IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 14, 2019, accepted October 20, 2019, date of publication October 25, 2019, date of current version November 6, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949409

Preserving Location Privacy in Spatial Crowdsourcing Under Quality Control

XIANG CHU[1], JUN LIU[2], DAQING GONG[3], AND RUI WANG[4]

[1]School of Maritime Economics and Management, Dalian Maritime University, Dalian 116026, China
[2]International Business School, Beijing Foreign Studies University, Beijing 100089, China
[3]School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China
[4]Faculty of Economics and Management, Dalian University of Technology, Dalian 116000, China

Corresponding author: Daqing Gong (dqgong@bjtu.edu.cn)

This work was supported in part by the National Science Foundation of China under Grant 71802037 and Grant J1824031, in part by the Fundamental Funds for the Ministry of Education in China (MOE) Project of Humanities and Social Sciences under Grant 19YJC630137 and Grant 19YJC630043, and in part by the Fundamental Research Funds for the Central Universities under Grant 3132019223 and Grant 3132019224.

ABSTRACT Emerging spatial crowdsourcing (SC) provides an approach for collecting and analyzing spatiotemporal information from intelligent transportation systems. However, the exposure of massive location privacy to potential adversaries for the purpose of quality control makes workers more vulnerable. To protect workers' location privacy, an obfuscation scheme is proposed to incorporate uncertainties into the SC quality control problem through obfuscating the standard location data in terms of both space and time. Two measures, location entropy and results accuracy, are used to evaluate the performance of location privacy protection. We theoretically and experimentally confirm the security and accuracy of the obfuscation approach. The results of experiments show that: a) hiding workers' location from the requester reduces the quality of SC; and b) obfuscation arithmetic with appropriate obfuscation coefficients protects workers' location privacy with little effect on SC quality. Under the protection of this obfuscation scheme, the new system provides better security and similar quality compared to the existing SC system.

INDEX TERMS Spatial crowdsourcing, obfuscation location privacy of workers, quality control, EM algorithm.

I. INTRODUCTION

The crowdsourcing model is frequently used to gather data in intelligent transportation systems (ITS) applications e.g., avoiding traffic congestion. The new mechanism for collecting and analyzing spatiotemporal information is spatial crowdsourcing (SC), and this mechanism exploits a large volume of vehicles and their mobility. In the SC platform, a requester outsources a set of spatiotemporal tasks to a set of workers with mobile devices, and workers perform tasks after physically traveling to places of interest [1], e.g. the outsourcer requests workers in cars to collect real-time traffic information on roads Moreover, to assist in checking workers' submissions, some existing SC platforms require workers to disclose their immediate locations along with the taskspecified submission to the requester, who may be a potential adversary seeking to attack the location privacy of individual workers [2]. In practice, some malicious requesters may collect private information on worker locations through deliberately designed SC tasks. In addition, a crowdsourcing task is a kind of micro task, so a worker may submit numerous tasks with location information in a short period of time. Revealing workers' precise locations may allow an adversary to infer sensitive information and even to stalk or mug workers [3]. Hence, protecting location privacy is an essential aspect of SC, since workers will not agree to participate in spatial tasks if there is a possibility of a privacy breach.

To illustrate how the location privacy of workers in SC might be exploited by an adversary, we consider a trafficinformation collecting task as shown in Fig. 1 In the example, a worker in a car submits traffic information separately at positions A and B on the road. Assume that the road's speed limit is 80 km/h, the distance from A to B is 100 km, and the worker submits the task in position A at 13:00 and in position B at 14:00 of the same day. With this precise location data, the requester knows the average speed of worker travelling

The associate editor coordinating the review of this manuscript and approving it for publication was Dalin Zhang.

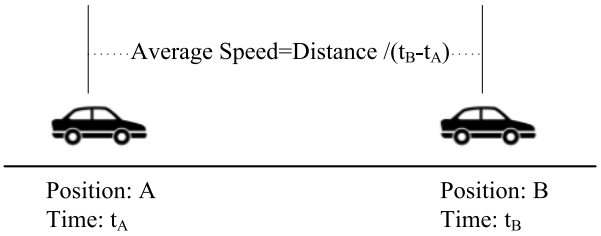


FIGURE 1. Traffic-information collecting task.

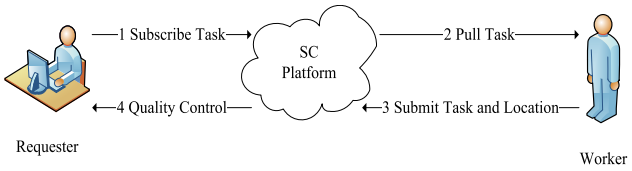


FIGURE 2. Workflow of SC system.

from A to B is 100 km/h and can infer that the worker exceeded the speed limit sometime during 13:00∼14:00. Speeding should be private driver information, but a series of corresponding attacks can be conducted to attack the driver once an adversary exploits speeding behavior.

Workers' location privacy is threatened in two phases of SC: tasking and reporting [4]. Existing works on the former focus on effectively and safely assigning spatial tasks to workers in terms of location [5]–[14]. In particular, Hien To and his coauthors made great contributions to preserving workers' location privacy during the task assignment process. However, our paper aims to protect workers' location privacy in the latter reporting phase. Related studies have put forward approaches to control crowdsourcing quality via evaluating the credibility of the contributed data for SC restricted by the given budget [15]–[22] However, notably, the methods designed for privacy and accuracy trade-off between protection and utility in general crowdsourcing scenarios are not suitable for SC. Few existing studies discuss how to achieve a beneficial trade-off between location privacy and accuracy in the reporting phase.

