Task Assignment in Spatial Crowdsourcing: Challenges and Approaches

Hien To
supervised by Prof. Cyrus Shahabi
Department of Computer Science, University of Southern California
Los Angeles, CA 90089
hto@usc.edu

ABSTRACT

Spatial crowdsourcing (a.k.a mobile crowdsourcing) is a new paradigm of data collection, which has been emerged in the last few years to enable workers to perform tasks in the physical world. The objective of spatial crowdsourcing is to outsource a set of location-specific tasks to a set of workers, in which the workers are required to physically be at the task locations to complete them, i.e., taking pictures or collecting air quality information at specified locations of interest. In this paper, we discuss the unique challenges of spatial crowdsourcing: task assignment, incentive mechanism, worker's location privacy and the absence of real-world datasets. Thereafter, we present our current approaches to those issues.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Database Applications—*Spatial databases and GIS*

1. INTRODUCTION

The increase in computational and communication performance of mobile devices, coupled with the advances in sensor technology lead to an exponential growth in data collection and sharing by smartphones. Exploiting this large volume of potential users and their mobility, a new mechanism for efficient and scalable data collection has emerged, named Spatial Crowdsourcing (SC) [5]. With SC, the requester issues a location-specific task (*spatial task*) to a spatial crowdsourcing server (SC-server). Consequently, the SC-server crowdsources the task among the available workers in the vicinity of the events. Once the workers document their events with their mobile phones, the results are sent back to the requester. SC has applications in numerous domains such as journalism, tourism, intelligence, emergency response and urban planning. Examples of real-world SC applications are TaskRabbit[1] (commercialized system) and gMission[2] (testbed system).

A recent survey [19] thoroughly discuss the unique challenges of SC, including task assignment, incentive mechanism, privacy protection, the absence of real-world datasets, scalability and quality of reported data, etc. Our work has been focusing on the first four issues. The first objective of the SC-server is to match many requesters' tasks to numerous workers in real-time given the dynamic arrivals of workers and tasks, i.e., new tasks and workers become available or as tasks are completed (or expired) and workers leave the system. Thus, the dynamism of the arriving tasks and workers renders an optimal solution infeasible in the online scenario. Toward this end, we study the complexity of offline task assignment where the server is clairvoyant about the future workers and tasks and then propose heuristics for the online scenario that exploit the spatial and temporal knowledge acquired over time.

The second issue that hinders the success of SC is workers' location privacy as mobile users may not accept to engage in spatial tasks if their privacy is violated. To illustrate, workers are often assigned to nearby tasks to minimize the travel cost; thus, matching must take into account the location of workers. However, disclosing individual locations (to the SC-server) has serious privacy implications. Without privacy protection, a malicious adversary can stage a broad spectrum of attacks such as physical surveillance and stalking, and breach of sensitive information such as an individual's health issue (e.g., presence in a cancer treatment center), alternative lifestyles, political and religious preferences (e.g., presence in a church). Thus, we propose a framework that enables the participation of the workers without compromising their privacy.

A major challenge in any crowdsourcing system is how to motivate people to participate, in which payment is a popular mean. However, little study has been done on the relationship between incentives and workers' participation in SC. Motivated by such relationship, we conduct two real-world SC campaigns utilizing our mobile app, named Genkii, which enables users to report moods at their locations and time (regarded as spatial task). We reward users for each performed task by means of Yahoo! Japan Crowdsourcing. We conduct various analysis on the reported data, revealing interesting mobility patterns. Last but not least, we mention the absence of real-world datasets in the research community

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

3rd ACM SIGSPATIAL PhD Workshop'16, October 31-November 03 2016, Burlingame, CA, USA
© 2016 ACM. ISBN 978-1-4503-4584-2/16/10…$15.00
DOI: http://dx.doi.org/10.1145/3003819.3003820

[1] https://www.taskrabbit.com/
[2] http://www.gmissionhkust.com/

and discuss our approach to generate realistic data workload for evaluating SC algorithms.

This paper describes our contributions to spatial crowdsourcing. We first present a taxonomy for SC (Figure 1); followed by our approaches to the challenges in SC (Section 3); Finally, Section 4 concludes the paper and provides future research directions.

2. TAXONOMY

In this section, we define a taxonomy of SC (Figure 1) and classify our studies based on the taxonomy (Table 1).

