Hindawi
Complexity
Volume 2021, Article ID 2581288, 15 pages
https://doi.org/10.1155/2021/2581288

Research Article

Efficient Sample Location Selection for Query Zone in Geo-Social Networks

Kejian Tang,[1] Shaohui Zhan,[1] Tao Zhan,[1] Hui Zhu,[2] Qian Zeng,[3] Ming Zhong,[3] Xiaoyu Zhu,[3] Yuanyuan Zhu,[3] Jianxin Li,[4] and Tieyun Qian[3]

[1]*Jiangxi Branch, State Grid Corporation of China, Beijing, Jiangxi, China*
[2]*Beijing Huitong Jincai Information and Technology Company Limited, Beijing, China*
[3]*School of Computer Science, Wuhan University, Wuhan, Hubei, China*
[4]*School of Information Technology, Deakin University, Melbourne, Australia*

Correspondence should be addressed to Ming Zhong; clock@whu.edu.cn

Received 13 May 2021; Accepted 23 November 2021; Published 17 December 2021

Academic Editor: Rosa M. Benito

While promoting a business or activity in geo-social networks, the geographical distance between its location and users is critical. Therefore, the problem of Distance-Aware Influence Maximization (DAIM) has been investigated recently. The efficiency of DAIM heavily relies on the sample location selection. Specifically, the online seeding performance is sensitive to the distance between the promoted location and its nearest sample location, and the offline precomputation performance is sensitive to the number of sample locations. However, there is no work to fully study the problem of sample location selection for 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 (query zone). Then, we propose two efficient location sampling approaches based on facility location analysis, which is one of the most well-studied areas of operations research, and these two approaches are denoted by Facility Location based Sampling (FLS) and Conditional Facility Location Based Sampling (CFLS), respectively. FLS conducts one-time sample location selection, and CFLS extends the one-time sample location selection to a continuous process, so that an online advertising service can be started immediately without sampling a lot of locations. Our experimental results on two real datasets demonstrate the effectiveness and efficiency of the proposed methods. Specifically, both FLS and CFLS can achieve better performance than the existing sampling methods for the DAIM problem, and CFLS can initialize the online advertising service in a matter of seconds and achieve better objective distance than FLS after sampling a large number of sample locations.

1. Introduction

1.1. Motivation. The widely used geo-position-enabled devices (mobile phone, tablets, laptops, etc.) and services (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 the 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.

There are two typical problem definitions for the above scenario. The first one is called location-aware influence maximization (LAIM) [1]. 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 the LAIM problem, how to select an appropriate query region for a given query location is unclear

[2]. If the query region is too large, most users influenced by the selected seeds may be distributed near the boundary of the region, thereby being far away from the given 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 definition called distance-aware influence maximization (DAIM) [3] has been proposed. For the DAIM problem, each user has a weight that is determined by the distance between the user and the query location no matter whether the user is in a query region, and the influence spread to users is adjusted according to their weights.

Typically, to address the DAIM problem, the existing approaches [2, 3] select a set of sample locations in the 2D space where the users are distributed, and precompute the influence spread of the sample locations. Then, for an arbitrary query location, its influence spread can be approximated according to the influence of its nearest sample location during the online seeding process. Note that, the shorter the distance between the given query location and its nearest sample 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 under simplified sampling methods like random sampling [2] or equal cell sampling [3]. They need to sample a large number of locations to achieve a reasonable objective distance. Thus, it is unlikely to achieve a good online seeding performance without incurring heavy precomputation overhead while using such simple sampling. Let us consider the following example.

Example 1. Figure 1(a) shows the spatial distribution of users in Brightkite, a real-world geo-social network. We can see that most users live in a few urban areas. The equal cell sampling ignores this fact and tries to reach an arbitrary point in the space within the minimum distance, as shown in Figure 1(b). However, it is unreasonable 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, in a "query zone" around the users. Figure 1(c) shows an example of query zone consisting of circles centered at each user with an identical radius r. Thereby, any query location in the query zone is no farther than r from at least a user. Then, we can use a more delicate sampling method to reduce the number of sample locations required for achieving the same objective distance. As shown in Figure 1(d), the query zone can be covered by the red circles centered at only a few samples, and the radius of circles is equal to the half-length of the diagonal line of equal cells in Figure 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 the users of real-world geo-social networks are usually distributed sparsely in a 2D space. In other words, 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 weight is 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 approaches for selecting sample locations.

We develop two sample location selection approaches denoted by Facility Location Based Sampling (FLS) [4] and Conditional Facility Location Based Sampling (CFLS), respectively. Compared with the existing sampling methods [2, 3], FLS conducts a one-time sample location selection based on the spatial distribution of users. Specifically, FLS exploits the existing techniques of the l-center problem, one subtype of classic facility location problem, which is to find l points in the space that can reach any user with the minimum distance. The objective distance achieved by FLS is much smaller than the existing sampling approaches, thereby improving the efficiency of online seeding. Moreover, to achieve the same objective distance with the existing sampling approaches, FLS only needs to select a much smaller number of sample locations, which can significantly reduce the precomputation overhead.

While it largely reduces the precomputation overhead, FLS still adopts the scheme that the precomputation must be completely done before starting online advertising. To further accelerate the process, we propose CFLS that achieves a quick start. Specifically, it only selects a small number of sample locations at first, and thus the initial objective distance may be very large. Then, it can reduce the objective distance and improve the quality of the online advertising service by adding sample locations in subsequent steps. Moreover, given the same number of sample locations, we have observed that CFLS can achieve better objective distance than FLS when the number of sampled locations is large enough.

1.2. Our Contributions. Our contributions are generalized as follows:

(i) We formalize a novel and important 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 that adheres to the geographical distribution of social network users.

(ii) We devise two efficient location sampling approaches. FLS can achieve much smaller objective distance and improve the efficiency of online seeding. Furthermore, CFLS extends the sample location selection to a continuous process, so that

