SIAM International Conference on DATA MINING [SDM18]
May 3-5, 2018
San Diego Marriott Mission Valley, San Diego, California, USA

SIAM International Conference on Date Mining(SDM'18),
May 3-5, 2018 San Diego, California, USA

Maximizing the Effect of Information Adoption: A general Framework

Tianyuan Jin, Tong Xu, Hui Zhong, Enhong Chen,
Zhefeng Wang, Qi Liu

Anhui Province Key Laboratory of Big Data Analysis and Application
University of Science and Technology of China

Reporter:Tianyuan Jin

jty123@mail.ustc.edu.cn

Outline

- **Backgroud**
- Model of adoption spread
- Adoption maximization problem
  - Algorithm
  - Proof of performance bound
- Experiments
  - Date and setting
  - Results
- Conclusion

Background

- Social network
  - Social networks like Facebook, Twitter, WeChat become more and more popular
    - Users share feelings, photos and chat with friends
- Information Diffusion
  - In the diffusion process, people influence each other
  - Many application
    - Viral Marketing
    - Recommender Systems
    - Feed Ranking

Traditional diffusion model

- Basic information diffusion models
  - Independent Cascade model
  - Linear Threshold model
  - Triggering model

- Two type of nodes
  - Active nodes
    - Get the information and immediately adopt it
  - Inactive nodes
    - Can not get the information and never adopt it

Motivation

- Motivation example: You *organized a party and tweeted the invitation. Your friend Irene retweeted this message, but unfortunately she had no time to attend. On the contrary, another friend Abel, though not retweeted the message, finally participated the party.*
  - Irene acted as information channel without participation(adoption)
  - Abel, participation happened without information spread

The information diffusion and adoption should be treaded separately!!!

Outline

- Backgroud
- **Adoption spread model**
- Adoption maximization problem
  - Algorithm
  - Proof of performance bound
- Experiments
  - Date and setting
  - Results
- Conclusion

Adoption spread

- Diffusion stage
  - Independent Cascade model
    - Given seed set **S**
    - When node v becomes active, it has a single chance of activating each currently inactive neighbor w
    - After diffusion process ends, get **Active(S)**
- Adopt stage
  - Each node get influence from its active neighbors and the conditions(active or inactive) of it itself
    - Define: $A_u = u \cup N^{in}(u) \cap \mathbf{Active}(S)$

$$\mathbf{Adopt}(S) = \bigcup_{u \in V} \boxed{[f_u(A_u)]}$$

Adoption maximization task

- Adoption spread of node set S: F(S)
  - expected number of adopt 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

Special cases

- Influence maximization problem[kdd 2003]:

$$f_v(A_v) = \begin{cases} 1 & \text{if } v \in A_v \\ 0 & \text{if } v \notin A_v. \end{cases}$$

- Information coverage maximization problem[IJCAI 2015]:

$$f_v(A_v) = \begin{cases} 1 & \text{if } A_v \neq \emptyset \\ 0 & \text{if } A_v = \emptyset. \end{cases}$$

Outline

- Backgroud
- Model of adoption spread
- Adoption maximization problem
  - Properties
  - Algorithm
- Experiments
  - Date and setting
  - Results
- Conclusion

