Targeted Influential Nodes Selection in Location-Aware Social Networks

Susu Yang, Hui Li*, and Zhongyuan Jiang

School of Cyber Engineering, Xidian University, Xi'an 710126, China

Correspondence should be addressed to Hui Li; hli@xidian.edu.cn

Received 2 September 2018; Accepted 10 October 2018; Published 1 November 2018

Guest Editor: Jianxin Li

Copyright © 2018 Susu Yang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Given a target area and a location-aware social network, the location-aware influence maximization problem aims to find a set of seed users such that the information spread from these users will reach the most users within the target area. We show that the problem is NP-hard and present an approximate algorithm framework, namely, TarIM-SF, which leverages on a popular sampling method as well as spatial filtering model working on arbitrary polygons. Besides, for the large-scale network we also present a coarsening strategy to further improve the efficiency. We theoretically show that our approximate algorithm can provide a guarantee on the seed quality. Experimental study over three real-world social networks verified the seed quality of our framework, and the coarsening-based algorithm can provide superior efficiency.

1. Introduction

In recent years, social networks have become prevalent platforms for the spread of product adoption, ideas, and news. Under this trend, influence maximization (IM) problem is becoming popular, which aims to seek k users (referred to as seeds) to maximize the number of influenced users (referred to as influence spread) in the network. Kempe et al. [1] proved that this problem is NP-hard and presented an \((1 − 1/e)\)-approximate algorithm by greedily selecting \(k\) seed users which has the maximum marginal gain of influence spread. Motivated by their work, a vast amount of studies then focused on improving the effect of influence spread and the efficiency, such as the heuristic-based algorithm PMIA [2] and the sketch-based algorithm IMM [3].

However, many real-world applications such as location-based word-of-mouth marketing recently have location-aware requirements in IM. In [4], Li et al. focused on a location-aware IM (LAIM) query, which aims to seek \(k\) users to maximize the expected influence in a query region. They assumed that the network and the location of users are given beforehand, where an index can be constructed offline, and the target region is submitted online as a query. However, such assumption may not always be satisfied as the network and IM request are given at the same time, which is exactly the scenario discussed in traditional IM works [2, 5, 6]. Besides, existing works in LAIM can only answer the problem towards simple regions such as a rectangle or a circle. However, locations of users are always complicatedly visualized and managed via maps, the atomic regions of which are not necessarily rectangles or circles. Instead, they always show up as various and complex polygons. Therefore, it is meaningful to find an efficient method to address the LAIM problem, by targeting at an arbitrary polygon region, from scratch. Below we provide a running example to elaborate this point.

Example. A company wants to sell a new product in a city. It is obvious that people in this city are potential buyers. In order to propagate this product to the public, we need to find several individuals who are the most influential in the network, and hope that, through their propagation, as many people as possible could know this product and then further purchase it.

In our paper, we present a novel algorithm framework, namely, TarIM-SF (Targeted Influence Maximization with Spatial Filtering), to deal with this problem. Given a location-aware social network and a target region, we firstly adopt a spatial filtering model (SFM) to identify the targeted users. Then we will utilize the elegant sampling approach in latest IM solutions [3] to find seed nodes. In order to further
improve the efficiency, we have coarsened the social network. In all, our contributions in this work are as follows:

(i) We relax the target region in LAIM to arbitrary polygons, which is more practical in real applications.

(ii) Secondly, our model can address LAIM from scratch without any assumption for offline processing.

(iii) To the best of our knowledge, we are the first to prove the hardness of LAIM theoretically in detail, and for the large-scale and complex networks, we propose a coarsening-based model that can further improve the efficiency with guaranteed seed quality.

Experiments on real-world datasets Gowalla, Tweets, and Weibo demonstrate that our framework could generate a seed set with theoretically guaranteed quality, which outperforms a series of baseline methods in terms of influence spread quality. Besides, the coarsening-based algorithm can provide superior efficiency.

The rest of the paper is organized as follows. Section 2 lists the related studies. Section 3 gives the definition of LAIM with proving its hardness and presents some fundamental knowledge. Afterwards, we discuss the proposed algorithm framework in Section 4. Section 5 shows the theoretical guarantee of seed quality. Section 6 reports the experimental results and some discussion. In Section 7, we conclude the paper.