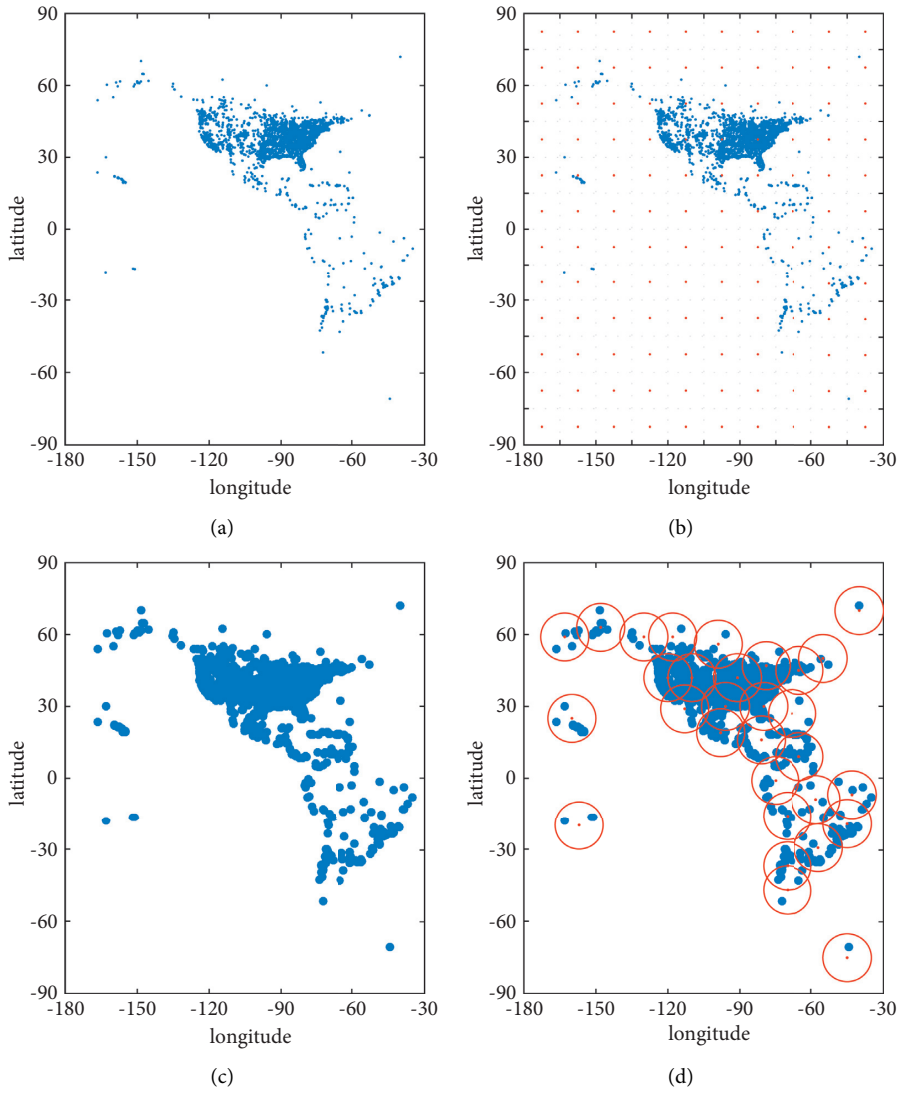

Figure 1: A motivation example on a geo-social network named Brightkite.

the online advertising service can be started immediately.

(iii) We perform comprehensive experiments on two real-world datasets. The experimental results confirm the effectiveness and efficiency of the proposed techniques.

The rest of this paper is organized as follows: We review the related work in Section 2. The formalized problem definition is given in Section 3. The FLS sample location selection approach is shown in Section 4, and we present the CFLS sample location selection approach in Section 5. The experiment results are demonstrated in Section 6, and we conclude our work in Section 7.

2. Related Work

2.1. Influence Maximization in Social Networks. Influence maximization problem is first defined by Kemple et al. [5]. The authors also define the independent cascade model and the linear threshold model, and they prove the hardness of the problem in the paper. Since then, there are a large number of literatures on influence maximization, like [6–11], and so on. In [6], the authors propose the CELF algorithm, which exploits the submodular property to significantly boost the traditional greedy approach. Chen et al. [7] propose the PMIA approach, and the influence is considered to propagate only through the maximum influence path between users. Cohen et al. [8] propose a bottom-k sketch-based approach to reduce the costs of influence estimation. The materialized sketch can be used as an oracle to evaluate the influence of any subset users. More recently, due to the popularity of network embedding [12, 13], many research works start to explore deep learning techniques in influence maximization.

2.2. Influence Maximization in Geo-Social Networks. With the appearance of geo-position-enabled devices and services, researchers begin to pay attention to the impact of

geographical location on influence maximization, like [1, 3, 14–16], etc. Zhu et al. [15] attempt to measure the influence between users by considering social relation and location information, while Li et al. [16] propose a novel network model and an influence propagation model, and they think the influence propagation should be conducted both in online social networks and physical world. Li et al. [1] attempt to maximize the influence spread in a query region. However, it is nontrivial to determine an appropriate query region when conducting a location-aware promotion. The works that are most relevant to ours is [2, 3], which proposes 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, while the RIS-DA approach comes up with an unbiased estimator for the distance-aware influence maximization. Both of these two approaches need to estimate the necessary size of network samples for any potential query location, and such process is very time-consuming. More recently, Cai et al. [17] devise a novel holistic influence diffusion model that takes into account both cyber and physical user interactions, and Haldar et al. [18] propose a method to infer the top activity location of social users using the implicit information available in the network.

2.3. Facility Location Problem. Researches in location theory were formally started in 1909 by Alfred Weber [19], who is 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 the existing techniques for facility location problem, like [20, 21]. Elzinga and Hearn give a geometric algorithm to solve the 1-center problem with Euclidean distances, and prove the correctness of the algorithm in [22]. Drezner [23] 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 algorithms and an optimal algorithm to solve the problem in [24]. Then Callaghan et al. [25] attempt to speed up the optimal method proposed in [24] by introducing neighbourhood reduction schemes and embedding a CPLEX policy.

2.4. Conditional Facility Location Problem. The conditional location problem was initially introduced by Minieka [26], who studied conditional centers and medians on a graph. Given the locations of p existing facilities, l additional facilities are needed to be located to minimize the maximum distance between any demand point and its nearest facility (whether existing or new) in the conditional l-center problem. Drezner [27] explains that the conditional l-center problem can be solved by solving $O(\log n)l$-center problem, where n represents the number of demand points. A method for solving both the conditional l-median and l-center problem is investigated by Berman and Drezner [28]. The method requires one-time solution of an unconditional l-median and l-center problem incorporating the shortest distance matrix. Chen and Chen [29] propose a relaxation-based algorithm for solving both the conditional discrete and continuous l-center problem. More recently, Zeng et al. [30] propose a novel business location planning approach for the emerging online-to-offline businesses. In the approach, online social network marketing is defined as an influence maximization process based on a particular diffusion model that is aware of offline factors such as competitive locations, target users, and geographic distance.

Compared to the existing works, our approach FLS selects sample locations based on facility allocation techniques, while CFLS is based on conditional facility allocation techniques. Both of our approaches can derive shorter objective distance than the existing sampling approaches. Moreover, CFLS extends FLS to a continuous process, so that the online advertising service can be started immediately.

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, and then we give a formal definition of the problem discussed in this paper.

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 $v \in V/S$ 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 distance-aware influence maximization (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.,

$$S^* = \arg\max_{S \subseteq V} \left\{ I_q(S) \middle\| S| = k \right\}, \tag{1}$$

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. [2, 3] 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 $1 - 1/e - \varepsilon$ with at least $1 - \delta$ probability. According to the comparison in [2], 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 into 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 overhead 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, and the number of sample locations be l. The time complexity of precomputation for MIA-DA is $O(n^2)$ and for RIS-DA is $O(l^2k^2n \log n)$. 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 companies need to advertise for their products through social networks, 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 $dis(q, v) \leq r$.

Intuitively, for a user, the area of activities is a circle centered at its location with a radius r, which is called the 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 Figure 1. We denote by $q \in Q$ that a point q is located in the query zone Q. Under this assumption, the problem to be addressed in this paper can be formalized as follows.

Problem 1. (sample location selection). 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, the total number $m (m = p + l)$ of sample locations, the p existing sample locations, and an additional location sampling budget l, find a set of l sample locations in the 2D space and denote the set of m sample locations by SL, 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 (whether existing or new), 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.

The total number m of sample locations is fixed, when $l = 0$, i.e., $p = m$, the selection of sample locations is a one-time process. Once $l > 0$, the process of sample location selection is continuous.

For example, as shown in Figure 2, there are two users v_1 and v_2, and the yellow circles comprise the corresponding query zone. If $p = 1$, as shown in Figure 2(a), the sample location s_1 is the middle point of line segment v_1v_2. Thus, the farthest query locations to s_1 are 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 $D(\{s_2\}, Q) = dis(q_2, s_2) = dis(v_1, s_2) + r > dis(v_1, s_1) + r = dis(q_1, s_1)$ It is obvious that the minimum objective distance $d_o = dis(q_1, s_1)$, so that the optimal set of sample location is $\{s_1\}$. If we have selected one sample location s_1, and we need to sample one more location, since the maximum distance between any query point in Q and its nearest sample location (whether existing or new) needs to be minimized, the two sample locations will be distributed as shown in Figure 2(b), and the objective distance $d_o = dis(q_1, s_1) = dis(q_3, s_2) = r$.

