

Protecting Privacy in Location-based Services Using K -anonymity without Cloaked Region

Zhenqiang Gong

School of Computer Science and
Technology
University of Science and Technology
of China
Hefei, China
gzqiang@mail.ustc.edu.cn

Guang-Zhong Sun

School of Computer Science and
Technology
University of Science and Technology
of China
Hefei, China
gzsun@ustc.edu.cn

Xing Xie

Microsoft Research Asia
Beijing, China
xingx@microsoft.com

Abstract—The emerging location-detection devices together with ubiquitous connectivity have enabled a large variety of location-based services (LBS). Unfortunately, LBS may threaten the users’ privacy. K -anonymity cloaking the user location to K -anonymizing spatial region (K -ASR) has been extensively studied to protect privacy in LBS. Traditional K -anonymity method needs complex query processing algorithms at the server side. SpaceTwist [8] rectifies the above shortcoming of traditional K -anonymity since it only requires incremental nearest neighbor (INN) queries processing techniques at the server side. However, SpaceTwist may fail since it cannot guarantee K -anonymity. In this paper, our proposed framework, called KAWCR (K -anonymity Without Cloaked Region), rectifies the shortcomings and retains the advantages of the above two techniques. KAWCR only needs the server to process INN queries and can guarantee that the users issuing the query is indistinguishable from at least $K-1$ other users. We formulate the communication costs of KAWCR, traditional K -anonymity and SpaceTwist under the assumptions that POIs and users are uniformly distributed. We also did extensive experiments to compare KAWCR with traditional K -anonymity and SpaceTwist in terms of communication costs. The experimental results show that the communication cost of KAWCR for k NN queries is lower than that of both traditional K -anonymity and SpaceTwist.

Keywords: K -anonymity, privacy, location-based services.

I. INTRODUCTION

Nowadays, location-detection devices—such as cellular phones, GPS-like devices and RFID, etc—are more and more widely used. These location-detection devices together with ubiquitous connectivity have enabled a large variety of location-based services (LBS) which are able to tailor services according to the location of the user requiring the services. Such services mainly rely on k -nearest-neighbor queries (k NN) [9][12], which retrieve k points-of-interest (POIs) closest to the user’s location.

Unfortunately, LBS may threaten our privacy. Malicious attacker may collude with LBS provider to steal users’ location information and query logs. Assume Bob is the user applying LBS and Alice is the malicious attacker who wants to disclose Bob’s privacy. After Alice colludes with the LBS provider to get Bob’s location information and his query logs, he can infer

Bob’s privacy as follows: First, Alice may relate the location information to Bob [1]. To do this, Alice may choose from a variety of techniques such as physical observation of Bob, triangulating his mobile phone’s signal, or consulting publicly available databases. If, for instance, Bob uses his phone within his residence, Alice can easily convert the coordinates to a street address (most online maps provide this service) and relate the address to Bob by accessing an online white pages service. Second, Alice can infer Bob’s privacy through Bob’s query logs. For example, if Bob issues a query “Where is the nearest AIDS clinic?” or “Where is the nearest Christ church?”, then Alice can infer that Bob may be infected with AIDS or he is possible to be a Christian, which may be Bob’s privacy.

K -anonymity, which is introduced from statistical database, has been widely studied to protect privacy in LBS. Its main idea is to make the user issuing the query indistinguishable from at least $K-1$ other users. Most existing works [1][2][3][4][5][6][13][14], adopt the framework shown in Fig.1. The user sends its location, query and K to the *anonymizer*, which is a trusted third party in centralized systems [1][2][4] or a peer in decentralized systems [5][6][13]. The anonymizer cloaks the exact user location to K -anonymizing spatial region (K -ASR) including at least $K-1$ other users. Then anonymizer sends the K -ASR and query to the LBS sever, which calculates the candidate results respect to the cloaked region and sends them back to the anonymizer. At last, the anonymizer calculates the actual results and sends them back to the user. Two serious drawbacks of this framework are: 1) high processing cost since the LBS server has to process range k -nearest-neighbor queries (Rk NN) [1][7], and 2) high communication cost since the number of candidate results can be large.

Different from K -anonymity, SpaceTwist [8] sends a false location to the server instead of a cloaked region. SpaceTwist requires only simple query processing algorithm on the server—namely, incremental nearest neighbor (INN) retrieval [9]. However, SpaceTwist may fail if the attacker already knows the locations of all the users. According to [8], the location of the user issuing the query can be bounded in a region \mathcal{R} . If only one user lies in the region \mathcal{R} , then attacker can easily infer that the query is issued by the user, which may threaten the user’s privacy. The reason why SpaceTwist may fail is that it does not guarantee K -anonymity.

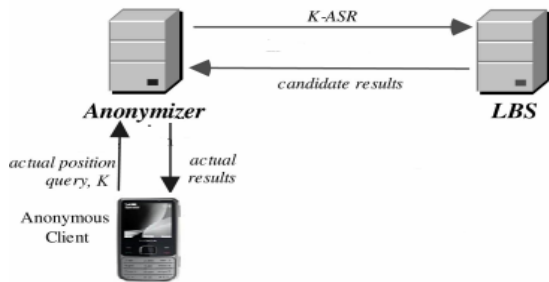


Figure 1. Architecture of traditional K -anonymity

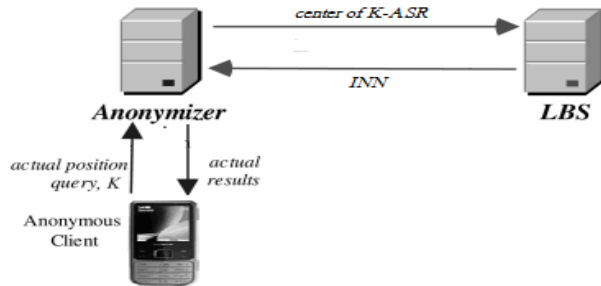


Figure 2. Architecture of KAWCR

In this paper, we present a framework, called KAWCR (K-anonymity Without Cloaked Region), that aims to improve on the above approaches. Its architecture is shown in Fig.2. The user sends its location, query and K to the anonymizer. The anonymizer cloaks the exact user location to K -ASR including at least $K-1$ other users. Then the anonymizer sends the center of K -ASR to the server, which performs INN retrieval respect to the center. This paper only discusses the case where K -ASR is a circle since it's similar to discuss the case where K -ASR is a rectangle. The main difference between KAWCR and traditional K -anonymity (we call K -anonymity that sends the K -ASR to the server as traditional K -anonymity) is that KAWCR only sends the center of K -ASR to the server while traditional K -anonymity sends the whole K -ASR to the server. KAWCR only requires server to process INN retrieval, which has been studied extensively [9] and readily implemented on the server. In contrast, traditional K -anonymity requires specialized query processing algorithms, which incur high processing costs. The essential difference between KAWCR and SpaceTwist is that KAWCR can guarantee K -anonymity while SpaceTwist cannot.