2. Related Works

Kempe et al. [1] first formulated the influence maximization problem and proved that it is NP-hard in general, but can be approximated with \((1 - 1/e - \epsilon)\) factor. They presented a greedy algorithm with a provable approximation guarantee to solve this problem. However, the greedy algorithm needs to perform the Monte Carlo simulation [7] to obtain the approximate ratio, which has a large time overhead. Furthermore, in order to improve the efficiency and the effect, there has been a large body of research works that can be divided into three types. Simulation-based methods accurately estimate influence by simulating the diffusion process repeatedly with a theoretical guarantee. Leskovec et al. [8] proposed a CELF method with the lazy-forward heuristic, which is originally designed to optimize submodular functions in [9], as well as [10–14]. Heuristic-based methods are developed to avoid using Monte Carlo simulation at the expense of solution quality. For example, Chen et al. [2] proposed to use local directed acyclic graphs to approximate the influence regions of nodes, while [15] restricting the spread of influence into communities and [6] approximating the influence spread using linear systems. Sketch-based methods resolved the inefficiency of Monte Carlo simulations without loss of accurate guarantees. Borges et al. [16] presented a nearly optimal time algorithm for IM under IC model. This method relies on reverse simulations of the diffusion process and builds sketches to estimate the influence function efficiently. In subsequential works, techniques for bounding the sketches’ size are developed [3, 17–21] and [3, 18, 19] are the representative ones that exhibit higher efficiency in all sketch-based methods. Moreover, Liu et al. [22] construct a community-level influence analysis model, instead of focusing on individual-influence, while [23] defining the outer influence of a community and aiming to find the most influential communities, as well as [24] constructing an influential propagation model considering the temporal-interaction between users in the social network.

Recently, more additional demands for IM problem emerged, such as considering the interests of users [25, 26], geographical factor, or some factors of time. Especially in order to meet the location-aware requirements in IM, Li et al. [4] proposed a method to solve location-aware IM, which works by seeking \(k\) users to maximize the expected influence spread of the query region. Wang et al. [27] considered the distance between two users and defined the distance-aware IM problem. They proposed a priority based algorithm with \((1−1/e)\)-approximation ratio. The authors in [28] also studied the DAIM problem, considering the distance between the locations and the users. Zhu et al. [29] proposed Gaussian based and distance based mobility models, to derive the location-aware propagation probability in LBSN. Zhou et al. [30] take users’ historical mobility behaviour into account and study the IM problem under O2O model. Li et al. [31] aim to find several seed users to maximize geographic spanning regions (MGSR) in the query region, while Li et al. [32] assume that users have their location preference and solve the IM problem for the targeted users. Furthermore, some works focus on spatial-temporal IM problem [33, 34], which aims to find \(k\) best trajectories to be attached with an advertisement and maximizes the number of influenced users. Besides the location, the interests/topics of users are also taken into consideration in IM, and [35] proposed an algorithm that returns top-\(k\) topics related to the query of a user. Su et al. [36] take not only users’ interests but also their preference for locations into account, to find the targeted users, and then seek seeds to maximize the influence for targeted users.

3. Problem Statements and Preliminaries

3.1. Problem Definition

Definition 1 (LAIM). Given a location-aware social network \(G = (V, E)\) where each node \(v \in V\) is associated with a location (denoted as \(v.\phi\)), a budget \(k\), and a target region \(Q\), the location-aware influence maximization (LAIM) aims to find \(k\) seed nodes (denoted as \(S\)) from \(V\), such that the influence spread from \(S\) can reach the most number of nodes in \(Q\).

We show the hardness of LAIM problem under Independence Cascade (IC) model, which is one of the most popular diffusion models [1]. Before that, we first define a problem called Subset Cover, which will be utilized in the following content.

Definition 2 (subset cover). There are an element set \(U = \{u_1, u_2, \ldots, u_n\}\), a subset \(U_q = \{u_{q1}, u_{q2}, \ldots, u_{qn}\}\) of \(U\), and a collection of subsets \(S = \{S_1, S_2, \ldots, S_m\}\) of \(U\), and we wish to know whether there exist \(k\) subsets in \(S\), whose union is equal to \(U_q\).

Theorem 3. The location-aware influence maximization problem is NP-hard for IC model.
which makes different elements to with this, we are able to construct a corresponding directed graph. Problem, Set Cover problem can also be solved. Based on corresponding to each taking the set of nodes in \( \{Z_4, 7, 12, 9, 0, 5, 1, 10, \ldots\} \), \( \{Z_2, 12, 4, 7, 5, 9, 6, 8, \ldots\} \), propagation probability for edges between two user nodes.