3.2.1. Problem Hardness. As problem 1 has described, the sample location selection process can be completed in several steps. Since the location sampling in each step can be simplified to the problem discussed in [4], and it has been proved in [4] that the location sampling in each step is NP-hard, thereby problem 1 is NP-Hard.

4. Facility Location-Based Sampling (FLS)

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 allocation techniques [23, 24].

4.1. Methodology. Due to the hardness of the 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 its nearest sample location within the minimum distance d_a. Note that, FLS selects all sample locations at once, i.e., $p = m$ and $l = 0$. Let d_z be the maximum distance between any point in the query zone and

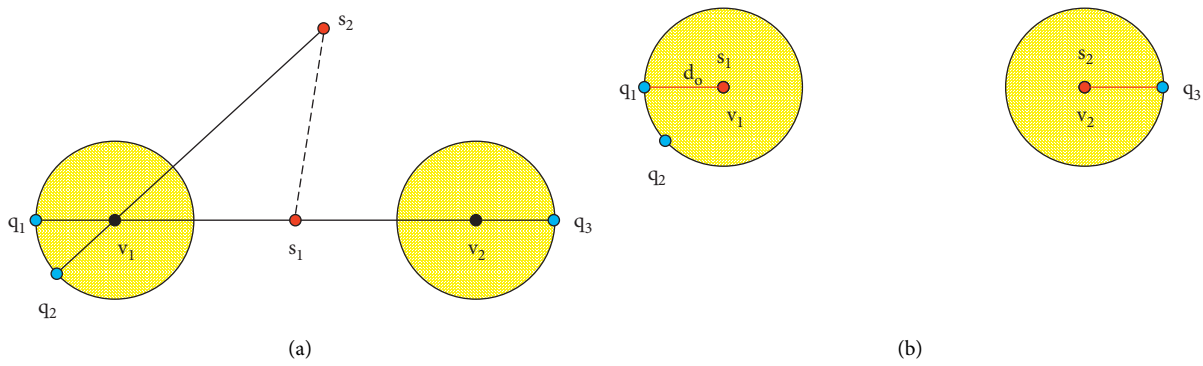

Figure 2: A simple example of sample location selection problem.

its nearest selected sample location. Although there could be some areas of query zone that cannot 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 certain 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.

Note that, two anchor point selection strategies have been proposed in [4], and we exploit the baseline anchor point selection strategy in this paper.

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

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

$$F(I) = \min_{x \in X} \max_{ap \in I} di\ s(ap, x), \tag{2}$$

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 m-partition. Thus, we have

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

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

4.2. Algorithm. The pseudo code of sample location selection algorithm is given in Algorithm 1. Initially, we choose m anchor points as centers (line 1), and assign each 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 $N(m-1)$ 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 [22], 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 $F(\alpha) = F(T(I))$, 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, by calling Algorithm 2, the center points of the optimal partition of N are 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 ap' from $B(I')$, and set the new I' as $T(I' \cup \{ap'\})$. 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 [23]. The idea is that, first check if any two points ap_1 and ap_2 define the solution. If so, let $x = (x_{ap_1} + x_{ap_2})/2$, $y = (y_{ap_1} + y_{ap_2})/2$, and distance between the other point ap_3 and (x, y) is no more than $di\ s(ap_1, ap_2)/2$, then (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.

Input: a set N of anchor points and a positive integer m;
Output: the m sample locations and F_α;
1: choose m 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 α;
5: randomly 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_α;

Algorithm 1: Sample location selection.

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 $x_0 = \sum_{ap \in I} x_{ap}/|I|$, $y_0 = \sum_{ap \in I} y_{ap}/|I|$;
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: $ap' \leftarrow$ the farthest point from $B(I')$;
5: $I' \leftarrow T(I' \cup \{ap'\})$;
6: end while
7: return $B(I')$ and $F(I')$;

Algorithm 2: 1-center problem algorithm.

4.3. Analysis. We have an observation that FLS may fall into local optimization when the total number of sample locations is large enough. As shown in Figure 3, the objective distance remains the same when the number of sample locations is 500, 1000, and 1500, respectively. We analyze the distribution of all sample locations and show the proportion of the number of sample locations within a certain radius in Table 1. Given a sample location and the anchor points assigned to this sample location, the radius represents the maximum distance among anchor points and the sample location. As the number of sample locations increases, the number of sample locations with lower radius increases, while the number of sample locations with higher radius decreases. We can conclude that the radius decreases on the whole, but the optimization is only local rather than global because the objective distance remains the same as the number of sampled locations increases. The reasons for local optimization of FLS are presented as follows.

First of all, FLS performs a one-time selection of all sample locations, and the objective distance is derived by refining the partition iteratively. When refining the partition, we remove points from the binding set of the extremal subset to other subsets only if the objective distance will be decreased. However, it is obvious that there will be no refinement if the extremal subset is far away from other subsets since no points in the extremal subset can be removed to other subsets. Moreover, the center in a densely populated location will not be refined to the sparsely populated location because these two locations are always far away from each other and the value of F_α will not be decreased. As the refinement occurs, the center of each subset may change, but the total number of centers remains the same, that is, the centers which are close to each other will not be combined into one even if the distance among them is very short and the corresponding radius is small.

Since the total number of sample locations is fixed, the distribution of sample locations will follow the spatial distribution of anchor points. In other words, the number of sample locations in densely populated areas is much larger than in sparsely populated areas.

Lemma 1. *The number of sample locations in densely populated areas is much larger than in sparsely populated areas. Let N represent the total anchor points, d_1 or d_2 represent the densely populated area (or the sparsely populated area), N_1 or N_2 represent the number of anchor points in d_1 or the number of anchor points in d_2, and SN_1 or SN_2 represent the number of sample locations in d_1 (or the number of sample locations in d_2), we have $SN_1 > SN_2$.*

Proof 1. Since the initial centers are randomly selected from the anchor points, the possibility that each anchor point is selected as a center is $1/N$. In general, the number of anchor points in densely populated areas is greater than in sparsely populated areas, thereby $N_1 > N_2$. Assuming the fixed number of sample locations is m, initially, we have $SN_1 = N_1 \cdot 1/N \cdot m$ and $SN_2 = N_2 \cdot 1/N \cdot m$. Since $N_1 > N_2$, we have $SN_1 > SN_2$. Once the initial m centers are selected,

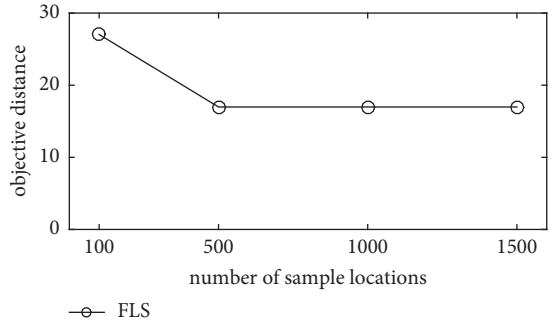

Figure 3: The objective distance of sample location selected by FLS in Brightkite.

Table 1: Distribution of sample locations selected by FLS.

Sample location #	Radius (%) ≤5	5 < radius (%) ≤10	10 < radius (%) ≤15	Radius (%) >15
500	81.4	12.6	5.8	0.2
1000	88.5	11	0.4	0.1
1500	92.2	7.53	0.2	0.07

other anchor points will be assigned to their nearest center and form a m-partition. Then, the partition will be refined until the value of F_α is no longer decreased. As mentioned above, the centers will not be combined during the refinement, and the center in a densely populated location will not be refined to the sparsely populated location because they are always far away from each other and the value of F_α will not be decreased. Therefore, SN_1 is always greater than SN_2.

