Influence Maximization

LI Guanyao, ZOU Zhuojin

Outline

- Introduction and motivation
- Stochastic diffusion models
- Influence maximization
- Algorithms
- Conclusion

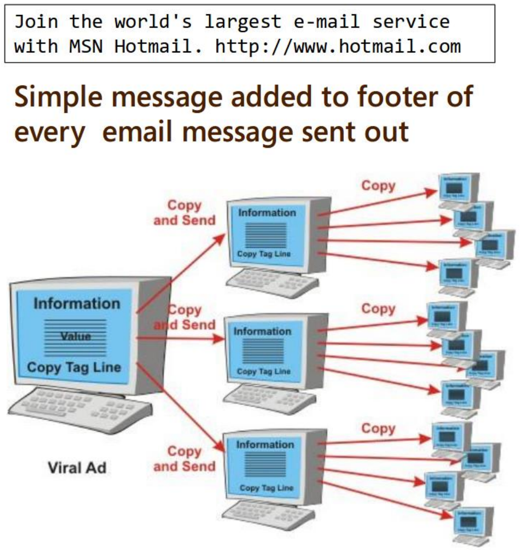
Social influence

- Wikipedia definition: Social influence occurs when one's opinions, emotions, or behaviors are affected by others, intentionally or unintentionally.

Booming of online social networks

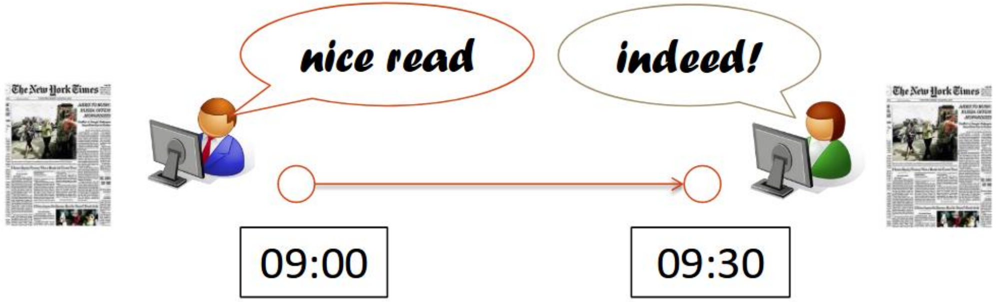
Hotmail: online viral marketing story

- Hotmail’s viral climbed to the top spot (90s): 8 million users in 18 months
- Boosted brand awareness
- Far more effective than conventional advertising by rivals
    - and far cheaper



Stochastic Diffusion Models

Influence propagation



People are **connected** and perform **actions**

friends, fans, followers, etc.

comment, link, rate, like, retweet, post a message, photo, or video, etc.

Terminologies

- Directed graph $G = (V, E)$
  - Node $v \in V$ represents an individual
  - Edge $(u, v) \in E$ represents the influence relationship
- Discrete time t: 0, 1, 2, …
- State of nodes: active or inactive
- S_t: set of active nodes at time t
  - S_0, seed set, initial nodes selected to start the diffusion

Stochastic diffusion models

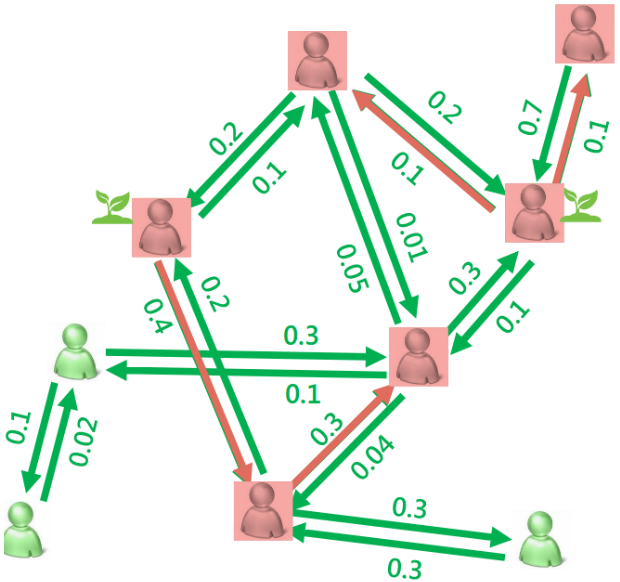
Definition: A *stochastic diffusion model (with discrete time steps)* for a social graph $G = (V, E)$ specifies the randomized process of generating active sets S_t for all $t > 0$ given the initial seed set S_0.

Progressive models: for all $t > 0$, $S_{t-1} \subseteq S_t$

- Once activated, always activated, e.g. once bought a product, cannot undo it.
- **Influence spread σ(S)** : expected number of activated nodes when the diffusion process starting from the seed set S ends

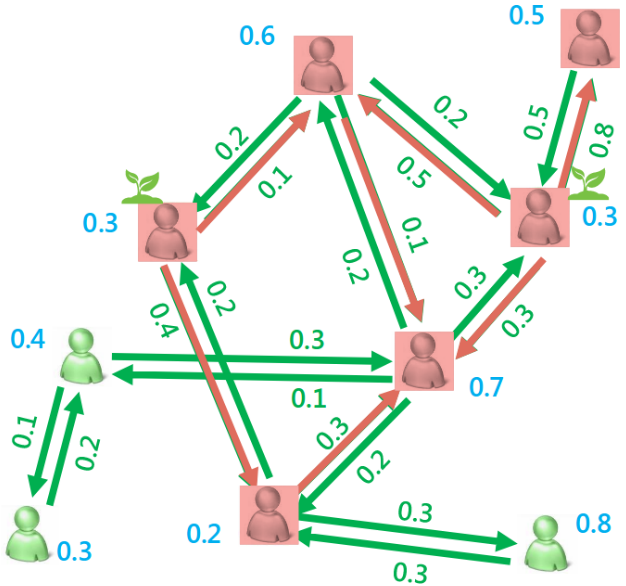
Independent cascade model

- Each edge (u, v) has a influence probability p(u, v)
- Initially seed nodes in S_0 are activated
- At each time step t, each node u activated at step t-1 activates its neighbor v independently with probability p(u, v)