3.2. Sampling Technique. Borgs et al. [16] introduced a sampling method called RIS, which first constructs a suitable number of sketches from different target nodes reversing DFS (referred to as RR sets \( \mathcal{R} \)) and then finds out \( k \) users as the seed nodes with the maximum coverage of \( \mathcal{R} \). The process of constructing RR sets is as follows.

Firstly, given an edge-weighted graph \( G = (V, E) \), we denote \( g = (V, E, p) \) as the influence graph of \( G \), where \( p \) is the propagation probability for edges between two user nodes. We delete every edge \( e \) in \( G \) with probability \( (1 - p_e) \). After that, we need to randomly choose one node in \( V \) and then construct a hypergraph and get the RR set for it.

**Definition 4** (RR set). Let \( v \) be a node in \( V \). A RR set for \( v \) is generated by firstly sampling a graph \( g' \) from \( g \) and then taking the set of nodes in \( g' \) that can reach \( v \).

For instance, from hypergraphs in Figure 1, we can get the RR sets as follows: \( R_1 = \{Z1, 6, 4, 0, 2, 5, 9, 8, \ldots\} \), \( R_2 = \{Z2, 12, 4, 7, 5, 9, 6, 8, \ldots\} \), \( R_3 = \{Z3, 2, 6, 4, 3, 5, 0, 8, \ldots\} \), \( R_4 = \{Z4, 7, 12, 9, 0, 5, 1, 10, \ldots\} \). Through the construction of \( \mathcal{R} \), we can seek \( k \) nodes as seeds which have the maximum coverage of \( \mathcal{R} \).

3.3. Coarsening Method

**Definition 5** ([37] coarsened influence graph). Given a social network \( G = (V, E) \), let \( g = (V, E, p) \) be an influence graph and \( P = \{C_j\}_{j \in [t]} \) be a partition of \( V \), where each \( C_j \) is strongly connected (SC). Then, a coarsened influence graph obtained from \( g \) is defined as a vertex-weighted influence graph \( H = (W, F, q, w) \), where

\[
W = \{C_j \mid j \in [t]\}, \\
F = \{(C_x, C_y) \mid C_x \neq C_y, \exists (u, v) \in E, u \in C_x, v \in C_y\}, \\
q_{C_x, C_y} = 1 - \prod_{(u, v) \in E \atop u \in C_x, v \in C_y} (1 - p_{uv}), \forall (C_x, C_y) \in F, \\
w(C_j) = |C_j|, \forall C_j \in W.
\]

The mapping \( V \rightarrow W \) is defined as \( \pi(v) = C_j \) such that \( v \in C_j \). For a vertex set \( S \), we let \( \pi(S) = \{\pi(v) \mid v \in S\} \) and \( F = \{\pi(u), \pi(v) \mid \pi(u) \neq \pi(v), (u, v) \in E\} \). We follow the same coarsening process as [37]. Given an influence graph \( g \), whose distribution is \( D_g \), we first construct subgraphs \( G_1 = (V_1, E_1), \ldots, G_t = (V_t, E_t) \), which are random graphs sampled from \( D_g \). In these subgraphs, we identify the vertex sets connected to each other in all subgraphs, so that we get a partition \( P = \{C_1, C_2, \ldots, C_t\} \) of \( g \) and corresponding \( H = (W, F, q, w) \) (see Figure 2).

3.4. Spatial Filtering Model. In order to figure out which nodes fall into region \( Q \), the most intuitive way is to compare the location of each node with the boundary of \( Q \), which is costly when \( Q \) is complex or \( |V| \) is very large. Herein, we will adopt an efficient method for this task. Our method works by comparing the convex hull of nodes with \( Q \) and iteratively removing those nodes falling out of \( Q \). Finally, we can end with a group of nodes whose convex hull is inside \( Q \). In this manner, we avoid enumerating all nodes in \( V \). Notably, for a point set \( T \), we could use Graham Scan method [38] to seek its convex hull. Let \( T = \{v_1, \phi, \ldots, v_n, \phi\} \) be the nodes’ locations and \( (q_1, \ldots, q_m, \phi) \) be the boundary of \( Q \) as a point sequence. Then the process of finding target nodes in \( Q \) is shown in Algorithm 1.
For instance, in Figure 3, given a location-aware social network $G$, $k = 1$, and a query region $Q$ (such as a triangle), we first use SFM (line 3) to identify the goal users \{3, 4, 6\}, then we coarsen the whole network, and we get the partition $P = \{C_1, C_2, C_3, C_4\}$, where $C_1 = \{1\}, C_2 = \{4, 6, 7\}, C_3 = \{5\}, C_4 = \{2, 3\}$ (line 4). Next we will sample the coarsened influence graph and construct enough RR sets for the goal users (line 5). In the coarsened influence graph, the probability of partition node $C_4$ being chosen to construct RR set is $1/3$, and there is $2/3$-probability for node $C_2$. As a result, we get RR sets: $R_1 = \{C_2, C_4\}, R_2 = \{C_4\}, R_3 = \{C_2, C_4\}$. We can see that node $C_4$ has the maximum coverage of $\mathcal{R}$, including $R_1, R_2, R_3$. In turn for the original network graph, we choose one user node randomly in $C_4$ as the seed, such as node 2 (lines 6-10).