The radius of centers in densely populated areas is very small because there are much more sample locations in such areas. The distance among anchor points in sparsely populated areas is always large, i.e., the radius of centers in such areas is large; thereby, the objective distance is determined by the radius in such areas according to equation (3). As shown in Table 1, as the number of sampled locations increases, the objective distance remains the same since the center in the most sparsely populated area cannot be refined to other areas. Thus, we can conclude that the optimization is mainly conducted in densely populated areas but not the most sparsely populated area. In other words, FLS may achieve a local optimal value when a large number of sample locations are selected at one time.

5. Conditional Facility Location-Based Sampling (CFLS)

In this section, we devise the CFLS sampling approach, which extends the FLS approach to a continuous process and thereby can make the online advertising service start immediately.

While the precomputation overhead has been largely reduced compared to the existing sampling approaches [2, 3], FLS still adopts the scheme that the precomputation must be completely done before starting online advertising. To further accelerate the process, we propose to conduct the sample location selection in several steps and only select a small number of sample locations in each step. Whenever the sample location selection in one step is completed, the online seeding process can be started. Since CFLS only selects a small number of sample locations at first, the selection can be completed quickly, but the objective distance may be very large. However, it can reduce the objective distance and improve the quality of the online advertising service by adding sample locations in subsequent steps. As mentioned in Section 4.3, the objective distance is determined by the radius of centers in sparsely populated areas, so CFLS attempts to invest sample locations in sparsely populated areas in subsequent steps, and thereby the objective distance can be effectively decreased.

Moreover, we have an observation that given the same number of sample locations, CFLS can achieve better objective distance than FLS when the number of sample locations is large enough. Since FLS selects all sample locations at once, many sample locations are distributed in densely populated areas based on Lemma 2, thereby the objective distance of FLS is larger because the objective distance is determined by the radius of centers in sparsely populated areas. In contrast, CFLS invests sample locations in sparsely populated areas in subsequent steps, so there are a large number of sample locations in sparsely populated areas, which leads to a better objective distance.

For example, as shown in Figure 4, the number of sample locations is the same in Figures 4(a) and 4(b). The red points represent sample locations, the black points represent anchor points, and the radius of red circles represents the objective distance. In Figure 4(a), we can find that there are much more sample locations in densely populated areas. However, these areas can be covered by even less sample locations as shown in Figure 4(b), and the extra sample locations can be invested in sparsely populated areas. Therefore, the objective distance can be reduced because there are much more sample locations in sparsely populated areas, as the objective distance in Figure 4(a) is 11.84, and the objective distance in Figure 4(b) is 11.27.

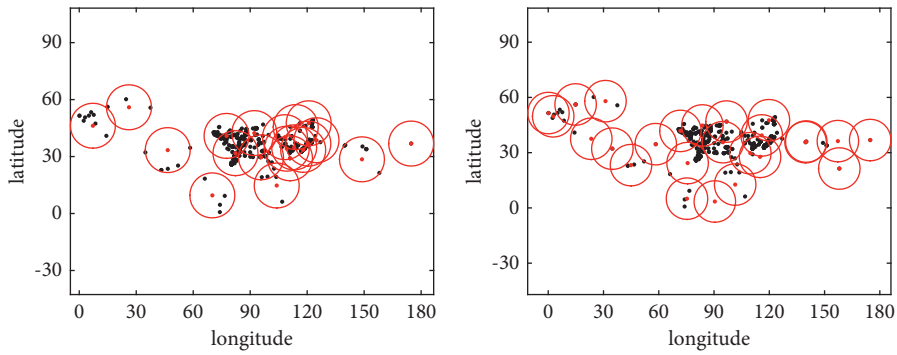

Figure 4: Explanation for the motivation of CFLS.

5.1. Methodology. We divide the sample location selection into continuous steps and select a small number of sample locations in each step. Formally, we denote by $\langle p, l_1, l_2, \ldots, l_n \rangle$ the sample location selection in continuous steps. Specifically, given the fixed number of sample locations $m (m = p + l_1 + l_2 + \cdots + l_n)$, we first select p sample locations, then we select another l_1 sample locations based on the results of the previous p sample locations, and select another l_2 sample locations based on the previous $p + l_1$ sample locations, and so on. Note that, each anchor point is assigned to its nearest sample location (whether existing or new). The idea is that every time we add sample locations, we calculate the distance between each anchor point and its nearest sample location, and rank the distance in decreasing order, then the optimal solution is to find the value of r^* so that the first r^* anchor points are assigned to the new l_1 or $l_2, l_3, \ldots, l_n$ sample locations, and the rest anchor points are assigned to the existing p or $p + l_1, p + l_1 + l_2, \ldots, p + l_1 + l_2 + \cdots + l_{n-1}$ sample locations. The key problem is to determine the value of r^*. Here, we perform a binary search to find r^*. Note that, we denote by l the number of adding sample locations in each step hereafter.

Let $\{Y_1, Y_2, \ldots, Y_{r^*}, \ldots, Y_N\}$ denote the N anchor points, Z represent the distance vector, where Z_i represents the distance between Y_i and its nearest sample location, and the decreasing order is $Z_1 \geqslant Z_2 \geqslant \ldots \geqslant Z_{r^*} \geqslant \ldots \geqslant Z_N$. We need to find the value of r^* to make sure $\{Y_1, Y_2, \ldots, Y_{r^*}\}$ are assigned to the new l sample locations, and $\{Y_{r^*+1}, Y_{r^*+2}, \ldots, Y_N\}$ are assigned to the existing p sample locations. Let us denote the optimal solution of FLS for $\{Y_1, Y_2, \ldots, Y_{r^*}\}$ by F_{r^*}, and denote the optimal solution of CFLS by t.

Lemma 2. *The optimal solution of CFLS is* $\min_{0 \leq r^* \leq N} \max\{F_{r^*}, Z_{r^*+1}\}$.

Proof 2. We give a simple proof referred to [29]. Assume to the contrary that there exists a solution to CFLS with value $t^* < \min_{0 \leq r^* \leq N} \max\{F_{r^*}, Z_{r^*+1}\}$, and let r^* satisfy $Z_1 \geqslant Z_2 \geqslant \ldots \geqslant Z_{r^*} \geqslant t^* \geqslant Z_{r^*+1} \geqslant \ldots \geqslant Z_N$, then $\{Y_1, Y_2, \ldots, Y_{r^*}\}$ are assigned to the new sample locations, $\{Y_{r^*+1}, Y_{r^*+2}, \ldots, Y_N\}$ are assigned to the existing sample locations. Since $t^* < t$, where t is the optimal solution to CFLS, and $t^* \geqslant Z_{r^*+1}$, then $t^* < F_{r^*}$. However, the distance of each anchor point in $\{Y_1, Y_2, \ldots, Y_{r^*}\}$ to its nearest sample location is less than or equal to t^*. That is to say the solution of FLS on $\{Y_1, Y_2, \ldots, Y_{r^*}\}$ must satisfy $F_{r^*} \leqslant t^*$, a contradiction to $t^* < F_{r^*}$.

5.2. Algorithm. The pseudo code of the continuous sample location selection algorithm is given in Algorithm 3. Since the sample location selection process is divided into continuous steps, we add certain a number of sample locations in every step. The algorithm is stopped until no more sample locations can be added. Every time we add sample locations, we exploit the conditional l-center [27] algorithm to determine the optimal solution. As for the sample location addition process, we first calculate the distance vector Z based on the distance between each anchor point and its nearest sample location (whether existing or new), then we sort Z in decreasing order (line 2–3). Note that the order of indices of anchor points is not important, so we refine them as $Z_1 \geqslant Z_2 \geqslant \ldots \geqslant Z_{r^*} \geqslant \ldots \geqslant Z_N$. As mentioned above, we need to find an appropriate value of r^* to make sure $\{Y_1, Y_2, \ldots, Y_{r^*}\}$ are assigned to the new l sample locations, and $\{Y_{r^*+1}, Y_{r^*+2}, \ldots, Y_N\}$ are assigned to the existing p sample locations. Thus, we perform a simple binary search to find r^* (line 4–14), and we call FLS to calculate the value of F_{r^*}. Note that, since we only select a small number of sample locations in each step, the local optimization of FLS can be ignored when the number of sample locations is small. According to Lemma 2, we refine the iterative conditions by comparing F_{r^*} with Z_{r^*+1} (line 8, 11), and the value of optimal solution is $t = \min\left\{\max\left\{F_{r^*_{\min}}, Z_{r^*_{\min}+1}\right\}, \max\left\{F_{r^*_{\max}}, Z_{r^*_{\max}+1}\right\}\right\}$ when the iteration is stopped. Note that, the complexity is dominated by $O(\log n)$ FLS, and n represents the number of anchor points.