In this paper, we propose two obfuscation schemes to protect workers' location privacy in SC, whereby the requester only has access to fuzzy location data. The two proposed schemes are spatial obfuscation arithmetic (SOA) and temporal obfuscation arithmetic (TOA). The former transforms the exact longitude and latitude of an individual's location to an area in the SC task map, reducing the probability of sensitive behavior inference The latter replaces the precise time element with an appropriate time slot, so that the requester does not know the execution timestamp of tasks. However, obfuscation impedes the requester's ability to distinguish between spammers and workers, since less information is reported. Therefore, this paper optimizes the tradeoff extending the research of [23]. The requester can infer a true task result from the submission of several workers under a repeated labeling technique [24].

The rest of this paper is organized as follows. Section 2 describes those problems to solve in protecting location privacy of workers. Section 3 proposes two protection schemes and the performance measurement. Section 4 theoretically guarantees the availability of these schemes. Section 5 experimentally evaluates the performance of these schemes in terms of entropy and accuracy. Finally, Section 6 concludes this paper and puts forward considerations for future work.

II. PROBLEM STATEMENT

We consider the tradeoff between task quality, a concern of the requestor, and location privacy, a concern of the workers, in SC. Therefore, this section describes three problems: quality control, privacy breaches, and the combination of these two elements.

A. SPATIAL CROWDSOURCING QUALITY CONTROL

SC tasks are usually tedious with low pay. Consequently, errors are common even among workers who work hard. Moreover, some workers are spammers, submitting arbitrary answers independent of the question to earn their fee [25]. Thus, SC requesters need to distinguish true answers based on all sorts of worker submissions. This estimation behavior by requesters functions as quality control. Most requesters employ the redundancy approach, outsourcing each task to multiple workers and aggregating their results by some method such as the *majority voting technique* [24].

Fig. 2 shows the typical workflow of an SC system. A requester subscribes several tasks $T_i(1 \leq i \leq m)$ to the SC platform, and workers $W_j(1 \leq j \leq n)$ pull tasks from the platform in which to engage. Workers do not know one another, nor can they share task results Each task is executed by n workers redundantly, and each worker executes m tasks. Then task result matrix $V_{m\times n}$ and workers' location matrix $L_{m\times n}$ are received by the requester. $V_{m\times n} = \{v_{ij}\}$ consists of all reports from workers about task content, while $L_{m\times n} = \{l_{ij}\}$ records locations at which tasks are submitted. Compared with the target address, location data helps the requester to detect spammers among a group of workers. If any one of m task positions submitted by a worker is different from the target address, the requester regards the corresponding worker as a spammer. All results submitted by spammers are excluded. The quality control problem of SC is defined as Problem 1.

Problem 1(SC Quality Control): the requester estimates the results $V_m = \{v_i\}$ of tasks from result matrix $V_{m\times n}$ and location matrix $L_{m\times n}$.

We use a measure, *result accuracy* (RA), to evaluate the performance of the estimation. Let $\tilde{V}_m = \{\tilde{v}_i\}$ be the set of true results and RA equal the ratio of same elements in V_m and $\tilde{V}_m$ For simplicity, we assume that each task instance is binary. A worker submits yes ($v_{ij} = 1$) or no ($v_{ij} = 0$) for a task.

Without loss of generality, findings can be extended to a wide range of task types (Hiroshi et al., 2014).

B. LOCATION PRIVACY BREACHES OF WORKERS

In Fig. 2, worker location data is forwarded to the requester without any intervention, which leads to privacy breaches. Workers may face undesirable effects from such a breach, such as threats to personal safety, location-based spam, and sensitive information inference. The location privacy protection problem is defined as follows.

Problem 2(Location Privacy Protection): The location privacy protection problem is to prevent workers from sharing their own precise location with the requester.

This paper does not explain how the attack is implemented, although it does measure how much private information is leaked. Thus, we first define the location data *loc*:

$$loc = (worker\ id, lng, lat, time)$$

where *worker id* is the unique identifier of a worker in the system, and *lng* and *lat* are, respectively, the longitude and latitude of the worker's physical location when the timestamp is *time*.

Uncertainty is a state of having limited knowledge where it is impossible to describe exactly the existing state, a future outcome, or more than one possible outcome. In the context of location privacy, uncertainty explains how difficult it is for adversaries to pinpoint workers' actual locations. Following existing literature, we describe uncertainty with the location entropy of worker W_j at time t:

$$I_t\left(W_j\right) = -\sum\nolimits_i Pr\{W_j \text{ in } A_i \text{ at } t\} log Pr\{W_j \text{ in } A_i \text{ at } t\} \quad (1)$$

where A_i is part of an area in which the worker may be, and $\bigcup_i A_i$ covers the whole task map. $Pr\{W_j \text{ in } A_i\}$ is the probability that W_j lies in A_i. Location entropy measures the degree of privacy protection. Higher entropy means more uncertainty, more imprecision and better security.

C. LOCATION OBFUSCATION PROBLEM

We consider that the location privacy of workers in SC system model consists of the following three dimensions: *worker*, *region*, *time*. Correspondingly, the goal of privacy protection is to hide these elements from the SC requester at the appropriate level. There are numerous studies on preserving location privacy that can be divided mainly into two categories: anonymity and obfuscation. However, the precise *worker* element is required for the requester to pay the worker his or her earned reward, and thus, the anonymity technique [26] is not suitable for the discussed SC scenario. Therefore, we adapt the obfuscation technique [27] However a general obfuscation technique is unable to control crowdsourcing quality. Different from the literature on information security, this paper focuses on developing a customized obfuscation technique compatible with crowdsourcing quality control.

