PrivGeoCrowd: A Toolbox for Studying Private Spatial Crowdsourcing

Hien To [1], Gabriel Ghinita [2], Cyrus Shahabi [3]

[1,3] *University of Southern California,* [2] *University of Massachusetts Boston*
{hto,shahabi}@usc.edu
Gabriel.Ghinita@umb.edu

***Abstract*—Spatial Crowdsourcing (SC) is a novel and transformative platform that engages individuals, groups and communities in the act of collecting, analyzing, and disseminating environmental, social and other spatio-temporal information. SC outsources a set of spatio-temporal tasks to a set of *workers*, i.e., individuals with mobile devices that perform the tasks by physically traveling to specified locations of interest. Protecting location privacy is an important concern in SC, as an adversary with access to individual whereabouts can infer sensitive details about a person (e.g., health status, political views). Due to the challenging nature of protecting worker privacy in SC, solutions for this problem are quite complex, and require tuning of several parameters to obtain satisfactory results. In this paper, we propose *PrivGeoCrowd*, a toolbox for interactive visualization and tuning of SC private task assignment methods. This toolbox is useful for several real-world entities that are involved in SC, such as: mobile phone operators that want to sanitize datasets with worker locations, spatial task requesters, and SC-service providers that match workers to tasks.**

I. Introduction

Latest-generation mobile phones are powerful devices that can collect and share various types of data, e.g., picture, video, location, movement speed and acceleration. *Spatial Crowdsourcing (SC)* [1] is emerging as a transformative platform that engages individuals, groups and communities in the act of collecting, analyzing, and disseminating information for which spatio-temporal features are relevant. With SC, *task requesters* outsource their spatio-temporal tasks to a set of *workers*, i.e., individuals with mobile devices that perform the tasks by physically traveling to specified locations of interest. The nature of tasks may vary from environmental sensing to capturing images at social or entertainment events. Typically, requesters and workers register with a centralized *spatial crowdsourcing server (SC-server)* that acts as a broker between parties, and often also plays a role in how to assign tasks to workers (i.e., scheduling according to some performance criteria). SC has numerous applications in environmental sensing, journalism, crisis response, etc.

SC requires workers and tasks to be matched effectively, i.e., tasks must be completed in a timely fashion, and workers need not travel long distances. Thus, matching at the SC-server must take into account worker locations. However, the SC-server may not be trusted, and disclosing worker locations may help an adversary learn sensitive details about an individual's health status, political views, etc [2], [3], [4].

In previous work [5] we proposed a differentially-private framework for spatial task assignment in SC, whereby the SC-server only has access to sanitized *Private Spatial Decompositions (PSDs)* [4]. A PSD is a sanitized spatial index, where each index node contains a noisy count of the workers rooted at that node. Every worker subscribes to a *cellular service provider (CSP)* that provides Internet connectivity. The CSP already has access to the worker locations (e.g., through cell tower triangulation), and signs a contract with its subscribers, which stipulates the terms and conditions of location disclosure. The CSP collects worker locations and releases them to third-party SC-servers in sanitized form. However, differential privacy (DP) introduces two difficult challenges, as discussed next.

First, the SC-server must match workers to tasks using noisy data, which requires complex strategies to ensure effective task assignment. Second, by the nature of the DP protection model, fake entries may need to be created in the PSD. Thus, the SC-server cannot directly contact workers, not even if pseudonyms are used, as merely establishing a network connection would allow the SC-server to learn whether an entry is real or not, and breach privacy. To address this challenge, the framework in [5] proposed the use of geocasting [6] to deliver task requests to workers. Geocast introduces overhead considerations that need to be carefully considered in the framework design.

Preserving worker location privacy in SC is a challenging and complex task, which requires a careful system analysis and tuning for several framework components, such as the shape and granularity of the PSD, the strategy for finding appropriate geocast regions, etc. In addition, the distribution of the worker and task datasets significantly impact the effectiveness and efficiency of the framework.