In this paper, our contributions are as follows:

- 1 We proposed a new framework called KAWCR to protect privacy in LBS. KAWCR can guarantee that the user issuing the query is indistinguishable from at least $K-1$ other users with low query processing costs and low communication costs.
- 1 We comprehensively analyzed the communication costs of KAWCR, traditional K -anonymity and SpaceTwist in the cases where the POIs and users are uniformly distributed.
- 1 We did extensive experiments to compare KAWCR with traditional K -anonymity and SpaceTwist in terms of communication costs both on synthetic datasets and real-world dataset. Our experimental results showed that the communication cost of KAWCR is lower than that of traditional K -anonymity and SpaceTwist.

The rest of this paper is organized as follows: Section II presents the related works. Next, in Section III, we introduce our approach, which mainly focuses on the interaction between anonymizer and LBS server. Section IV estimates the performance of KAWCR, traditional K -anonymity and SpaceTwist. Our extensive experimental results are illustrated

in Section V. Finally, in Section VI, we conclude this paper and figure out our future works.

II. RELATED WORKS

In this section, we review previous works. Although there have existed a variety of approaches to protect privacy in LBS, we focus our mind in traditional K -anonymity and SpaceTwist since they are much related to our work. In section II-A, we discuss traditional K -anonymity, followed by section II-B, where SpaceTwist is discussed.

A. Traditional K -anonymity

K -anonymity in relational database K -anonymity was first discussed in relational databases, where published data (for example, medical or census) should not be linked to specific persons. K -anonymity in relational database is defined as follows [10][11]: A relation satisfies K -anonymity if every tuple is indistinguishable from at least $K-1$ other tuples with respect to a set of *quasi-identifier* attributes. Quasi-identifiers are attributes—such as date of birth, gender, and zip code—that can be linked to publicly available data to identify individuals.

K -anonymity in LBS Most existing works on LBSs adopt K -anonymity by using the framework illustrated in Fig. 1. This framework works as follows: A user sends its location, query and K to the anonymizer, which is a trusted third party [1][2][4] in centralized systems or a peer in decentralized systems [5][6][13]. The anonymizer removes the ID of the user and cloaks the exact user location to K -ASR including at least $K-1$ other users. Then anonymizer sends the K -ASR and query to the LBS sever, which calculates the candidate results respect to the cloaked region and sends them back to the anonymizer. At last, the anonymizer which knows the locations of all the users calculates the actual results and sends them back to the user.

k NN query processing A k NN query needs the server to retrieval k POIs closest to a specific point p . In traditional K -anonymity, the LBS server receives the K -ASR and parameter k . Since the server does not know the parameter p , it has to return all the k closest POIs of all points in K -ASR, which is a range k -nearest-neighbor query (R k NN). H.Hu and D.L.Lee proposed an R k NN algorithm denoted as R-R k NN for the situation where the region is a rectangle [7]. P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias proposed another algorithm denoted as C-R k NN for the case where the region is

a circle [1]. The main idea of both R- $RkNN$ and C- $RkNN$ is to split the boundary of the region, which is a rectangle in R- $RkNN$ or a circle in C- $RkNN$, into sufficient segments satisfying that all the points in the same segment have the same k nearest POIs. These algorithms go beyond well-known point kNN algorithms [9][12] and introduce complexity. In addition, the communication costs of kNN queries may be high since the size of candidate results can be large.

In summary, traditional K -anonymity incurs expensive processing cost and high communication cost for kNN queries. The essential reason is that the anonymizer sends a cloaked region K -ASR to the server. To avoid the above short comings, our proposal KAWCR only sends the center of K -ASR to the server. KAWCR only requires INN query processing at the server side, which has been studied extensively [9] and readily implemented on the server.

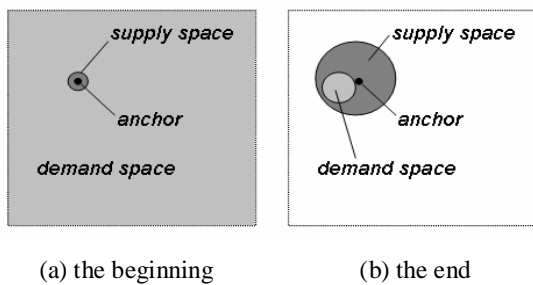


Figure 3. Supply Space and Demand Space (cited from [8])

B. SpaceTwist

Fig.3 (This Fig. is cited from [8]) provides an overview of SpaceTwist [8]. SpaceTwist works as follows: A user specifies an *anchor* and iteratively requests POIs from the LBS server in ascending order respect to the anchor. A circle which is centered at the anchor and includes all the POIs retrieved is called *supply space*. A circle whose center is the user's location and radius is the distance between the user and its current k^{th} nearest neighbor is called *demand space*. In the beginning, the supply space is empty while the demand space is initialized to be the domain space. As POIs are retrieved incrementally from the server, the supply space expands and the demand space shrinks. When the supply space covers the demand space, the algorithm halts and the user is guaranteed to produce accurate kNN results. In order to decrease the communication costs, the server accumulates multiple POIs, packs them into the same packet, and then sends the packet to the user.

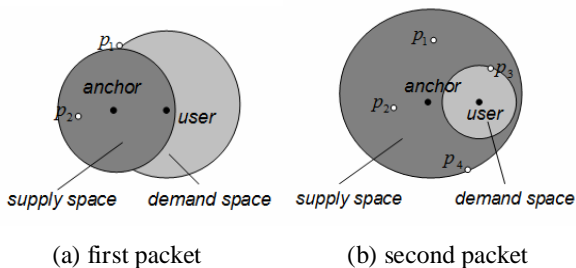


Figure 4. An example illustrating SpaceTwist

The following example illustrates how SpaceTwist works.