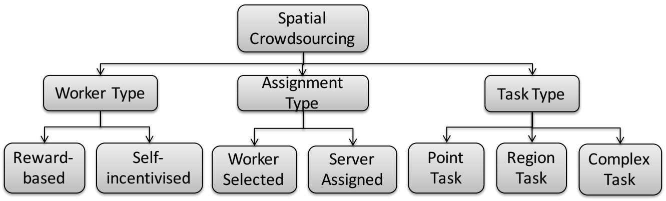

Figure 1: The taxonomy of spatial crowdsourcing.

Crowdsourcing can be classified based on the motivation of the workers into two classes: *self-incentivised* and *reward-based* (Figure 1). With self-incentivised SC, workers volunteer to perform the tasks or usually have other incentives rather than receiving a reward such as promoting their cultural or religious views. An example of this class is traffic.berkeley.edu, in which more than 5000 users voluntarily install traffic software onto their phones and report traffic information. With reward-based SC, every spatial task has a price (assigned by a requester) and workers will receive a certain reward for every task they perform satisfactorily. Examples of this class include fieldagent.net and gigwalk.com. Another example is [4], a reward-based SC platform.

Next, we divide the spatial tasks into three groups based on their geospatial coverages, i.e., *point* task, *region* task and *complex* task. Most of our studies assume *point* task [13, 14, 15, 16], such as taking pictures at particular locations of interest, which require workers to physically present at the task location to perform the task. Meanwhile, our other studies [11, 10] consider *region* task, for example, collecting environmental information. Since the requested data such as air quality, temperature and precipitation exhibit spatial/temporal continuity in measurement, workers available in close vicinity of the task location and request time are sufficient to fulfill that task. Furthermore, one may need to crowdsource a spatial *complex* task consisting of some spatial sub-tasks [2]. An example of a *complex* task is to obtain pictures of ten specific buildings; none of them is allowed to be missed. The *complex* task is completed only when all of the ten pictures, each of a particular building, are captured.

Finally, we define two task assignment modes in SC, *Worker Selected (WS)* and *Server Assigned (SA)* [13, 11, 14, 15, 16, 2]. With the WS mode, the SC-server publishes the spatial tasks and online workers can choose any spatial task in their vicinity without the need to coordinate with the server. An example of this mode is TaskRabbit, where the workers browse for available spatial tasks, and pick the ones in their neighborhood. One drawback of the WS mode is that the server does not have control over the allocation of spatial tasks; thus, workers can choose tasks based on their own objectives (e.g., choosing the k closest spatial tasks to minimize their travel cost). This may result in some spatial tasks never be assigned, while others are assigned redundantly. With the SA mode, online workers send their locations to the SC-server, which then assigns to every worker his nearby tasks. The advantage of SA is that unlike WS, the SC-server has the global picture, and therefore, can assign to every worker his nearby tasks while maximizing the overall task assignment. However, the drawback is that the server knows the locations of all workers, which can pose a privacy threat.

3. FOCUS STUDY

3.1 Task Assignment

With server-assigned SC, requesters and workers typically register with a centralized SC-server that acts as a broker between parties, and often also plays a role in how tasks are assigned to workers. Therefore, the main challenge with SC is to devise an efficient approach to assign tasks to workers given the large scale and dynamism of the environment. In our prior work [16], we propose a framework where the server assigns to every worker tasks in proximity to his location with the aim of maximizing the overall number of assigned tasks. We formally define this maximum task assignment problem and propose alternative solutions by exploiting the spatial properties of the problem space, including the spatial distribution and the travel cost of the workers. The heuristic that obtains the maximum number of assigned tasks is Least Popular Priority (LPP). The idea of LPP is to prioritize tasks during the assignment, i.e., which tasks should be assigned now and which tasks can be deferred to a future time period. In [16], the “popularity” of a task’s location indicates whether the task can be assigned to future workers. The reason is that tasks situated in a “worker-sparse” area are less likely to be performed in the future since there may be no worker visiting such area. Thus, those tasks should have higher priority at the current time period. LPP provides 20% increase in the number of assigned tasks when compared to a baseline algorithm.