The basic idea of the obfuscation approach is to degrade the quality of workers' location data by reducing precision.

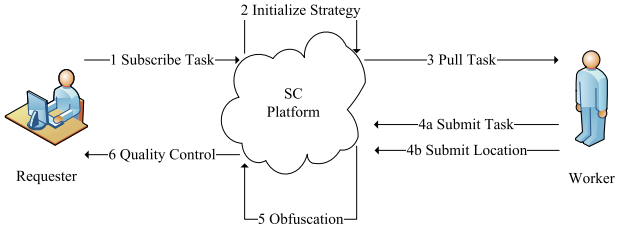


FIGURE 3. SC system preserving worker location privacy.

Inevitably, a less precise location decreases the accuracy of task results estimated by the requester. Location obfuscation introduces a challenge to optimize the trade-off between privacy and accuracy. The *location obfuscation problem*(LOP) in SC is defined as follows.

Problem 3(Location Obfuscation Problem): Find a location privacypreserving algorithm to transform precise location $L_{m\times n}$ of workers into fuzzy location. With the fuzzy location, the requester can still estimate task results at a high level of accuracy.

III. OBFUSCATION ARITHMETIC

The obfuscation scheme incorporates uncertainties into the precise location data in terms of both *region* and *time* guaranteeing high-quality results. Section 3.1 redesigns the system model and workflow for better preservation of privacy. Section 3.2 describes the obfuscation arithmetic. Section 3.3 discusses associated performance measures.

A. SYSTEM WORKFLOW

The existing system in Fig. 2 allows precise location information to be exposed to the requester. Workers become vulnerable with as the amount of privacy information exposed accumulates over time. The chosen approach to preserving privacy is to change the mechanism of the system, i.e. to obfuscate location data.

To implement the obfuscation mechanism, we redesign the workflow of the SC system as shown in Fig. 3. Compared with the system in Fig. 2, the new system makes the following changes:

1) As soon as the requester subscribes a task to the platform, the trustworthy SC platform initializes the strategy of location obfuscation in activity 2. The strategy could be determined by either the platform or the requester, depending on the task content.
2) Activity 4 separates location data from task data. After a worker submits the task, task data is forwarded to the requester without intervention by the platform.
3) Activity 5 executes the preprocessing of obfuscation on location data, following the strategy set forth in activity 2, before sending the data to the requester.

The implementation of this strategy includes two parts: arithmetic and arithmetic parameter setting. The former is discussed in Section 3.2, while the latter depends on specific tasks.

B. OBFUSCATION ARITHMETIC

All tasks of SC stay within a geographical map, which we call a task map. Divide the map into m sections of areas $A_i(1 \le i \le m)$. Each area A_i contains one and only one task position $P_i(1 \le i \le m)$. The division satisfies the condition that any point P in the map belongs to the area A_i whose task point P_i is the closest to P. That is,

$$|P - P_i| \le |P - P_p|, \quad \forall P \in A_i, \text{ for all } p = 1, 2, \ldots, m$$

Here are two ways to reduce the precision of submitted location data: spatial obfuscation and temporal obfuscation. Both ways increase the uncertainty of worker location when location data is sent to the requester.

1) SPATIAL OBFUSCATION ARITHMETIC

Spatial obfuscation arithmetic (SOA) reduces the precision of position data that the worker submits, i.e. longitude and latitude, abstracted as

$$f_{\text{SOA}} : P \longrightarrow \{A'_1 A'_2, \ldots, A'_k\}, \tag{2}$$

where $A'_p \in \{A_i\}_{i=1,\ldots,n}$ and $p = 1, \ldots, k$.

SOA transforms a point with exact longitude and latitude into k sections of an area in the task map. Obfuscation coefficient k indicates the degree of spatial obfuscation. The higher k is, the safer the worker's location privacy is. If $k = m$, the worker may be located at any position on the map, given the location information sent to the requester. The security level of the $k = m$ scheme is same as the scheme in which worker location is not forwarded to the requester along with task results. The detailed steps of SOA are as follows.

Step 1: Identify in which area the worker is located, i.e., $P \in A'_k$.

Step 2: Randomly pick $k - 1$ areas from the task map excluding A'_k. The $k - 1$ areas and A'_k comprise a new set of k elements, i.e.,$\{A'_1 A'_2, \ldots, A'_k\}$.

Step 3: Reorder the new set, and replace the precise position P with the new data.

Recall the example of real-time traffic information collection task mentioned in Fig. 1. As shown in Fig. 4, the road length is 200 km, which the requester divides into five sections, i.e., $A_i(1 \le i \le 5)$, each of which is 40 km. With the support of the SC platform, the requester wants to know the respective traffic status of these five sections. A worker submits tasks, respectively, at position A and at position B. Now obfuscated areas A_2 and A_4 are transmitted along with traffic information instead of precise positions. Given the obfuscated location information, the requester can infer the average speed of the worker ranges between [40, 120] km/h. The worker's speeding behavior cannot be precisely determined.