5.3. Analysis. As shown in Algorithm 3, we sort the anchor points in decreasing order based on the distance to their nearest sample location, and add sample locations in the area where the distance between anchor points and their nearest

Input: a set N of anchor points, p existing sample locations, and a positive integer l;
Output: the value of optimal solution t;
1: repeat
2: Calculate the distance vector Z;
3: Sort Z in decreasing order, and refine the index of anchor points to make $Z_1 \geqslant Z_2 \geqslant \ldots \geqslant Z_{r^*} \geqslant \ldots \geqslant Z_N$;
4: Set $r^*_{\min} = 1$, $r^*_{\max} = N$;
5: while $r^*_{\max} - r^*_{\min} > 1$ do
6: Set $r^* = (r^*_{\min} + r^*_{\max})/2$;
7: Calculate F_{r^*};
8: if $F_{r^*} \leqslant Z_{r^*+1}$ then
9: Set $r^*_{\min} = r^*$;
10: end if
11: if $F_{r^*} > Z_{r^*+1}$ then
12: Set $r^*_{\max} = r^*$;
13: end if
14: end while
15: Calculate the value t of optimal solution of CFLS: $t = \min\left\{\max\left\{F_{r^*_{\min}}, Z_{r^*_{\min}+1}\right\}, \max\left\{F_{r^*_{\max}}, Z_{r^*_{\max}+1}\right\}\right\}$;
16: until no more sample locations are added
17: return the value of optimal solution t;

Algorithm 3: Continuous sample location selection.

sample location is very far. In other words, the sample locations are invested in sparsely populated areas. We have already mentioned that the objective distance is always determined by the radius of centers in sparsely populated areas, thereby the objective distance attained by CFLS can be effectively reduced in each iteration. As shown in Table 2, the radius of all centers can be reduced to a small value compared to the results of FLS in Table 1, that is, the radius of the center in the most sparsely area is optimized. Therefore, the objective distance of CFLS will outperform FLS when the total number of sample locations is large enough. Then we can conclude that by conducting sample location selection in a continuous process, not only can the online advertising service be started immediately but also a better objective distance than FLS can be attained.

6. Experiments

Our experiments are conducted on a PC with Intel Core 3.2 GHz CPU and 16G memory. The algorithms are implemented in *C++* with TDM-GCC 4.9.2.

6.1. Setup. Algorithms. There are four algorithms to be compared in our experiments. (1) CFLS is the continuous facility location based sampling method, and the sample location selection is completed in a continuous process. (2) FLS is the facility location based sampling method, which adopts the baseline anchor point selection strategy proposed in [4], and all sample locations are selected at once. (3) K-means simply clusters the users to a given number of groups with respect to distance, and selects the center of each cluster to make up the final sample locations. (4) RSQ extends RS to filter out the sample locations that are outside of the query zone, and RS is the random sampling method in [2].Datasets. In our experiments, we use two real-world geo-social networks where users can share their check-ins, as shown in Table 3. The check-ins represent users' locations, and the datasets are obtained from http://snap.stanford.edu/data/. Note that just 88.6% and 54.4% users have check-ins in Brightkite and Gowalla, respectively, and we pretreat the datasets as follows. Since there are a few users who do not have location information in Brightkite, we randomly generate a location for them based on the spatial distribution of other users. As for Gowalla, since almost half of the users have no check-ins, we delete those users who do not have location information, and the actual points used in Gowalla are 100K.

6.1.1. Parameters. Both FLS and CFLS utilize the baseline anchor point selection strategy proposed in [4], and the total number of selected sample locations is set as the same for FLS and CFLS.

6.2. Effectiveness Analysis. The effectiveness of sample location selection algorithms can be evaluated by four metrics [4]; as the other three metrics are directly influenced by the metric of objective distance, we only present the evaluation of the metric of objective distance. We select 100 and 1000 sample locations, respectively, and we show four different methods to select a certain number of sample locations by adopting CFLS. The first value on the x axis represents the existing number of sample locations, while the others represent the current number of sample locations after addition in previous steps.

As shown in Figure 5, the total number of sample locations is 100, and Figures 5(a), 5(b), 5(c), and 5(d) represent four different methods of selecting 100 sample locations in Brightkite. For example, Figure 5(a) shows that CFLS completes the selection of 100 sample locations with $\langle 20, 20, 20, 20, 20\rangle$. Specifically, we first select 20 sample locations, then add another 20 sample locations in every step

Table 2: The distribution of sample locations selected by CFLS.

Sample location #	Radius ≤5	5 < radius ≤10	10 < radius ≤15	Radius >15
500	76.2%	22%	1.8%	0
1000	97.2%	2.8%	0	0
1500	1	0	0	0

Table 3: Experimental datasets.

Dataset	Node number (K)	Edge number	Avg. in-degree	Avg. out-degree
Brightkite	58	428K	7	7
Gowalla	100	1.9 M	13	13

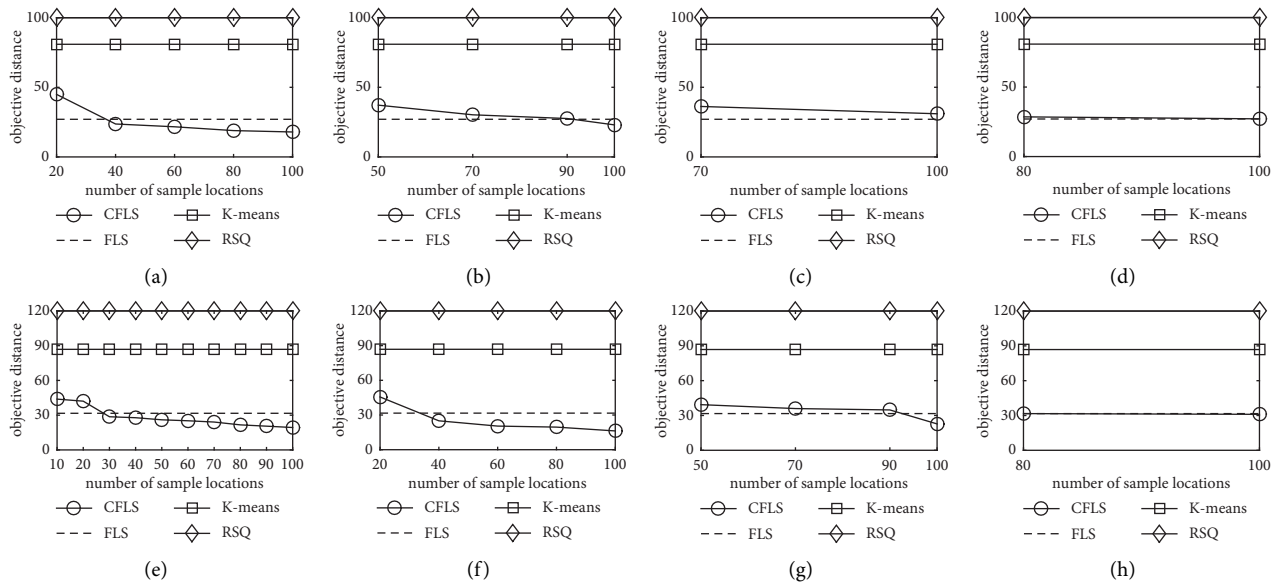

Figure 5: Effectiveness: total sample size = 100.