The aforementioned approach maximizes task assignment at every time snapshot; such local optimization may not result in a globally optimal answer over the entire campaign. In our recent study [11], we propose a problem of maximizing the number of assigned tasks under the constraint on the number of workers to activate referred to as “budget”. When a budget is specified for the entire campaign, we devise an adaptive strategy to dynamically allocate the given budget to a number of time periods. We propose an online algorithm based on the contextual bandit to captures the arriving patterns of workers and tasks. The empirical results show that the proposed solution increases the task coverage by 40% over the prior heuristic (LPP).

3.2 Workers’ Location Privacy

Thus far, our assignments take as input exact worker locations to minimize the travel cost of workers to the task locations [11, 16, 2]. However, leaking location information may lead to serious consequences. For example, a security flaw in a gay dating app named Grindr reveals precise location of 90% users [1], leading to serious issues in countries where homosexuals faced extreme dangers, such as Egypt and Russia. A recent study [18] shows that a single device with limited resources can report false congestion and

Paper	Worker Type		Task Type			Assignment Type	
	Reward-based	Self-incentivised	Point	Region	Complex	Worker Selected	Server Assigned
[14][15][16][13]		x	x				x
[2]		x	x		x		x
[11]	x	x		x			x
[12]	x		x			x	
[10]	N/A	N/A	x	x		N/A	N/A

Table 1: Our contributions to spatial crowdsourcing.

accidents and automatically reroute user traffic on Waze–a crowdsourced mapping service. Thus, there is a need to protect worker location privacy while still use SC-server as a broker to assign tasks to workers.

Our prior work [14] shows that privacy-preserving worker-task assignment is a challenging task; existing solutions in the context of location-based services and outsourced databases neither offer satisfactory results nor apply in SC. One may claim that simply removing workers' identity by using fake identity (i.e., pseudonymity) would achieve privacy. However, we argue that hiding users' identity without hiding their locations does not provide privacy. This is because a user's location information can be tracked through several stationary connection points (e.g., cell towers). The user's location trace can be easily associated with a certain residence home or office, which reveal the user's identity. This has been referred to as inference attack [6].

Unlike another study [8] that uses cloaking technique to tackle the privacy issue, we propose a framework for protecting the privacy of worker locations, whereby the SC-server only has access to data sanitized according to *differential privacy (DP)* [3]. With this framework, worker locations are first pooled together by the data owner (i.e., cell service provider) and sanitized according to DP. In practice, every worker subscribes to a cellular service provider (CSP), which already has access to the worker locations, e.g., through cell tower triangulation. Thereafter, the SC-server only has access to the sanitized data. However, using DP techniques introduces three challenges.

First, the SC-server must match workers to tasks using noisy data, which requires complex strategies to ensure effective task assignment. Worker location data are sanitized at the CSP using a private spatial decomposition (PSD), named Adaptive Grid [9]. PSD is a sanitized spatial index, where each index node contains a noisy count of the workers rooted at that node. On top of the noisy data, to ensure that task assignment has a high success rate, we developed analytical models and task assignment strategies that consider task completion rate, worker travel distance and system overhead. Second, by the nature of DP protection model, fake entries may need to be created in the worker PSD. Thus the SC-server cannot directly contact workers, not even if pseudonyms are used, as establishing a network connection to an entity would allow the SC-server to learn whether an entry is real or not, and breach privacy. To address this challenge, a geocast mechanism was introduced for the task request dissemination. Geocast is a routing and addressing method, which is used to deliver information to all devices situated within a geographical area. Once a PSD partition is identified by the analytical model outlined above, the task request is geocast to all the workers within that partition. Third, protecting worker locations across multiple timestamps is notoriously difficult. As workers move, new snapshots of sanitized worker locations must be disclosed, to maintain task assignment effectiveness. However, access to sequential releases gives an adversary more powerful attack opportunities. To counter such threats, differential privacy requires more noise injection, which in the worst case may reach amounts that are proportional to the length of the released location history (i.e., number of disclosed snapshots). Clearly, such large noise would render the data useless, since SC is likely to be a continuously offered service in practice. To address the challenge of moving workers, we investigate privacy budget allocation techniques across consecutive releases, and we employ post-processing techniques based on Kalman filters to reduce the inaccuracy introduced by noise addition [13]. Our experimental results show that workers' location privacy is protected without compromising performance and the extra travel cost is tolerable (20% increase when compared to the non-private case).