In this demo, we propose *PrivGeoCrowd*, an interactive visualization and tuning toolbox for privacy-preserving SC. *PrivGeoCrowd* helps system designers investigate the effect of varying system parameters, as well as the behavior of different choices of geocasting strategy. *PrivGeoCrowd* includes several modules which can be deployed at different system components. For instance, the CSPs may use *PrivGeoCrowd* to evaluate the effect that PSD characteristics and privacy budget allocation have on assignment quality. To that extent, the CSP can use *PrivGeoCrowd* to generate several sanitized datasets, and evaluate the impact on assignment quality using previous histories of task requests. On the other hand, the SC-server can investigate, given an instance of a sanitized dataset,

978-1-4799-7964-6/15/$31.00 © 2015 IEEE

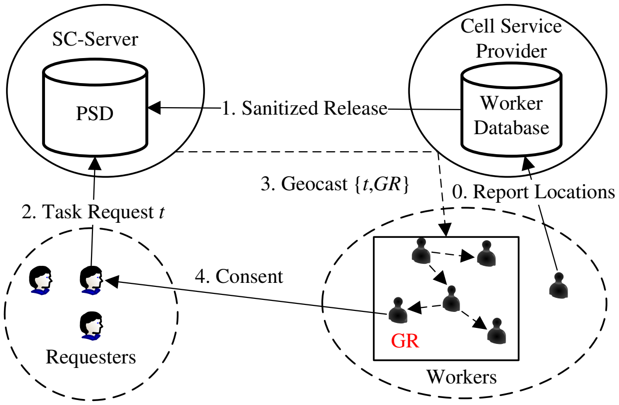


Fig. 1: Privacy framework for spatial crowdsourcing

what is the impact of different geocast region generation and task matching strategies. Finally, task requesters can use *PrivGeoCrowd* to visualize how their previously submitted tasks have been dispatched and completed, and thus decide whether the billing from their SC-server is fair.

The remainder of this paper is organized as follows. Section II provides a technical overview of the private SC framework from [5], which is the basis for the interactive visualization tool. Next, in Section III we present the *PrivGeoCrowd* demonstration plan. Finally, we conclude with directions for future work in Section IV.

II. SC Privacy Framework

A. Overview

Fig. 1 summarizes the differentially-private SC framework proposed in [5]. SC workers subscribe to a *cellular service provider (CSP)* and periodically report their location to the CSP (Step 0 in the diagram). In Step 1, the CSP sanitizes worker locations according to differential privacy with privacy budget ϵ using *Private Spatial Decompositions (PSD)* [4]. Each PSD node has a noisy count of the data points enclosed by that node's extent (instead of actual data points in the case of traditional indexes). Various PSD types such as grids, quad-trees or k-d trees can be used.

The SC framework in [5] uses the *Adaptive Grid (AG)* PSD approach [7]. *AG* is a two-level grid, where the granularity of the second-level grid is chosen based on the noisy counts obtained in the first-level. *AG* is a hybrid technique which inherits the simplicity and robustness of space-partitioning (i.e., data independent) index structures, but still uses a certain amount of data-dependent information when choosing the granularity for the second level.

When the SC-server receives a task request t (Step 2), it queries the PSD to determine a *geocast region (GR)* that encloses with high probability workers in relative proximity to t. Next, the SC-server initiates a *geocast* communication [6] process (Step 3) to disseminate t to all workers within GR. According to DP, sanitizing a dataset requires creation of fake locations in the PSD. If the SC-server is allowed to directly contact workers, then failure to establish a communication channel would breach privacy, as the SC-server is able to distinguish fake workers from real ones. Using geocast is a unique feature of the framework, and a necessary step to achieve protection. Geocasting [6] is a routing and addressing method used to communicate with nodes in a network identified by their geographical locations. Geocasting can be performed either with the help of the CSP infrastructure, or through a mobile ad-hoc network where the message is disseminated on a hop-by-hop basis to the entire GR.