Linear threshold model

- Each edge (u, v) has an influence weight w(u, v):
  - if $(u, v) \notin E$, $w(u, v)=0$
  - $\Sigma_u\, w(u, v) \leq 1$
- Each node v selects a threshold $\theta_v \in [0,1]$ uniformly at random
- Initially seed nodes in S_0 are activated
- At each step, node v checks if the weighted sum of its activated in-neighbors is greater than or equal to its threshold θ_v, if so v is activated



Interpretation of IC and LT models

- IC model reflects simple contagion, e.g. information, virus
- LT model reflects complex contagion, e.g. product adoption, innovations

  (activation needs social affirmation from multiple sources [1])

[1] Centola, Damon, and Michael Macy. "Complex contagions and the weakness of long ties." *American journal of Sociology 2007.*

General threshold model

- Each node v has a threshold function
$$f_v: 2^V \to [0,1]$$
- Each node v selects a threshold $\theta_v \in [0, 1]$ uniformly at random
- If the set of active nodes at the end of step t-1 is S, and $f_v(S) \geq \theta_v$, v is activated at step t
- Reward function r(A(S)): if A(S) is the final set of active nodes given seed set S, r(A(S)) is the reward from this set
- Generalized influence spread:
$$\sigma(S) = E[r(A(S))]$$

Submodularity in the general threshold model

- Properties: submodularity and monotonicity
- Theorem
  - In the general threshold model,
  - if for every v ∈ V, f is monotone and submodular with f(Ø)=0
  - and the reward function r is monotone and submodular
  - then the general influence spread funnction σ is monotone and submodular

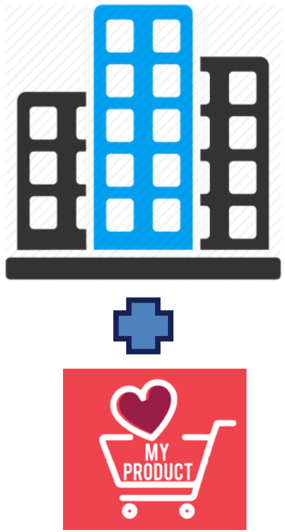
[2] Mossel, Elchanan, and Sebastien Roch. On the submodularity of influence in social networks. STOC, 2007.

Influence Maximization

Problem formulation

- Given a social network, a diffusion model with given parameters, and a number k, find a seed set S of at most k nodes such that the influence spread of S is maximized.
- May be further generalized:
  - Instead of k, given a budget constraint and each node has a cost of being selected as a seed
  - Instead of maximizing influence spread, maximizing a (submodular) function of the set of activated nodes.

Example



Hardness of influence maximization

- Influence maximization under both IC and LT models are NP hard
  - IC model: reduced from k-max cover problem
  - LT model: reduced from vertex cover problem
- Need approximation algorithms

Greedy-based Algorithm

Greedy-based Algorithm

- Influence Spread
  - Given a seed set S, the influence spread of S means that
    - the expected number of influenced users after the propagation terminates if the users in S were selected as seed users
  - $f(S)$
- Marginal gain
  - The influence from users has overlap
    - Some of the users influenced by user u may be the same as the users influenced by the seed set S, where $u \notin S$
  - The Marginal gain of a user u means that
    - If user u was selected into the seed set, how much extra influence spread it would contribute
  - $\Delta(u|S) = f(S \cup \{u\}) - f(S)$

Greedy-based Algorithm

- Iteratively selects a node that provides the maximum marginal gain in each round

Algorithm 1 Greedy(k, f)

```
initialize S = ∅
for i = 1 to k do
    select u = arg max_{w∈V\S} (f(S ∪ {w}) − f(S))
    S = S ∪ {u}
end for
output S
```

$$\text{select } u = \arg\max_{w \in V \setminus S} \underbrace{(f(S \cup \{w\}) - f(S))}_{\text{marginal gain}}$$

