Sample Location Selection for Efficient Distance-Aware Influence Maximization in Geo-Social Networks

Ming Zhong[1](✉), Qian Zeng[1], Yuanyuan Zhu[1], Jianxin Li[2], and Tieyun Qian[1]

[1] School of Computer, Wuhan University, Wuhan 430072, China
{clock,wennie,yyzhu,tyqian}@whu.edu.cn
[2] School of Computer Science and Software Engineering,
University of Western Australia, Crawley, WA 6009, Australia
jianxin.li@uwa.edu.au

Abstract. In geo-social networks, the distances of users to a location play an important role in populating the business or campaign at the location. Thereby, the problem of Distance-Aware Influence Maximization (DAIM) has been investigated recently. The efficiency of DAIM computation heavily relies on the sample location selection, because the online seeding performance is sensitive to the distance between sample location and promoted location, and the offline precomputation performance is sensitive to the number of samples. However, there is no work to fully study the problem of sample location selection w.r.t. DAIM in geo-social networks. To do this, we first formalize the problem under a reasonable assumption that a promoted location always adheres to the distribution of users. Then, we propose an efficient location sampling approach based on the heuristic anchor point selection and facility allocation techniques. Our experimental results on two real datasets demonstrate that our approach can improve the online and offline efficiency of DAIM approach like [9] by orders of magnitude.

1 Introduction

Motivation. The widely-used geo-position enabled devices (e.g., mobile phone, tablets, laptops, etc.) and services (e.g., geolocation, geocoding, geotagging, etc.) allow social networks to connect users with local places and events that match their interests. For example, there are currently a lot of popular geo-social network applications like Yelp, Gowalla, Facebook Places and Foursquare. Due to the obvious implication, many researches turn to focus on taking location information into account in the influence maximization problem of geo-social networks. Different from the traditional influence maximization, a typical scenario of influence maximization in geo-social networks is to promote a specific location like a newly opened restaurant or an upcoming sale activity, which is called query location. In that case, the users near the query location are more valuable to be influenced, because they are more likely to visit the location.

© Springer International Publishing AG, part of Springer Nature 2018
J. Pei et al. (Eds.): DASFAA 2018, LNCS 10827, pp. 355–371, 2018.
https://doi.org/10.1007/978-3-319-91452-7_24

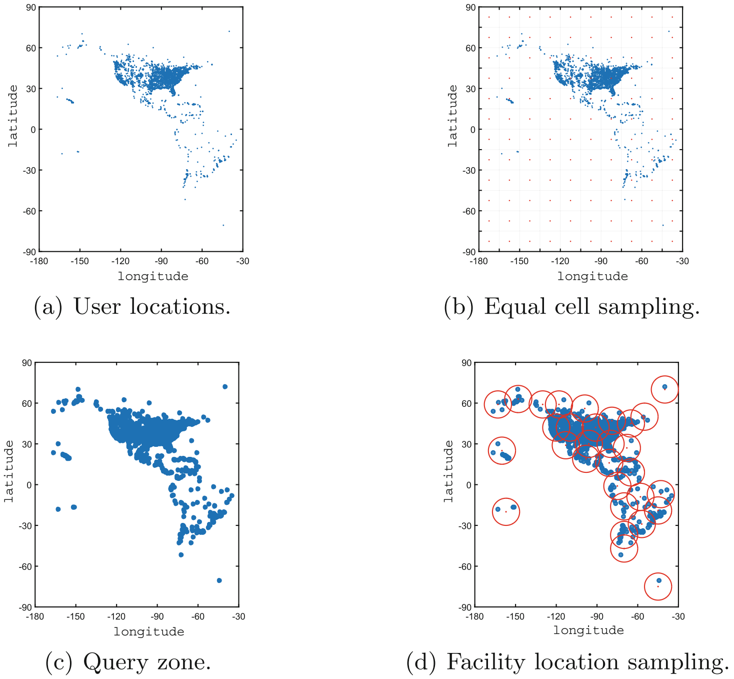
There are two typical problem definitions for the above scenario. The first one is called location-aware influence maximization (LAIM) [8]. The LAIM problem is to maximize the influence to only the users in a given query region, which is a rectangle containing the query location. As a shortcoming of LAIM problem, how to select an appropriate query region for a given query location is unclear. If the query region is too large, most users influenced by the selected seeds may distribute near the boundary of region, thereby being far away from the exact query location. If the query region is too small, many potential users near the query location but outside of the region will be neglected. To overcome the shortcoming of LAIM, the second one called distance-aware influence maximization (DAIM) [9] is proposed. For the DAIM problem, each user has a weight that is determined by the distance between it and the query location no matter whether it is in a query region, and the influence spread to users is adjusted according to their weight.

Typically, to address the DAIM problem, the existing approaches select a set of sample locations in the 2D space where the users are distributed, and precompute the influence spread with respect to the sample locations, which can be leveraged by the online seeding algorithms. Note that, the shorter the distance between sample locations and the given query location, the better the performance of online seeding algorithms. Moreover, the precomputation is very time-consuming for a sample location. Thus, given a budget of location sampling, we hope to minimize the objective distance between any possible query location and its nearest selected sample location.

However, the existing DAIM approaches focus on the seeding algorithms and only use naive sampling like random sampling [9] or equal cell sampling [10]. The naive sampling needs to sample a large number of locations to achieve a promising objective distance. Thus, it is unlikely to achieve a good online seeding performance without spending heavy precomputation overhead while using such naive sampling. Let us consider the following example.