Upon receiving request t, a worker w decides whether to perform the task or not. If yes (Step 4), s/he sends a *consent* message to the SC-server confirming w's availability (alternatively, the consent can be directly sent to the requester). If w is not willing to participate in the task, then no consent is sent, and no information about the worker is disclosed.

B. Challenges and Performance Metrics

Protecting worker locations complicates significantly task assignment, and may reduce the effectiveness of worker-task matching. Given task t, the GR construction algorithm must balance two conflicting requirements: determine a region that *(i)* contains sufficient workers such that task t is accepted with high probability, and *(ii)* the size of the GR is small. The willingness of a worker to accept a received task is modeled as the *acceptance rate* AR, which can vary as either a linear or a Zipfian function of the worker-task distance. The input to the algorithm is task t as well as the worker PSD, consisting of the two-level *AG* with noisy worker counts.

The algorithm chooses as initial GR the level-2 cell that covers the task, and keeps expanding the GR by adding neighboring cells until the GR *utility*, measured as the probability that at least one worker inside the GR accepts the task, exceeds a required threshold. GRs are expanded using several heuristic criteria. The *distance*-based heuristic selects the nearest cell to the task. The *compactness*-based heuristic chooses the cell that forms the most compact GR, i.e., a GR with a shape that is more suitable for geocast.

Three important metrics need to be considered:

- *Assignment Success Rate (ASR)* measures the ratio of tasks accepted for execution to the total number of task requests. The ideal ASR is 100%.
- *Worker Travel Distance (WTD)* is an indicator of travel cost. The challenge is to keep WTD low, even when exact worker locations are not known.
- Due to the noisy data, redundant messages must be sent, increasing overhead. The average number of notified workers (*ANW*) measures the *communication overhead* of disseminating tasks. Another important metric is the number of hops required to disseminate the tasks when ad-hoc routing is used.

To obtain practical values for the above performance metrics, the privacy framework needs to be carefully tuned with respect to several parameters, such as privacy budget and its allocation strategy, granularity of *AG*, utility threshold settings, choice of GR construction strategy, etc. To evaluate the impact of these factors, and to support a comprehensive exploration of the effects of each parameter, we developed *PrivGeoCrowd*, an interactive visualization toolbox for privacy-preserving SC that we introduce next.

III. PrivGeoCrowd

In Section III-A we outline the software architecture of the tool, followed by a presentation of the main GUI elements in Section III-B and the demonstration scenario in Section III-C.

A. System Architecture

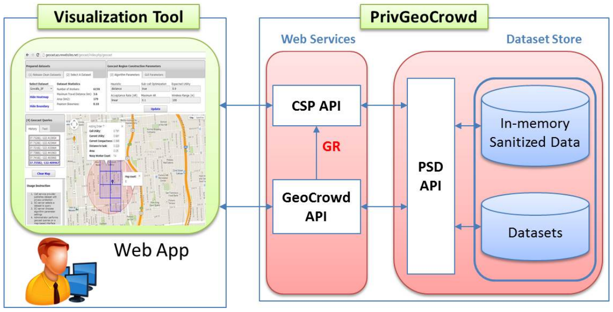


Fig. 2: Architecture of PrivGeoCrowd

Fig. 2 presents the architecture of *PrivGeoCrowd*, which is integrated with MediaQ [8], a geospatial crowdsourcing system developed at USC. The GUI is provided to end-users (i.e., CSPs, SC-server administrators or task requesters) in web-based form, and does not require any additional software to be installed at the client. The web application is written using Javascript, Python, PHP5, and Google Maps API for map rendering. The GeoCrowd API provides SC services [8] such as profile management, etc. The worker location datasets are stored in a MySQL database, which can be queried using the CSP API to generate sanitized PSDs.

B. Graphical User Interface

Fig. 3 gives an overview of the main GUI, which comprises several component modules:

CSP Panel: it allows the CSP to sanitize and publish datasets according to privacy budget $\varepsilon \in [0.1, 1]$. The system ships with three datasets: Gowalla-San Francisco (SF), Gowalla-Los Angeles (LA) and Yelp-Phoenix (Yelp), but the CSP has the option of uploading additional datasets to the repository. The CSP can also specify the budget allocation split between *AG* levels (0.5 means equal split between levels 1 and 2). The CSP can customize the granularity of the level-1 *AG* grid used for constructing the worker PSD. Once the "Publish Data" button is clicked, the corresponding PSD of the dataset is generated and made available to other modules through the PSD API.