Example 1 Fig.4 shows how SpaceTwist works for the case where each packet contains 2 POIs and $k=1$. After the user receives the first packet (illustrated in Fig.4 (a)), the supply space does not cover the demand space. So the user continues retrieving the second packet from the server. After the user receives the second packet, the supply space covers the demand space (illustrated in Fig.4 (b)), which means that the closest neighbor of the user has been retrieved. In this example, the communication cost of SpaceTwist is two packets.

As stated in [8], the location of the user issuing the query can be bounded in an inferred privacy region. If, unfortunately, the malicious attacker knows all the users' locations and only one user lies in the privacy region, then the attacker can infer that it is the user issuing the query, which may lead to the leakage of her/his privacy.

Our proposal KAWCR guarantees that the user issuing the query is indistinguishable from at least $K-1$ other users. Even if the attacker knows all the users' locations, the probability that the attacker infers the user issuing the query does not exceed $1/K$. In addition, our experimental results show that if we set K to be the number of users in the privacy region—that is to say, KAWCR provides the same level of privacy as SpaceTwist provides under the situation that the attacker knows all the users' locations—the communication cost of KAWCR is significantly lower than that of SpaceTwist.

III. K-ANONYMITY WITHOUT CLOAKED REGION

We propose anonymizer-side kNN algorithm that compute kNN queries in incremental fashion and guarantee K -anonymity. Our proposal only requires the LBS server to support incremental nearest neighbor retrieval, which has been studied extensively [9] and readily implemented on servers. In the following, we first describe the anonymizer-side kNN algorithm in section III-A. Then, in section III-B, we prove the K -anonymity of our proposal, that is to say we prove that the user issuing the query is indistinguishable from at least $K-1$ other users.

A. Anonymizer-side kNN Algorithm

The framework we adopt is illustrated in Fig.2. A user sends his location, query and K to the anonymizer, which calculates K -ASR to include at least $K-1$ other users. In order to calculate the K -ASR, the anonymizer can use any kind of cloaking algorithm, such as *Interval Cloak* [2], *Casper* [4], *CloakP2P* [6], *NNC* [1], and so forth. Then the anonymizer sends an INN query with the center of K -ASR to the LBS server. The server executes INN query processing, which means it iteratively sends POIs back to the anonymizer in ascending distance order respect to the center. After the supply space covers the demand spaces of all the users in the K -ASR, the anonymizer halts the INN query on the server. Considering the communication cost, the server accumulates multiple POIs, packs them into the same packet, and sends them back to the anonymizer. Let parameter α denote the number of POIs in a packet. We measure the communication cost of different approaches as the overall number of packets retrieved from the server. In the following, we first discuss anonymizer-side kNN

algorithm (ASkNNA), then prove its correctness.

The anonymizer-side kNN algorithm (ASkNNA) is shown in Fig.5 (the same terms have the same meanings as in SpaceTwist). A max-heap $H_k(q_i)$, initialized with k virtual objects, maintains the k nearest POIs of q_i seen so far. Let $dist(p, q)$ denote the distance between point p and q , β_i denote the current radius of the demand space of user q_i , K' denote the actual number of users in the K -ASR (K' may be larger than K for some cloaking algorithms[1][2][4]), and τ denote the current radius of the supply space. For each $i=1,2,\dots,K'$, if $\beta_i + dist(C, q_i) > \tau$, then the anonymizer continues requesting packets from the LBS server and update $H_k(q_i)$, $H_k(q_{i+1})$, ..., $H_k(q_{K'})$ and $\beta_i, \beta_{i+1}, \dots, \beta_{K'}$ until $\beta_i + dist(C, q_i) \leq \tau$, which means the supply space covers the demand spaces of all the users. At last, the anonymizer terminates the INN query on the server and sends POIs in $H_k(q_i)$ to the q_i (assume q_1 issues the query).

Anonymizer-side kNN Algorithm(ASkNNA)

```

1:  $K'$  the number of users in  $K$ -ASR
2:  $C$  the center of  $K$ -ASR, i.e. the anchor
3: let  $q_1, q_2, \dots, q_{K'}$  represent the  $K'$  users in  $K$ -ASR
4: for  $i = 1$  to  $K'$  do
5:    $H_k(q_i)$  new max-heap of tuples  $(q, dist(p, q))$ 
6:   initialize  $H_k(q_i)$  with  $k$  tuples of  $(NULL, \infty)$ 
7:    $\beta_i = \infty$  denotes the radius of the demand space of  $q_i$ 
8: end for
9:  $\tau = 0$  > current radius of the supply space
10: send INN query with  $C$  to the LBS server
11: for  $i = 1$  to  $K'$  do
12:   while  $\beta_i + dist(C, q_i) > \tau$  do
13:      $T$  receive the next packet of POIs from server
14:      $\tau = \max_{p \in T} dist(C, p)$ 
15:     for all  $p \in T$  do
16:       for  $j = i$  to  $K'$  do
17:         if  $dist(p, q_j) < \beta_j$  then
18:           update  $H_k(q_j)$  and  $\beta_j$  using  $p$ 
19:         end if
20:       end for
21:     end for
22:   end while
23: end for
24: stop the INN query on the server
25: return  $H_k(q_1)$  to  $q_1$  > assume  $q_1$  issues the query

```

Figure 5. Anonymizer-side kNN algorithm

The following theorem shows us that ASkNNA finds exact kNN results for all the users in K -ASR.

Theorem 1 ASkNNA finds the exact kNN results for each user in K -ASR.

Proof: For each $i=1,2,\dots,K'$, we need to prove that for any

POI p outside $H_k(q_i)$, inequity $dist(p, q_j) \geq \beta_j$ is satisfied, where β_j is the maximum distance in $H_k(q_i)$. There are two cases depending on whether p is seen by the anonymizer or not.

Case 1: p is seen by the anonymizer. Since $H_k(q_i)$ maintains the current k nearest neighbors of q_i , we can easily have $dist(p, q_j) \geq \beta_j$, as desired.