In our framework, the first step is to identify the users in the target region $Q$, whose complexity depends on the location distribution of all nodes. In case that all users’ locations are uniformly randomly distributed, the time complexity for identifying users within the target region is $O(n^{1/3} m_q)$. 

4. TarIM-SF Framework

Here we describe our TarIM-SF framework in detail. In our LAIM problem, target users change from the whole network in classic IM problem into users in a target region. As the construction of RR sets starts from the target users, it is reasonable for us to construct enough RR sets $\mathcal{R}$ over the whole network for users in the target region within our problem and then choose the node set $s$ which maximizes the coverage of $\mathcal{R}$ (referred to as $F_{\mathcal{R}}(s)$). Based on this idea, we first need to identify the users in the target region before constructing RR sets, which is addressed using the method proposed in Section 3.4. Moreover, in order to improve the efficiency of constructing RR sets, we have coarsened the network using the method proposed in Section 3.3. More details are given in Algorithm 2.

For instance, in Figure 3, given a location-aware social network $G$, $k = 1$, and a query region $Q$ (such as a triangle), we first use SFM (line 3) to identify the goal users \{3, 4, 6\}, then we coarsen the whole network, and we get the partition $P = \{C_1, C_2, C_3, C_4\}$, where $C_1 = \{1\}, C_2 = \{4, 6, 7\}, C_3 = \{5\}, C_4 = \{2, 3\}$ (line 4). Next we will sample the coarsened influence graph and construct enough RR sets for the goal users (line 5). In the coarsened influence graph, the probability of partition node $C_4$ being chosen to construct RR set is $1/3$, and there is $2/3$-probability for node $C_2$. As a result, we get RR sets: $R_1 = \{C_2, C_4\}, R_2 = \{C_4\}, R_3 = \{C_2, C_4\}$. We can see that node $C_4$ has the maximum coverage of $\mathcal{R}$, including $R_1, R_2, R_3$. In turn for the original network graph, we choose one user node randomly in $C_4$ as the seed, such as node 2 (lines 6-10).

In our framework, the first step is to identify the users in the target region $Q$, whose complexity depends on the location distribution of all nodes. In case that all users’ locations are uniformly randomly distributed, the time complexity for identifying users within the target region is $O(n^{1/3} m_q)$.

**Algorithm 1: SFM ($T, Q$).**

**Input:**
A point set $T = \{v_1, \ldots, v_n, \phi\}$, a polygon $Q = (q_1, \ldots, q_m)$;

**Output:**
The points of $T$ within $Q$.

1: Initialize $i = 1$, $T_i = T$, $\mathcal{C}_i = \emptyset$, $\mathcal{C}_i' = \emptyset$;
2: repeat
3: Compute the convex hull $\mathcal{C}_i$ of $T_i$, and compare $\mathcal{C}_i$ with $Q$;
4: if $((q_1, \ldots, q_m) \in \mathcal{C}_i)$ then
5:   return $\{\mathcal{C}_i, \mathcal{C}_i', \ldots, \mathcal{C}_{i-1}, T_i\}$.
6:   else if $((q_1, \ldots, q_m) \notin \mathcal{C}_i)$ then ($\forall c \in \mathcal{C}_i$, $\exists c' \in Q$)
7:     return 0.
8: end if
9: until $T_i = \emptyset$.
10: end if
11: if $\mathcal{C}_i \in Q$, $1 \leq j \leq i$;
12: $T_{i+1} = T_i - \mathcal{C}_i$, $i = i + 1$;
13: until $T_i = \emptyset$.
14: return $\{\mathcal{C}_1', \mathcal{C}_2', \ldots, \mathcal{C}_{i-1}'\}$.