Example 1. Figure 1(a) shows the geographical distribution of users in Brightkite, a real-world geo-social network. We can see that most users live in a few urban areas[1]. The naive sampling like equal cell sampling ignores this fact and try to reach an arbitrary point in the space within the minimum distance, as shown in Fig. 1(b). However, it is nonsense to promote a place that is far away from the users, since the users will hardly visit the place under the settings of DAIM. Instead, the possible query location in reality should adhere to the users, namely, be in a "*query zone*" around the users. Figure 1(c) shows an example of query zone comprised of circles centered at each user with an identical radius r. Thereby, any query location in the query zone is no farther from at least a user than r. Then, we can use more delicate sampling method to reduce the number of sample locations that is necessary for achieving the same objective distance. As shown in Fig. 1(d), the query zone can be covered by the red circles centered

[1] The 2D space in this example is the surface of earth. In reality, we only consider a city or even smaller district for location promotion. However, the situation of sparse user distribution remains the same.



(a) User locations.

(b) Equal cell sampling.

(c) Query zone.

(d) Facility location sampling.

Fig. 1. A motivation example on a geo-social network named Brightkite. (Color figure online)

at only a few samples, and the radius of circles is equal to the half length of diagonal line of equal cells in Fig. 1(b), namely, both sampling methods have the same objective distance.

Therefore, we focus on the selection of sample locations in this paper. Our work is based on an important observation. That is, the users of real-world geo-social networks are usually distributed sparsely in a 2D space. It means, the real query location must not be an arbitrary point in the space, and should be close enough to some users. Otherwise, given a query location that has no user nearby, trying to maximize the influence to users with very low weights are actually meaningless. Consequently, we reasonably assume that the potential query location is not farther than a specific distance from the users. Under this assumption, we try to find a set of sample locations such that the maximum distance between any qualified query location and its nearest sample location is minimized. As a result, a DAIM approach can generally improve the online seeding performance with less offline precomputation overhead by using our approach for selecting sample locations.

Our Contributions. To address the above problem, we propose a novel sample location selection approach. Due to the NP-hardness of this problem, our approach deals with it heuristically as follows. Firstly, we select a set of *anchor* *points* from the query zone in various ways. Then, given a set of anchor points, we address an l-center problem, one subtype of classic facility location problem, which is to find l points in the space that can reach any anchor point with the minimum distance. Since the l-center problem is still NP-hard, we offer a heuristic algorithm to return l points as the final sample locations. Moreover, we estimate the upper bound of objective distance between a possible query location in the query zone and the selected sample locations, according to the anchor point selection strategy and the result of l-center problem.

Our contributions are generalized as follows.

- We formalize a novel and interesting sample location selection problem for distance-aware influence maximization in geo-social networks. In this problem, the query location must be in a particular query zone but not the whole 2D space under a reasonable assumption.
- We address the problem by simplifying it to an l-center problem for a set of anchor points selected from a given query zone. For that, we propose a flexible strategy for selecting anchor points that can improve the tightness of objective distance bound by investing more sampling time, and develop an efficient heuristic algorithm for dealing with l-center problem. Then, we can derive a safe and tight upper bound of the objective distance.
- We perform comprehensive experiments on two real-world datasets. Compared with [9], our approach can reduce the precomputation overhead significantly, and meanwhile, improve the efficiency of online seeding significantly.

The rest of this paper is organized as follows. We review the related work in Sect. 2. The formalized problem definition is given in Sect. 3. We present the sample location selection approach in Sect. 4. The experiment results are demonstrated in Sect. 5. Lastly, we conclude our work in Sect. 6.

2 Related Work

Influence Maximization in Social Network. Influence maximization problem is first defined by Kempe et al. [1], the authors also define the independent cascade model and the linear threshold model and prove the hardness of the problem in the paper. Since then, there are a large number of literatures on influence maximization, like [2–5] etc. In [2], the authors propose the CELF algorithm, which exploits the submodular property to significantly boost the traditional greedy approach. Chen et al. [3] propose the PMIA approach, the influence is considered to propagate only through the maximum influence path between users. While Cohen et al. [5] propose a bottom-k sketch based approach to reduce the cost influence estimation. The materialized sketch can be used as an oracle to evaluate the influence of any subset users.

Influence Maximization in Geo-Social Network. With the appearance of geo-position enabled devices and service, researchers begin to pay attention to the impact of geographical location on influence maximization. Zhu et al. attempt to measure the influence between users by considering social relation and location information in [6], while Cai et al. propose a novel network model and an influence propagation model in [7], they think the influence propagation should conduct both in online social networks and physical world. Li et al. [8] attempt to maximize the influence spread in a query region. However, it is non-trivial to determine an appropriate query region when conducting a location-aware promotion. The work which is most related to ours is by Wang et al. in [9,10], they propose the MIA-DA and RIS-DA approaches. The MIA-DA approach gives a priority based algorithm which compromises three pruning rules and a novel index structure. The RIS-DA comes up with an unbiased estimator for the distance-aware influence maximization, both of them need to estimate the necessary size of network samples for any potential query, but such process is very time-consuming.