until the selection of 100 sample locations is completed, while FLS, K-means, and RSQ select 100 sample locations all at once. We can find that the objective distance of simple sampling approaches such as RSQ is much worse than K-means, FLS, and CFLS. Note that, the objective distance of RSQ will be much greater than 100 and 120 in Brightkite and Gowalla, respectively, when the number of sample locations is 100; so, we set a limit of 100 and RSQ always achieves the largest value in Figures 5 and 6, and the setting is the same when the number of sample locations is 1000. Although K-means can reduce the objective distance significantly, it is still not as effective as FLS and CFLS. Moreover, the objective distance of CFLS outperforms FLS when the number of the added sample locations is larger than the number of the existing sample locations. Since CFLS adds sample locations in sparsely populated areas, the objective distance is definitely reduced after addition in each subsequent step. Once the number of the added sample locations is larger, the improvement is more significant as shown in Figures 5(a) and 5(b). However, Figure 5(c) shows that the objective distance of CFLS can be worse than FLS because the number of the added sample locations is so small. Note that the objective distance of FLS remains the same when the number of sample location is 80 and 100 because of the problem of local optimization. Figures 5(e), 5(f), 5(g), and 5(h) show the results of four different methods of selecting 100 sample locations in Gowalla, and the results are almost the same with that in Brightkite.

We select 1000 sample locations in Figure 6, and we can find that K-means and RSQ perform worse than the proposed approaches FLS and CFLS, and CFLS performs better than FLS both in Brightkite and Gowalla. As shown in Figures 6(a), 6(b), 6(c), and 6(d), the objective distance of CFLS is almost half of that of FLS. Since FLS selects all sample locations at once, the number of sample locations in sparsely populated areas is much less when the total number of sample locations is large based on Lemma 2. However, the objective distance is always determined by the centers in sparsely populated areas. In other words, CFLS outperforms FLS because CFLS focuses on investing more sample locations in sparsely populated areas. Moreover, we can find that FLS falls into local optimization as shown in Figures 6(d)

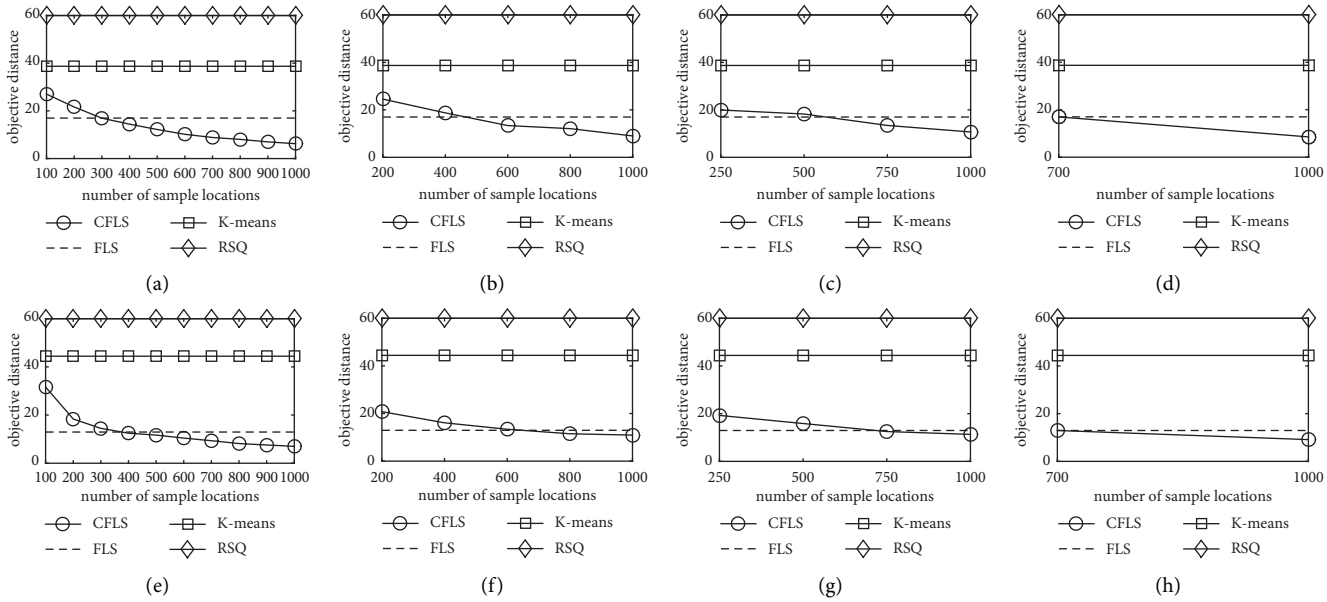

Figure 6: Effectiveness: total sample size = 1000.

and 6(h) since the objective distance of selecting 700 sample locations and 1000 sample locations is the same. Note that FLS falls into local optimization because the objective distance of CFLS selecting 700 sample locations is the same as the objective distance of FLS selecting 1000 sample locations, and CFLS exploits the techniques of FLS to select sample locations in every step.

We can conclude that the sampling approaches proposed in this paper can achieve better results than K-means and the simple sampling approaches such as random sampling. As for our own approaches, we can conclude that when the total number of sample locations is large enough, CFLS always outperforms FLS and is independent of the methods of selecting sample locations. Note that the improvement of objective distance in CFLS is limited to the distribution of anchor points. We have conducted CFLS on some synthetic datasets, where the anchor points are evenly distributed, and we find that the objective distance of FLS is better than CFLS. However, since the distribution of users in social networks is always sparse, CFLS can achieve better performance in real-world applications.

6.3. Efficiency Analysis. We evaluate the efficiency of sample location selection algorithms by focusing on the response time of selecting certain number of sample locations. Since CFLS completes the selection in a continuous process, we present both the time costs of each step and the cumulative time costs. Figure 7 shows the response time of each step of CFLS when the total number of sample locations is 100. We can find that RSQ runs the fastest among all the algorithms, and K-means always outperforms FLS and CFLS. Since RSQ just randomly selects sample locations in the 2D space and filters the sample locations outside of the query zone, there is little time consumption during the process. As for FLS and CFLS, we can find that the response time of CFLS adding sample locations in each step may be longer than the total response time of FLS. Since the number of sample locations is 100, FLS can complete the selection of such small number of sample locations in a few seconds. While CFLS needs to perform binary search to find the appropriate value of r^* in each step, the response time will increase as the number of calling FLS increases. However, the aim of CFLS is to provide a continuous sample location selection and make the online advertising service start immediately. CFLS can select a small number of sample locations quickly, and thereby users can get the seeding results immediately because they do not need to wait until the selection of all sample locations is completed. As shown in Figure 7(a), the selection of FLS takes almost 8 seconds, while CFLS completes the selection of 20 sample locations in 2 seconds and then the online seeding algorithms can be started.

Since it consumes little time for FLS to select 100 sample locations, the advantages of CFLS are not obvious. As shown in Figure 8, 1000 sample locations are selected, and we can find that the response time of FLS is 49.46 and 129.34 seconds in Brightkite and Gowalla, respectively. Users need to wait nearly a minute or two for the online seeding because in FLS, the online process cannot be started until the selection of all sample locations is completed. However, as for CFLS is concerned, the selection of the first part of sample locations is completed in less than 10 seconds, then the online seeding algorithms can be conducted. Note that the response time of each step of CFLS is much less than the total response time of FLS. When the online seeding process based on the selected sample locations in previous steps has been done, the added sample location selection for the current step is also completed offline; thereby, users can get instant and continuous online advertising service.