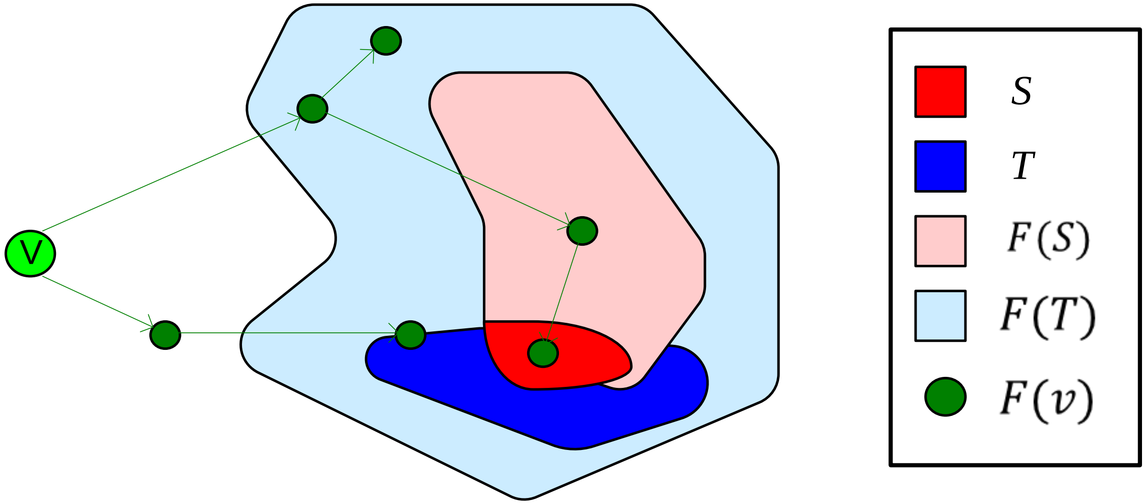
F(S): properties

- Non-negative (obviously)
- Monotone: $F(S+v) \geq F(S)$
- Submodular:
  - Let N be a finite set
  - A set function $F: 2^N \rightarrow R$ is submodular *iff*

$$\forall S \subset T \subset N, \quad \forall v \in N \backslash T,$$
$$F(S+v) - F(S) \geq F(T+v) - F(T)$$

F(S): properties

- Diminishing returns



- *Theorem: In IC model, F is submodular iff f is submodular*

Bad news 1

- For a submodular function F, if F only takes non-negative value, and is monotone, finding a k-element set S for which $F(S)$ is maximized is an NP-hard optimization problem

- It is NP-hard to determine the optimum for influence maximization

Good news

- We can use Greedy Algorithm!
  - 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?
  - 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 adopt nodes that any size-k set S could influence.

Bad news 2

- Theorem: computing $F(S)$ is #p-hard
  - Monte Carlo method is used to estimate $F(S)$

Outline

- Backgroud
- Model of adoption spread
- Adoption maximization problem
  - Properties
  - Algorithm
- Experiments
  - Date and setting
  - Results
- Conclusion

Polling based algorithm

- Construct the unbiased estimation of $F(S)$
  - For each $v \in U_v$, simulating information diffusion in G^T to find its potential influencers.
- Sampling method ← Consistent sampling

Algorithm 1 RR sampling

1: Initialize $\mathcal{L}$=$\{\mathcal{R}_1, \mathcal{R}_2, \cdots, \mathcal{R}_M\}$
2: **for** $\lambda = 1$ to M **do**
3: Choose a node v from G uniformly at random
4: Let $U_v = \{v_1, \cdots, v_j\}$.
5: **for** $i = 1$ to j **do**
6: Simulate information spread, starting from v_i in G^T and keep the edge results (live edges and dead edges) for the next simulations.
7: Let $R_v(v_i)$ be the set of nodes discovered in the simulation process
8: Add $R_v(v_i)$ to the $\mathcal{R}_\lambda$
9: **end for**
10: **end for**
11: **return** $\mathcal{L}$.

Greedy strategy

- Selecting seed set
  - Choose the seed set greedily
    - In each iteration, we add the node that has the largest marginal gain of adoption

Algorithm 3 optimal seed set selection

Initialize a set $S = \emptyset$
for $j = 1$ to k **do**
  find the node u such that

$$u = \underset{u \in (V \setminus S)}{arg\ max}\, W(S \cup u)$$

  add u into S.
end for
return S

Time complexity

- *Theorem: If we can compute* f_v *in* $O(1)$ *time, then we can get a* $1-\frac{1}{e}-\varepsilon$ *approximate solution with at least* $1-\frac{1}{n}$ *probability in* $O(knm\log n/OPT\varepsilon^2)$ *time*

- *More generally, for triggering function* f *we have better results, the time complexity will reduce to* $O(k(n+m)\log n/\varepsilon^2)$

- **The triggering adoption function** Each node v independently choose a random "triggering set" T_v according to some distribution over the set U_v. In each sampled graph g, if $A_v \cap T_v \neq \emptyset$, then $f_v = 1$, i.e., the node v will adopt the information.