Facility Location. Research in location theory formally started in 1909 by Weber [11], known as the father of modern location theory. He studies the problem of locating a single warehouse to minimize the total travel distance between the warehouse and a set of customers. Since then, many researchers have observed this problem in different areas, and there are some surveys about techniques for facility location problem, like [12,13]. Elzinga and Hearn give a geometric algorithm to solve the 1-center problem with Euclidean distances, and prove the correctness of the algorithm in [14]. Drezner and Wesolowsky [15] discusses the problem of locating a new facility among n given demand points by taking the l_p-norm distance into consideration, and proposes two heuristic and an optimal algorithms to solve the problem for a given l in time polynomial in n in [16]. Then Callaghan et al. [17] attempt to speed up the optimal method of Drezner in [16] by introducing neighbourhood reduction schemes and embedding an CPLEX policy.

Compared to these work, our approach focuses on combine facility allocation techniques into sample location selection, so that the objective distance derived by our approach can be much shorter. Thus, our approach can reduce the pre-computation overhead, and meanwhile, improve the efficiency of online seeding significantly.

3 Preliminary and Problem Definition

In this section, we first introduce the definition of DAIM problem and analyze the existing DAIM approaches of sample location selection, then give a formal definition of the problem proposed by us.

3.1 Distance-Aware Influence Maximization

We consider a geo-social network as a directed graph $G = (V, E)$, where V represents a set of users and $E = V \times V$ represents the relationships between users.

Each user $v \in V$ has a geographical location (x, y), where x and y represent the latitude and longitude respectively. We denote by $I(S, v)$ the probability that a node set $S \subseteq V$ can activate v under a specific propagation model. The traditional influence maximization problem is to find S with $|S| = k$ that maximizes $\sum_{v \in V} I(S, v)$. However, influence maximization in geo-social networks normally considers the promotion of a query location (like a restaurant). Intuitively, the users near the location are more likely to visit the location. We denote by $w(v, q)$ the weight of a user v with respect to a location q, and the weight depends on the distance between v and q. Thus, the definition of $\underline{\text{d}}$istance-$\underline{\text{a}}$ware $\underline{\text{i}}$nfluence $\underline{\text{m}}$aximization (DAIM) is given as follows.

Definition 1 *(Distance-Aware Influence Maximization)*. *Given a geo-social network $G = (V, E)$, a query location q and a positive integer k, the problem of distance-aware influence maximization is to find a set S^* of k nodes in G which has the largest distance-aware influence spread, i.e.,*

$$S^* = \underset{S \subseteq V}{arg\,max} \{I_q(S) || S| = k\} \tag{1}$$

where $I_q(S) = \sum_{v \in V} I(S, v) w(v, q)$ is the distance-aware influence propagation of a node set S.

To address the DAIM problem, Wang et al. [9,10] propose two approaches, namely, MIA-DA and RIS-DA under the independent cascade model. MIA-DA extends the maximum influence arborescence model, and can achieve an approximation ratio of $1 - 1/e$. RIS-DA extends the reverse influence sampling model, and can achieve an approximate ratio of $1 - 1/e - \epsilon$ with at least $1 - \delta$ probability. According to the comparison in [9], RIS-DA is more precise but less efficient than MIA-DA.

Such DAIM approaches need to precompute the influence spread with respect to some sample locations. Then, based on the precomputed influence spread, they can derive the bounds of influence spread for any query location by investigating the relationship between the query location and the sample locations. Since the query location could be any point in the 2D space, they select sample locations distributed uniformly over the space. For example, MIA-DA partitions the space in to a number of equal cells, and selects the center of each cell as samples. While, RIS-DA selects sample locations randomly, and then partitions the space into Voronoi cells based on the set of samples. Therefore, there is surely a nearby sample location for an arbitrary promoted location, no matter which cell it is in.

3.2 Problem Definition

The above sample location selection methods result in heavy precomputation overheads and large index spaces in order to guarantee a good estimation of influence bounds. Let the number of user points be n, the number of seeds be k, the number of sample locations be l. The time complexity of precomputation for MIA-DA is $O(n^2)$ and for RIS-DA is $O(l^2 k^2 n \log n)$. Moreover, to derive tight bounds, the distance between the query location and its nearest sample location needs to be short enough. Since the sample locations are distributed uniformly over the space, the number of sample locations increases dramatically with the decrease of distance between sample locations and potential query locations.

In this paper, we argue that the query location in DAIM problem should consider the spatial distribution of users and should not be an arbitrary point in the 2D space. The possible query locations always follow the distribution of users in reality. For example, when a company needs to advertise for their products through the social network, they are more likely to select a query location which is in a densely populated location, but not far away from the crowd. Otherwise, there are no potential consumers with respect to the distance between them to the query location, and thereby addressing the DAIM problem is meaningless. So we have the following reasonable assumption of the query location distribution.

Assumption 1. *The given query location should follow the spatial distribution of users. Formally, given a positive real number r, there exists at least a user $v \in V$ for a query location q such that $dis(q, v) \leqslant r$.*

Intuitively, for a user, the area of activities is a circle centered at its location with a radius r, which is called *user circle*. Thus, only the query location in this circle can attract the user. All user circles compose a query zone Q, as shown in Fig. 1(c). We denote by $q \in Q$ that a point q located in the query zone Q. Under this assumption, the problem to be addressed in this paper can be formalized as follows.

Problem 1. Given a geo-social network $G = (V, E)$, a query zone Q defined by the locations of V and the radius r of user activities, and a location sampling budget l, find a set SL of l sample locations in the 2D space, such that the objective distance $D(SL, Q)$ is minimized. The objective distance is the maximum distance between any query location in Q and its nearest sample location in SL, namely, $D(SL, Q) = \max_{q \in Q} \min_{s \in SL} dis(q, s)$. For convenience, we denote by d_o the optimal objective distance.