Greedy-based Algorithm

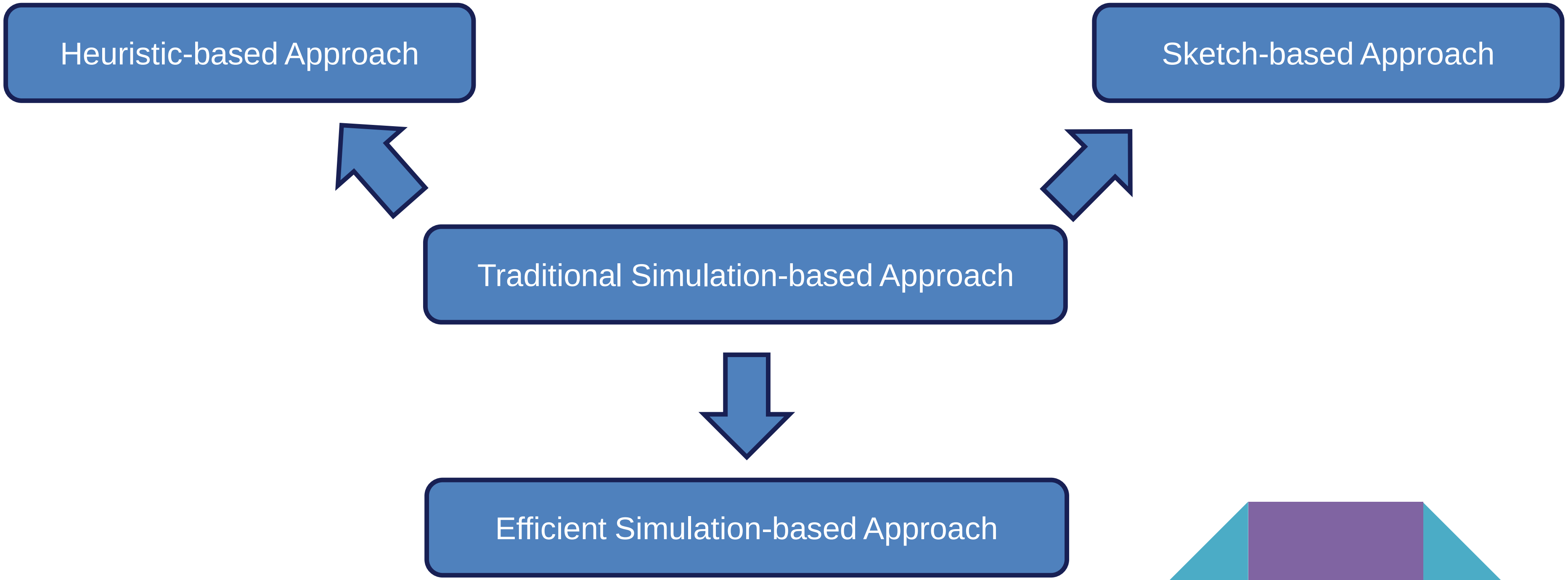
- Submodular
  - $f(S_j \cup \{u\}) - f(S_j) \leq f(S_i \cup \{u\}) - f(S_i)$ $where\ S_i \subseteq S_j$
- Monotone
  - $f(S_i) < f(S_j)$
- The greedy-based approach solves the influence maximization problem with an approximation ratio of 1 − 1/e.[2]

[2] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. KDD 2003.

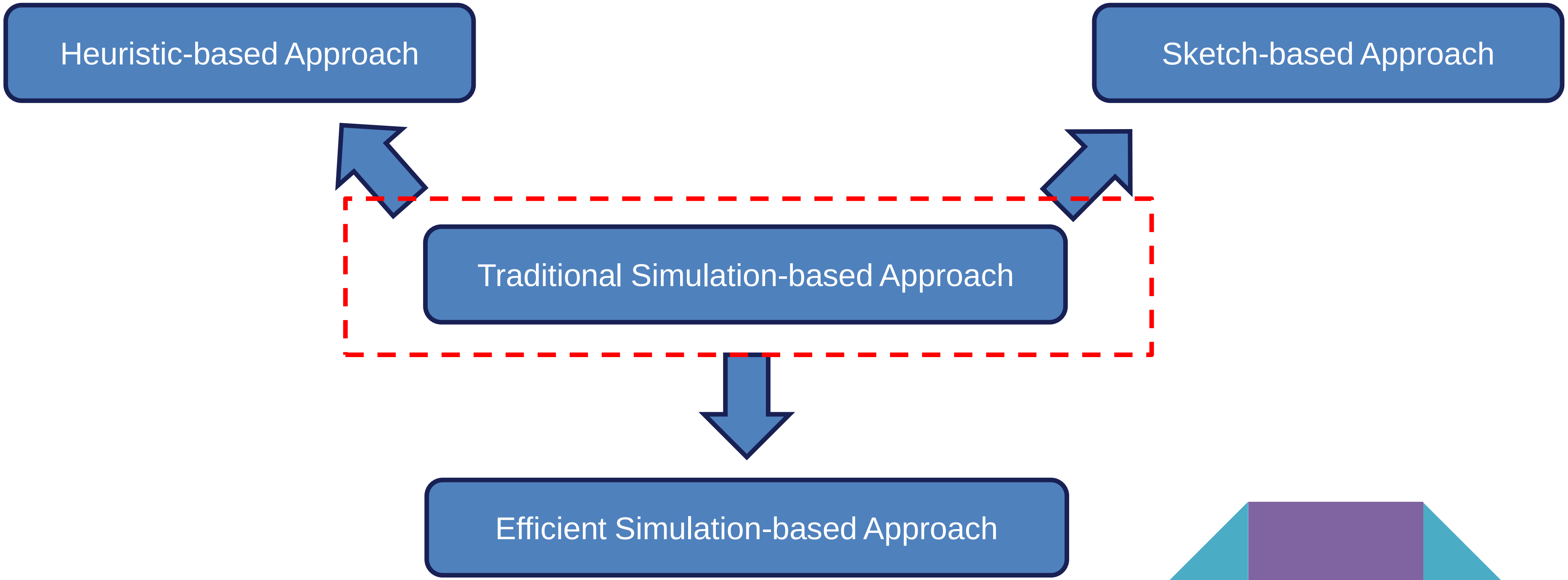
Greedy-based Algorithm

How to estimate the influence spread?