2) TEMPORAL OBFUSCATION ARITHMETIC

In addition to spatial obfuscation, we introduce another approach, called temporal obfuscation arithmetic (TOA), which protects privacy by obfuscating the timestamp of location data with an appropriate time slot. The arithmetic can be abstracted as

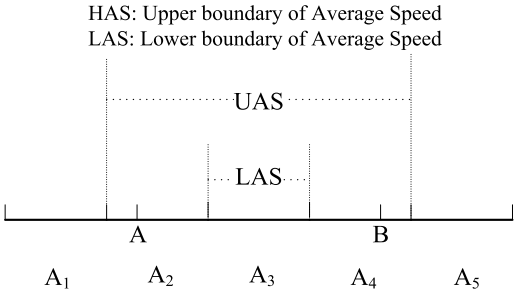


FIGURE 4. Spatial obfuscation in collecting traffic information.

$$f_{TOA} : t \longrightarrow [t - \Delta t_1 t + \Delta t_2], \quad \Delta t_1, \Delta t_2 > 0. \tag{3}$$

TOA may appear to be a simple trick transforming a time point into an interval. However, the implementation of TOA is somewhat complex due to the following two steps:

Step 1: Check whether a precise timestamp is necessary for the requester. If yes, TOA will not be executed on the worker's location data, and original time data t is sent to the requester. If no, go to step 2.

Step 2: Determine temporal obfuscation parameters Δt_1 and Δt_2, which are respectively lead time and lag time on the actual time tolerated by the requestor.

Recall the example of real-time traffic information collection task again. Let obfuscation parameters be $\Delta t_1 = \Delta t_2 = 10min$; then, the submitted location is obfuscated to $(W_j, A, [12 : 50, 13 : 10])$ and $(W_j, B, [13 : 50, 14 : 10])$ under the protection of TOA. From the obfuscated time slot, the requester only knows that the average speed ranges from 75 km/h to 150 km/h but cannot perceive the speeding behavior of an individual worker, even if the worker drives at a speed of 100 km/h from A to B.

TOA is a potential technique to preserve the driver's behavioral privacy. However, the arithmetic works only if the requester accepts nonreal-time traffic information, i.e., $\Delta t_1, \Delta t_2 > 0$. For traffic information consumers, a several-minute delay of traffic report is usually tolerable, making TOA reliable for location privacy protection. As a result, a delay tolerance check is the first step of TOA. After that process, the requester determines Δt_1 and Δt_2 according to the degree of consumers' tolerance in the second step.

C. MEASURES OF THE PERFORMANCE

Both spatial and temporal obfuscation approaches reduce the certainty of location data, which significantly shields workers' location privacy from the requester. Due to the increased location uncertainty, it is more difficult for the requester to pick out spammers from a mass of workers. Consequently, more arbitrary answers may sneak into the result set, which may reduce the accuracy of estimation implemented by majority voting technique.

Therefore, we use the following two measures, location entropy and results accuracy, to evaluate the performance of location privacy protection.

1) LOCATION ENTROPY (LE)

Location entropy measures the degree of privacy protection. Higher entropy means more uncertainty, more imprecision and better security. We follow the definition of workers' location entropy in Section 2 and measure the LE of W_j at time t as

$$I_t\left(W_j\right) = -\sum\nolimits_i Pr\{W_j \text{ in } A_i \text{ at } t\} log Pr\{W_j \text{ in } A_i \text{ at } t\}. \quad (4)$$

The goal of the platform is to maintain LE at a high level in the process of preserving privacy.

2) RESULTS ACCURACY (RA)

Results accuracy is the accuracy of the estimated tasks. Obfuscation schemes obfuscate workers' location information, which results in less information received by the requester. Inevitably, less input impedes the requester's task of quality control and may decrease the estimated accuracy of the tasks. Thus, the goal of the platform is to maintain the RA under the obfuscation scheme close to the RA achieved under the precise location scheme.

Based on above two measures, location privacy protection is successful if and only if it significantly increases LE but does not significantly decrease RA.

IV. QUALITY CONTROL OF SC

A. TASK RESULT ESTIMATION METHOD

The LC (latent class) method (Dawid and Skene, 1979) is the classical approach used in quality control problems. Regarding the error rate of workers as the latent class, the EM algorithm provides a way of obtaining the maximum likelihood estimation of task results in SC. Let n-dimensional vector $\eta_n = \{\eta_j\}$ be the error rates of n workers. Worker W_j completes a task the result of which is wrong in the probability of η_j, $\eta_j = (\eta_j^1 \eta_j^2)$. η_j^1 and η_j^2 are independent, representing the error rate of W_j when the true result is 1 and 0, respectively.

$$\eta_j^1 = P\left(V_{ij} = 0 \,|\, \tilde{v}_i = 1\right), \quad \eta_j^2 = P(V_{ij} = 1 | \tilde{v}_i = 0). \quad (5)$$

The EM algorithm [28] repeats the expectation step and the maximization step alternately until convergence. The expectation step estimates true results using the *majority voting technique* weighted by workers' accuracy rate (equals 1 – error rate). The maximization step updates the error rate using estimated true results. Repeating these two steps increases the value of likelihood function Q, which gradually converges. Finally, low-quality workers are eliminated from the results. Each step of EM is described in the following.