SC Panel - Dataset Selection: The SC-server administrator can select among available sanitized datasets from the drop-down list. Several statistics of the selected dataset are automatically provided on the right-hand side, and the user has the option to visualize the dataset density heatmap, as well as the dataset boundary.

SC Panel - GR Construction Tuning: The SC-server administrator is able to select one of the supported heuristics (i.e., distance-based, compactness-based), as well as the parameters of the GR construction algorithm. The user can choose the threshold for acceptance success rate (ASR). The maximum task acceptance rate AR threshold (i.e., when worker and task are co-located) can be varied between 0.1 and 1.0. The wireless communication range for the geocast is customizable between 25 and 100 meters.

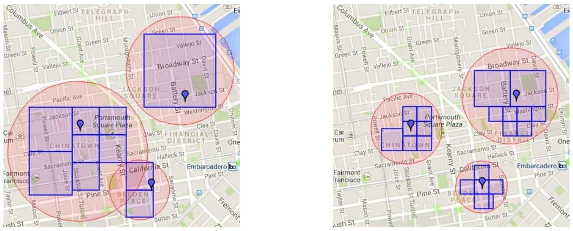
Task Requester Panel - Geocast Region Rendering: There are three ways one can submit a task request: (1) by double clicking on the map, (2) by providing latitude/longitude values in the task text box, or (3) by selecting a particular task in the history tab. The latter task list is extracted from MediaQ [8]. When one specifies a task, its GR is computed and rendered in real-time. In Fig. 3, the pop-up dialog of the last visited cell presents some statistics. Typically, the current utility (i.e., accumulated utility) measures the probability that at least one worker accepts the task if it is geocast.

Mobility Panel: The administrator can generate new datasets from existing ones by having workers move according to a pre-defined mobility model. The heat map of the updated dataset illustrates the movement. When the stop button is clicked, the current snapshot of the worker locations are uploaded to the server, and a new PSD is constructed.

C. Demonstration Plan

During the demonstration, we will highlight the role of *PrivGeoCrowd* in evaluating the effectiveness and efficiency of private spatial crowdsourcing in several prominent scenarios. Namely, we study the effect of: *(i)* varying *AG* granularity during sanitization, *(ii)* varying dataset density and *(iii)* varying the heuristic used in GR construction.

Customized AG granularity: Fig. 4 presents the effect of AG granularity on GR size. The visualization highlights the GRs obtained by our customized granularity AG method [5], which are significantly more compact than the ones obtained with the original AG from [7].