Overview


Heuristic-based Approach
Sketch-based Approach
Traditional Simulation-based Approach
Efficient Simulation-based Approach

Overview


Heuristic-based Approach
Sketch-based Approach
Traditional Simulation-based Approach
Efficient Simulation-based Approach

Simulation-based approach

- The approach utilizes Monte Carlo(MC) simulations for estimating influence spread.
  - Starts from the seed set S
  - Simulates the activation process wrt. the corresponding diffusion model
  - Outputs the number of activated users

Traditional Simulation-based Approach

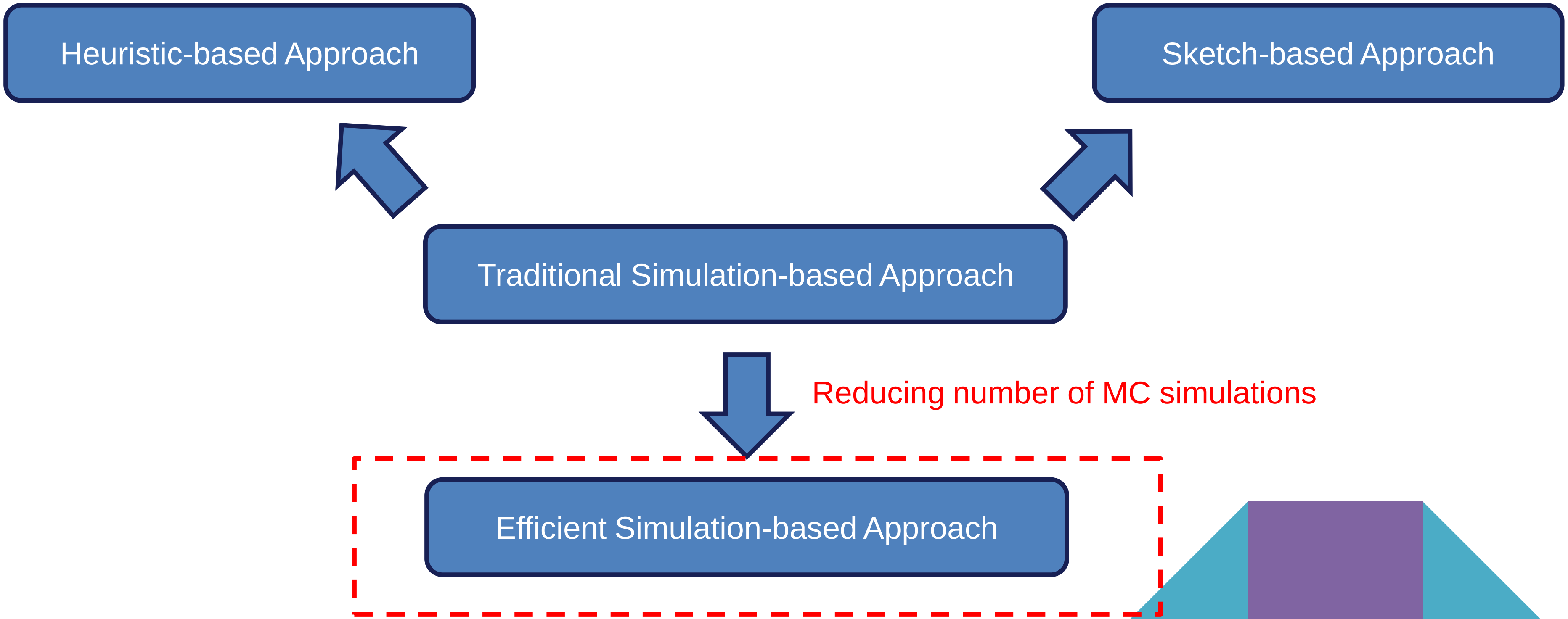
Algorithm 1 GeneralGreedy(G, k)

```
initialize S = ∅ and R = 20000
for i = 1 to k do
    for each vertex v ∈ V \ S do
        s_v = 0.
        for i = 1 to R do
            s_v += |RanCas(S ∪ {v})|
        end for
        s_v = s_v / R
    end for
    S = S ∪ {arg max_{v ∈ V \ S} {s_v}}
end for
output S.
```

Time complexity: O(knRm)

[2] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. KDD 2003.

Overview


Heuristic-based Approach
Sketch-based Approach
Traditional Simulation-based Approach
Reducing number of MC simulations
Efficient Simulation-based Approach

Traditional Simulation-based Approach

Algorithm 1 GeneralGreedy(G, k)

initialize $S = \emptyset$ and $R = 20000$
for $i = 1$ to k **do**
  for each vertex $v \in V \setminus S$ **do**
    $s_v = 0.$
    for $i = 1$ to R **do**
      $s_v \mathrel{+}= |RanCas(S \cup \{v\})|$
    end for
    $s_v = s_v / R$
  end for
  $S = S \cup \{\arg\max_{v \in V \setminus S}\{s_v\}\}$
end for
output S.

Update the marginal gain of all unselected nodes

Time complexity: O(knRm)

[2] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. KDD 2003.

Traditional Simulation-based Approach

Algorithm 1 GeneralGreedy(G, k)

initialize $S = \emptyset$ and $R = 20000$
for $i = 1$ to k **do**
  for each vertex $v \in V \setminus S$ **do**
    $s_v = 0.$
    for $i = 1$ to R **do**
      $s_v \mathrel{+}= |RanCas(S \cup \{v\})|$
    end for
    $s_v = s_v / R$
  end for
  $S = S \cup \{\arg\max_{v \in V \setminus S} \{s_v\}\}$
end for
output S.

Time complexity: O(knRm)

Update the marginal gain of all unselected nodes

Necessary?

[2] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. KDD 2003.