1) EXPECTATION STEP

Define an m-dimensional vector μ. $\mu_i (1 \le i \le m)$ is the a posteriori probability that the true result of task T_i equals 1. That is,

$$\mu_i = P(\tilde{v}_i = 1 | V_{m\times n}, \eta). \quad (6)$$

Initialize vector μ using *majority voting* weighted by accuracy rate:

$$\mu_i^{(t)} = \frac{pa_i}{pa_i + (1-p)b_i}. \quad (7)$$

where t means the t-th iteration, p is the probability that the true result of task T_i equals 1, $1-p$ is the probability that the true result of task T_i equals 0, a_i is weight of probability when the true result equals 1, and b_i is weight of probability when the true result equals 0.

$$a_i = \prod\nolimits_{j=1}^{n} (1-\eta_j^1)^{V_{ij}} (\eta_j^1)^{1-V_{ij}},$$
$$b_i = \prod\nolimits_{j=1}^{n} (1-\eta_j^2)^{1-V_{ij}} (\eta_j^2)^{V_{ij}}. \quad (8)$$

2) MAXIMIZATION STEP

With expectation of μ in the previous step, we can estimate p:

$$p = \frac{1}{m} \sum\nolimits_{i=1}^{m} \mu_i. \quad (9)$$

By calculating the approximation of the maximum likelihood estimator, we estimate the error rate:

$$\eta_j^1 = 1 - \frac{\sum_{i=1}^{m} \mu_i V_{ij}}{\sum_{i=1}^{m} \mu_i}, \quad \eta_j^2 = 1 - \frac{\sum_{i=1}^{m} (1-\mu_i)(1-V_{ij})}{\sum_{i=1}^{m} (1-\mu_i)}. \quad (10)$$

The following likelihood function Q is adopted:

$$Q\left(p, \eta_j\right) = \sum\nolimits_{i=1}^{m} [\mu_i \log pa_i + (1-\mu_i) \log (1-p) b_i]. \quad (11)$$

The algorithm is considered to converge if the following condition holds:

$$\left| Q\left(p^{(t+1)}, \eta_j^{(t+1)}\right) - Q\left(p^{(t)}, \eta_j^{(t)}\right) \right| / |Q\left(p^{(t)}, \eta_j^{(t)}\right)| < \varepsilon. \quad (12)$$

where ε is the threshold of algorithm convergence.

If function Q has not converged, return to the expectation step and start the next iteration, i.e., $t = t+1$; otherwise, end the algorithm and return estimated results μ.

B. THEORETICAL PROOF

We assume workers are uniformly distributed in the task map. A spammer has $1/m^m$ chance to evade detection by the requester who knows workers' precise locations. By contrast, a spammer has k^m/m^m chance to evade detection by the requester who receives the workers' obfuscated location information.

Although the obfuscation scheme increases the probability of spammer evasion by k^m times, the evasion probability is still close to 0 when $k \ll m$ or m is relatively large. In practice, however, the obfuscation scheme results in very few spammers escaping from detection. This very small number of spammers is tolerable based on the following observation.

Lemma 1: Inserting a small number of wrong task results into the original result set does not significantly impact the accuracy of estimation.

Proof: For a binary SC task, n workers submit the task, and workers' average probability of error is η, $\eta < 0.5$. Applying the majority voting technique to the result set received from n workers, the a posteriori probability that estimation of the task result is wrong (the true result is 1, while the estimated is 0, or 1 is estimated when the true is 0) will be in error with probability:

$$P_e = \sum_{i=\frac{n+1}{2}}^{n} C_n^i \eta^i (1-\eta)^{n-i}. \tag{13}$$

The probability of error decreases exponentially with the number of workers n and drops to zero for large numbers of workers. If more Δn wrong results are added into the original results set, $\Delta n \ll n$, the new probability of error based on the set of $n + \Delta n$ results is

$$P'_e = \sum_{i=\frac{n-\Delta n+1}{2}}^{n} C_n^i \eta^i (1-\eta)^{n-i}. \tag{14}$$

Subtracting formulation (13) from (14), we obtain the increment of error probability

$$\begin{aligned} \Delta P_e &= P'_e - P_e = \sum_{i=\frac{n-\Delta n+1}{2}}^{\frac{n-1}{2}} C_n^i \eta^i (1-\eta)^{n-i} \\ &\leq (\frac{\Delta n}{2} - 1) C_n^{\frac{n-1}{2}} \eta^{\frac{n-\Delta n+1}{2}} (1-\eta)^{\frac{n+1}{2}}. \end{aligned} \tag{15}$$

The right-hand side of formulation (15) decreases exponentially and ultimately drops to zero with n. Hence, a small number of wrong task results does not impact the accuracy of the estimation. □

Although the obfuscation scheme may shield spammers from detection, lemma 1 guarantees that this small number of spammers will not impede the requester's estimation. Therefore, based on the analysis above, we achieve main conclusion of this paper.

Theorem 1: Proposed obfuscation schemes do not impede SC quality control.

V. NUMERICAL EXPERIMENTS

The existing SC system forwards workers' location data directly to the requester without protecting workers' location privacy. The obfuscation arithmetic changes the performance of the SC system in terms of location entropy and results accuracy, as described in Section 3.3. The results accuracy measure of obfuscation arithmetic has been theoretically guaranteed acceptable by Theorem 1. Furthermore, this section validates proposed approaches experimentally.

A. DESIGN AND SYNTHETIC DATASET

This experiment is designed to assess the effect on location entropy and results accuracy, varying the number of tasks m, the number of workers n, the error rate of workers η, the percentage of spammers r, and the obfuscation coefficient k.

We use the LC model to generate synthetic datasets. First, for all tasks, we generate the vector of true results $\tilde{V}_m$ from a Bernoulli distribution with parameter $p = 0.5$, where p is the probability that true result of a task is 1. Then, for all workers, we set a worker to be a spammer at the rate of r. If worker W_j is a spammer, we generate his task results $v_{ij}(1 \leq i \leq m)$ from Bernoulli distribution $B(1, 0.5)$. Otherwise, we generate $v_{ij}(1 \leq i \leq m)$ from Bernoulli distribution $B(1, 1-\eta_j^1)$ in case of $\tilde{v}_i = 1$ and $B(1, 1-\eta_j^2)$ in case of $\tilde{v}_i = 0$. Finally, for all tasks, we generate workers' locations. If worker W_j is a spammer, randomly pick one area from set $\{A_i\}$, and set it as the submitted location of one task by the worker. Otherwise, the location is set as the area in which the task is located.