For example, as shown in Fig. 2, there are two users v_1 and v_2, and the yellow circles comprise the corresponding query zone. The sample location s_1 is the middle point of line segment v_1v_2. Thus, the farthest query location to s_1 is q_1 and q_3, and we have $D(\{s_1\}, Q) = \max_{q \in Q} dis(q, s_1) = dis(q_1, s_1)$ or $dis(q_3, s_1)$. For any other sample location s_2, suppose s_2 is closer to v_2, we have $D(\{s_2\}, Q) = dis(q_2, s_2) = dis(v_1, s_2) + r > dis(v_1, s_1) + r = dis(q_1, s_1)$. Suppose we only select one sample location, namely, $l = 1$. It is obviously that the minimum objective distance $d_o = dis(q_1, s_1)$, so that the optimal set of sample location is $\{s_1\}$. Given this set of sample locations, no matter which point in the query zone needs to be promoted, we can find a sample location within the optimal distance d_o.

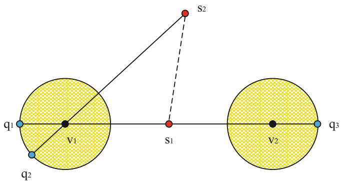


Fig. 2. A simple example of sample location selection problem. (Color figure online)

Problem Hardness. We briefly discuss the hardness of our problem as follows. Consider a query zone as an infinite set of points. If we only keep a fixed number of points in a query zone, the Problem 1 is reduced to the classic l-center problem [13]. It has been proved that the exact solution of l-center problem is NP-hard [18]. Approximation to the problem is also NP-hard when the error is small. Due to the hardness of l-center problem, the Problem 1 is also NP-hard where l is an infinite number.

4 Sample Location Selection

In this section, we firstly present a heuristic methodology to select sample locations for a given query zone in a 2D space, and develop efficient algorithms based on the studies of facility location problems.

4.1 Methodology

Due to the hardness of sample location selection problem defined above, we propose a heuristic approach to address it. The main idea is that, we select a set of discrete *anchor points* from the query zone, and find a given number of sample locations in the 2D space, such that each anchor point can reach the nearest sample location within the minimum distance d_a. Let d_z be the maximum distance between any point in the query zone and its nearest selected sample location. Although there could be some areas of query zone that can not be reached by selected sample locations within the distance d_a, namely, $d_a \leqslant d_z$, we can guarantee that $d_z - d_a$ is no more than $f(r)$ by selecting the anchor points with a particular strategy, where the function $f : r \mapsto (0, r]$ is determined by the strategy. It is certainly that any point in the query zone can reach its nearest selected sample location within a distance $d_a + f(r)$. Thus, we safely use the upper bound $d_a + f(r)$ of d_z (and of course d_o) as the final objective distance.

In the followings, we introduce the strategy of selecting anchor points and the heuristics of selecting sample locations for a given set of anchor points.

Anchor Point Selection Strategy. We propose two strategies of anchor point selection, the baseline and the improved. The improved strategy can achieve a tighter bound of d_z than the baseline strategy.

The baseline strategy is to select the user points as anchor points. Let us consider a set of *result circles* with an identical radius $d_a + f(r)$ whose centers are the sample locations selected by our approach. The result circles of baseline strategy can cover the whole query zone when $f(r) = r$.

Lemma 1. *Given a set of sample locations selected by the baseline strategy, for any query location in the query zone, its distance to the nearest sample location is no more than* $d_a + r$.

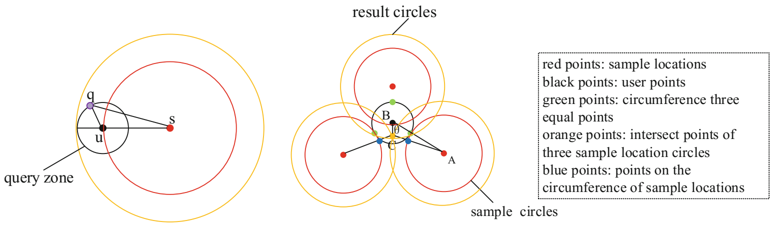
Proof. We denote by u a user point, s the nearest sample location to u selected by the baseline strategy, and q a query location in the user circle of u, as shown in Fig. 3(a). We have $dis(s, q) \leqslant dis(s, u) + dis(u, q)$ according to the triangle inequality. Since $dis(s, u) \leqslant d_a$ and $dis(u, q) = r$, $dis(s, q) \leqslant d_a + r$. Thus, for any query location in the query zone, there exists at least a sample location like s such that the distance between them is no more than $d_a + r$.

Usually, the value of r is relatively very small, so that $d_a + r$ could be a tight bound of d_z. While, if the value of r is not that small, we can use an improved strategy to get a tighter bound, which selects more anchor points other than the user points. For each user circle, we divide its circumference into a number of equal arcs, and use the end points of these arcs as the additional anchor points. Here we only discuss about dividing the circumference into three equal arcs, as shown in Fig. 3(b). We have four anchor points, the black user point and the three green points on the circumference. In this case, the result circles of improved strategy can cover the whole query zone when $f(r) = \frac{r(2d_a+r)}{3d_a+r}$.