Case 2: p is not seen by the anonymizer. In this case, we have $dist(p, C) \geq \tau$, where C is the center of K -ASR and τ is the furthest distance respect to C seen so far. Since $\beta_i + dist(C, q_i) \leq \tau$ when ASkNNA halts, we have $dist(p, C) \geq \tau \geq \beta_i + dist(C, q_i)$, that is, we have $dist(p, C) - dist(C, q_i) \geq \beta_i$. Combined with $dist(p, q_i) \geq dist(p, C) - dist(C, q_i)$, we have $dist(p, q_i) \geq \beta_i$, as desired.

The following example shows how ASkNNA works.

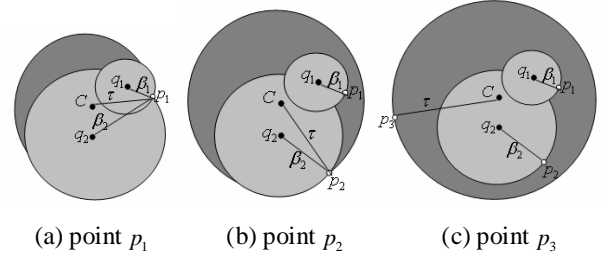


Figure 6. An example illustrating ASkNNA

Example 2 Fig.6 illustrates how the algorithm ASkNNA works for the case $\alpha = 1$, $k=1$ and $K=2$. When point p_1 is returned to the anonymizer (see Fig.6 (a)), the current nearest neighbor of both q_1 and q_2 is set to p_1 , τ is set to the distance between p_1 and C . Since $\beta_1 + dist(C, q_1) > \tau$, the anonymizer continues requesting the next point p_2 . Then the current nearest neighbor of q_2 is updated to p_2 and τ is set to the distance between p_2 and C (see Fig.6 (b)). Since $\beta_2 + dist(C, q_2) > \tau$, the anonymizer continues requesting the next point p_3 . Parameter τ is updated to the distance between p_3 and C (see Fig.6 (c)). Since $\beta_i + dist(C, q_i) \leq \tau$ for $i=1,2$, ASkNNA halts. In this example, the communication cost of ASkNNA is three packets.

B. K-anonymity of KAWCR

This section analyzes the privacy protecting provided by the KAWCR. We assume that the malicious attacker knows: 1) all the users' locations, 2) the center C , parameter k , and all the packets retrieved from the server, and 3) the algorithms used in the anonymizer side. The first assumption is motivated by the fact that users may often issue queries from the same locations e.g. at home or in the office, which may be easily identified through telephone directories, public databases, and so on. Furthermore, users may reveal their locations by issuing queries without privacy requirements. For more discussion about the first assumption, take a look at [1]. The second assumption actually states that either the LBS server or the

communication channel between the LBS server and the anonymizer is not trusted. The third assumption is common in the security literature.

Under the above assumptions, the location of the user issuing the query can be bounded in a region. However, we will prove that there are at least $K-1$ other users within this region, which guarantees the K -anonymity of KAWCR. Let m be the number of packets received by the anonymizer and let the POIs received (in their retrieval order) be $p_1, p_2, \dots, p_{m\alpha}$. Let q_c denotes a possible user's location which can be inferred by the attacker, and ϕ denotes the region consisting of all the possible locations q_c . According to the termination condition of ASkNNA, the possible user location q_c satisfies:

$$\text{dist}(C, q_c) + \min_{1 \leq i \leq m\alpha}^k \text{dist}(p_i, q_c) \leq \text{dist}(C, p_{m\alpha}) \quad (1)$$

Where the middle term represents the k^{th} smallest distance of all POIs retrieved respect to q_c . The region ϕ consists of all possible locations q_c satisfying the inequality (1). According to ASkNNA, if we replace q_c with any user q_i within K -ASR, inequality (1) is also satisfied, which means that all the users within K -ASR are also within ϕ . So the user issuing the query is indistinguishable from at least $K-1$ other users.

For further discussion, we give out another inequality as follows:

$$\text{dist}(C, q_c) + \min_{1 \leq i \leq (m-1)\alpha}^k \text{dist}(p_i, q_c) > \text{dist}(C, p_{(m-1)\alpha}) \quad (2)$$

Let ϕ' denote the region consisting of all possible locations q_c satisfying inequality (1) and (2). Since ASkNNA does not halt when receiving $(m-1)$ packets, the attacker can infer that at least one user within K -ASR belongs to ϕ' . But this inference is insufficient to disclose the privacy of the user issuing the query.

IV. PERFORMANCE ANALYSIS

In this section, we deduce some formulations to estimate the communication costs of KAWCR, traditional K -anonymity and SpaceTwist in the cases where the POIs and users are uniformly distributed in a 2D space whose area is S .

TABLE I. MEANINGS OF BASIC SYMBOLS USED

| | |
|----------|---|
| N | Number of POIs |
| M | Number of users |
| α | Packet capacity |
| k | Number of required results |
| K | User-specified level of privacy |
| S | Area of the 2D space |
| R_i | Distance between a point and its i^{th} nearest POI |
| r_i | Distance between a point and its i^{th} nearest user |
| d | Distance between the anchor and the user |

Lemma 1 Given a point p , the distance R_i between p and its i^{th} nearest POI can be estimated as $R_i = \sqrt{\frac{iS}{\pi N}}$.

Proof: The number of POIs in a circle with radius R_i can be estimated as $\frac{\pi R_i^2}{S} N$. Let $\frac{\pi R_i^2}{S} N = i$, then we have $R_i = \sqrt{\frac{iS}{\pi N}}$.

Similarly, we can have the following lemma 2.

Lemma 2 Given a point p , the distance r_i between p and its i^{th} nearest user can be estimated as $r_i = \sqrt{\frac{iS}{\pi M}}$.

The following theorem estimates the communication cost of KAWCR in the cases where the K -ASR are circles.

Theorem 2 If the K -ASR is the minimal circle containing and the K users, then the communication cost of KAWCR can be estimated as $N(\sqrt{\frac{K}{M}} + \sqrt{\frac{k}{N}})^2 / \alpha$.