B. EXPERIMENTS ON RESULTS ACCURACY

We experimentally evaluate the accuracy of the results by comparing three settings. The settings $k = 1$ and $k = m$ serve as benchmarks, and the setting $k = 5$ illustrates the performance of the obfuscation schemes. In the setting of $k = 1$, workers submit tasks with data on the area in which they are located. This process leads to the same accuracy as the existing SC system. In the setting of $k = 5$, 5 sections of areas are submitted. The requester knows the worker is located within one of these 5 areas. In the setting of $k = m$, it is same as the case in which no location is submitted. For the requester, the worker may be located anywhere on the map. An EM algorithm [29] is used to estimate true results from the standpoint of the requester.

1) EXPERIMENTAL METHODOLOGY

We examine the accuracy of the estimated results by varying parameters of the datasets. For given parameters m, n, η, and r, we randomly generate 1000 instances. Each instance has m tasks and n workers, and the answers follow error rate η and the percentage of spammers r. Results accuracy under the parameter setting is averaged over these 1000 replications with respect to $k = 1, 5, m$.

It is necessary to mention that the threshold of algorithm convergence ε is is set as 10^{-5} for the EM algorithm. Since time cost and accuracy change little with sufficiently small ε in experiments, we will not discuss the effect of ε.

2) RESULTS

Fig. 5 shows the accuracy of estimated results averaged over 1000 replications under varying parameter settings. Figs. 5a and 5b are experimented with low error rate $\eta = (0.2, 0.2)$ and low spammer rate $r = 0.2$. The results show that the accuracy of three settings ($k = 1, 5, m$) are close to one another, regardless of the numbers of tasks and workers. Nevertheless, more tasks and workers improve the accuracy of these approaches. From Figs. 5c and 5d, we see that accuracy remains close under low error and spammer rates, whereas when error rate and spammer rate are high, the accuracy of settings $k = 1, 5$ are significantly higher than the setting $k = m$.

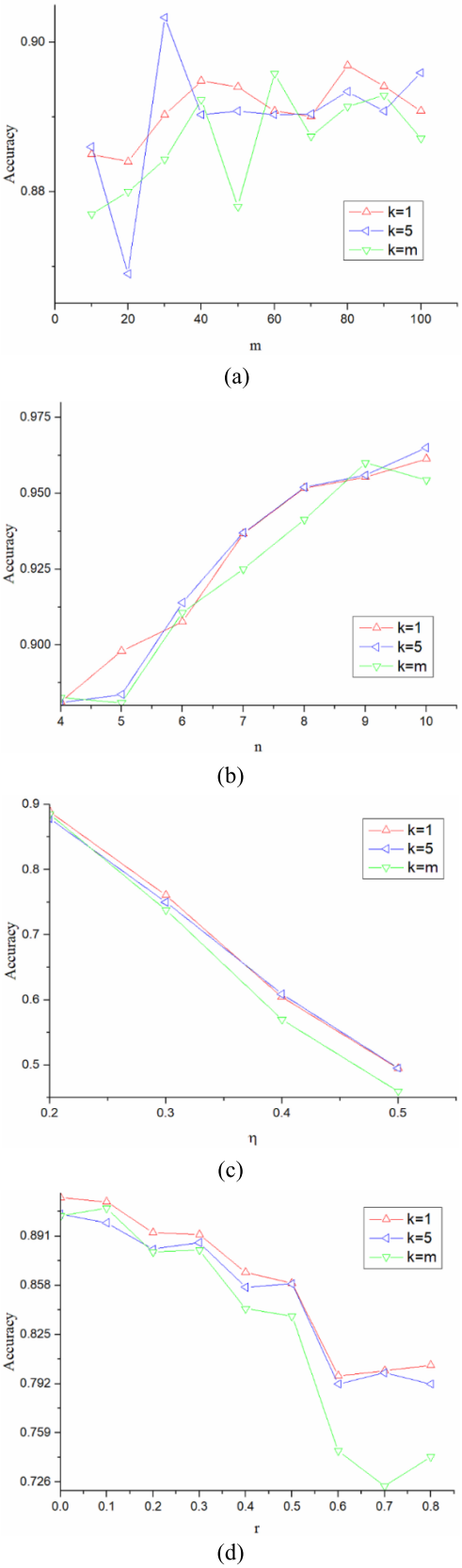


FIGURE 5. Comparing accuracy when k = 1,5,m under various parameter combinations.

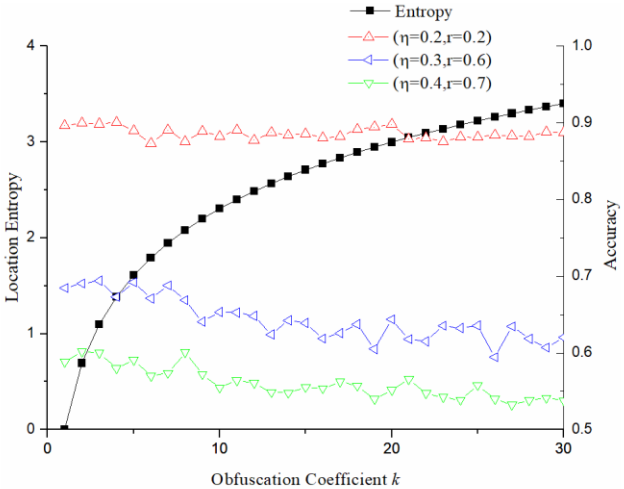


FIGURE 6. Effect of k on entropy and accuracy.