Traditional Simulation-based Approach

Algorithm 1 GeneralGreedy(G, k)

initialize $S = \emptyset$ and $R = 20000$
for $i = 1$ to k **do**
  for each vertex $v \in V \setminus S$ **do**
    $s_v = 0.$
    for $i = 1$ to R **do**
      $s_v \mathrel{+}= |RanCas(S \cup \{v\})|$
    end for
    $s_v = s_v / R$
  end for
  $S = S \cup \{\arg\max_{v \in V \setminus S}\{s_v\}\}$
end for
output S.

Time complexity: O(knRm)

Update the marginal gain of all unselected nodes

Necessary?

No!!!

[2] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. KDD 2003.

CELF algorithm

- CELF[3] estimates an upper bound of the marginal gain to avoid updating the marginal gain of all nodes

[3] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance. Cost-effective outbreak detection in networks. KDD 2007

CELF algorithm

- Submodularity
  - $f(S_j \cup \{u\}) - f(S_j) \leq f(S_i \cup \{u\}) - f(S_i)$ where $S_i \subseteq S_j$
  - $\Delta(u|S_i) = f(S_i \cup \{u\}) - f(S_i)$ is an upper bound for any $\Delta(u|S_j)$ where $S_i \subseteq S_j$

- If the upper bound of node u is less than the updated marginal gain of node v,
  - the updated marginal gain of u is less than $v's$
  - no need to update the marginal gain of u

CELF algorithm

- Updates the marginal gain of nodes in descending order according to their upper bound
- Early terminates whenever the upper bound of a not updated node is less than the marginal gain of the updated node

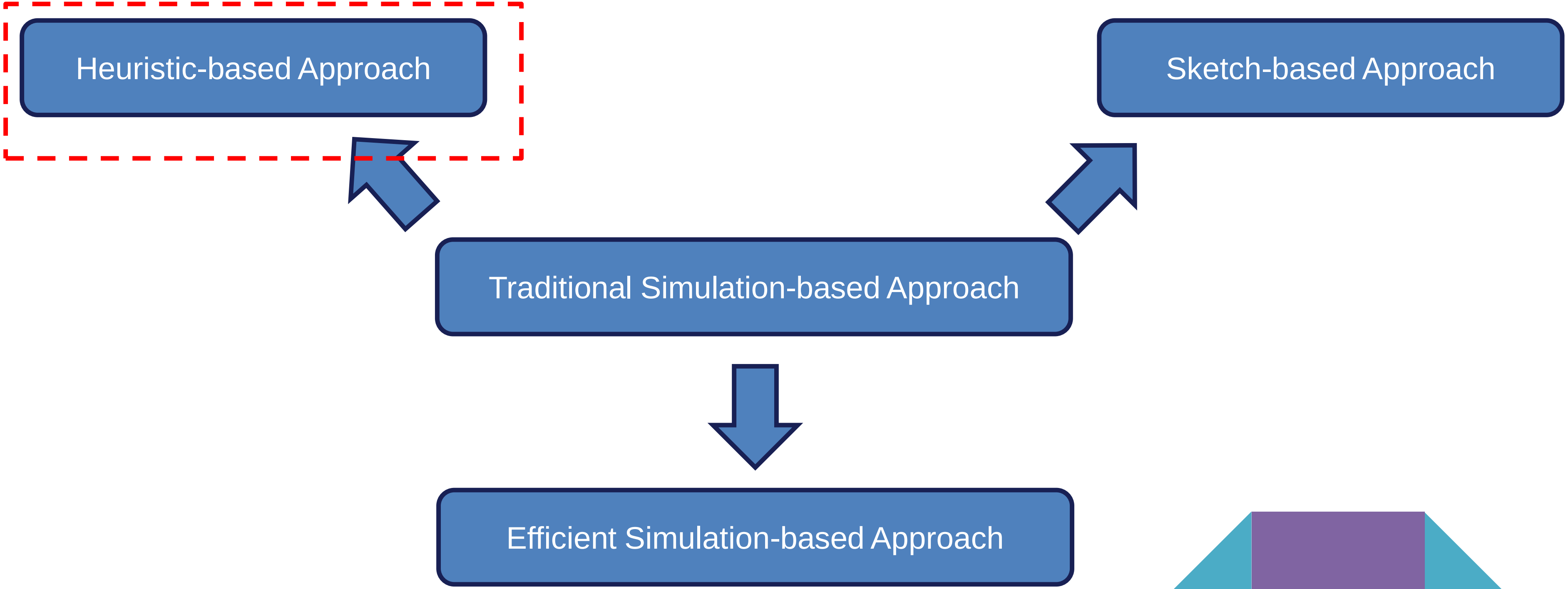
CELF algorithm

- Updates the marginal gain of nodes in descending order according to their upper bound
- Early terminates whenever the upper bound of a not updated node is less than the marginal gain of the updated node

Enables an up to 700 times improvement compared with the traditional algorithm.

Overview

Estimate the influence spread in a heuristic way



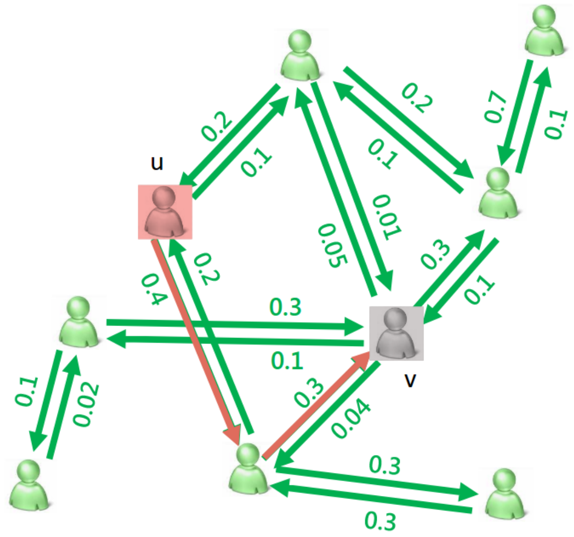
Heuristic approaches

- Exact influence computation is NP hard, for both IC and LT models -- computation bottleneck [4][5]
- Design new heuristics
  - MIA for general IC model [4]
  - LDAG for LT model [5]
  - IRIE for IC model [6]