Figure 9 presents the cumulative time costs of four sampling methods. As shown in Figures 9(a) and 9(b), in

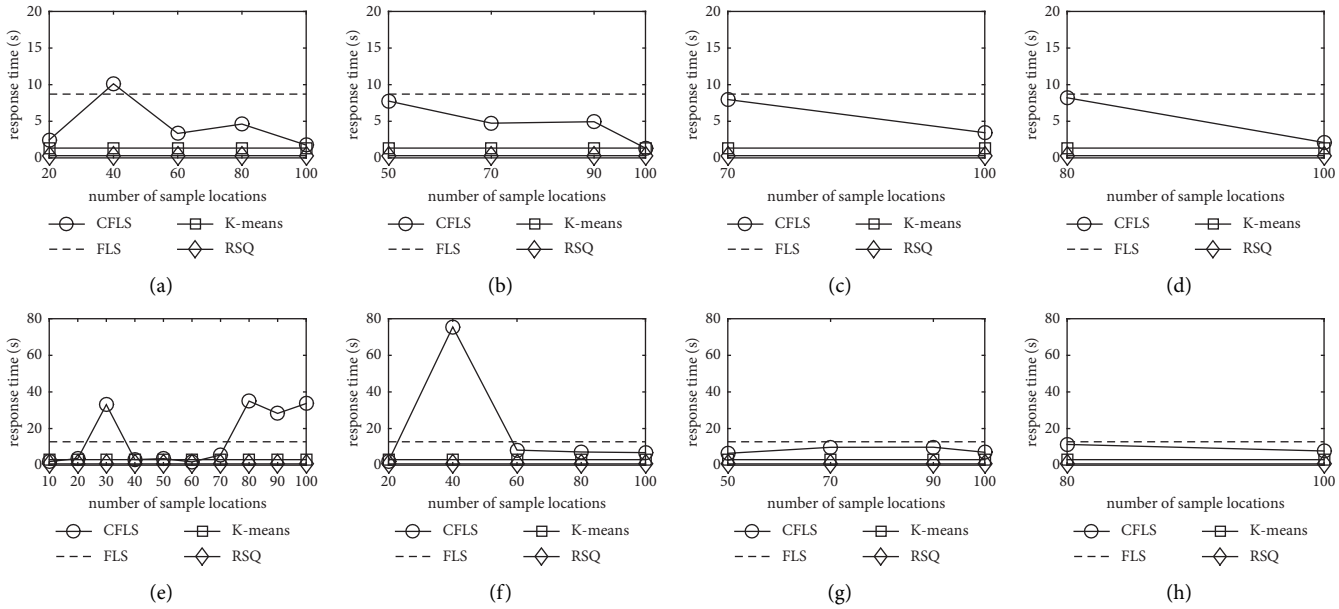

Figure 7: Response time: total sample size = 100.

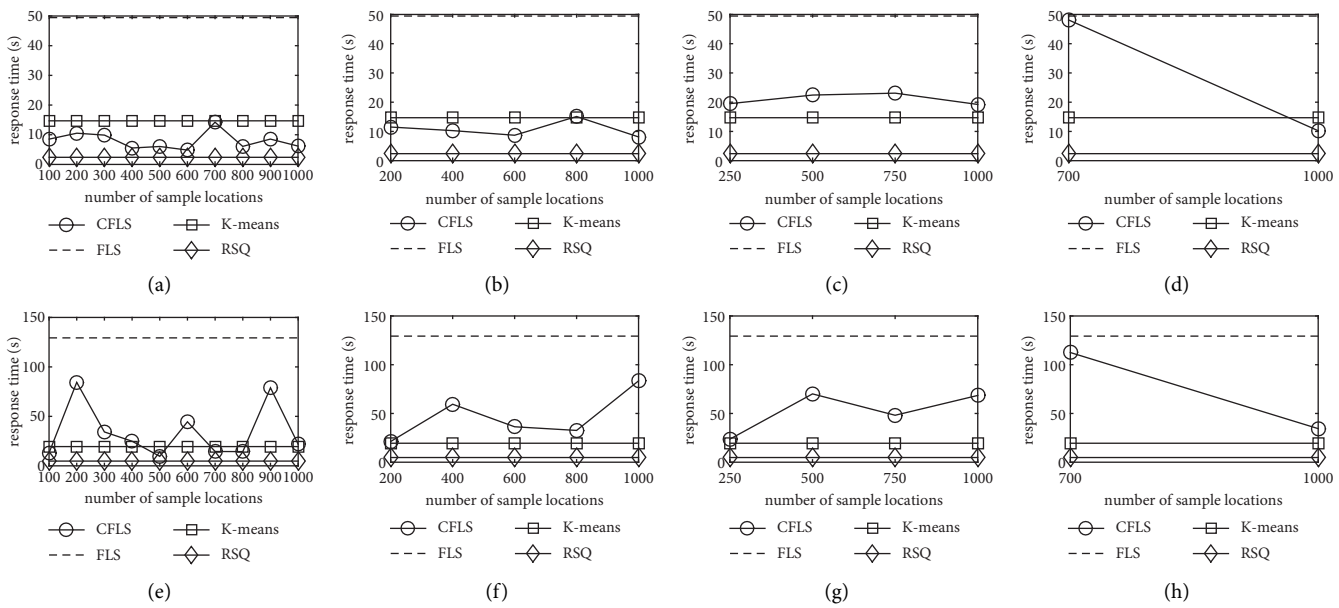

Figure 8: Response Time: total sample size = 1000.

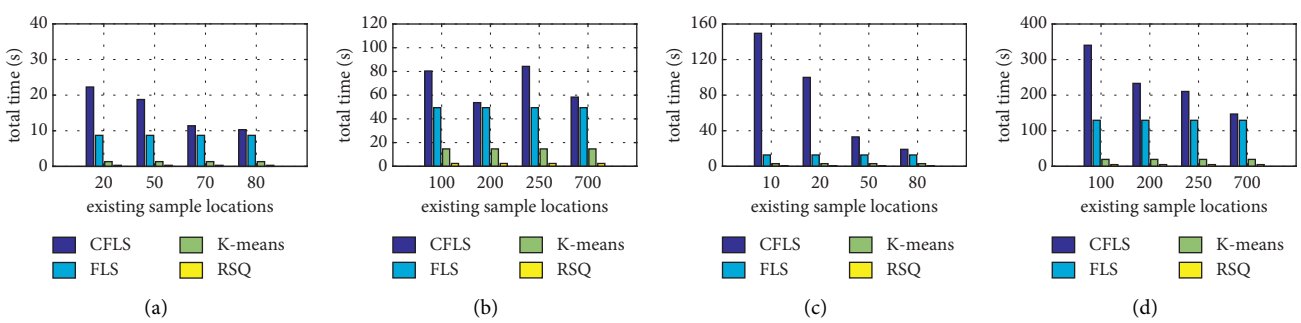

Figure 9: Total response time.

Brightkite, the time costs of RSQ are the smallest, and K-means runs faster than FLS and CFLS, and the time costs of FLS are almost half of CFLS. As for Gowalla, the results are almost the same when the number of sample locations is 1000 as Figure 9(d) shows, while they are worse when the number of sample locations is 100 as Figure 9(c) shows. The reason is that CFLS needs to perform binary search and call FLS to determine the value of r^*, and the response time depends on the efficiency of finding r^*. Note that the total response time of CFLS is just a little higher than FLS when the number of existing sample locations is larger than the number of added sample locations.

