges	1768149
Nodes in largest WCC	81306 (1.000)
Edges in largest WCC	1768149 (1.000)
Nodes in largest SCC	68413 (0.841)
Edges in largest SCC	1685163 (0.953)
Average clustering coefficient	0.5653
Number of triangles	13082506
Fraction of closed triangles	0.06415
Diameter (longest shortest path)	7
90-percentile effective diameter	4.5

J. McAuley and J. Leskovec. Learning to Discover Social Circles in Ego Networks. NIPS, 2012.

Simulation results



- Using small social network including about 250 users and each user has about 1,000 features.

- In the first few time slices, the One-hop greedy algorithm is in the lead.

- In latter time slice, the Two-hop greedy surpasses.

Analyses

Problems:

Due to the randomness of dynamic network model, its great approximation ratio is in the sense of average case. We need

Algorithms designed for static network may not perform stably in dynamic network settings.

Possible solutions:

Run the simulation program for several times, and compute the average results. (Actually not that practical)

Consider the much more refined algorithms. Such as, consider the case where the current graph is also relevant to the former several latent spaces instead of just one latent space before it.

THANK YOU