[4] Chen, Wei, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. KDD 2010
[5] Chen, Wei, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. ICDM 2010
[6] Jung, Kyomin, Wooram Heo, and Wei Chen. Irie: Scalable and robust influence maximization in social networks. ICDM 2012

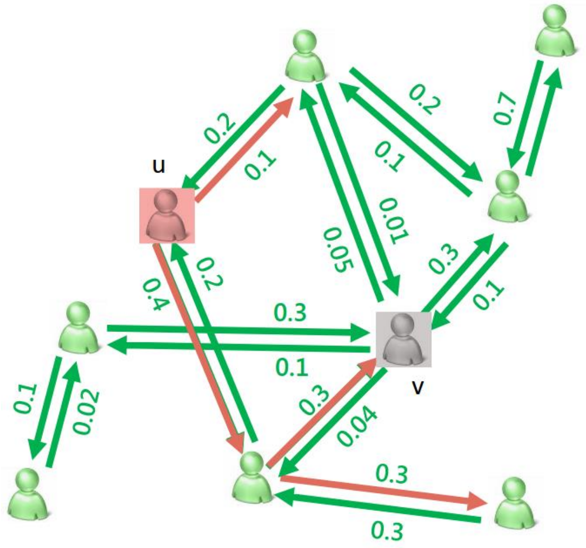
Maximum Influence Arborescence (MIA)

- For any pair of nodes u and v, find the maximum influence path (MIP) from u to v
- Ignore MIPs with too small probability (< parameter θ)



MIA (cont'd)

- Local influence regions
  - for every node v, all MIPs to v form its maximum influence in-arborescence (MIIA)
  - for every node u, all MIPs from u form its maximum influence out-arborescence (MIOA)
  - computing MIAs and the influence through MIAs is fast



MIA (cont’d)

- Recursive computation of activation probability ap(u) of a node u in its in-arborescence, given a seed set S

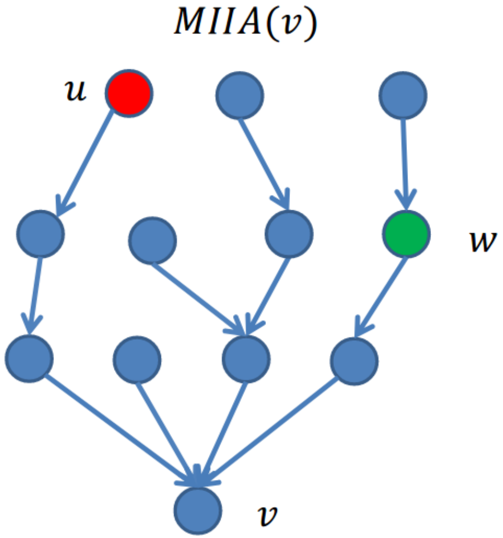
Algorithm 2 $ap(u, S, MIIA(v, \theta))$

if $u \in S$ **then**
    $ap(u) = 1$
else if $Ch(u) = \emptyset$ **then**
    $ap(u) = 0$
else
    $ap(u) = 1 - \Pi_{w \in Ch(u)}(1 - ap(w) \cdot pp(w, u))$
end if

- Can be us … eds, but not efficient enough

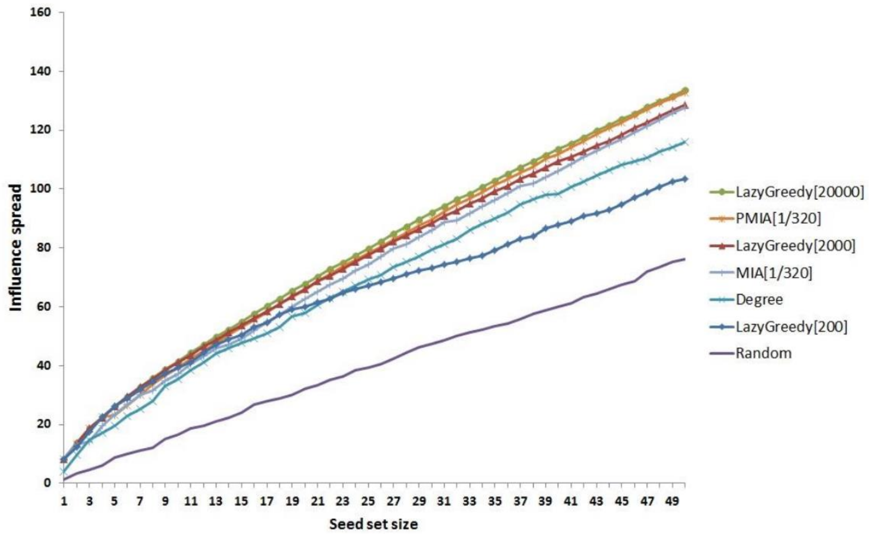
MIA (cont’d)

- u is the new seed in MIIA(v)
- Naive update: for each candidate W, redo the computation in the previous page to compute w’s incremental influence to v
  - O(|MIIA(v)|^2)
- Fast update: based on linear relationship of activation probabilities between any node w and root v, update incremental influence of all w’s to v in two passes
  - O(|MIIA(v)|)