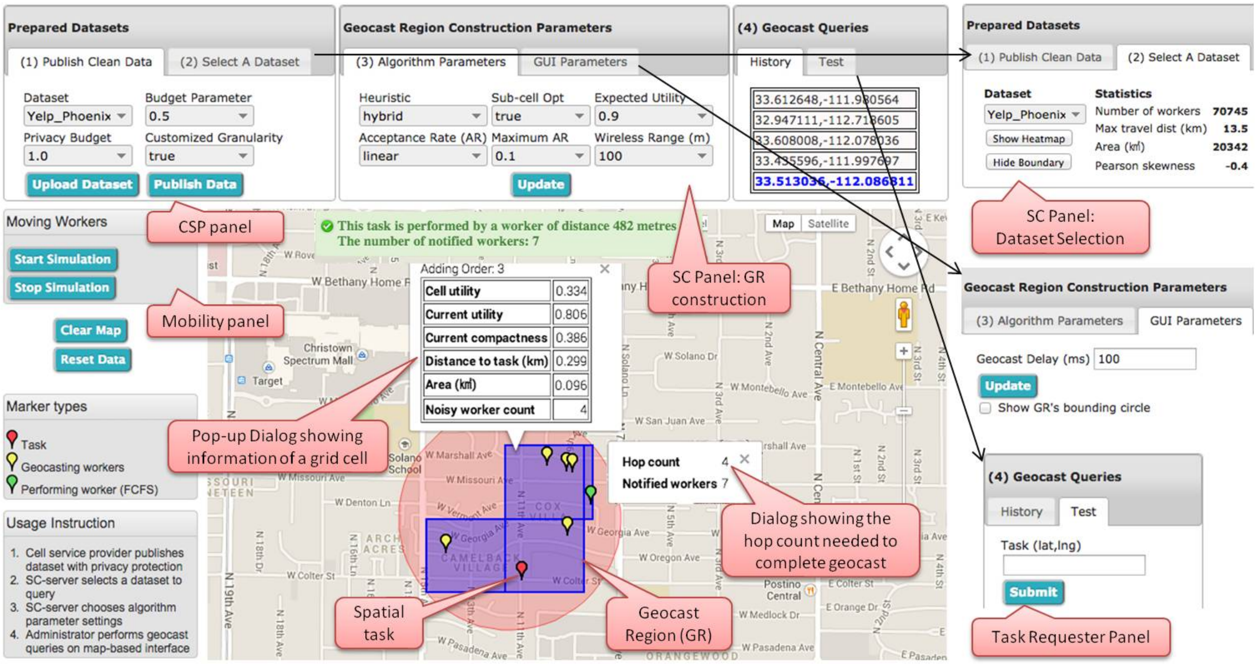


(a) Original *AG* (b) Modified *AG*

Fig. 4: The effect of customized granularity on GR (SF data)

Dense vs Sparse Area: Fig. 5 shows the effect of worker density on the size of obtained GRs. One can observe that the GR obtained in the left-hand side of the figure, corresponding to a sparse suburb area of Phoenix, AZ (Yelp dataset) is much larger than the GR obtained in a denser downtown area of the same dataset. On the other hand, the denser the population, the higher granularity of the *AG*. This confirms that our customized *AG* [5] adapts well to worker density.

Effect of Alternative GR Construction Heuristics: Fig. 6 demonstrates the behavior of different heuristics on the size and shape of the GR. In Fig. 6a we illustrate the case where GR construction is guided solely by the expected probability of task acceptance success rate (ASR). In this case, populated grid cells tend to be selected first. Fig. 6b demonstrates the



Prepared Datasets

(1) Publish Clean Data (2) Select A Dataset

Dataset: Yelp_Phoenix
Budget Parameter: 0.5
Privacy Budget: 1.0
Customized Granularity: true

Geocast Region Construction Parameters

(3) Algorithm Parameters GUI Parameters

Heuristic: hybrid
Sub-cell Opt: true
Expected Utility: 0.9
Acceptance Rate (AR): linear
Maximum AR: 0.1
Wireless Range (m): 100

(4) Geocast Queries

History Test

33.612648,-111.980564
32.947111,-112.718605
33.608008,-112.078036
33.435596,-111.997697
33.513036,-112.086811

This task is performed by a worker of distance 482 metres
The number of notified workers: 7

Adding Order: 3

Cell utility	0.334
Current utility	0.806
Current compactness	0.386
Distance to task (km)	0.299
Area (㎢)	0.096
Noisy worker count	4

Hop count 4
Notified workers 7

Moving Workers

Marker types
- Task
- Geocasting workers
- Performing worker (FCFS)

Usage Instruction
1. Cell service provider publishes dataset with privacy protection
2. SC-server selects a dataset to query
3. SC-server chooses algorithm parameter settings
4. Administrator performs geocast queries on map-based interface

Prepared Datasets

(1) Publish Clean Data (2) Select A Dataset

Dataset: Yelp_Phoenix

Statistics

Number of workers	70745
Max travel dist (km)	13.5
Area (㎢)	20342
Pearson skewness	-0.4