**Figure 2:** Two subgraphs of $g$ and the coarsened influence graph $H$. 

- **Subgraph 1**
- **Subgraph 2**
- **Coarsened influence graph**

4 Complexity

For instance, in Figure 3, given a location-aware social network $G$, $k = 1$, and a query region $Q$ (such as a triangle), we first use SFM (line 3) to identify the goal users \{3, 4, 6\}, then we coarsen the whole network, and we get the partition $P = \{C_1, C_2, C_3, C_4\}$, where $C_1 = \{1\}, C_2 = \{4, 6, 7\}, C_3 = \{5\}, C_4 = \{2, 3\}$ (line 4). Next we will sample the coarsened influence graph and construct enough RR sets for the goal users (line 5). In the coarsened influence graph, the probability of partition node $C_4$ being chosen to construct RR set is $1/3$, and there is $2/3$-probability for node $C_2$. As a result, we get RR sets: $R_1 = \{C_2, C_4\}, R_2 = \{C_4\}, R_3 = \{C_2, C_4\}$. We can see that node $C_4$ has the maximum coverage of $\mathcal{R}$, including $R_1, R_2, R_3$. In turn for the original network graph, we choose one user node randomly in $C_4$ as the seed, such as node 2 (lines 6-10).

In our framework, the first step is to identify the users in the target region $Q$, whose complexity depends on the location distribution of all nodes. In case that all users’ locations are uniformly randomly distributed, the time complexity for identifying users within the target region is $O(n^{1/3} m_q)$.
Input:
A location-aware social network \( G = (V, E) \), a query polygon \( Q \), and a budget \( k \);

Output:
A seed set \( S \).

1: Initialize \( S = \emptyset \), a set of goal users \( q_s = \emptyset \);
2: \( T \leftarrow \text{location points of } V \);
3: \( q_s \leftarrow \text{SFM}(T, Q) \);
4: \( G' \leftarrow \text{Coarsening}(G) \), according to Algorithm 1 in [37];
5: \( R \leftarrow \text{Sampling}(G', q_s, k, \epsilon) \), according to Algorithm 2 in [3];
6: for \( i = 0 \) to \( k - 1 \) do
7: \( s' \leftarrow \text{seek a node in } G' \) with the maximum marginal coverage for \( R \);
8: \( s \leftarrow \text{choose a vertex in } s' \) randomly;
9: \( S \leftarrow S \cup \{s\} \);
10: end for
11: return \( S \).

Algorithm 2: TarIM-SF \((G, Q, k, \epsilon)\).
of $E(\sigma(S))/n_q$, where $n_q$ is the number of vertices in target region $Q$. In TIM$^*$ [17], it has been proved that the solution which covers the maximum number of RR sets provides a $(1−1/e−\epsilon)$-approximation with at least $(1−1/n^t)$ probability, but the number of RR sets is at least $\lambda/OPT$, where $OPT$ is the maximum expected influence of any size-$k$ nodes set in $G$ and $\lambda$ is a function of $k, \epsilon, n_t,$ and $\epsilon$. In IMM, it seeks a tighter lower bound $LB$ of $OPT$ than TIM$^*$. Next, based on the analysis of the performance guarantee in the IMM, we will describe the parameter settings and performance guarantee in our framework.

Let $R_1, R_2, \ldots, R_q$ be the sequence of generated RR sets for nodes in the target region $Q$. Let $S_i$ be any size-$k$ seed set in $G$ and $X_i$ be random variable that equals 0 if $S_i \cap R_j = \emptyset$ and 1 otherwise; then based on Lemma 7, we have

$$E[\sigma(S_i)] = \frac{n_q}{\theta} \cdot E\left[\sum_{i=1}^{\theta} X_i\right].$$

Consider $S_i^*$ is the size-$k$ node set with the maximum expected influence; let $OPT = E[\sigma(S_i^*)]$. From (2), we can get that $n_q \cdot F_{R}(S_i^*)$ is an unbiased estimator of $OPT$. By Corollary 2 in which we set $p = OPT/n_q$ and Lemma 3 in [3], we have the following.