Proof: According to lemma 2, among the K users in K -ASR, the distance between the furthest user and the center of the K -ASR can be estimated as $r_k = \sqrt{\frac{KS}{\pi M}}$. According to lemma 1, the radius of every user's demand space can be estimated as $R_k = \sqrt{\frac{kS}{\pi N}}$. So the radius of supply space can be estimated as $r_k + R_k$. That is to say, the number of POIs retrieved can be estimated as $\frac{\pi(r_k + R_k)^2 N}{S}$. After simplifying it, we have $\frac{\pi(r_k + R_k)^2 N}{S} = N(\sqrt{\frac{K}{M}} + \sqrt{\frac{k}{N}})^2$. So the communication cost of KAWCR can be estimated as $N(\sqrt{\frac{K}{M}} + \sqrt{\frac{k}{N}})^2 / \alpha$, as desired.

Similarly, we have the following theorem.

Theorem 3 If the K -ASR is the minimal circle containing the K users, then the communication cost of traditional K -anonymity can be estimated as $N(\sqrt{\frac{K}{M}} + \sqrt{\frac{k}{N}})^2 / \alpha$.

According to theorem 2 and theorem 3, if the users and POIs are uniformly distributed and the K -ASR is the minimal circle containing the K users, then the estimated average communication cost of KAWCR is the same as that of traditional K -anonymity. However, in our experiments, the communication cost of KAWCR is lower than that of traditional K -anonymity over real dataset. Furthermore, KAWCR only needs the server to process INN queries while traditional K -anonymity needs the server to process RkNN queries, which incurs complexity.

The following theorem estimates the communication cost of SpaceTwist.

Theorem 4 Let d denote the distance between the anchor and the user issuing the query, the communication cost of SpaceTwist can be estimated as $\frac{\pi(d + R_k)^2 N}{\alpha S}$.

Proof: According to lemma 1, the radius of the user's demand space can be estimated as $R_k = \sqrt{\frac{kS}{\pi N}}$. So the radius of the supply space can be estimated as $d + R_k$. Hence, the communication cost of SpaceTwist can be estimated as $\frac{\pi(d + R_k)^2 N}{\alpha S}$.

Theorem 4 tells us that the estimated average communication cost of SpaceTwist is a quadratic function of the distance between the anchor and the user issuing the query.

According to [8], the user issuing the query can be bounded in an inferred privacy region, the following theorem estimates the number of users in.

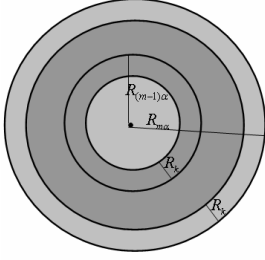


Figure 7. Estimated inferred privacy region

Theorem 5 If the communication cost of SpaceTwist is m packets, then the number of users in can be estimated as $\frac{M}{N}(\alpha - \sqrt{\frac{k\alpha}{m}})$, where α is the packet capacity.

Proof: Let m denote the number of packets retrieved from the server. When the server has retrieved $(m-1)$ packets, the radius of supply space can be estimated as $R_{(m-1)\alpha} = \sqrt{\frac{(m-1)\alpha S}{\pi N}}$ according to lemma 1. When the server has retrieved m packets, the radius of supply space can be estimated as $R_{m\alpha} = \sqrt{\frac{m\alpha S}{\pi N}}$ according to lemma 1. The radius of the user's demand space can be estimated as $R_k = \sqrt{\frac{kS}{\pi N}}$. So the region can be estimated as the hoop with inner radius $(R_{(m-1)\alpha} - R_k)$ and outer radius $(R_{m\alpha} - R_k)$ (the dark gray area in Fig.7).

Hence, the number of users in can be estimated as $\frac{\pi((R_{m\alpha} - R_k)^2 - (R_{(m-1)\alpha} - R_k)^2)M}{S} = \frac{\pi M(R_{m\alpha}^2 - R_{(m-1)\alpha}^2 - 2R_k(R_{m\alpha} - R_{(m-1)\alpha}))}{S}$
 $= \frac{M}{N}(\alpha - 2\sqrt{k\alpha}(\sqrt{m} - \sqrt{m-1})) = \frac{M}{N}(\alpha - \frac{2\sqrt{k\alpha}}{\sqrt{m} + \sqrt{m-1}}) \approx \frac{M}{N}(\alpha - \sqrt{\frac{k\alpha}{m}})$, as desired.

The following theorem estimates the number of users in using the distance between the anchor and the user issuing the query.

Theorem 6 If the distance between the anchor and the user issuing the query is d , then the number of users in can be estimated as $\frac{\alpha d M}{N(d + R_k)}$.

Proof: We can get the result just combining theorem 4 and theorem 5.

Theorem 6 tells us that the estimated number of users in is an increasingly monotone function of d . And according to $\lim_{d \rightarrow \infty} \frac{\alpha d M}{N(d + R_k)} = \frac{\alpha M}{N}$, we know that the estimated number of users in tends to be $\frac{\alpha M}{N}$, which is a negative result.

The following theorem compares SpaceTwist with KAWCR.

Theorem 7 If the number of users in K-ASR equates the estimated number of users in, that is $K = \frac{M}{N}(\alpha - \sqrt{\frac{k\alpha}{m}})$, and the K-ASR is a circle, then the estimated communication cost of KAWCR is lower than that of SpaceTwist when $m \geq m_c$, where m is the communication cost of SpaceTwist and $m_c = \frac{4k\alpha}{(\sqrt{k^2 + 4k\alpha} - k)^2}$.

Proof: Let function $f(m) = N(\sqrt{\frac{K}{M}} + \sqrt{\frac{k}{N}})^2 / \alpha - m$, where the first term is the communication cost of KAWCR and the second term is the communication cost of SpaceTwist. After replacing K in function f with $K = \frac{M}{N}(\alpha - \sqrt{\frac{k\alpha}{m}})$, we have

$f(m) = (\sqrt{\alpha - \sqrt{\frac{k\alpha}{m}}} + \sqrt{k})^2 / \alpha - m$. Let $\gamma = \alpha - \sqrt{\frac{k\alpha}{m}}$, then $m = \frac{k\alpha}{(\alpha - \gamma)^2}$. So, $f(m) = (\sqrt{\gamma} + \sqrt{k})^2 / \alpha - \frac{k\alpha}{(\alpha - \gamma)^2}$. We have the following equivalent equalities:

$$\begin{aligned} f(m) &= (\sqrt{\gamma} + \sqrt{k})^2 / \alpha - \frac{k\alpha}{(\alpha - \gamma)^2} \leq 0 \\ \Leftrightarrow (\sqrt{\gamma} + \sqrt{k})^2 &\leq \frac{k\alpha^2}{(\alpha - \gamma)^2} \Leftrightarrow \sqrt{\gamma} + \sqrt{k} \leq \frac{\alpha}{\alpha - \gamma} \sqrt{k} \\ \Leftrightarrow \gamma + \sqrt{k}\gamma - \alpha &\geq 0 \Leftrightarrow \gamma \geq \frac{k + 2\alpha - \sqrt{k^2 + 4k\alpha}}{2} \\ \Leftrightarrow m &\geq \frac{4k\alpha}{(\sqrt{k^2 + 4k\alpha} - k)^2} \Leftrightarrow m \geq m_c \end{aligned}$$

Hence, we have the correctness of theorem 7.