7. Conclusion

Sample location selection is crucial for the DAIM problem in geo-social networks. The previous works mainly select sample locations by simple methods such as random sampling or equal cell sampling, which can achieve a good online seeding performance within a moderate precomputation overhead. We propose the conception of query zone and reasonably formulate a novel problem of sample location selection for a given query zone, and we devise two methods to select sample locations, denoted by Facility Location Based Sampling (FLS) and Conditional Facility Location Based Sampling (CFLS), respectively. As for FLS, the problem is solved by selecting some anchor points from the query zone and developing a heuristic partition refining algorithm to find a number of centers of the anchor points as the sample locations, and all sample locations are selected at once. FLS can achieve a specific objective distance by selecting much less sample locations than the existing sampling methods, thereby balancing the online performance and precomputation overhead effectively. While CFLS selects sample locations continuously, that is, the selection of sample locations is divided into several steps, and the objective distance can be effectively reduced in each step, thereby CFLS can start the online advertising service immediately and the quality of the online advertising service can also be guaranteed. Moreover, the objective distance of CFLS can even outperform FLS when the number of the sampled locations is large enough.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Additional Points

The main difference is that, the Facility Location Based Sampling (FLS) method is proposed in our previous conference paper, and the Conditional Facility Location Based Sampling (CFLS) method is proposed in this extension paper. FLS conducts a one-time sample location selection based on the spatial distribution of users before online advertising. In contrast, CFLS can incrementally select sample locations, so that we can start online advertising quickly with a small set of sample locations and then improve the effectiveness of online advertising by selecting more sample locations. Moreover, given the same large number of sample locations, CFLS can achieve better objective distance than FLS.

Disclosure

The earlier version of the manuscript has been presented as a conference paper in "Database Systems for Advanced Applications" by [4].

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Study conception and design were proposed by Kejian Tang, Qian Zeng, Ming Zhong, Yuanyuan Zhu, Jianxin Li, and Tieyun Qian; data collection was done by Tao Zhan, Hui Zhu, Qian Zeng, and Ming Zhong; analysis and interpretation of results were supervised by Shaohui Zhan, Qian Zeng, and Ming Zhong; draft manuscript preparation was performed by Kejian Tang, Qian Zeng, Ming Zhong, and Xiaoyu Zhu. All authors reviewed the results and approved the final version of the manuscript. All authors of this manuscript meet the ICMJE criteria.

References

[1] G. Li, S. Chen, J. Feng, K.-l. Tan, and W.-s. Li, "Efficient location-aware influence maximization," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, pp. 87–98, ACM Library, New yark, June 2014.

[2] X. Wang, Y. Zhang, W. Zhang, and X. Lin, "Efficient distance-aware influence maximization in geo-social networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 3, pp. 599–612, 2017.

[3] A. Wang, A. Zhang, A. Zhang, and A. Lin, "Distance-aware influence maximization in geo-social network," in *Proceedings of the IEEE International Conference on Data Engineering*, pp. 1–12, IEEE, Helsinki, Finland, May 2016.

[4] M. Zhong, Q. Zeng, Y. Zhu, J. Li, and T. Qian, "Sample location selection for efficient distance-aware influence maximization in geo-social networks," in *Proceedings of the Database Systems for Advanced Applications,Lecture Notes in Computer Science*, pp. 355–371, Springer, Cham, Switzerland AG, May 2018.

[5] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146, ACM Library, New yark, August 2003.

[6] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 420–429, ACM Library, New yark, August 2007.

[7] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proceedings of the ACM SIGKDD International*

Conference on Knowledge Discovery and Data Mining, pp. 1029–1038, ACM Library, New yark, July 2010.

[8] E. Cohen, D. Delling, T. Pajor, and R. F. Werneck, "Sketch-based influence maximization and computation," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 629–638, ACM Library, New yark, November 2014.

[9] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 88–97, IEEE, Sydney, NSW, Australia, December 2010.

[10] J. Li, C. Liu, J. X. Yu, Y. Chen, T. Sellis, and J. S. Culpepper, "Personalized influential topic search via social network summarization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 7, pp. 1820–1834, 2016.

[11] J. Li, X. Wang, K. Deng, X. Yang, T. Sellis, and J. X. Yu, "Most influential community search over large social networks," *2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, IEEE, in *Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE)*, pp. 871–882, April 2017.

[12] J. Chen, M. Zhong, J. Li, D. Wang, T. Qian, and H. Tu, "Effective deep attributed network representation learning with topology adapted smoothing," *IEEE Transactions on Cybernetics*, pp. 1–12, 2021.

[13] G. Xue, M. Zhong, J. Li, J. Chen, C. Zhai, and R. Kong, *Dynamic Network Embedding Survey*, Elsevier, Wuhan China, 2021.

[14] J. Li, T. Sellis, J. S. Culpepper, Z. He, C. Liu, and J. Wang, "Geo-social influence spanning maximization," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1653–1666, 2017.

[15] W.-Y. Zhu, W.-C. Peng, L.-J. Chen, K. Zheng, and X. Zhou, "Modeling user mobility for location promotion in location-based social networks," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1573–1582, ACM Library, New yark, August 2015.

[16] J. L. Z. Cai, M. Yan, and Y. Li, "Using crowdsourced data in location-based social networks to explore influence maximization," in *Proceedings of the IEEE INFOCOM 2016 - the IEEE International Conference on Computer Communications*, pp. 1–9, IEEE, San Francisco, CA, USA, April 2016.

[17] T. Cai, J. Li, A. S. Mian, R. li, T. Sellis, and J. X. Yu, "Target-aware holistic influence maximization in spatial social networks," *IEEE Transactions on Knowledge and Data Engineering*, p. 1, 2020.

[18] N. A. H. Haldar, M. Reynolds, Q. Shao, C. Paris, J. Li, and Y. Chen, "Activity location inference of users based on social relationship," *World Wide Web*, vol. 24, no. 4, pp. 1165–1183, 2021.

[19] A. Weber, "Ueber den standort der industrieni," *Translated as Alfred Wber's Theory of Location of Industries*, vol. 3, 1929.

[20] A. Boloori Arabani and R. Z. Farahani, "Facility location dynamics: an overview of classifications and applications," *Computers & Industrial Engineering*, vol. 62, no. 1, pp. 408–420, 2012.

[21] C. Irawan and S. Salhi, "Aggregation and non aggregation techniques for large facility location problems: a survey," *Yugoslav Journal of Operations Research*, vol. 25, no. 3, pp. 313–341, 2015.

[22] J. Elzinga and D. W. Hearn, "Geometrical solutions for some minimax location problems," *Transportation Science*, vol. 6, no. 4, pp. 379–394, 1972.

[23] Z. Drezner and G. O. Wesolowsky, "Single facility $l_p $-Distance minimax location," *SIAM Journal on Algebraic and Discrete Methods*, vol. 1, no. 3, pp. 315–321, 1980.

[24] Z. Drezner, "The p-centre problem-heuristic and optimal algorithms," *Journal of the Operational Research Society*, vol. 35, no. 8, pp. 741–748, 1984.

[25] B. Callaghan, S. Salhi, and G. Nagy, "Speeding up the optimal method of Drezner for the p -centre problem in the plane," *European Journal of Operational Research*, vol. 257, no. 3, pp. 722–734, 2017.

[26] E. Minieka, "Conditional centers and medians of a graph," *Networks*, vol. 10, no. 3, pp. 265–272, 1980.

[27] Z. Drezner, "Conditionalp-center problems," *Transportation Science*, vol. 23, no. 1, pp. 51–53, 1989.

[28] O. Berman and Z. Drezner, "A new formulation for the conditional -median and -center problems," *Operations Research Letters*, vol. 36, no. 4, pp. 481–483, 2008.

[29] D. Chen and R. Chen, "A relaxation-based algorithm for solving the conditional -center problem," *Operations Research Letters*, vol. 38, no. 3, pp. 215–217, 2010.

[30] Q. Zeng, M. Zhong, Y. Zhu, T. Qian, and J. Li, "Business location planning based on a novel geo-social influence diffusion model," *Information Sciences*, vol. 559, pp. 61–74, 2021.