**Lemma 8.** Let $\epsilon_1 > 0$, $\delta_1 \in (0, 1)$, and

$$\theta_1 = \frac{2n_q \cdot \log(1/\delta_1)}{OPT \cdot \epsilon_1^2}.$$

if $\theta \geq \theta_1$, then $n_q \cdot F_R(S_i^*) \geq (1 - \epsilon_1) \cdot OPT$ holds with at least $(1 - \delta_1)$ probability.

Assuming that our framework without coarsening returns a solution $S_1$, and if $n_q \cdot F_R(S_i^*) \geq (1 - \epsilon_1) \cdot OPT$ holds, according to the properties of the greedy approach, then

$$n_q \cdot F_R(S_i^*) \geq \left(1 - \frac{1}{e}\right) \cdot n_q \cdot F_R(S_i^*) \geq \left(1 - \frac{1}{e}\right) \cdot (1 - \epsilon_1) \cdot OPT.$$

**Lemma 9.** Let $\epsilon_1 < \epsilon$, $\delta_2 \in (0, 1)$, and

$$\theta_2 = \frac{(2 - 2/e) \cdot n_q \cdot \log((1 - \epsilon_1)/\delta_2)}{OPT \cdot (1 - 1/e - \epsilon_1)^2}.$$

if (4) holds and $\theta \geq \theta_2$, then $E[\sigma(S_i^*)] \geq (1 - 1/e - \epsilon) \cdot OPT$ holds with at least $(1 - \delta_2)$ probability.

Based on Lemmas 8 and 9, we have the following.

**Theorem 10.** Given any $\epsilon_1 \leq \epsilon$ and any $\delta_1, \delta_2 \in (0, 1)$ with $\delta_1 + \delta_2 \leq 1/n^t$, if $\theta \geq \max(\theta_1, \theta_2)$, the result which IMM returns is a $(1 - 1/e - \epsilon)$-approximate solution with $(1 - 1/n^t)$ probability.

For the parameters in Theorem 10, we set $\delta_1 = \delta_2 = 1/(2n^t)$, and under this setting, $\theta$ is minimized when $\theta_1 = \theta_2$, and $\epsilon_1 = \epsilon / \alpha((1 - 1/e) \cdot \alpha \cdot \beta)$, where

$$\alpha = \sqrt{\ell \log n + \log 2},$$

$$\beta = \sqrt{\left(1 - \frac{1}{e}\right) \cdot \left(\log (n/k) + \ell \log n + \log 2\right)}.$$

In this case, $\theta = (2n_q \cdot (1 - 1/e) \cdot \alpha + \beta)^2)/(OPT \cdot e^3)$, and conversely if we set $\theta = \lambda^*/OPT$, where

$$\lambda^* = \frac{2n_q \cdot (1 - 1/e) \cdot \alpha + \beta)^2}{e^3}.$$

the seed set $S_i^*$ that covers the maximum number of RR sets is a $(1 - 1/e - \epsilon)$-approximation. However, as $OPT$ is unknown in advance, we will find a tight lower bound $LB$ of $OPT$ as IMM method. In the sampling phase in IMM, Lemma 6, Lemma 7, and Lemma 8 in [3] proved that $LB \leq OPT$ and $LB$ is close to $OPT$. And based on that, we can also prove that $LB \leq OPT$ and $LB$ is a tight lower bound of $OPT$ with a high probability by changing $n$ to $n_q$.

**Theorem 11.** With at least $(1 - 1/n^t)$ probability, sampling algorithm in our framework returns a set $S$ of RR sets with $|S| \geq \lambda^*/OPT$, where $\lambda^*$ is as defined in (7).

Combining Theorem 10 and Theorem 11, our algorithm framework without coarsening can get a solution $S_{out}$ which is a $(1 - 1/e - \epsilon)$-approximation with a high probability.

5.2. Seed Quality for Coarsening Method. In this part, we will show the seed set $S_{out}$ that can activate in the coarsened influence graph $G(V, E, \omega, \rho)$ for $\omega \epsilon \sim \tau$. After coarsening the network, we define $\omega \epsilon \sim \tau$, and $\omega \epsilon \sim \tau$. If $\omega \epsilon \sim \tau$, we also have $\omega \epsilon \sim \tau$. It denotes the sum of weights for nodes in $Q$ that $\omega \epsilon \sim \tau$ can activate in $H$. After coarsening the network, we define $I = (V, E, \rho')$, where $\rho'_{uv} = 1$ if $u, v \in C_j$ for all $j \in [\tau]$; otherwise $\rho'_{uv} = \rho_{uv}$, and $\text{inf}_{H}(Q, S)$ is the number of users in target region $Q$ that $S$ can activate in $I$.