What is interesting is that the critical value m_c is only related to k and α . Considering the typical cases where $k=10$ and $\alpha=67$ (the reason why the typical value of α is 67 is shown in section V-A), we have $m_c \approx 1.4681$, which is a very small value. So, in most typical cases, the communication cost of KAWCR may be lower than that of SpaceTwist.

V. EXPERIMENTS

This section shows our extensive experimental results. Our algorithms are implemented in C++. We perform our algorithms on Dual Core 2.13GHz PC with 2GB memory. In section V-A, we show the descriptions of the experimental setup. Section V-B shows the experimental results that compare

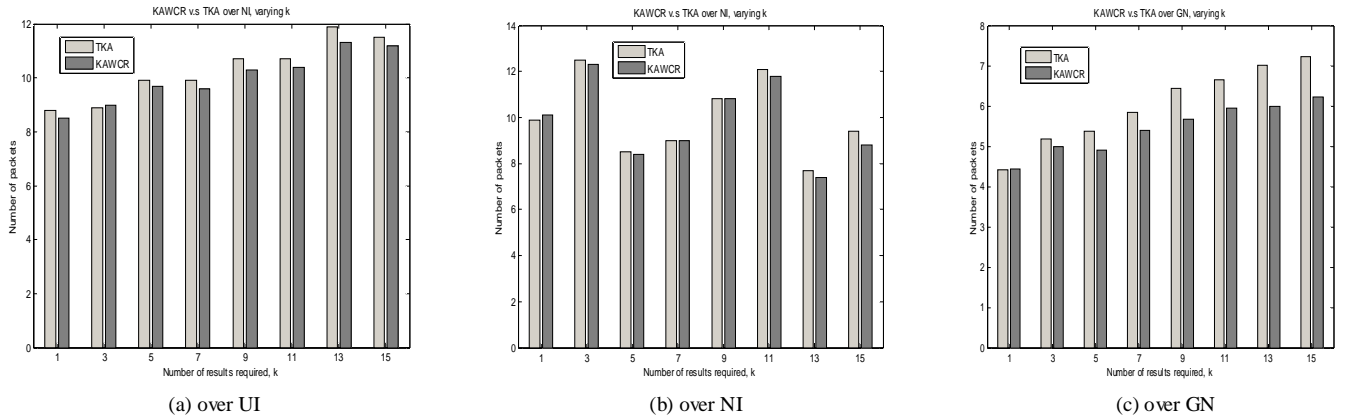


Figure 8. KAWCR v.s TKA, $K=100$ and k varies from 1 to 15

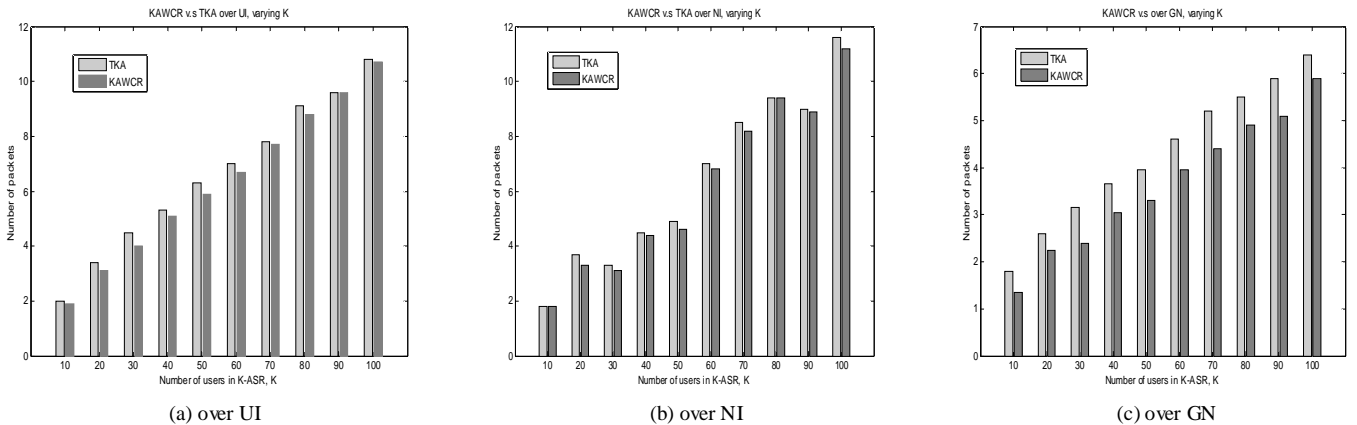


Figure 9. KAWCR v.s TKA, $k=10$ and K varies from 10 to 100

KAWCR and traditional K -anonymity (denoted as TKA) and SpaceTwist in terms of communication costs. We summarize our experimental results in section V-C.

A. Experimental setup

In our experiments, we use both synthetic datasets and one real dataset. The synthetic datasets are UI and NI. UI means that the coordinates of the POIs are uniformly and independently distributed. NI means that the coordinates of the POIs are normally and independently distributed. The real dataset we use is GN. GN contains 398,958 POIs, which are extracted from U.S. Board on geographic names (<http://geonames.usgs.gov/index.html>). We assume that there are 200,000 users and they are uniformly distributed. In our experiments, the coordinates of POIs and users are normalized to the square with extent 10,000 meters.