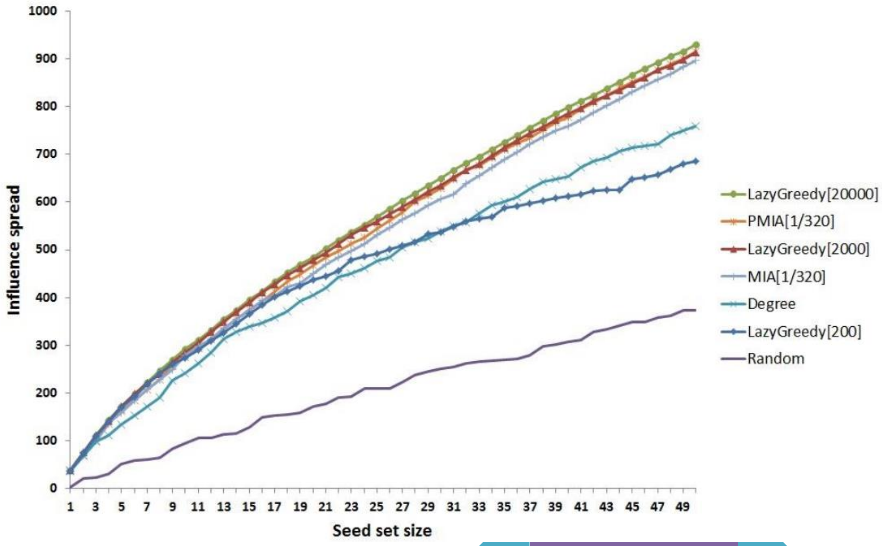


Experiment result

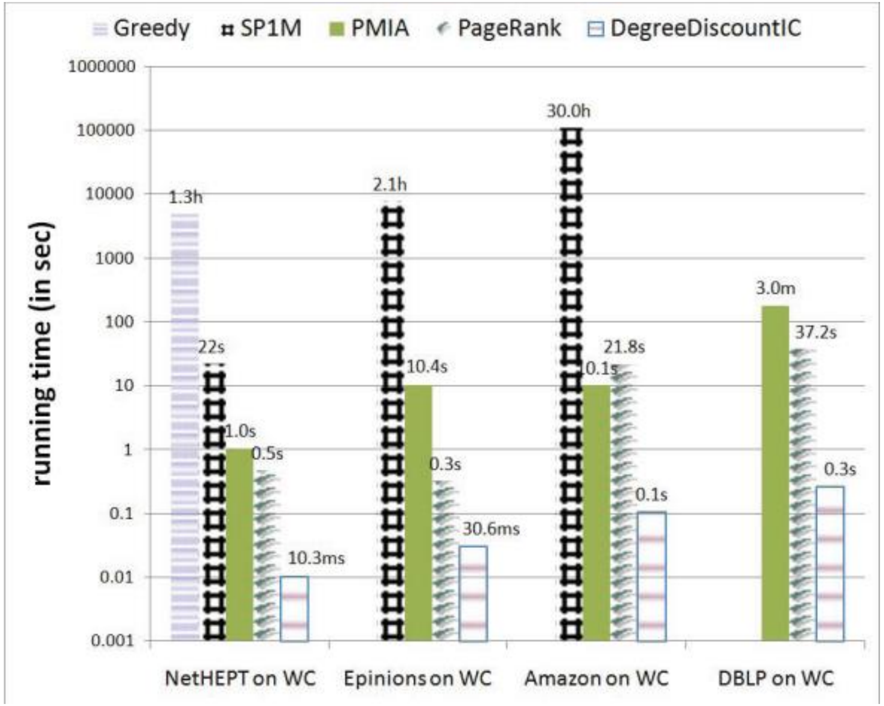
Influence spread in IC-UP[0.01] model



Influence spread in IC-WC model

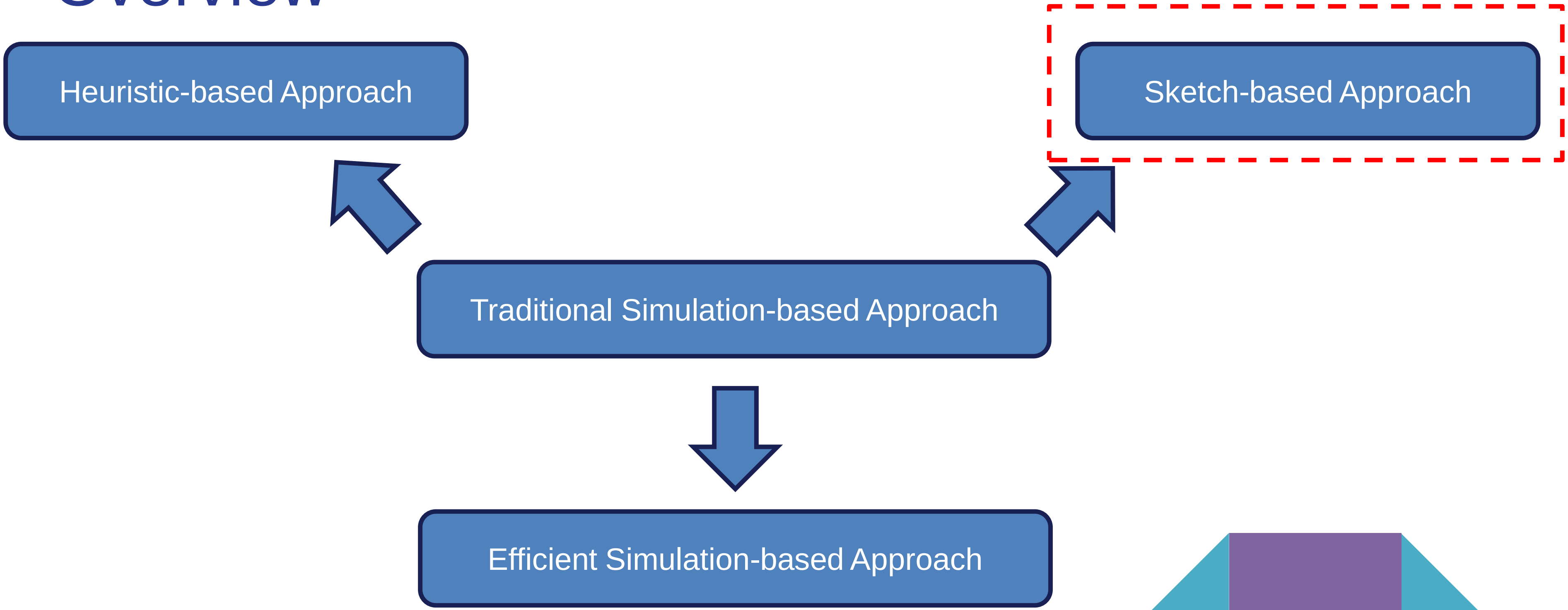


Experiment result



Estimate the influence spread based on the pre-computed sketch

Overview



Sketch-based Approach

- Let $v \in V$ be a node in G, and g be a graph obtained by removing each edge e in G with $1 - p(e)$ probability
  - The reverse reachable(RR) set
    - $RR(v)$: contains all nodes in g that can reach v
  - The Radom RR set
    - For $RR(v)$, v is randomly picked from V

C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in SODA, 2014
Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Nearoptimal time complexity meets practical efficiency," in SIGMOD, 2014, pp. 75–86.

Sketch-based Approach

- If a seed set S^* covers the maximal number of the RR set, S^* is likely to be the optimal seed set.

Algorithm 2: RR-Sketch (G, k, θ) **[100]**

Input : $G = (V, E)$: A social graph k: A number;
θ: Number of RR Sets.

Output: S: Seed Set.

$\mathcal{R} \leftarrow \emptyset, S \leftarrow \emptyset$
Generate θ random RR sets and insert them into $\mathcal{R}$
for $i = 1, \ldots, k$ **do**
    Pick node v_i that covers the most RR sets in $\mathcal{R}$
    Add v_i into S
    Remove from $\mathcal{R}$ all RR sets that are covered by v_i
return S

Conclusion

Pros: Efficient
Cons: Without theoretical guarantees

Heuristic-based Approach

Pros:
1. Efficient
2. Theoretical guarantees

Cons: Not general

Sketch-based Approach

Traditional Simulation-based Approach

Efficient Simulation-based Approach

Pros:
1. General
2. Theoretical guarantees

Cons: Inefficient

Future directions(I)

- Context-aware influence maximization
  - Topic-aware influence maximization[9]
  - Location-aware influence maximization[10][11]

[9] S. Chen, J. Fan, G. Li, J. Feng, K. Tan, and J. Tang, “Online topic-aware influence maximization,” PVLDB, 2015

[10] G. Li, S. Chen, J. Feng, K. Tan, and W. Li, “Efficient location-aware influence maximization,” in SIGMOD, 2014

[11] X. Wang, Y. Zhang, W. Zhang, and X. Lin, “Distance-aware influence maximization in geo-social network,” in ICDE, 2016