Here, we show the relationship between $H$ and $g$ in terms of influence function through $I$, which has the same structure as $g$. Complexity
Lemma 12. For any $Q$ and any $S \subseteq V$, $\inf_f(Q, S) = \inf_f(H, \pi(S))$.

Proof. For any $u$ and $v$ in $V$, “$u$ can reach $v$ through the edges in $E$ with $p^v_0$” and only if “$\pi(u)$ can reach $\pi(v)$ through the edges in $F$ with $q^v_0$,” since every subgraph $(C_j, E_j)$, $j \subseteq [r]$, is strongly connected. Therefore, it holds that $R_{(V,E)}(Q, S) = R_{(V,Y)}(Q, \pi(S))$. Thus,

$$\inf_f(Q, S) = \sum_{E \subseteq E} p^v_0 e^v \cdot R_{(V,E)}(Q, S)$$

$$= \sum_{Y \subseteq Y} q^v_0 e^v \cdot R_{(V,Y)}(Q, \pi(S))$$

$$= \inf_H(Q, \pi(S))$$

□

For $g$ and $I$, we also can find the relationship between them as follows.

Lemma 13. For any $Q$ and any $S \subseteq V$, $\inf_g(Q, S) \leq \inf_f(Q, S)$.

Proof. $g$ and $I$ are the influence graph with the same structure, but $p_e \leq q^v_0 e^v$ for every edge $e$. So $\inf_g(Q, S) \leq \inf_f(Q, S)$ for any $S \subseteq V$.

For any subgraph $g[V'] = (V', E')$ of $g$, $V' \subseteq V$, its strongly connected reliability, denoted as $Rel(g[V'])$, is defined the same as in Equation 14 in [37] and indicates the probability that $g[V']$ is strongly connected. Thus

$$\inf_g(Q, S) \leq \prod_{j \in [r]} Rel(g[C_j])^{-1} \cdot \inf_f(Q, S),$$

for any $Q$ and $S \subseteq V$.

Proof. For each $j \in [r]$, we define $E_j = \{u, v\} \in E | u, v \in C_j$ and $E_0 = E \cup (\bigcup_{j \in [r]} E_j)$. For $X_j \subseteq E_0 (j \in [r])$ and $X_{E,0} \subseteq E_0$,

$$R_{(V,X_0,U|U)}(S) = R_{(V,X_0,U|U)}(S)$$

holds if every subgraph $(C_j, X_j)$ for $j \in [r]$ is strongly connected. Thus

$$\inf_g(Q, S) = \sum_{X_0 \subseteq E_0, X_{E,0}} p_{X_0|E_0} \cdot R_{(V,X_0,U|U)}(S)$$

$$\geq \sum_{E_j \subseteq E_j} p_{X_j|E_j} \cdot \prod_{(C_j, X_j) \text{ is SC}} p_{X_j|E_j}$$

$$= \prod_{j \in [r]} Rel(g[C_j]) \cdot \inf_f(Q, S)$$

□

Theorem 15. For any $Q$ and any $S \subseteq V$, $\inf_g(Q, S) \leq \inf_H(Q, \pi(S)) \leq \prod_{j \in [r]} Rel(g[C_j])^{-1} \cdot \inf_f(Q, S)$.

Let $S^*$ and $T^*$ be the optimal solutions of size $k$ for $g$ and $H$, respectively. Based on Lemma 12 and Lemma 13, we can sure that $\inf_f(Q, T^*) \geq \inf_g(Q, S^*)$, and have

$$\inf_H(Q, S_{out}) \geq \left(1 - \frac{1}{e} - \epsilon\right) \cdot \inf_f(Q, T^*)$$

$$\geq \left(1 - \frac{1}{e} - \epsilon\right) \cdot \inf_g(Q, S^*)$$

Then applying Lemma 14, we have

$$\inf_H(Q, S_{out}) \geq \prod_{j \in [r]} Rel(g[C_j]) \cdot \left(1 - \frac{1}{e} - \epsilon\right) \cdot \inf_g(Q, S^*)$$

Therefore, our coarsening-based algorithm achieves a $(1 - 1/e - \epsilon)\alpha_g$-approximate solution for $g$, where $\alpha_g$ refers to $\prod_{j \in [r]} Rel(g[C_j])$.