TABLE II shows the default values of the parameters used in our experiments. In most previous papers, the default value of k is set to be 1, such as [1][8]. However, this setting is inappropriate since users may also concern other aspects of the POIs apart from the distances. For example, if a user issues a

query to retrieve the interested hotels, he/she may also concern the prices and services of the hotels apart from the distances. So we should return several candidate hotels for the user to choose. In our experiments, the default value of k is set to be 10. We set the packet capacity α as $(576-40)/8=67$. As explained in [8], the typical size of a Maximum Transmission Unit (MTU) over a network is 576 bytes, a 2D data point takes 8 bytes, and a packet has a 40-byte header.

TABLE II. DEFAULT VALUES OF PARAMETERS USED

| Parameters | Default Values |
|---------------------------|----------------|
| K | 100 |
| k | 10 |
| d | 200m |
| α | 67 |
| Number of users, i.e. M | 200,000 |
| Number of POIs, i.e. N | 1,000,000 |

In each experiment, we use workload with 50 uniformly random generated queries and measure the average communication cost, which is the number of packets retrieved from the LBS server.

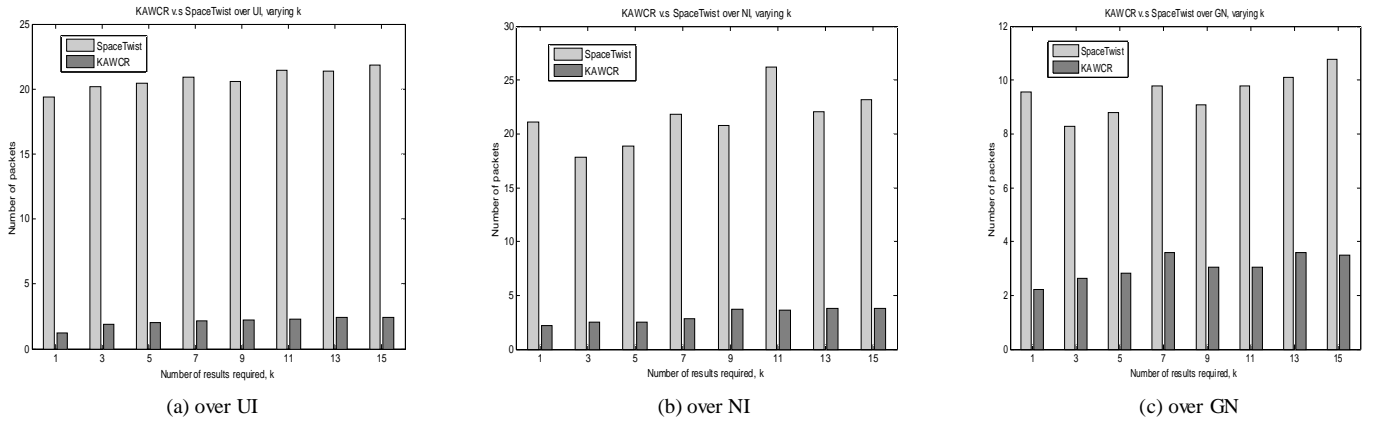


Figure 10. KAWCR vs. SpaceTwist, $d=200$ and k varies from 1 to 15

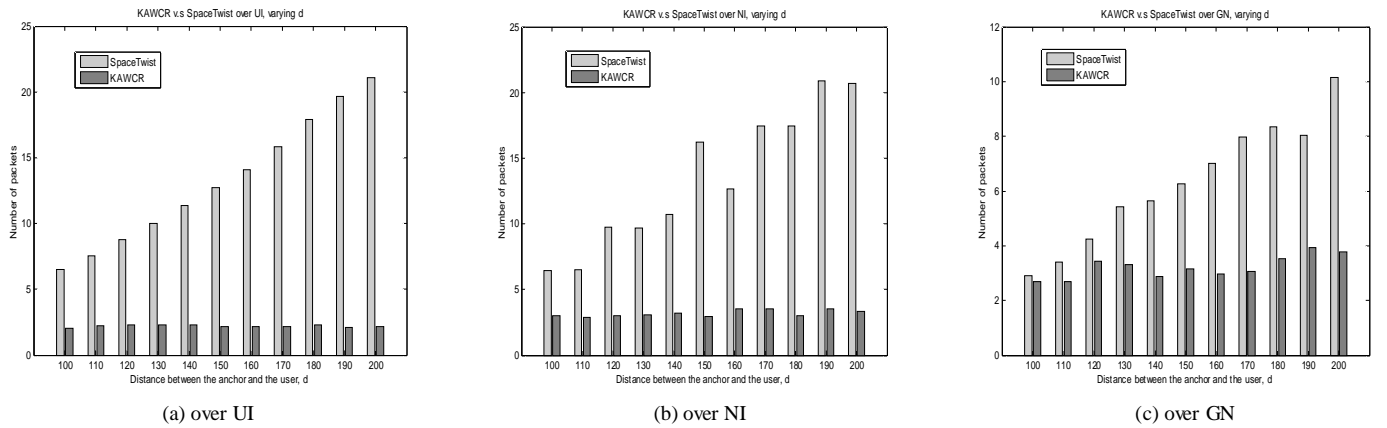


Figure 11. KAWCR vs. SpaceTwist, $k=10$ and d varies from 100 to 200

B. Experimental results

We compare KAWCR with TKA and SpaceTwist in terms of communication costs. In the following, we first show the experimental results of the comparison between KAWCR and TKA, then the results of the comparison between KAWCR and SpaceTwist.

Comparing KAWCR with TKA In our experiments, we use *NNC* algorithm [1] to cloak the user location to K -ASR both in KAWCR and TKA. *NNC* algorithm cloaks the user location to a circle including at least $K-1$ users.

Fig.8 shows the experimental results when k , the number of results required, varies from 1 to 15 with step 2 and other parameters are set as shown in TABLE II. From the results, we know that the communication costs of KAWCR are lower than those of TKA in most cases. For UI and GN datasets, the communication costs of both KAWCR and TKA slightly increase with k . However, it's different for NI dataset. The communication costs of both KAWCR and TKA fluctuate with k over NI dataset. As k increases, the advantage of KAWCR over TKA increases. The predominance of KAWCR over TKA is more apparent over GN than over UI and NI. For UI dataset,

although KAWCR outperforms TKA, the difference between their communication costs is small, which conforms to the formulations deduced in section IV.