From the experiments on results accuracy, we prove that:

1) Hiding workers' location from the requester reduces the quality of SC.
2) Obfuscation arithmetic with appropriate obfuscation coefficients protects workers' location privacy with little effect on the quality of SC.

C. SENSITIVITY ANALYSIS FOR k

In this subsection, we examine the tradeoff between location entropy and results accuracy. As stated in Sections 3 and 4, obfuscation arithmetic improves location privacy protection, but it may reduce the quality of SC to a lesser or greater extent. We experimentally evaluate the effect of obfuscation on location entropy and results accuracy at the same time and try to find the appropriate obfuscation coefficient as mentioned in Section 5.2.2.

1) EXPERIMENTAL METHODOLOGY

To understand the pros and cons of obfuscation schemes, we assess their effect on the location entropy and results accuracy, varying the obfuscation coefficient k. According to the definition in Section 2.2, the location entropy of a worker depends only on the parameter k, irrespective of m, n, η or r. Figs. 5a and 5b imply that parameter k may change the accuracy only if the error rate and spammer rate are both high. Therefore, we examine results accuracy under two parameter settings with high η and r, with the other setting with low η and r serving as a benchmark.

For every setting, we randomly generate 1000 instances, each of which has $m = 30$ tasks and $n = 5$ workers. The results accuracy under the parameter setting is averaged over these 1000 replications with respect to k.

2) RESULTS

Fig. 6 shows location entropy and results accuracy under varying obfuscation coefficient k. We have three findings

from Fig. 6. First, with respect to k, location entropy grows fast in the left part of the horizontal axis, while the growth slows down in the right part. Second, the value of k has little effect on accuracy when error rate ($\eta = (0.2, 0.2)$) and spammer rate ($r = 0.2$.) are low. Part of the conclusion has been evaluated in Section 5.2. Finally, from curve ($\eta = 0.3$, r $= 0.6$) and ($\eta = 0.4$, r $= 0.7$), we find that the accuracy declines significantly respect to k, but the decline is nonlinear. Accuracy remains high in the low-k area and starts to drop significantly when $k > 8$.

Based on the above, the best value of k ranges in the interval where entropy grows slowly, and accuracy has not started to drop.

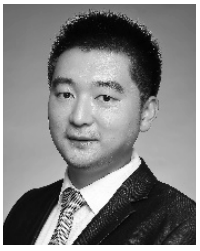
VI. CONCLUSION

In the SC system, sharing workers' precise location with requesters creates vulnerability. However, hiding workers' location information from requesters weakens the quality control of SC. Therefore, we propose obfuscation schemes to solve the trade-off between security and quality by obfuscating spatiotemporal data. Under the protection of this obfuscation scheme, the new system provides better security and the same quality compared to the existing SC system.

This paper identifies a privacy problem in SC and establishes a simple research framework. Nevertheless, every protection scheme has its limit. Obfuscated location information of workers still leaves space for adversaries to attack privacy by analyzing the probability of a worker's location. In future work, we will extend our proposed method to develop a more sophisticated obfuscation scheme that provides better performance.