6. Results and Discussion

In this section, we conduct experiments on several real-world datasets to test the performance of the proposed algorithm framework. All algorithms are implemented in C++ and run on Ubuntu 16.10 machine with Intel Core i5-6500 quad-core, 3.20GHz, 16GB RAM.

In the following experiments, we use three location-aware social networks, namely, Gowalla, Tweets, and Weibo. The statistics for the datasets are listed in Table 1 ($n$ represents the number of vertices and $m$ represents the number of edges). By default, we use a randomly selected $Q$ for all datasets, and the number of user nodes falling in $Q$ is denoted as $n_q$. We conduct our experiments on WC model, which is widely used for information diffusion. For the weight of every edge, we set the probability of an edge $(u, v)$ as $1/N_v$, where $N_v$ denotes the in-degree of user node $v$. In TarIM-SF, we set $\epsilon = 0.1$.

6.1. Comparison with Baseline. We evaluate the performance of our algorithm framework TarIM-SF compared with method Assembly in [4] under WC model. In order to estimate the performance in general, we selected three regions with fixed $n_q$ for each dataset ($n_q \approx 20K$ for Gowalla, $n_q \approx 120K$ for Tweets, and $n_q \approx 120K$ for Weibo) and reported the average performance, while varying $k$ for Gowalla from 10 to 50 and $k$ for Tweets and Weibo from 100 to 500. In Figure 4, we can see that the target influence spread of seeds in our framework is obviously superior to that of Assembly, especially on Tweets and Gowalla.

6.2. Effect of Coarsening. As mentioned above, we adopt coarsening technique so as to improve the efficiency for large-scale social networks. In this part, we did a series of experiments on Weibo, a large-scale network with about a million users. In order to justify how the parameter $r$ in coarsening method will affect the target influence spread of seeds, we report the Relative Error, which measures the gap between the real influence spread of seed set we get and its
Table 1: Datasets statistic.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>n</th>
<th>m</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gowalla</td>
<td>196,585</td>
<td>351,452</td>
</tr>
<tr>
<td>Tweets</td>
<td>554,372</td>
<td>2,402,720</td>
</tr>
<tr>
<td>Weibo</td>
<td>1,020,730</td>
<td>16,490,916</td>
</tr>
</tbody>
</table>

Figure 4: Influence spread under WC model.

Figure 5: Effect when using coarsening.

estimated influence spread. We set $k = 300$ and the target region as the whole network. Figure 5(a) shows the relative error is decreasing when $r$ becomes bigger. When $r = 16$, the estimated influence spread of seeds using our algorithm with coarsening is nearly equal to the eventual influence spread of seeds without coarsening. Figures 5(b) and 5(c) indicate that the running time for coarsening approach is significantly less than that of noncoarsening one, without loss of seed quality.

6.3. Varying the Size and Shape of $Q$. We also conducted another group of experiments by varying $Q$ in terms of both size and shape. Specifically, we vary $Q$ as triangle, tetragon, and pentagon, respectively. For each shape, we vary the size at several different levels and report the performance of our algorithm (shown in Figure 6, $|Q| = n_q$). It justifies that our algorithm can work on target region with arbitrary polygon shapes. Besides, the time spent on the first phase in our algorithm framework increases slightly as the shape of $Q$ varied from pentagon to triangle at the same scale, because nodes in convex hull need to be compared with the polygon queried and the fewer the number of polygon edges, the fewer the comparison times and the less time.

7. Conclusions

In this study, we present a novel model that can address LAIM with a target region that can be an arbitrary polygon. Our framework uses a spatial filtering model to initially figure out the nodes falling into the target polygon. Afterwards, a coarsening process is conducted over the network. Then, the state-of-the-art sampling algorithm adopted in traditional IM is used to find the solution. We theoretically prove the
influence spread guarantee in both noncoarsened and coarsened cases. Empirical study over three real-world datasets demonstrates that our framework outperforms the baseline algorithm in terms of influence spread and is efficient in large-scale networks.

Data Availability

This study is based on the datasets Gowalla, Tweets, and Weibo provided in [4], and they are available at http://dbgroup.cs.tsinghua.edu.cn/ligl/laim/.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The work is supported by the National Natural Science Foundation of China (Grant No. 61672408), Fundamental Research Funds for the Central Universities (No. JB181505), Natural Science Basic Research Plan in Shaanxi Province of China (No. 2018JM6073), and China 111 Project (No. B16037).

References


Submit your manuscripts at
www.hindawi.com