3.3 Incentive Mechanism

Incentive mechanism plays an important role in maximizing the number of performed tasks in SC. However, little research has been done to understand worker's motivation in SC markets using real systems. To fill this gap, in a recent work [12], we study the workers' behavior in two new paid SC campaigns in Japan. We develop an Android app named Genkii to collect users' moods and use Yahoo! Japan Crowdsourcing as the payment platform. To receive a reward, a worker needs to use Genkii to report his/her mood (i.e., Happy, Ok, Dull) at the right time and at the right place. Subsequently, the participating users' behaviors are analyzed through spatial and temporal analysis. Our findings in this study are three-fold.

We first study the relationship between incentives and participation by analyzing the impact of offering a fixed reward versus an increasing reward scheme. Note that the total rewards were the same in both campaigns. We observe that users tend to stay in a campaign longer when the provided incentives gradually increase over time, showing that workers are motivated by growing incentives. Second, we report the worker performance during the two campaigns. We obtain a total of 1059 reports from both campaigns, out of which 436 reports were from the first campaign and 623 reports from the second. We observe a cyclic pattern in the number of reports per hour during a day. Particularly, 4, 12 and 20 are the hours with peak numbers of reports. Interestingly, they are pastimes in Japan. Moreover, 1, 9 and 17 are the hours with the least number of reports. Not surprisingly, these are common commute times in Japan. Third, we study worker mobility from the reporting locations. Each worker has a certain degree of mobility defined as the area of the minimum bounding rectangle that encloses all the reporting locations. We found that the degree of mobility is correlated with the reported information. For example, users who travel more are observed to be happier than the ones who travel less.

3.4 Lack of Real-world Datasets

One of the biggest challenges for the research community is the absence of real-world SC datasets. Thus, most SC algorithms are evaluated using proprietary data, users' historical call records [7] or synthetic datasets [5, 2, 11, 15, 17]. The call records (or mobility data) do not represent actual datasets needed for the SC studies, in which workers respond to requesters' tasks. Hence, the generated synthetic datasets from these studies may not represent the properties of real-world SC datasets.

To fill this void, we propose a realistic workload generator for the SC applications, namely *SCAWG* (SC for Adaptive Workload Generator). *SCAWG* considers realistic spatiotemporal properties and behaviors for both workers and tasks. The generated data are either purely synthetic or adapted from geosocial datasets with users' check-in information, such as Gowalla[3] and Yelp[4]. These datasets exhibit the *task locality* property in SC, i.e., workers tend to perform nearby tasks [5]. With purely synthetic data, since a single spatial distribution may not be able to simulate spatial distribution of some geosocial phenomena, *SCAWG* introduces the mixed distributions to mimic complex behaviors observed in the real world. The mixed distribution (i.e., either spatial or temporal) combines multiple primitive distributions (e.g., Uniform, Gaussian). With *SCAWG*, researchers can reuse our common, well-structured and extensible datasets in their evaluation studies, e.g., algorithms can be compared on the same workload, facilitating the reproducibility of research findings.

As SC aims to provide a generic platform for different campaigns, our basic workload includes only the core requirements of spatial workers and tasks. In particular, each task (or worker) comprises an initial location, a start time, i.e., arrival (online) time and an end time, i.e., expiry (offline) time. The locations of workers or tasks follow a spatial distribution (e.g., 2D Gaussian) while their start times follow a temporal distribution (e.g., Poisson). Furthermore, we observe that both workers and tasks may have application-specific constraints or properties. First, a worker may set the maximum number of tasks he can perform within a time period, or a *spatial region* within which he is willing to travel while a task may require a certain number of times it needs to be performed. Also, there are different kinds of tasks, e.g., tasks that require trustful workers, need a specific set of skills [16] or are associated with rewards [4, 12]. Second, the advanced generator takes into account worker *identity*, which can be used to either enhance task allocation [16, 11, 5] or avoid repetitive activations of the same workers [11]. Each worker also has an *activeness* value, in which active workers are likely to be available more often and perform more tasks.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we summarized our contributions to spatial crowdsourcing–a new paradigm of community data collection at large scale. We discussed our approaches toward addressing some of its challenges, with emphasis on the new privacy-preserving and efficient techniques for online task assignment. As future work, we will extend our framework to also protect the privacy of task locations. We also plan to conduct SC campaigns with more complex reward strategies to incentivize workers to accept SC tasks, e.g., dynamically update rewards for each task based on a given target, e.g., each task may require a number of responses to be considered fulfilled. Hence, an effective reward strategy is required to monitor the number of responses per task and the workers' locations over time.