Geocast Region Construction Parameters

(3) Algorithm Parameters GUI Parameters

Geocast Delay (ms): 100

□ Show GR's bounding circle

(4) Geocast Queries

History Test

Task (lat,lng)

Fig. 3: *PrivGeoCrowd* main GUI integrates several component module panels

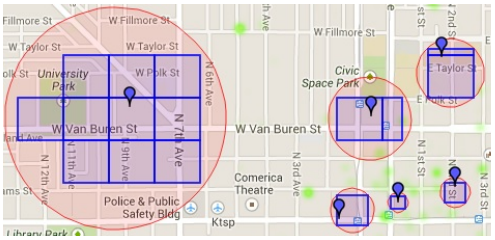
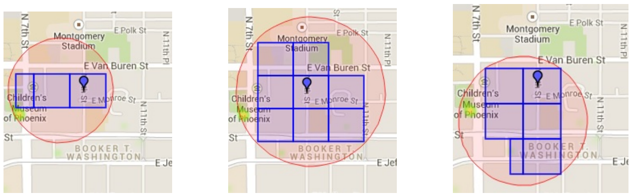


Fig. 5: The effect of worker density on GR size (Yelp data)

strategy of adding the nearest cell to the GR, which results in GRs that are centered at the task request. Finally, Figure 6c shows the third strategy that attempts to form GRs with a balanced shape, in order to obtain a low hop count when using ad-hoc geocast communication.



(a) ASR-centric (b) Distance (c) Compactness

Fig. 6: The effect of different heuristics on GR geometry

IV. Conclusion

We showcase *PrivGeoCrowd*, an interactive visualization and tuning toolbox for privacy-preserving spatial crowdsourcing. *PrivGeoCrowd* helps system designers investigate the effect of parameters such as privacy budget and allocation strategy, GR construction heuristics, dataset density, etc., on the effectiveness of private SC task matching. In future work, we will extend *PrivGeoCrowd* to support multiple types of PSD (in addition to adaptive grids). We will also integrate *PrivGeoCrowd* with a network simulator component that can provide precise geocasting overhead measurements when routing in a realistic mobile ad-hoc environment.

Acknowledgment

H. To and C. Shahabi have been supported in part by NSF grant IIS-1320149, the USC Integrated Media Systems Center (IMSC), and unrestricted cash gifts from Google and Northrop Grumman. G. Ghinita was supported by NSF award CNS-1111512. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the sponsors such as NSF.

References

[1] L. Kazemi and C. Shahabi, “Geocrowd: enabling query answering with spatial crowdsourcing,” in *ACM SIGSPATIAL GIS*, 2012, pp. 189–198.

[2] M. F. Mokbel, C.-Y. Chow, and W. G. Aref, “The New Casper: Query Processing for Location Services without Compromising Privacy,” in *Proc. of VLDB*, 2006.

[3] G. Ghinita, P. Kalnis, A. Khoshgozaran, C. Shahabi, and K.-L. Tan, “Private queries in location based services: anonymizers are not necessary,” in *SIGMOD*, 2008, pp. 121–132.

[4] G. Cormode, C. Procopiuc, D. Srivastava, E. Shen, and T. Yu, “Differentially private spatial decompositions,” in *ICDE*, 2012, pp. 20–31.

[5] H. To, G. Ghinita, and C. Shahabi, “A framework for protecting worker location privacy in spatial crowdsourcing,” *Proceedings of the VLDB Endowment*, vol. 7, no. 10, 2014.

[6] J. C. Navas and T. Imielinski, “Geocast:geographic addressing and routing,” in *Proc. of ACM/IEEE international conference on Mobile computing and networking*. ACM, 1997, pp. 66–76.

[7] W. Qardaji, W. Yang, and N. Li, “Differentially private grids for geospatial data,” in *International Conference on Data Engineering (ICDE)*, 2013.

[8] S. H. Kim, Y. Lu, G. Constantinou, C. Shahabi, G. Wang, and R. Zimmermann, “Mediaq: Mobile multimedia management system,” in *ACM Multimedia Systems Conference*, 2014.