Lemma 2. *Given a set of sample locations selected by using the users points and three equal points on the circumference of user circles as anchor points, for any query location in the query zone, its distance to the nearest sample location is no more than* $d_a + \frac{r(2d_a+r)}{3d_a+r}$.

Proof. Consider the worst case that makes the value of $f(r)$ maximized (the trivial proof is omitted). As shown in Fig. 3(b), two sample circles are tangent to the query zone, the other one is intersect with the query zone in the original user point, so that the four anchor points are exactly covered by the sample circles. Since some area of the query zone are not reached, then in order to cover the whole query zone, these sample circles with a radius d_a must be extended to the result circles with a radius $d_a + f(r)$ and intersect in a same point, like the orange point shows. In Fig. 3(b), the blue points and the black point are all remove a same distance d and intersect in the orange point, consider the triangle, $dis(A, B) = r + d_a$, $dis(A, C) = d + d_a$, $dis(B, C) = d$, and $\theta = 60°$. According to the cosine theorem, $\cos 60° = \frac{d^2+(r+d_a)^2-(d+d_a)^2}{2d(d_a+r)}$. Then we can attain $d = \frac{r(2d_a+r)}{3d_a+r}$, due to $\frac{2d_a+r}{3d_a+r} < 1$, so $d < r$.

When r is varying in a certain region, the returned d_a of improved strategy is approximately the same with the returned d_a of baseline strategy. Since $f(r) = \frac{r(2d_a+r)}{3d_a+r}$ of improved strategy is smaller than $f(r) = r$ of baseline strategy, the


q
s
u
query zone
result circles
B
θ
C
A
sample circles
red points: sample locations
black points: user points
green points: circumference three equal points
orange points: intersect points of three sample location circles
blue points: points on the circumference of sample locations

(a) Basic strategy.

(b) Improved strategy.

Fig. 3. An example of anchor point selection. (Color figure online)

upper bound $d_a + f(r)$ is tighter in the improved strategy compared to the baseline strategy. Note that, selecting more anchor points will make the bound tighter, however the cost of sample location selection algorithm will increases, the detailed comparison of objective distance and efficiency will be presented in Sect. 5.

Sample Location Selection Heuristics. Given a set N of anchor points, we aim to find a set SL of l sample locations in the space to minimize the maximum distance between any anchor point and its nearest sample location, namely, $\max_{p \in N} \min_{s \in SL} dis(p, s)$, which is called l-center problem. In a nutshell, the heuristics of our solution to l-center problem is as follows. Let $\alpha = \{I_1, I_2, \ldots, I_l\}$ be a l-partition of N, namely, $\cup_{i=1}^{l} I_i = N$, where $I_i \subset N$. Given an optimal l-partition α, we find a center point for each $I_i \in \alpha$ by addressing a 1-center problem for I_i, and select the l center points as the final sample locations.

To get the optimal l-partition, we need to define an objective function. Let $F(I)$ be the optimal objective distance of 1-center problem for I. We have

$$F(I) = \min_{x \in X} \max_{p \in I} dis(p, x) \tag{2}$$

where X is the set of all points in the space. For convenience, let $B(I)$ be the optimal point of 1-center problem for I. Then, let $F(\alpha)$ be the objective function for an l-partition. Thus, we have

$$F(\alpha) = \max_{i=1}^{l} F(I_i) \tag{3}$$

Obviously, the optimal l-partition with respect to $F(\cdot)$ leads to the sample locations with the minimum objective distance. In particular, for $I_i \in \alpha$, if $F(I_i) = F(\alpha)$, then I_i is called *extremal subset.*

4.2 Algorithm

The pseudo code of sample location selection algorithm is given in Algorithm 1. Initially, we choose l anchor points as centers (line 1), and assign each other

Algorithm 1. sample location selection

```
Input: a set $N$ of anchor points and a positive integer $l$
Output: $l$ sample locations and $F_\alpha$
 1: choose $l$ center points out of $N$;
 2: assign each other anchor point to the subset of its nearest center by using Voronoi
    diagram;
 3: repeat
 4:     $i \leftarrow$ a point from $T(I_i)$, where $I_i$ is the extremal set of $\alpha$;
 5:     choose a subset $I_j$ other than the extremal subset $I_i$;
 6:     if $F(I_j \cup \{i\}) < F(\alpha)$ then
 7:         $I_j \leftarrow I_j \cup \{i\}$, $I_i \leftarrow I_i - \{i\}$;
 8:     end if
 9: until the value of $F(\alpha)$ does not change anymore
10: return the optimal point $B(I_i)$ of each subset $I_i \in \alpha$ and $F_\alpha$;
```

Algorithm 2. 1-center problem algorithm

```
Input: a set $I$ of anchor points
Output: the optimal center point $B(I)$, and $F(I)$
 1: choose the initial center point $(x_0, y_0)$, where $x_0 = \sum_{p \in I} x_p/|I|$, $y_0 = \sum_{p \in I} y_p/|I|$;

 2: $I' \leftarrow$ the three points that are farthest from $(x^{(0)}, y^{(0)})$;
 3: while there exists a point in $I - I'$ such that the distance between it and $B(I')$ is
    larger than $F(I')$ do
 4:     $p' \leftarrow$ the farthest point from $B(I')$;
 5:     $I' \leftarrow T(I' \cup \{p'\})$;
 6: end while
 7: return $B(I')$ and $F(I')$;
```