Acknowledgement

This research has been funded by NSF grants IIS-1320149 and CNS-1461963, the USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Google, Northrop Grumman, Microsoft, and Oracle. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any of the sponsors such as NSF.

5. REFERENCES

[1] J. Cook. Security flaw in gay dating app grindr reveals precise location of 90% of users. August 2014.

[2] H. Dang, T. Nguyen, and H. To. Maximum complex task assignment: Towards tasks correlation in spatial crowdsourcing. In *IIWAS*, page 77:81, NY, USA, 2013. ACM.

[3] C. Dwork. Differential privacy. In *Automata, languages and programming*, pages 1–12. Springer, 2006.

[4] T. Kandappu, N. Jaiman, R. Tandriansyah, A. Misra, S.-F. Cheng, C. Chen, H. C. Lau, D. Chander, and K. Dasgupta. Tasker: Behavioral insights via campus-based experimental mobile crowd-sourcing. *UbiComp*, 2016.

[5] L. Kazemi and C. Shahabi. GeoCrowd: enabling query answering with spatial crowdsourcing. In *ACM SIGSPATIAL*. ACM, 2012.

[6] J. Krumm. Inference attacks on location tracks. In *Pervasive Computing*, pages 127–143. Springer, 2007.

[7] Y. Liu, B. Guo, Y. Wang, W. Wu, Z. Yu, and D. Zhang. TaskMe: Multi-task allocation in mobile crowd sensing. *arXiv preprint arXiv:1608.02657*, 2016.

[8] L. Pournajaf, L. Xiong, V. Sunderam, and S. Goryczka. Spatial task assignment for crowd sensing with cloaked locations. In *MDM*, volume 1, pages 73–82. IEEE, 2014.

[9] W. Qardaji, W. Yang, and N. Li. Differentially private grids for geospatial data. *ICDE*, 2013.

[10] H. To, M. Asghari, D. Deng, and C. Shahabi. SCAWG: A toolbox for generating synthetic workload for spatial crowdsourcing. In *PerCom Workshops*, pages 1–6. IEEE, 2016.

[11] H. To, L. Fan, L. Tran, and C. Shahabi. Real-time task assignment in hyperlocal spatial crowdsourcing under budget constraints. In *PerCom*, pages 1–8. IEEE, 2016.

[12] H. To, R. Geraldes, C. Shahabi, S. H. Kim, and H. Prendinger. An empirical study of workers' behavior in spatial crowdsourcing. *GeoRich Workshop*, 2016.

[13] H. To, G. Ghinita, L. Fan, and C. Shahabi. Differentially private location protection for worker datasets in spatial crowdsourcing. In *TMC*. IEEE, 2016.

[14] H. To, G. Ghinita, and C. Shahabi. A framework for protecting worker location privacy in spatial crowdsourcing. *Proceedings of the VLDB Endowment*, 7(10):919–930, 2014.

[15] H. To, G. Ghinita, and C. Shahabi. PrivGeoCrowd: A toolbox for studying private spatial crowdsourcing. In *ICDE*, pages 1404–1407, April 2015.

[16] H. To, C. Shahabi, and L. Kazemi. A server-assigned spatial crowdsourcing framework. *ACM TSAS*, 1(1):2, 2015.

[17] Y. Tong, J. She, B. Ding, L. Chen, T. Wo, and K. Xu. Online minimum matching in real-time spatial data: Experiments and analysis. *Proceedings of the VLDB Endowment*, 9(12), 2016.

[18] G. Wang, B. Wang, T. Wang, A. Nika, H. Zheng, and B. Y. Zhao. Defending against Sybil devices in crowdsourced mapping services. In *MobiSys*, pages 179–191. ACM, 2016.

[19] Y. Zhao and Q. Han. Spatial crowdsourcing: current state and future directions. *IEEE Communications Magazine*, 2016.

[3] snap.stanford.edu/data/loc-gowalla.html.

[4] yelp.com/dataset_challenge