Outline

- Backgroud
- Model of adoption spread
- Adoption maximization problem
  - Properties
  - Algorithm
- Experiments
  - Date and setting
  - Results
- Conclusion

Experiments Settings

- Datasets

dataset	#node	#link	#Average degree
Wiki	7,115	103,689	14.57
Epinions	75,879	508,837	6.71
Slashdot	82,168	948,464	11.54
Weibo	76,491	9,572,897	125.15

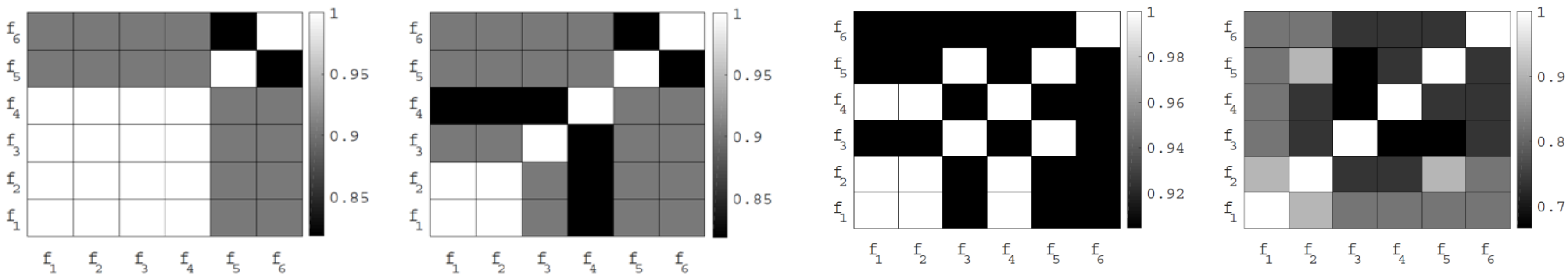
- Adoption functions
  - $f_{1,2,3,4} \rightarrow$ real life adoption maximization with parameter $\theta = 0.1, 0.3, 0.5, 0.7$
  - $f_5 \rightarrow$ influence maximization
  - $f_6 \rightarrow$ information converge maximization

Experiments Settings

- Baseline
  - LFG: a greedy algorithm for adoption maximization problem with lazy forward strategy
  - Simulate the process 10000 times for each targeted set