anchor point to the subset of its nearest center by leveraging the principle of Voronoi diagram (line 2). Then we refine the partition α of anchor points iteratively until the value of $F(\alpha)$ cannot be decreased (line 3–9). At each iteration, we try to move a point from a subset to another to get a better value of $F(\alpha)$. Straightforwardly, we can reallocate each anchor point to another subset, and choose the best plan. However, there are $N(l-1)$ possible plans, and not all of them can decrease the value of $F(\alpha)$. Thus, we give an efficient repartition method as follows. According to the study of minimum covering circle problem in [14], the value of $F(I)$ can be determined by no more than three points in I, the set of which is denoted by $T(I)$, namely, $F(I) = F(T(I))$. Given an extremal set I_i of α, we have $F(\alpha) = F(T(I_i))$, so that the value of $F(\alpha)$ will be changed if we remove a point $i \in T(I_i)$ from I_i. As a result, we only consider to reallocate the anchor points in $T(I_i)$ to achieve a better value of F_α. Lastly, the center points of the optimal partition of N is returned as the sample locations.

Algorithm 2 gives a solution to 1-center problem and the complexity is $O(n)$. Initially, for a subset I of anchor points, we choose a point (x_0, y_0) in the space as the center of I (line 1). Since $F(I) = F(T(I))$, we choose the three farthest points from (x_0, y_0) to compose a set I' as the possible $T(I)$ (line 2). Then we begin to update I' iteratively unless there is no point in $I - I'$ outside of the circle determined by I', namely, $I' = T(I)$ (line 3–6). At each iteration, we choose the farthest point p' from $B(I')$, and set the new I' as $T(I' \cup \{p'\})$. Lastly, we return $B(I')$ as the optimal center of I since $F(I) = F(T(I)) = F(I')$.

To get the center $B(I')$ of I' that has exact three points, the *three-point problem* is studied in [15]. The idea is that, first check if any two points p_1 and p_2 define the solution. If so, let $x = (x_1 + x_2)/2$ and $y = (y_1 + y_2)/2$, the distance between the other point p_3 and (x, y) is no more than $dis(p_1, p_2)/2$, and thus (x, y) is the center of these three points. Otherwise, we find a point inside the triangle of these three points as the center $B(I')$, which possesses equal distances to the three vertices of the triangle.

5 Experiments

Our experiments are conducted on a PC with Intel Core 3.2 GHz CPU and 16 GB memory. The algorithms are implemented in C++ with TDM-GCC 4.9.2 (Table 1).

Table 1. Experimental datasets.

Dataset	Node number	Edge number	Average in-degree	Average out-degree
Brightkite	58K	428K	7	7
Gowalla	100K	1.9M	13	13

5.1 Setup

Algorithms. There are four algorithms to be compared in our experiments. (1) RSQ extends RS to filter out the sample locations outside of the query zone, while RS is the original random sampling in [9]. The distance between query location and sample location is calculated by using Voronoi diagram. (2) K-means simply clusters the user points to a given number of groups with respect to distance, and select the cluster center as sample locations. (3) FLS is our facility-location-based sampling method with the baseline anchor point selection strategy. (4) FLS-3 is our facility location sampling method with an improved anchor point selection strategy which adds three points on each query zone circle into the anchor points.

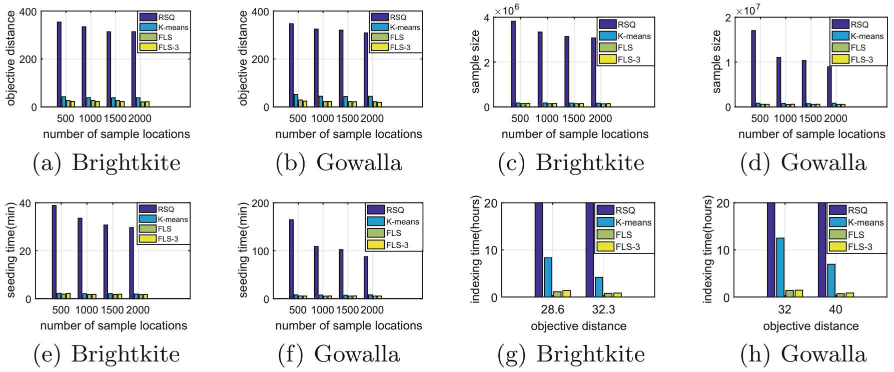
Datasets. In our experiments, we use two real-world geo-social networks where users can share their check-ins. This check-ins represent users' locations, and the datasets are obtained from http://snap.stanford.edu/data/. Note that there are just 88.6% and 54.4% users have check-ins in the Brightkite and the Gowalla respectively, we need to pre-treat the datasets as follows: since there are a few users who don't have location information in Brightkite, so we randomly generate a location for them according to other users' location distribution. While in Gowalla, almost half of them have no check-ins, so we delete those users who don't have location information, in fact, there are 100K points in Gowalla used by us. The location information of added anchor points are calculated according to the original user points.

Parameters. The probability of edge (u, v) is set as $\frac{1}{N_{in}(v)}$, where $N_{in}(v)$ represents the number of incoming neighbours of v. The independent cascade model is used in influence spread, the size of seed set varies from 10 to 50, the interval is 10, and we run 10000 round for each returned seed set, we evaluate the average influence propagation of the returned seed set. The radius of each query zone is set to 10, for each query zone, the entire circumference is divided into 3 equal arcs, and add the end points of these equal arcs into the anchor points, the number of sample locations is set to 500, 1000, 1500 and 2000 relatively.