Future directions(II)

- Influence probability inference
  - Influence probability between users is fundamental for influence spread estimation[12][13]

[12] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Learning influence probabilities in social networks. In ACM WSDM, 2010.

[13] W. Zhu, W. Peng, L. Chen, K. Zheng, and X. Zhou. Modeling user mobility for location promotion in location-based social networks. KDD 2015

Reference

[1] Centola, Damon, and Michael Macy. "Complex contagions and the weakness of long ties." American journal of Sociology 2007.

[2] D. Kempe, J. M. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. KDD 2003.

[3] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen and N. S. Glance. Cost-effective outbreak detection in networks. KDD 2007

[4] Chen, Wei, Chi Wang, and Yajun Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. KDD 2010

[5] Chen, Wei, Yifei Yuan, and Li Zhang. Scalable influence maximization in social networks under the linear threshold model. ICDM 2010

[6] Jung, Kyomin, Wooram Heo, and Wei Chen. Irie: Scalable and robust influence maximization in social networks. ICDM 2012

[7] C. Borgs, M. Brautbar, J. Chayes, and B. Lucier, "Maximizing social influence in nearly optimal time," in SODA, 2014

[8] Y. Tang, X. Xiao, and Y. Shi, "Influence maximization: Nearoptimal time complexity meets practical efficiency," in SIGMOD, 2014.

[9] S. Chen, J. Fan, G. Li, J. Feng, K. Tan, and J. Tang, "Online topic-aware influence maximization," PVLDB, 2015

[10] G. Li, S. Chen, J. Feng, K. Tan, and W. Li, "Efficient location-aware influence maximization," in SIGMOD, 2014

[11] X. Wang, Y. Zhang, W. Zhang, and X. Lin, "Distance-aware influence maximization in geo-social network," in ICDE, 2016

[12] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan. Learning influence probabilities in social networks. In ACM WSDM, 2010

[13] W. Zhu, W. Peng, L. Chen, K. Zheng, and X. Zhou. Modeling user mobility for location promotion in location-based social networks. KDD 2015

[14] Li Y, Fan J, Wang Y, et al. Influence Maximization on Social Graphs: A Survey. TKDE 2018

Thanks!