REFERENCES

[1] Y. Zhao and Q. Zhu, "Evaluation on crowdsourcing research: Current status and future direction," *Inf. Syst. Frontiers*, vol. 16, no. 3, pp. 417–434, 2014.
[2] Y. Zhao and Q. Han, "Spatial crowdsourcing: Current state and future directions," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 102–107, Jul. 2016.
[3] M. Wernke, P. Skvortsov, F. Dürr, and K. Rothermel, "A classification of location privacy attacks and approaches," *Pers. Ubiquitous Comput.*, vol. 18, no. 1, pp. 163–175, 2014.
[4] Z. He, Z. Cai, and J. Yu, "Latent-data privacy preserving with customized data utility for social network data," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 665–673, Jan. 2018.
[5] H. To and C. Shahabi, "Location privacy in spatial crowdsourcing," in *Handbook of Mobile Data Privacy*, A. Gkoulalas-Divanis and C. Bettini, Eds. Cham, Switzerland: Springer, 2018.
[6] P. Cheng, X. Lian, L. Chen, J. Han, and J. Zhao, "Task assignment on multi-skill oriented spatial crowdsourcing," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2201–2215, Aug. 2016.
[7] J. Cui, J. Wen, S. Han, and H. Zhong, "Efficient privacy-preserving scheme for real-time location data in vehicular ad-hoc network," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3491–3498, Oct. 2018.
[8] D. Deng, C. Shahabi, U. Demiryurek, and L. Zhu, "Task selection in spatial crowdsourcing from worker's perspective," *GeoInformatica*, vol. 20, no. 3, pp. 529–568, 2016.
[9] Y. Gong, C. Zhang, Y. Fang, and J. Sun, "Protecting location privacy for task allocation in ad hoc mobile cloud computing," *IEEE Trans. Emerg. Topics Comput.*, vol. 6, no. 1, pp. 110–121, Jan./Mar. 2015.
[10] H. ul Hassan and E. Curry, "Efficient task assignment for spatial crowdsourcing: A combinatorial fractional optimization approach with semi-bandit learning," *Expert Syst. Appl.*, vol. 58, pp. 36–56, Oct. 2016.
[11] H. To, G. Ghinita, and C. Shahabi, "A framework for protecting worker location privacy in spatial crowdsourcing," *Proc. VLDB Endowment*, vol. 7, no. 10, pp. 919–930, Jun. 2014.
[12] H. To, G. Ghinita, L. Fan, and C. Shahabi, "Differentially private location protection for worker datasets in spatial crowdsourcing," *IEEE Trans. Mobile Comput.*, vol. 16, no. 4, pp. 934–949, Apr. 2017.
[13] H. To, C. Shahabi, and L. Xiong, "Privacy-preserving online task assignment in spatial crowdsourcing with untrusted server," in *Proc. IEEE 34th Int. Conf. Data Eng.*, Paris, France, Apr. 2018, pp. 833–844.
[14] D. Wu, Y. Zhang, L. Bao, and A. C. Regan, "Location-based crowdsourcing for vehicular communication in hybrid networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 14, no. 2, pp. 837–846, Jun. 2013.
[15] H. Ye, K. Han, C. Xu, J. Xu, and F. Gui, "Toward location privacy protection in spatial crowdsourcing," *Int. J. Distrib. Sensor Netw.*, vol. 15, no. 3, 2019. doi: 10.1177/1550147719830568.
[16] C. Miao, H. Yu, Z. Shen, and C. Leung, "Balancing quality and budget considerations in mobile crowdsourcing," *Decis. Support Syst.*, vol. 90, pp. 56–64, Oct. 2016.
[17] W. S. Lasecki, Y. C. Song, H. Kautz, and J. P. Bigham , "Real-time crowd labeling for deployable activity recognition," in *Proc. CSCW*, San Antonio, TX, USA, Feb. 2013, pp. 1203–1212.
[18] H. Kajino, H. Arai, and H. Kashima, "Preserving worker privacy in crowdsourcing," *Data Mining Knowl. Discovery*, vol. 28, pp. 1314–1335, Sep. 2014.
[19] C. Shahabi, "Towards a generic framework for trustworthy spatial crowdsourcing," in *Proc. MobiDE*, New York, NY, USA, Jun. 2013, pp. 1–4.
[20] L. R. Varshney, "Privacy and reliability in crowdsourcing service delivery," in *Proc. Annu. SRII Global Conf.*, Jul. 2012, pp. 55–60.
[21] L. R. Varshney, A. Vempaty, and P. K. Varshney, "Assuring privacy and reliability in crowdsourcing with coding," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2014, pp. 1–6.
[22] G. Zhang and H. Chen, "Quality control for crowdsourcing with spatial and temporal distribution," in *Proc. 6th Int. Conf. Internet Distrib. Comput. Syst. (IDCS)*, Hangzhou, China, 2013, pp. 169–182.
[23] G. Zhang and H. Chen, "Quality control of massive data for crowdsourcing in location-based services," in *Proc. 13th Int. Conf. Algorithms Archit. Parallel Process. (ICA3PP)*, Sorrento, Italy, 2013, pp. 112–121.
[24] X. Chu and Q. Zhong, "Crowdsourcing quality control model protecting location privacy of workers," *Syst. Eng.-Theory Pract.*, vol. 36, no. 8, pp. 2047–2055, 2016.
[25] V. S. Sheng, F. Provost, and P. G. Ipeirotis, "Get another label? Improving data quality and data mining using multiple, noisy labelers," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 614–622.
[26] D. R. Karger, S. Oh, and D. Shah, "Budget-optimal task allocation for reliable crowdsourcing systems," *Oper. Res.*, vol. 62, no. 1, pp. 1–24, 2014.
[27] A. R. Beresford and F. Stajano, "Location privacy in pervasive computing," *IEEE Pervasive Comput.*, vol. 2, no. 1, pp. 46–55, Jan. 2003.
[28] C. Lee, Y. Guo, and L. Yin, "A framework of evaluation location privacy in mobile network," *Procedia Comput. Sci.*, vol. 17, pp. 879–887, Dec. 2013.
[29] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *J. Roy. Statist. Soc. C (Appl. Statist.)*, vol. 28, no. 1, pp. 20–28, 1979.
[30] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc., B (Methodol.)*, vol. 39, no. 1, pp. 1–38, 1977.



XIANG CHU received the B.S. and M.S. degrees in mathematics and the Ph.D. degree in management information system from the Dalian University of Technology, China, in 2006, 2008, and 2015, respectively. From 2008 to 2011, he was a System Engineer with the Citi Group. From 2016 to 2018, he held a postdoctoral position at Tsinghua University, China. Since 2018, he has been an Associate Professor with the School of Economics and Management, Dalian Maritime University, China. His research interests include supply chain and logistics management, scheduling, and management information systems.

JUN LIU received the B.S. and M.S. degrees in finance from the Dongbei University of Finance and Economics, China, in 2012 and 2015, respectively, and the Ph.D. degree in economics from Tsinghua University, China, in 2019. Since 2019, she has been an Assistant Professor with the International Business School, Beijing Foreign Studies University, China. Her research interests include information economics and technology management.

DAQING GONG received the Ph.D. degree in management science from Beijing Jiaotong University, Beijing, China. He is currently a Lecturer of management science with Beijing Jiaotong University, and a Research Assistant of management science with Tsinghua University. His current research interests include data mining, big data, intelligent transportation, and simulation.

RUI WANG received the B.S. degree from Zhejiang University, China, in 2006, and the M.S. degree from the Dalian University of Technology, China, in 2019. At the same time, she is the CEO of a start-up focusing on emerging information technology.