5.2 Effectiveness Analysis

We evaluate the effectiveness of sample location selection algorithms by four metrics. The first one is of course the objective distance, the other two, namely, the necessary size of network samples (simplified as sample size) and the response time of seed selection (simplified as seeding time), and the last one is the time consumption of offline index construction (simplified as indexing time), are used to demonstrate the impact of location sampling on a specific DAIM approach. Once we get the sample locations, we run the offline index construction and then the online seeding algorithm of RIS-DA [9]. RIS-DA needs to generate a set of network samples, the number of which is determined by the objective distance, in order to guarantee the $1-1/e-\epsilon$ approximate ratio of influence maximization. During online seeding, RIS-DA needs to deal with each network sample. Thus, the response time of seed selection also depends on the objective distance finally. Since the objective distance can be reduced by increasing the number of sample locations, but the efficiency of offline index construction will be influenced if the number of sample locations gets larger, so the indexing time indirectly depends on the objective distance.

As shown in Fig. 4(a) and (b), our facility-location-based sampling approach can achieve the best objective distance on both datasets with varying numbers of selected samples. Note that, the objective distance of RSQ shown in Fig. 4(a) and (b) is the average of results of 100 repeated tests, thereby avoiding the bias of random sampling. We can see that, the objective distance of naive sampling approaches such as RSQ is much worse than K-means, FLS and FLS-3. Although K-means can reduce the objective distance significantly, it is still not as effective as FLS and FLS-3. Compared with FLS, FLS-3 is even more effective when the number of selected samples is small, due to the improved anchor point selection strategy. Moreover, the objective distance decreases with the increase of the number of selected samples generally for all algorithms. While, the decrease of FLS and FLS-3 is not significant. Therefore, we can actually select a relatively



(a) Brightkite (b) Gowalla (c) Brightkite (d) Gowalla

(e) Brightkite (f) Gowalla (g) Brightkite (h) Gowalla

Fig. 4. Effectiveness evaluation.

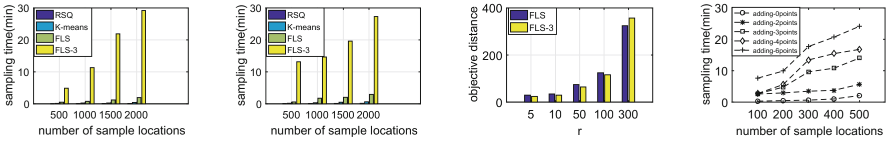
small number of sample locations to get a good objective distance by using our facility-location-based sampling approach.

Due to the decrease of objective distance, FLS and FLS-3 improve the online performance of RIS-DA dramatically. As shown in Fig. 4(c) and (d), the necessary size of network samples of FLS and FLS-3 is orders of magnitude less than naive sampling. Further, the response time of seed selection is shown in Fig. 4(e) and (f). With the decrease of sample size, the response time is reduced to only a few minutes, so that we can quickly find a set of seed users for promotion of a given query location. In conclusion, given a fixed budget of location sampling, FLS and FLS-3 are quite effective to improve the online performance of DAIM approaches by reducing the objective distance.

We keep the objective distance same to evaluate the response time of offline index construction. As shown in Fig. 4(g) and (h), we can find the time consumption of K-means is longer than FLS and FLS-3, this is because for the same objective distance, the number of sample locations in K-means is greater than that in FLS and FLS-3. For example, to achieve the objective distance of 28.6, there are 1000 sample locations needed in K-means, while just 80 in FLS and 100 in FLS-3. Note that, the number of sample locations of RSQ will be greater than 2000 in order to achieve the same objective distance, then the response time will be several times larger than K-means, so we set a time limit of 20 h, once the running time goes beyond the time limit, the algorithm will be stopped. Due to the efficiency of offline index construction is sensitive to the number of sample locations, the offline time-consuming of RSQ and K-means will be enormous to achieve the same online performance compared with FLS and FLS-3. We can conclude that FLS and FLS-3 can well balance the overheads of offline precomputation and online performance.

5.3 Efficiency Analysis

We evaluate the efficiency of sample location selection algorithms by focusing on the response time of selecting certain number of sample locations. As shown



(a) Brightkite (b) Gowalla (c) Brightkite (d) Brightkite

Fig. 5. Efficiency and anchor point selection evaluation.

in Fig. 5(a) and (b), RSQ runs the fastest among all the algorithms, K-means always outperforms FLS and FLS-3, and FLS-3 performs the worst. For RSQ, since it is just randomly selecting sample locations in the 2D space and filtering the sample locations outside of the query zone, there is little time consumption during the process. Compared with K-means, FLS and FLS-3 need to refine the partition and recalculate the center locations until the value of objective function is minimized. Since FLS and FLS-3 need to call the 1-center problem algorithm for all sample locations even there is a very tiny variety of the value of objective function in each iteration, such process will directly influence the response time of FLS and FLS-3. Furthermore, the number of anchor points in FLS-3 is three times more than FLS, when the number of sample locations is fixed, the number of points in each subset of α in FLS-3 will be larger than FLS, then the response time of 1-center problem algorithm in FLS-3 will be larger. Thus, we can find that the improved anchor point selection strategy may improve the objective distance, but the efficiency is decreased.