- Edge from v to w has probability $\frac{1}{indegree(w)}$ of activating w

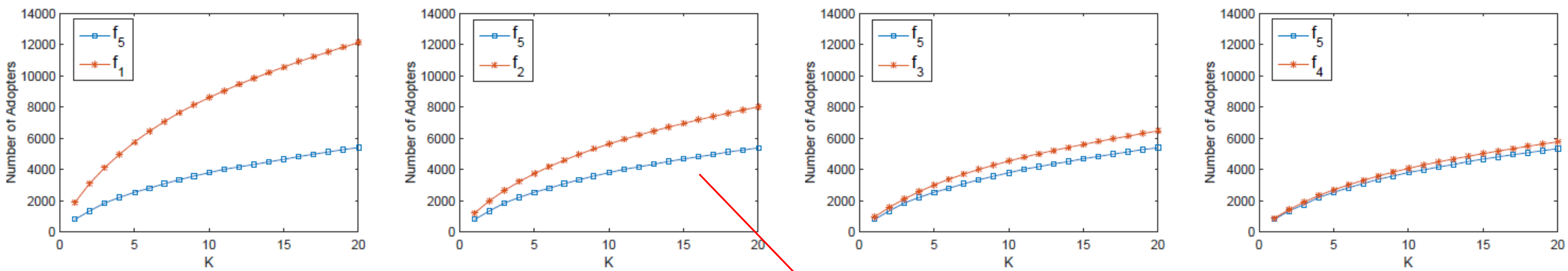
Seed set comparison



(a) Wiki (b) Epinions (c) Slashdot (d) Weibo

- Seed set similarity
  - Method: for k=20, compere the seed sets obtained by different adoption functions
  - Metric: Jaccard similarity coefficient
  - Results: different f select different seed set

Adoption comparison

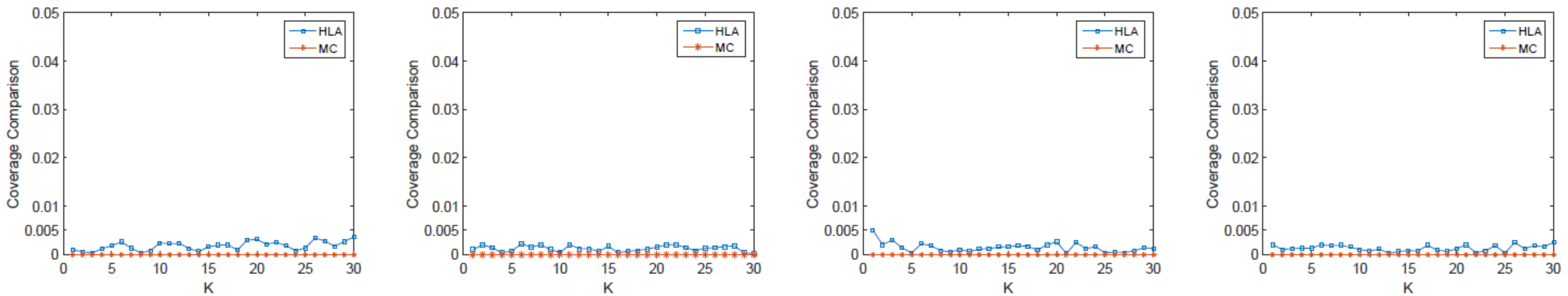


(a) $f_{5,1}$ (b) $f_{5,2}$ (c) $f_{5,3}$ (d) $f_{5,4}$

Corresponds to Influence maximization problem

- Adoption spread comparison
  - Method: for k=1-20, compere the adoption spread of different functions
  - Results: it is possible to achieve more adopt nodes than classical prediction

Effectiveness validation



(a) Wiki (b) Epinions (c) Slashdot (d) Weibo

- Effectiveness validation
  - Method: compare the adoption spread of different computing method
  - Results: our estimation for F(S) is indeed unbiased

Efficiency comparison

Table 4: Efficiency with $f = f_3$(in seconds)

Data set	Wiki	Epinions	Slashdot	Weibo
HLA	586	742	1,822	11,641
LFG	15,334	20,567	49,568	/

- General adoption function: f_3

Table 5: Efficiency with $f = f_6$(in seconds)

Data set	Wiki	Epinions	Slashdot	Weibo
HLA	13	174	204	6700
LFG	881	940	3200	/

- Triggering adoption functions: f_6

Outline

- Backgroud
- Model of adoption spread
- Adoption maximization problem
  - Properties
  - Algorithm
- Experiments
  - Date and setting
  - Results
- Conclusion

Conclusion

- **Motivation**: Two basic functions of social spread, i.e., information propagation and information adoption
- **Problem**: Adoption maximization problem
- **Contributions**:
  - Some properties
  - Polling based algorithm
  - extensive experiments demonstrates the effectiveness and efficiency of proposed algorithms

Q & A

Thanks !

Reporter: Tianyuan Jin
jty123@mail.ustc.edu.cn