Fig.9 shows the experimental results when K , the number of users in K -ASR varies from 10 to 100 with step 10 and other parameters are set as shown in TABLE II. From the results, we know that KAWCR outperforms TKA over all the datasets considered. And the predominance of KAWCR over TKA is more apparent over GN dataset, which is a real dataset. The reason why KAWCR outperforms TKA is that TKA retrieves the k NN results of all the points in K -ASR while KAWCR retrieves the k NN results of all the users in K -ASR in incremental fashion.

Comparing KAWCR with SpaceTwist According to [8], the location of the user issuing the query in SpaceTwist can be bounded in a privacy region. In order to compare SpaceTwist with our approach fairly, we let the number of users in K -ASR equate the number of users in K -ASR, which means SpaceTwist provides the same level of privacy as KAWCR provides under the assumption that the attacker knows the locations of all the users. And we use *NNC* algorithm [1] to cloak the user location to K -ASR.

Fig.10 shows the experimental results when k , the number of results required, varies from 1 to 15 with step 2 and other parameters are set as shown in TABLE II. From the results, we know that the communication costs of KAWCR are significantly lower than that of SpaceTwist in all cases considered. The communication costs of KAWCR slightly increase with k over UI and NI while slightly fluctuate with k over GN. And the communication costs of SpaceTwist slightly increase with k over UI while fluctuates with k over NI and GN.

Fig.11 shows the experimental results when d , the distance between the anchor and the user issuing the query, varies from 100 to 200 with step 10 and other parameters are set as shown in TABLE II. The results show us that the communication costs of KAWCR are significantly lower than that of SpaceTwist in all cases considered. The communication costs of KAWCR almost keep the same as d increases over UI and NI while slightly fluctuates with d over GN. And the communication costs of SpaceTwist almost increase with d over all datasets considered. As d becomes larger and larger, KAWCR outperforms SpaceTwist more and more. The reason why KAWCR performs much better than SpaceTwist is that SpaceTwist may retrieve the k NN results of other users which are not in the privacy region while KAWCR only retrieves the k NN results of the users in K -ASR.

C. Summary

Our experiments illustrate that the communication costs of our proposal KAWCR are lower than those of TKA, both on synthetic and real datasets. The reason why KAWCR outperforms TKA is that TKA retrieves the k NN results of all the points in K -ASR while KAWCR retrieves the k NN results of all the users in K -ASR in incremental fashion. From the experimental results, we also know that the communication costs of KAWCR are significantly lower than those of SpaceTwist when they provide the same level of privacies, both on synthetic and real datasets. The reason why KAWCR performs much better than SpaceTwist is that SpaceTwist may retrieve the k NN results of other users which are not in the privacy region while KAWCR only retrieves the k NN results of the users in K -ASR.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a new framework called KAWCR to protect privacy in location-based services. KAWCR can guarantee that the user issuing the query is indistinguishable from at least $K-1$ other users. Compared with traditional K -anonymity, KAWCR only needs INN query processing algorithm while traditional K -anonymity needs complex processing algorithm at the server side, and the communication cost of KAWCR is lower than that of traditional K -anonymity on some datasets. Compared with SpaceTwist, both KAWCR and SpaceTwist only need INN query processing algorithm at the server side, but the communication cost of KAWCR is significantly lower than that of SpaceTwist on some datasets

when they provide the same level of privacies. TABLE III summarizes the three techniques.

TABLE III. SUMMARY OF THREE TECHNIQUES

| | KAWCR | Traditional K -anonymity | SpaceTwist |
|-----------------------|-------|----------------------------|------------|
| K -anonymity | Yes | Yes | No |
| Query processing cost | Low | High | Low |
| Communication cost | Low | High | High |

Our proposal considers snapshot k -nearest-neighbor queries. It's interesting to extend it to support continuous queries [15]. Furthermore, we are interested in extending our proposal to support queries in road networks [16].

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China under the grant No. 60873210.

REFERENCES

- [1] P. Kalnis, G. Ghinita, K. Mouratidis, and D. Papadias, "Preventing Location-Based Identity Inference in Anonymous Spatial Queries." In *IEEE TKDE*, 2007.
- [2] M. Gruteser and D. Grunwald, "Anonymous Usage of Location-Based Services through Spatial and Temporal Cloaking." In *MobiSys*, 2003.
- [3] B. Gedik and L. Liu, "Location Privacy in Mobile Systems: A Personalized Anonymization Model." In *ICDCS*, 2005.
- [4] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, "The new casper: Query Processing for Location Services without Compromising Privacy." In *VLDB*, 2006.
- [5] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "Priv'ée: Anonymous location-based queries in distributed mobile systems." In *WWW*, 2007.
- [6] C.-Y. Chow, M. F. Mokbel, X. Liu, "A Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-based Services." In *ACM GIS*, 2006.
- [7] H. Hu and D. L. Lee, "Range Nearest-Neighbor Query." In *TKDE*, 2005.
- [8] M. L. Yiu, C. Jensen, X. Huang, and H. Lu, "Spacetwist: Managing the trade-offs among location privacy, query performance, and query accuracy in mobile services." In *ICDE*, 2008.
- [9] G. R. Hjaltason and H. Samet, "Distance Browsing in Spatial Databases." In *TODS*, 24(2): 265-318, 1999.
- [10] P. Samarati, "Protecting Respondents' Identities in Microdata Release." In *IEEE TKDE*, 2001.
- [11] L. Sweeney, "K-Anonymity: A Model for Protecting Privacy." In *Int'l J. on Uncertainty, Fuzziness and Knowledge-Based Systems*, Vol. 10, No. 5, pp. 557-570, 2002.
- [12] N. Roussopoulos, S. Kelley, and F. Vincent, "Nearest Neighbor Queries." In *SIGMOD*, pp. 71-79, 1995.
- [13] G. Ghinita, P. Kalnis, and S. Skiadopoulos, "MOBIHIDE: A Mobile Peer-to-Peer System for Anonymous Location-Based Queries." In *SSTD*, 2007.
- [14] C. Zhang and Y. Huang, "Cloaking Locations for Anonymous Location Based Services: A Hybrid Approach." In *GeoInformatica*, Vol.3, No.2, pp.159-182, 2009.
- [15] C.-Y. Chow and M. F. Mokbel, "Enabling Private Continuous Queries For Revealed User Locations." In *SSTD*, pp. 258-275, 2007.
- [16] T. Wang and L. Liu, "Privacy-aware mobile services over road networks." In *VLDB*, 2009.