5.4 Anchor Point Selection Strategy Analysis

We evaluate the effectiveness of anchor point selection strategy by the objective distance. As shown in Fig. 5(c), we focus on the objective distance of baseline strategy and improved strategy when r is varying. Note that, considering the problem of efficiency, we choose 200 sample locations here to evaluate the objective distance. We can find the objective distance of FLS-3 is smaller than FLS when r is under a certain value like 100 in Fig. 5(c). When r is too large, the value of d_a in FLS-3 will be several times greater than FLS, and the value of $f(r)$ is increasing when r and d_a both increase. Thus, the objective distance of FLS-3 will be larger than FLS if r is too large. However, considering the users' regular activity area won't be too large in reality, so the objective distance of the improved strategy is better than the baseline strategy. As a result, we can achieve a tight upper bound by adding extra anchor points.

We evaluate the efficiency of anchor point selection strategy by comparing the response time of sampling when adding different number of anchor points. As shown in Fig. 5(d), except for the baseline strategy and improved strategy, we add extra 2, 4 and 6 anchor points. From the results, we can find the time consumption of each strategy is increasing when the number of sample locations increases, this is because the number of calling 1-center problem algorithm is increasing when the number of sample locations increases. For a fixed number of sample locations, since the number of points in each subset of α is increasing when we add more anchor points, the response time of sampling will be increasing. Thus, in order to balance the effectiveness and efficiency, it is important to determine the number of anchor points to add.

6 Conclusion

Sample location selection is crucial for the DAIM problem in geo-social network, but there is no work to fully study such a problem. The previous work mainly selects sample locations by naive methods such as random sampling or equal cell sampling, which can hardly achieve a good objective distance even when a large number of samples are selected. While, the online seeding performance is sensitive to the objective distance, and the precomputation overhead is sensitive to the sample number. In this paper, we propose the conception of query zone and reasonably formulate a novel problem of sample location selection for a given query zone. Due to the hardness of this problem, we solve our problem by selecting some anchor points from the query zone and finding a number of centers of the anchor points as the sample locations. Specifically, we propose a flexible strategy of anchor point selection and develop a heuristic partition refining algorithm to select centers. According to the experimental results, our approach can improve the efficiency of DAIM approach like [9] significantly. Since our approach can achieve a specific objective distance by selecting much less sample locations, we can balance the online performance and precomputation overhead effectively.

Acknowledgments. This paper was supported by National Natural Science Foundation of China under Grant No. 61202036, 61502349 and 61572376.

References

1. Kempe, D., Kleinberg, J.M., Tardos, E.: Maximizing the spread of influence through a social network. In: SIGKDD, pp. 137–146 (2003)
2. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: ACM KDD (2007)
3. Chen, W., Wang, C., Wang, Y.: Scalable influence maximization for prevalent viral marketing in large-scale social networks. In: SIGKDD, pp. 1029–1038 (2010)
4. Cohen, E., Delling, D., Pajor, T., Werneck, R.F.: Sketch-based influence maximization and computation: scaling up with guarantees. In: CIKM, pp. 629–638 (2014)
5. Chen, W., Yuan, Y., Zhang, L.: Scalable influence maximization in social networks under the linear threshold model. In: International Conference on Data Mining, pp. 88–97 (2010)
6. Zhu, W., Peng, W., Chen, L., Zheng, K., Zhou, X.: Modeling user mobility for location promotion in location-based social networks. In: SIGKDD, pp. 1573–1582 (2015)

7. Cai, J.L.Z., Yan, M., Li, Y.: Using crowdsourced data in location-based social networks to explore influence maximization. In: IEEE International Conference on Computer Communications, pp. 1–9 (2016)
8. Li, G., Chen, S., Feng, J., Tan, K., Li, W.: Efficient location-aware influence maximization. In: SIGMOD, pp. 87–98 (2014)
9. Wang, X., Zhang, Y., Zhang, W., Lin, X.: Efficient distance-aware influence maximization in geo-social network. IEEE Trans. Knowl. Data Eng. **29**(3), 599–612 (2017)
10. Wang, X., Zhang, Y., Zhang, W., Lin, X.: Distance-aware influence maximization in geo-social network. In: ICDE, pp. 1–12 (2016)
11. Weber, A.: Über den Standort der Industrien 1. Reine theorie des standordes, Tübingen, Germany, Teil (1909)
12. Arabani, A.B., Farahani, R.Z.: Facility location dynamics: an overview of classifications and applications. Comput. Ind. Eng. **62**(1), 408–420 (2012)
13. Irawan, C.A., Salhi, S.: Aggregation and non aggregation techniques for large facility location problems: a survey. Yugosl. J. Oper. Res. **25**(3), 313–341 (2015)
14. Elzinga, J., Hearn, D.W.: Geometrical solutions for some minimax location problems. Transp. Sci. **6**, 379–394 (1972)
15. Drezner, Z., Wesolowsky, G.O.: Single facility l_p-distance minimax location. SIAM J. Algebr. Discret. Methods **3**, 315–321 (1980)
16. Drezner, Z.: The *p*-centre problem-heuristic and optimal algorithm. J. Oper. Res. Soc. **35**(8), 741–748 (1984)
17. Callaghan, B., et al.: Speeding up the optimal method of Drezner for the p-centre problem in the plane. Eur. J. Oper. Res. **257**(3), 722–734 (2017)
18. Fowler, R.J., Paterson, M.S., Tanimoto, S.L.: Optimal packing and covering in the plane are NP-complete. Inf. Process. Lett. **12**(3), 133–137 (1981)